Fragments d'Optimisation Différentiable - Théories et Algorithmes

Jean Charles Gilbert

▶ **To cite this version:**

Jean Charles Gilbert. Fragments d'Optimisation Différentiable - Théories et Algorithmes. Master. France. 2021, pp.979. hal-03347060

HAL Id: hal-03347060
https://inria.hal.science/hal-03347060

Submitted on 16 Sep 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Fragments d'Optimisation Différentiable

—

Théories et Algorithmes

Jean Charles GILBERT[†]

16 septembre 2021

[†] INRIA Paris, 2 rue Simone Iff, CS 42112, F-75589 Paris Cedex 12, France; Jean-Charles.Gilbert@inria.fr.

Préface

There are hardly any speculations in geometry more useful or more entertaining than those which relate to maxima and minima.

C. MacLaurin (1742). A treatise of Fluxions.

Voyager, c'est bien utile, ça fait travailler l'imagination. Tout le reste n'est que déceptions et fatigues. Notre voyage à nous est entièrement imaginaire. Voilà sa force.

Céline (1952). Voyage au bout de la nuit.

Je vais donc reprendre ma pauvre vie si plate et tranquille, où les phrases sont des aventures et où je ne recueille d'autres fleurs que des métaphores.

G. Flaubert. Correspondance.

Telle est, bien sûr, l'ambition secrète et démesurée de tout auteur d'anthologie. S'il la commence pour lui-même, c'est pour d'autres qu'il la termine et la publie. Choisir, dans un domaine déterminé, tout ce qui lui paraît digne et capable de provoquer chez le lecteur le choc de la beauté, voilà l'objet de son effort.

G. Pompidou (1961). Anthologie de la poésie française

Si vous voulez savoir ce qu'une démonstration démontre, regardez la démonstration.

L. Wittgenstein, cité par J. Bouveresse le 17 juin 1998 dans une conférence sur l'affaire Sokal et ses conséquences.

Cette difficulté de la science contemporaine est-elle un obstacle à la culture ou est-elle un attrait ? Elle est, croyons-nous, la condition même du dynamisme psychologique de la recherche. Le travail scientifique demande précisément que le chercheur se crée des difficultés. [...] *L'orgueil de savoir réclame le mérite de vaincre la difficulté de savoir.*

G. Bachelard (1971). Épistémologie.

This paper is too long ... please add information on topics X, Y, and Z.

B. Stroustrup (1994), raillant les commentaires contradictoires des rapports d'arbitrage sur son papier résumant la partie historique de son livre [574].

Il s'étonnait de souffrir autant. Profondément éloignée des catégories chrétiennes de la rédemption et de la grâce, étrangère à la notion même de liberté et de pardon, sa vision du monde en acquérait quelque chose de mécanique et d'impitoyable. [...] *La nuit* [*il*] *rêvait d'espaces abstraits, recouverts de neige.*

M. Houellebecq (1998). Les particules élémentaires.

L'optimisation numérique a ceci d'amusant que toute personne ayant une formation minimale en calcul scientifique et s'initiant à la discipline s'estime rapidement suffisamment compétente pour introduire de nouveaux algorithmes ! Il faut dire que le problème qui se pose en optimisation numérique différentiable sans contrainte renvoie à une image familière, celle où il s'agit de descendre au plus bas dans une vallée, et y trouver un chemin conduisant à destination semble bien être à la portée de tous. Prendre en compte des contraintes d'inégalité ne paraît guère plus difficile puisque cela revient à restreindre la descente à un enclos à la frontière bien définie. Il n'est donc pas rare de rencontrer des ingénieurs ou des chercheurs bricolant de « nouveaux » algorithmes aux propriétés incertaines, à la convergence aléatoire, au temps de calcul démesuré, mais qui satisfont pleinement leurs auteurs..., jusqu'au jour où ceux-ci se rendent compte que le problème n'est peut-être pas si simple, que leur algorithme a bien quelques faiblesses rédhibitoires et qu'il serait sage de consulter un spécialiste. Il est donc sans doute utile de rappeler ici, qu'il existe en effet une petite communauté de chercheurs dont c'est le métier d'étudier et d'améliorer l'algorithmique en optimisation, ayant leurs revues spécialisées, leurs vedettes, leurs maîtres vénérés et ne manquant pas leurs conférences internationales régulières. Il reste probablement beaucoup d'algorithmes à découvrir en optimisation et certainement beaucoup doivent être mieux compris, rendus plus efficaces ou adaptés à des situations nouvelles ou particulières. Ceci est d'autant plus vrai qu'aucun algorithme n'est entièrement satisfaisant, entraînant des déceptions que l'on cherche naturellement à adoucir par de nouveaux remèdes, ou mieux, par une nouvelle interprétation des algorithmes et de leur fondement permettant leur amélioration. L'exploration du champ des possibles est donc souhaitable, pourvu que ceci ne conduise pas à raviver de vieilles recettes périmées. La partie numérique de cet ouvrage rend compte des principales solutions que les numériciens ont apportées à quelques problèmes classiques de l'optimisation différentiable. Les impasses et les écueils à éviter y sont aussi décrits. Nous espérons ainsi être utiles à ceux qui se sentent une âme d'algorithmicien.

Si un objectif important de cet ouvrage est l'algorithmique en optimisation différentiable, le chemin pour y arriver pourra paraître long à certains. Les méthodes numériques efficaces reposent en effet sur une bonne compréhension de la structure et des propriétés des problèmes d'optimisation qu'elles cherchent à résoudre ; c'est une raison suffisante pour étudier ces derniers. Par ailleurs, même si la démonstration

de la convergence d'un algorithme ne doit pas être la motivation première lors de sa conception, les algorithmes ne sont vraiment acceptés que si l'on parvient à en décrire les propriétés de convergence globale et locale, voire de complexité. Les numériciens y consacrent une grande partie de leurs efforts. Ces différents aspects requièrent le développement d'une théorie solide, ce qui explique la première partie du sous-titre de cet ouvrage.

Ce livre est long ; il ne contient pourtant qu'une introduction aux différents sujets qu'il aborde, que des *fragments* de ceux-ci. Presque tous ses chapitres, parfois de simples sections, ont été développés en d'épaisses monographies par d'autres auteurs. Le spécialiste sera donc parfois frustré par l'absence de certains concepts ou de leur développement, par l'ignorance de certains algorithmes ou le côté superficiel de leur étude ; par ailleurs, le néophyte pourra être découragé par le foisonnement des sujets traités ou par la difficulté de certains passages. C'est donc dans la recherche d'un équilibre, tout subjectif, entre ces deux pôles que s'est constitué cet ouvrage. Nous avons essayé d'aborder de nombreux problèmes et algorithmes de résolution, en nous efforçant chaque fois de les contenir dans un chapitre de quelques dizaines de pages. Des notes de fin de chapitre invitent le lecteur à poursuivre son exploration, à enrichir ses connaissances, le long de pistes que nous avons voulu variées.

Ce manuel a été le support de cours d'enseignements délivrés à l'ENSTA (Paris, puis Saclay), à l'ENSAE (Paris) et aux universités Paris I, Paris VI et Paris Saclay. Les élèves et étudiants de ces écoles et universités ont donc été directement ou indirectement impliqués dans la réalisation de cet ouvrage et ont à leur manière contribué à son amélioration. L'école d'ingénieurs ENSTA qui m'a accepté comme enseignant en optimisation numérique durant de nombreuses années est particulièrement remerciée pour la confiance qu'elle m'a ainsi accordée.

Le cours d'optimisation numérique de l'ENSTA avait été précédemment assuré par Christian Saguez et Frédéric Bonnans et cet ouvrage a été dans un premier temps fortement inspiré du style et du contenu de leur support de cours. Qu'ils soient ici chaleureusement remerciés pour ce passage de flambeau sans lequel cet ouvrage n'aurait pas existé. Puis, le manuscrit a évolué en fonction des intérêts de l'auteur, des sujets rencontrés, des progrès de la discipline. Et il a aussi beaucoup enflé...

J'aimerais également remercier Paul Armand, Ibtihel Ben Gharbia, Kengy Barty, Benoît Bonnet, Pierre Carpentier, Laurent Chauvier, Alice Chiche, François Clément, Frédéric Delbos, Anna Désilles, Jean-Pierre Dussault, Émilien Flayac, Nidham Gazagnadou, Pierre Girardeau, Frédéric Jean, Xavier Jonsson, Elizabeth Karas, Michel Kern, Vincent Leclere, Claude Lemaréchal, Arthur Le Rhun, Eugénie Marescaux, Joëlle Mélotte, Quentin Mérigot, Luca Nenna, Baptiste Plaquevent-Jourdain, Manuel Ruiz, Claudia Sagastizábal, Filippo Santambrogio, Delphine Sinoquet, pour leurs remarques et suggestions, parfois pour leur lecture attentive du document.

Table des Matières

Les chapitres et sections marqués du signe « ▲ » sont inachevés, en travaux.

Partie I – Cadre — Outils théoriques — Concepts algorithmiques

1 Introduction . . . 3

1.1 Définition d'un problème d'optimisation . . . 3
1.1.1 Un cadre très général . . . 3
1.1.2 Vocabulaire . . . 4
1.1.3 Restrictions . . . 6
1.2 Existence de solution . . . 7
1.3 Problèmes d'optimisation équivalents . . . 10
Notes . . . 15
Exercices . . . 17

2 Ensembles convexes . . . 19

2.1 Définition et premières propriétés . . . 21
2.2 Aspects géométriques . . . 23
2.2.1 Enveloppe affine . . . 23
2.2.2 Enveloppe convexe . . . 24
2.2.3 Enveloppe conique . . . 27
2.2.4 Cône asymptotique . . . 28
2.2.5 Faces et points extrêmes . . . 30
2.3 Aspects topologiques . . . 31
2.4 Polyèdre convexe . . . 37
2.4.1 Représentations primale et duale . . . 37
2.4.2 Image linéaire . . . 38
2.4.3 Optimisation linéaire . . . 40
2.4.4 Faces et sommets . . . 41
2.4.5 Équivalence des représentations . . . 43
2.5 Opérations . . . 45
2.5.1 Projection . . . 45
2.5.2 Cône normal et hyperplan d'appui . . . 48
2.5.3 Séparation . . . 51
2.5.4 Enveloppe convexe fermée . . . 54
2.5.5 Cône dual . . . 55

2.5.6 Cône tangent . . . 62
Notes . . . 65
Exercices . . . 65

3 Fonctions convexes . . . 73

3.1 Définition . . . 73
3.2 Exemples . . . 77
3.2.1 Indicatrice . . . 77
3.2.2 Fonction affine et minorante affine . . . 77
3.2.3 Fonction convexe polyédrique . . . 79
3.2.4 Fonction sous-linéaire . . . 80
3.2.5 Fonction d'appui . . . 81
3.3 Régularité . . . 83
3.3.1 Continuité lipschitzienne . . . 83
3.3.2 Différentiabilité . . . 84
3.3.3 Reconnaître une fonction convexe par ses dérivées . . . 89
3.3.4 Fonction asymptotique . . . 97
3.4 Opérations . . . 101
3.4.1 Composition . . . 101
3.4.2 Enveloppes supérieure et inférieure . . . 103
3.4.3 Fonction marginale . . . 104
3.4.4 Inf-convolution . . . 105
3.4.5 Inf-image sous une application linéaire . . . 106
3.4.6 Adhérence . . . 108
3.5 Conjugaison . . . 110
3.5.1 Conjuguée . . . 110
3.5.2 Biconjuguée . . . 112
3.5.3 Règles de calcul . . . 115
3.6 Sous-différentiabilité . . . 122
3.6.1 Définitions . . . 123
3.6.2 Quelques propriétés . . . 128
3.6.3 Règles de calcul . . . 134
3.7 Proximalité . . . 139
3.7.1 Opérateur proximal . . . 139
3.7.2 Régularisée de Moreau-Yosida . . . 142
3.8 Point-selle et convexité-concavité . . . 144
3.8.1 Point-selle . . . 144
Notes . . . 144
Exercices . . . 145

4 Conditions d'optimalité . . . 157

4.1 Une condition nécessaire d'optimalité géométrique . . . 158
4.1.1 Cônes tangent et normal . . . 159
4.1.2 Condition nécessaire de Peano-Kantorovitch . . . 162
4.1.3 Problème avec convexité . . . 163
4.2 Problème sans contrainte . . . 163
4.2.1 Condition de Fermat . . . 164

4.2.2 Conditions d'optimalité du second ordre . . . 164
4.3 Problème avec contraintes d'égalité . . . 165
4.3.1 Conditions de Lagrange . . . 167
4.3.2 Conditions d'optimalité du second ordre . . . 173
4.3.3 Calcul pratique des solutions de (P_E) . . . 176
4.4 Problème avec contraintes d'égalité et d'inégalité . . . 177
4.4.1 Conditions de Karush, Kuhn et Tucker . . . 179
4.4.2 Qualification des contraintes . . . 186
4.4.3 Ensemble des multiplicateurs optimaux . . . 195
4.4.4 Conditions d'optimalité du second ordre . . . 196
4.4.5 Calcul pratique des solutions de (P_{EI}) . . . 205
4.5 Problème avec contraintes générales . . . 206
4.5.1 Conditions d'optimalité du premier ordre . . . 208
4.5.2 Condition de Robinson . . . 212
4.5.3 Qualification de Robinson . . . 224
4.6 Analyse de sensibilité . . . 228
4.6.1 Interprétation marginaliste des multiplicateurs optimaux . . . 229
4.6.2 Construction de chemins réguliers dans l'ensemble perturbé . . . 234
4.6.3 Continuité directionnelle de la fonction valeur . . . 237
4.6.4 Étude des sous-suites convergentes de solutions . . . 238
Notes . . . 241
Exercices . . . 242

5 Prolégomènes à l'algorithmique . . . 249

5.1 Vitesse de convergence des suites . . . 250
5.1.1 Vitesses de convergence en quotient . . . 250
5.1.2 Vitesses de convergence en racine ▲ . . . 258
5.2 Notions de complexité ▲ . . . 259
5.2.1 Famille de problèmes, machine de Turing . . . 259
5.2.2 Classes P et NP . . . 261
5.2.3 Problèmes NP-complets et NP-ardus . . . 262
5.3 Conditionnement d'un problème d'optimisation ▲ . . . 264
5.3.1 Notions de conditionnement . . . 264
5.4 Calcul de dérivées par état adjoint . . . 265
5.4.1 Position du problème . . . 266
5.4.2 Calcul de gradients . . . 268
5.4.3 Exemples . . . 271
5.4.4 Calcul de produits hessienne-vecteur . . . 275
5.5 Différentiation automatique . . . 276
5.5.1 Modèle de programme . . . 277
5.5.2 Différentiation en mode direct . . . 278
5.5.3 Différentiation en mode inverse . . . 281
5.6 Développement de codes d'optimisation . . . 286
5.6.1 Communication directe et inverse ▲ . . . 286
5.6.2 Profils de performance . . . 288
5.7 Pourquoi étudier l'optimisation numérique ? . . . 290
Notes . . . 291

Exercices 291

Partie II – Méthodes de l'optimisation sans contrainte

6 Méthodes à directions de descente 297

6.1 Principes généraux 298
6.2 Exemples de méthodes à directions de descente 301
6.2.1 Algorithme du gradient (ou de la plus profonde descente) 301
6.2.2 Algorithme du gradient conjugué 302
6.2.3 Algorithme de Newton 302
6.2.4 Algorithmes de quasi-Newton 303
6.2.5 Algorithme de Gauss-Newton 303
6.3 La recherche linéaire 304
6.3.1 Vue d'ensemble 304
6.3.2 Recherches linéaires « exactes » 305
6.3.3 Règles d'Armijo et de Goldstein 306
6.3.4 Règle de Wolfe 309
6.3.5 Mise en œuvre 311
6.4 Convergence des méthodes à directions de descente 313
6.4.1 Condition de Zoutendijk 313
6.4.2 Suites minimisantes spéciales 319
6.5 Propriétés asymptotiques 320
6.5.1 Admissibilité asymptotique du pas unité 320
6.5.2 Conditions de convergence superlinéaire ▲ 321
Notes 322
Exercices 323

7 Algorithmes du premier ordre 325

7.1 Algorithme du gradient 325
7.1.1 Définition 326
7.2 Algorithme proximal 327
7.2.1 Définition 327
7.2.2 Convergence 329
7.3 Méthode de Gauss-Seidel 333
7.3.1 En algèbre linéaire 333
7.3.2 Pour les systèmes non linéaires 335
7.3.3 En optimisation 336
Notes 339
Exercices 340

8 Optimisation quadratique 341

8.1 Sous-espaces de Krylov 342
8.2 Algorithme du gradient conjugué 346
8.2.1 Notion de directions conjuguées 347
8.2.2 Algorithme des directions conjuguées 347
8.2.3 Algorithme du gradient conjugué 349

8.2.4 Propriétés de l'algorithme du gradient conjugué 350
8.2.5 Mise en œuvre de la méthode du gradient conjugué 353
8.2.6 Méthode du gradient conjugué non linéaire 356
8.3 Algorithme du résidu minimal généralisé (GMRES) 359
8.3.1 Principe général 360
8.3.2 Construction d'une base orthonormale de K_{k+1} 361
8.3.3 Calcul de l'itéré x_k 362
8.3.4 L'algorithme GMRES 363
8.3.5 L'algorithme GMRES/QR 363
8.3.6 Algorithme GMRES avec redémarrage 365
Notes 365
Exercices 366

9 Méthodes à régions de confiance 369

9.1 L'algorithme de référence 369
9.1.1 Présentation de l'algorithme 369
9.1.2 Préconditionnement des méthodes à régions de confiance 373
9.2 Exemples de modèles quadratiques 374
9.2.1 Modèle linéaire 374
9.2.2 Modèle newtonien 375
9.2.3 Modèle gauss-newtonien 375
9.3 Analyse du sous-problème quadratique 375
9.3.1 Conditions d'optimalité 376
9.3.2 Le chemin de recherche 378
9.3.3 Discussion des solutions 382
9.4 Résolution numérique du sous-problème quadratique 384
9.4.1 Méthodes fondées sur le point de Cauchy 385
9.4.2 Méthodes de résolution fines 389
9.5 Résultats de convergence 392
9.5.1 Convergence avec modèle du premier ordre 392
9.5.2 Convergence avec modèle du second ordre 396
9.6 Comparaison avec la recherche linéaire 401
Notes 401
Exercices 402

10 Algorithmes de Newton 403

10.1 Méthodes locales 404
10.1.1 Systèmes d'équations 404
10.1.2 Optimisation 408
10.1.3 Défauts et remèdes 409
10.2 Méthodes inexactes ▲ 411
10.2.1 Systèmes d'équations 411
10.3 Globalisation de la convergence 412
10.3.1 Recherche linéaire 413
10.3.2 Régions de confiance ▲ 426
10.3.3 Autres méthodes 426
Notes 428

Exercices 431

11 Algorithmes de quasi-Newton 433

11.1 Système d'équations 435
11.1.1 Formules de mise à jour 435
11.1.2 Convergence linéaire locale 435
11.2 Optimisation 435
11.2.1 Formules de mise à jour 435
11.2.2 L'algorithme de BFGS 442
11.2.3 Propriétés de l'algorithme de BFGS 443
11.2.4 Mise en œuvre de l'algorithme de BFGS 448
11.2.5 L'algorithme ℓ-BFGS 450
Notes ▲ 453
Exercices 454

Partie III – Méthodes de l'optimisation avec contraintes

12 Projection et activation 459

12.1 Méthode de projection 460
12.1.1 Méthode du gradient projecté 461
12.2 Méthodes différentiation automatique ▲ 465
12.2.1 Motivation et schéma des méthodes 465
Notes 470
Exercices 470

13 Pénalisation 473

13.1 Vue d'ensemble 473
13.2 Pénalisation extérieure 477
13.2.1 Définition et exemples 477
13.2.2 Propriétés 479
13.2.3 Schéma algorithmique 482
13.2.4 Pénalisation quadratique 484
13.3 Pénalisation intérieure ▲ 486
13.4 Le lagrangien augmenté 488
13.4.1 Conditions d'exactitude du lagrangien 490
13.4.2 Le lagrangien augmenté de (P_E) 490
13.4.3 Le lagrangien augmenté de (P_{EI}) 492
13.4.4 Méthode du lagrangien augmenté 496
13.4.5 Méthode du lagrangien augmenté à directions alternées ▲ 499
13.5 Pénalisation exacte non différentiable ▲ 500
Notes 504
Exercices 505

14 Dualité 507

14.1 Dualité min-max 510
14.1.1 Introduction d'un problème dual 510
14.1.2 Liens entre problèmes primal et dual, point-selle 512
14.1.3 Existence de point-selle ▲ 515
14.1.4 Schéma algorithmique 517
14.2 Dualité par perturbation 518
14.2.1 Perturbation du problème primal 518
14.2.2 Le problème dual 519
14.2.3 Le lagrangien associé aux perturbations 520
14.2.4 Perturbation du problème dual 523
14.3 Dualité de Fenchel 524
14.4 Dualité non convexe de Toland 526
14.4.1 Perturbation du problème primal 526
14.4.2 Le problème dual 527
14.4.3 Existence de solutions primale et duale 529
14.4.4 Suites minimisantes primales et duales 530
14.5 Dualisation lagrangienne 531
14.5.1 Dualisation de contraintes d'égalité et d'inégalité 531
14.5.2 Dualisation de contraintes générales 537
14.6 Dualisation lagrangienne augmentée 539
14.6.1 Dualisation de contraintes d'égalité et d'inégalité 539
14.6.2 Dualisation de contraintes générales 544
14.7 Méthodes numériques ▲ 545
14.7.1 Minimisation de la fonction duale 545
14.7.2 Minimisation de la fonction duale régularisée 549
14.7.3 L'algorithme d'Arrow-Hurwicz 550
Notes 551
Exercices 552

15 Optimisation quadratique successive 559

15.1 L'algorithme OQS et sa convergence locale 560
15.1.1 L'algorithme OQS 560
15.1.2 Convergence locale 561
15.2 L'algorithme local 565
15.2.1 Un algorithme non convergent 565
15.2.2 L'algorithme OQS 566
15.2.3 Convergence locale ▲ 569
15.3 Globalisation de la convergence par recherche linéaire 572
15.3.1 Fonction de mérite 573
15.3.2 Condition de décroissance de la fonction de mérite 573
15.4 Versions quasi-newtoniennes ▲ 574
15.5 Le diable se cache dans les détails ▲ 576
15.5.1 Incompatibilité des contraintes 576
15.5.2 Troncature du pas : l'effet Maratos 577
15.5.3 Problèmes de commande optimale 577
Notes 578

Exercices 579

16 Méthodes de points intérieurs 581

16.1 Méthodes de points intérieurs en optimisation convexe 581
16.1.1 Vue d'ensemble 581
16.1.2 Fonction autoconcordante 582
16.1.3 Barrière autoconcordante 594
16.1.4 Trois approches algorithmiques ▲ 601
16.2 Méthodes de points intérieurs en optimisation non linéaire ▲ 608
16.2.1 Un algorithme de PI primal-dual 608
16.2.2 Analyse asymptotique 611
Notes 615
Exercices 616

Partie IV – Étude de problèmes particuliers

17 Optimisation linéaire : théorie et algorithme du simplexe 619

17.1 Introduction 619
17.1.1 Le problème à résoudre 619
17.1.2 Formulations canoniques 620
17.1.3 Exemples : problèmes d'optimisation dans des réseaux 621
17.2 Étude du problème 622
17.2.1 Structure de l'ensemble admissible 622
17.2.2 Existence de solution et conditions d'optimalité 624
17.3 Dualité 629
17.3.1 Dualité en optimisation linéaire 629
17.3.2 Une relation entre le primal et le dual 632
17.4 Algorithmes du simplexe 633
17.4.1 Algorithme du simplexe primal 634
17.4.2 Règles d'anti-cyclage ▲ 639
17.4.3 Énoncé de l'algorithme 639
17.4.4 Démarrage de l'algorithme du simplexe 640
17.4.5 Complexité 642
Notes 643
Exercices 643

18 Optimisation linéaire : algorithmes de points intérieurs 647

18.1 Le chemin central primal-dual 649
18.2 Éléments constitutifs des algorithmes 656
18.3 Algorithmes avec itérés admissibles 660
18.3.1 Préliminaires 660
18.3.2 Algorithme des petits déplacements 662
18.3.3 Algorithme des grands déplacements 665
18.3.4 Un algorithme prédicteur-correcteur 667
18.4 Un algorithme sans admissibilité forcée 670

18.5 Mise en œuvre . . . 677
18.5.1 Calcul du déplacement de Newton . . . 677
18.5.2 Logiciels . . . 678
Notes . . . 678
Exercices . . . 679

19 Systèmes non déterminés . . . 681

19.1 Moindres-carrés linéaire . . . 681
19.1.1 Définition du problème . . . 681
19.1.2 Exemple : régression linéaire . . . 682
19.1.3 L'ensemble des solutions . . . 683
19.1.4 Résolution numérique . . . 685
19.1.5 Estimation numérique de la qualité de la solution ▲ . . . 688
19.2 Moindres-carrés polyédrique . . . 688
19.2.1 Définition du problème . . . 688
19.2.2 Existence et unicité de solution . . . 690
19.2.3 Condition d'optimalité . . . 692
19.3 Moindres-carrés non linéaire . . . 693
19.3.1 Définition du problème . . . 693
19.3.2 Algorithme de Gauss-Newton . . . 694
19.3.3 Algorithme de Levenberg-Morrison-Marquardt . . . 697
19.4 Recherche de solution parcimonieuse ▲ . . . 706
19.4.1 Motivation . . . 706
19.4.2 Poursuite de base . . . 706
19.4.3 Problème LASSO ▲ . . . 711
Notes . . . 713
Exercices . . . 714

20 Optimisation semi-définie positive . . . 717

20.1 Définition du problème d'optimisation SDP . . . 717
20.1.1 Les problèmes primal et dual . . . 717
20.1.2 Réalisabilités primale et duale . . . 722
20.2 Exemples de modélisation SDP . . . 725
20.2.1 Optimisation linéaire . . . 726
20.2.2 Optimisation quadratique convexe . . . 726
20.2.3 Minimisation de la valeur propre maximale . . . 727
20.2.4 Minimisation de la norme matricielle ℓ_2 . . . 728
20.2.5 Minimisation globale de polynômes . . . 728
20.2.6 Relaxation SDP de problèmes non convexes . . . 731
20.3 Existence de solution et CNS d'optimalité . . . 733
20.4 Algorithmes de points intérieurs . . . 736
20.4.1 Le chemin central . . . 736
20.4.2 Remarques générales sur les algorithmes . . . 739
Notes . . . 743
Exercices . . . 743

Annexes

A Analyse ▲ 747

A.1 Topologie 747
A.2 Compacité 748
A.3 Fonction injective, surjective, continue 748
A.4 Espace métrique 750
A.5 Espace normé 751
A.6 Espace de Hilbert 754
A.7 Multifonction 757
Notes 758
Exercices 758

B Algèbre linéaire ▲ 761

B.1 Espaces vectoriels 761
B.1.1 Orthogonalité 762
B.2 Applications linéaires 763
B.3 Analyse spectrale 767
B.4 Matrices complexes 770
B.4.1 Nombre, vecteur et matrice complexes 770
B.4.2 Matrice hermitienne 771
B.5 Factorisations 772
B.5.1 Factorisation QR 772
B.5.2 Factorisation gaussienne (ou LU) 773
B.5.3 Factorisation de Cholesky 776
B.5.4 Factorisation en valeurs singulières (SVD) 782
Notes 784
Exercices 785

C Calcul différentiel ▲ 789

C.1 Dérivées directionnelle et au sens de Gâteaux 789
C.2 Dérivée au sens de Fréchet 793
C.2.1 Dérivée première 793
C.2.2 Dérivée seconde 802
Notes 807
Exercices 807

D Solutions des exercices ▲ 809

D.1 Indications 810
D.2 Solutions 818

Lexiques 923

Notations 929

Bibliographie 933

Index 963

Il reste toujours au fond de moi une part de mon expérience
qui n'est pas transmissible.
Jean-Christophe Lafaille et Benoît Heimermann [371].

Partie I

Cadre
Outils théoriques
Concepts algorithmiques

1 Introduction

Quand on me contredit, on éveille mon attention, mais non ma colère : je m'avance vers celui qui me contredit, qui m'instruit. La cause de la vérité devrait être la cause commune de l'un et de l'autre.

Montaigne (1572-1592). Essais, III, 8 [426].

*Selon la méthode cartésienne, pour rendre raison d'un phénomène complexe, il faut le décomposer rationnellement en éléments plus simples, par là même plus faciles à appréhender, et ainsi de suite, jusqu'à ce qu'on parvienne aux éléments fondamentaux : c'est l'*analyse. *Puis, partant de ces éléments, on remonte l'ensemble de la machine, en prenant soin de ne faire aucun « saut », afin d'être toujours à même d'expliquer l'opération qui suit par celles qui précèdent : c'est la* synthèse.

Jean-Marc Mandosio (2010). Présentation de la *Grammaire Générale et Raisonnée* d'Antoine Arnauld et Claude Lancelot, 1660 [404].

Ideally, mathematics should be seen as a thought process, rather than just as a mass of facts to be learned and remembered, which is so often the common view. [...] *It's the excitement of discovering new properties and relationships—ones having the intellectual beauty that only mathematics seems able to bring—that keeps me going. I never get tired of it. This process builds its own momentum. New flashes of insight stimulate curiosity more and more.*

R.T. Rockafellar sur le site Wikimization.

1.1 Définition d'un problème d'optimisation

1.1.1 Un cadre très général

Dans cet ouvrage, nous nous intéresserons à l'étude et à la résolution des problèmes qui s'énoncent de la manière suivante :

« Trouver $x_* \in X$ tel que, pour tout $x \in X$, on ait $f(x_*) \leqslant f(x)$ ».

Dans cet énoncé, X est un ensemble et f est une application définie sur X à valeurs dans la *droite achevée* $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, +\infty\}$. Il s'agit donc de trouver un point x_* de l'ensemble X qui donne à f sa plus petite valeur sur X. C'est ce que l'on appelle un *problème d'optimisation*. On notera également ce problème comme suit :

$$(P_X) \quad \left\{ \begin{array}{l} \inf\ f(x) \\ x \in X \end{array} \right.$$

ou encore

$$\inf_{x \in X}\ f(x) \qquad \text{ou} \qquad \inf\{f(x) : x \in X\}.$$

L'*optimisation* est la discipline qui étudie ces problèmes. Elle traite des questions d'existence et d'unicité de solution de ce problème, de l'établissement de ses conditions d'optimalité, de sa dualisation, *etc.* Par ailleurs, une grande partie de cette discipline, et de cet ouvrage, est consacrée aux méthodes numériques qui ont été conçues pour résoudre les problèmes d'optimisation.

1.1.2 Vocabulaire

L'optimisation a son propre vocabulaire, dont nous allons dévoiler maintenant les premières bribes. On dit que X est l'*ensemble admissible* du problème et un point de X est dit *admissible*. Ces deux notions sont surtout pertinentes lorsque X est une partie d'un autre ensemble $\mathbb{E}$ (souvent un espace vectoriel réel), si bien que tout point de $\mathbb{E}$ n'est pas nécessairement admissible. Lorsque X est non vide, on dit que le problème est *réalisable*. La fonction f est appelée *critère*, *fonction coût* ou *fonction objectif* du problème. On appelle *valeur optimale* de (P_X) la borne inférieure

$$\mathrm{val}(P_X) := \inf\{f(x) : x \in X\}$$

des valeurs prises par f sur X. On dit que le problème (P_X) est *borné* si sa valeur optimale ne vaut pas $-\infty$. Dans le cas contraire, on dit qu'il n'est *pas borné* ou qu'il est *non borné*. On a alors

$$\inf_{x \in X} f(x) = -\infty,$$

ce qui se produit s'il existe une suite $\{x_k\} \subseteq X$ (éventuellement *stationnaire*, c'est-à-dire avec tous les x_k égaux pour k grand) telle que $f(x_k) \to -\infty$. Par ailleurs

$$\inf_{x \in X} f(x) = +\infty,$$

si $f(x) = +\infty$ pour tout $x \in X$. Il en sera donc ainsi si $X = \varnothing$ (l'ensemble vide). On a donc

$$\inf_{x \in \varnothing} f(x) = +\infty \qquad \text{et} \qquad \sup_{x \in \varnothing} f(x) = -\infty. \tag{1.1}$$

Ayant défini un problème d'optimisation, il faut maintenant préciser ce qu'en est une solution. On dit qu'un point x_* est une *solution* ou un *minimum* ou encore un *minimiseur* du problème (P_X) si

$$x_* \in X \qquad \text{et} \qquad \forall x \in X,\ f(x_*) \leqslant f(x). \tag{1.2}$$

Il faut donc *deux* conditions : que x_* soit admissible et qu'il donne à f une valeur qui n'excède pas (strictement) celle donnée à f par tout autre point admissible. On dit aussi qu'un tel x_* est solution/minimum/minimiseur *global* de (P_X) pour distinguer cette notion des autres qui vont suivre. On note indifféremment par

$$\mathrm{Sol}(P_X) \qquad \text{ou} \qquad \arg\min_{x\in X} f(x)$$

l'ensemble des solutions de (P_X).

Remarque 1.1 Lorsque le problème (P_X) a une solution, on écrit

$$\min_{x\in X} f(x),$$

donc avec l'opérateur 'min' plutôt que 'inf'.

Lorsque X est inclus dans un espace topologique $\mathbb{E}$, on peut définir la notion plus faible de *minimum local* de (P_X). Il s'agit d'un point x_* tel qu'il existe un voisinage V de x_* dans $\mathbb{E}$ tel que

$$x_* \in X \qquad \text{et} \qquad \forall x \in X \cap V,\ f(x_*) \leqslant f(x).$$

Une solution globale est aussi une solution locale (on prend $V = \mathbb{E}$). On parlera de solution ou de minimum global ou local *strict* si $f(x_*) < f(x)$, pour tout $x \in X \setminus \{x_*\}$ ou pour tout $x \in (X \cap V) \setminus \{x_*\}$, respectivement.

Il est important de pouvoir prendre en compte des problèmes d'optimisation dans lesquels le critère f peut prendre des valeurs infinies, $-\infty$ ou $+\infty$, parce que ces fonctions sont parfois générées par des procédures qui ne leur assurent pas nécessairement que des valeurs finies (c'est le cas de la dualité au chapitre 14, par exemple). Le *domaine effectif* (ou simplement *domaine*) de f est l'ensemble des points de X où elle ne prend pas la valeur $+\infty$ (mais elle peut y prendre la valeur $-\infty$, pour une raison qui sera vue au chapitre 3 sur les fonctions convexes). On le note

$$\mathrm{dom}\, f := \{x \in X : f(x) < +\infty\}.$$

Il est clair que les problèmes

$$\inf_{x\in X} f(x) \qquad \text{et} \qquad \inf_{x\in \mathrm{dom}\, f} f(x) \tag{1.3}$$

ont les mêmes valeurs optimales et les mêmes solutions. Si $\mathrm{dom}\, f = \varnothing$, les valeurs optimales sont $+\infty$ (à gauche parce que $f(x) = +\infty$ pour tout $x \in X$, à droite parce que $\mathrm{dom}\, f = \varnothing$ et que l'on a adopté la convention (1.1) qui s'avère donc essentielle ici) ; si $\mathrm{dom}\, f \neq \varnothing$, alors on ne modifie pas le problème de gauche en excluant de son ensemble admissible les points où f prend la valeur $+\infty$, comme on le fait à droite. L'équivalence entre les problèmes de (1.3) sera souvent utilisée.

1.1.3 Restrictions

> *In our opinion, the main fact, which should be known to any person dealing with optimization methods, is that in general* optimization problems are unsolvable. *This statement, which is usually missing in standard optimization courses, is very important for an understanding of optimization theory and its development in the past and in the future.*
>
> Y. Nesterov [452].

Présenté comme ci-dessus, le problème (P_X) peut être très général, mais les méthodes théoriques et algorithmiques étudiées dans ce manuel ne seront efficaces que sur un petit sous-ensemble de ces problèmes, qui contient toutefois beaucoup de ceux qui se posent en pratique, mais en écarte aussi beaucoup d'autres. Les restrictions présentes dans (P_X) ou que nous nous imposerons, faute de pouvoir tout faire, sont les suivantes.

- L'ensemble d'arrivée de f est un espace vectoriel de dimension un, ce qui veut dire que l'on ne cherche à minimiser qu'un seul critère. Lorsque l'espace d'arrivée est de dimension supérieure, on parle d'*optimisation multicritère*. Nous n'aborderons pas ces problèmes dans cet ouvrage, bien qu'il soit fait allusion à la notion d'optimalité au sens de Pareto à l'exercice 3.14.
- Même si cela n'apparaît pas dans la formulation générale de (P_X), on éliminera également de nos préoccupations une autre classe importante de problèmes, ceux pour lesquels les variables sont entières : X est une partie de $\mathbb{N}^n$. Ces problèmes d'*optimisation en nombres entiers* ou *combinatoire* sont d'une autre nature que ceux qui peuvent être résolus par les algorithmes que nous étudierons (pour une introduction à ces problèmes, voir par exemple [476]). Ces derniers auront besoin d'une certaine régularité des données. Il faudra une topologie sur X et un critère f au moins continu, si possible différentiable (éventuellement dans un sens généralisé). Cette restriction à des classes de fonctions particulières permet aussi d'échapper au verdict étonnant et fâcheux, selon lequel tous les algorithmes sont équivalents, si on moyenne leur performance sur l'ensemble des fonctions [627 ; 1997], affirmation qui doit d'ailleurs être nuancée [26 ; 2007].
- Dans le même ordre d'idée, les algorithmes que nous étudierons ne seront efficaces que pour trouver des *minima locaux*. Trouver un minimum global d'une fonction qui a beaucoup de minima locaux s'apparente en effet souvent à un problème combinatoire (par exemple lorsque l'ensemble des minima locaux est discret), lequel a été éliminé de nos préoccupations ci-dessus. Les algorithmes locaux, c'est-à-dire ceux trouvant des minima locaux, sont toutefois utiles, car ils sont souvent très efficaces et sont d'ailleurs parfois utilisés dans certaine méthode d'optimisation globale. On peut en fait souvent montrer que plus les fonctions à minimiser sont régulières et plus rapide sera la convergence *locale* des itérés générés par un algorithme sachant utiliser cette régularité de manière idoine.
- Quand il sera question d'algorithmes, nous travaillerons toujours *en dimension finie*, c'est-à-dire que l'ensemble admissible X sera une partie d'un espace vectoriel de dimension finie sur $\mathbb{R}$ (par exemple, l'espace vectoriel $\mathbb{R}^n$ des n-uplets

$(x_1, \ldots, x_n)$ ou l'espace vectoriel $\mathcal{S}^n$ des matrices réelles d'ordre n symétriques). Les composantes de x peuvent alors être vues comme des paramètres servant à rendre optimal un système. En pratique, les problèmes de dimension infinie se rencontrent fréquemment, par exemple lorsqu'il s'agit de déterminer une trajectoire optimale ou une forme optimale. Il s'agit alors de déterminer une fonction plutôt qu'un vecteur. Lorsqu'on veut résoudre ces problèmes, il est nécessaire de passer par une *phase de discrétisation* qui, en utilisant une technique adéquate, construit un problème approché en dimension finie qui pourra être résolu sur ordinateur par les algorithmes vus dans cet ouvrage.

1.2 Existence de solution

Si l'ensemble admissible X de (P_X) est non vide, que ce problème soit borné ou non, il existe ce que l'on appelle une *suite minimisante*. C'est une suite $\{x_k\}$ vérifiant les propriétés suivantes :

$$x_k \in X \text{ et } f(x_k) \text{ converge vers } \mathrm{val}(P_X).$$

Il suffit en effet de se donner une suite de réels $\varepsilon_k \downarrow 0$ (qui converge vers zéro par des valeurs strictement positives) et d'observer que, par définition de la borne inférieure, on peut trouver un point $x_k \in X$ tel que $\mathrm{val}(P_X) \leqslant f(x_k) \leqslant \mathrm{val}(P_X) + \varepsilon_k$. On peut même supposer que la suite $\{f(x_k)\}$ est décroissante et, lorsque f est $\mathcal{C}^{1,1}$, que $f'(x_k) \to 0$ (voir le lemme 6.14).

Il faut se garder de confondre la notion d'existence de solution, existence de x_* vérifiant (1.2), et celle de l'existence d'une borne inférieure $\mathrm{val}(P_X)$ finie. Par exemple, si f est la fonction définie sur $X = \mathbb{R}$ par $f(x) = e^x$, le problème (P_X) n'a pas de solution alors que sa borne inférieure est nulle.

La suite de cette section est formée de variations autour du théorème de Weierstrass ci-dessous. On rappelle que $f : X \to \overline{\mathbb{R}}$ est *semi-continue inférieurement*, s.c.i. en abrégé, si pour tout $x \in X$ et toute suite $\{x_k\}$ convergeant vers x, on a $f(x) \leqslant \liminf_k f(x_k)$ (voir la section A.3).

Théorème 1.2 (Weierstrass, existence d'un minimum) *Si X est un compact non vide et si $f : X \to \overline{\mathbb{R}}$ est semi-continue inférieurement, alors (P_X) a au moins une solution.*

Démonstration. Soit $\{x_k\}$ une suite minimisante : $x_k \in X$ et $f(x_k) \to \mathrm{val}(P_X)$. Comme X est compact, on peut en extraire une sous-suite, encore notée $\{x_k\}$, convergente, disons vers $\bar{x} \in X$. Par le caractère s.c.i. de f, on a $f(\bar{x}) \leqslant \liminf f(x_k) = \mathrm{val}(P_X)$. Dès lors $\bar{x}$ est solution de (P_X). □

Comme $\inf f = -\sup(-f)$ (identité rappelée à la proposition 1.5 ci-dessous), on déduit de ce résultat qu'une fonction semi-continue supérieurement atteint sa borne supérieure sur un compact. Une fonction continue étant à la fois semi-continue inférieurement et semi-continue supérieurement, elle atteint ses bornes inférieure et supérieure sur un compact.

En dimension finie un ensemble est compact s'il est fermé et borné. La propriété de compacité est plus difficile à obtenir en dimension infinie, mais le théorème 1.2 est si important que l'on a été amené à affaiblir la topologie des espaces normés de manière à avoir plus de compacts, tout en gardant suffisamment de fonctions s.c.i. (voir [91 ; 1983]).

Supposons que X soit donné par des contraintes fonctionnelles d'égalité et d'inégalité au sens large, c'est-à-dire qu'il est de la forme :

$$X = \{x \in \mathbb{E} : c_E(x) = 0,\ c_I(x) \leqslant 0\},$$

où $\mathbb{E}$ est une espace topologique et $c_E : X \to \mathbb{R}^{m_E}$ et $c_I : X \to \mathbb{R}^{m_I}$ sont deux applications (m_E et m_I sont deux entiers). L'inégalité $c_I(x) \leqslant 0$ signifie que toutes les composantes de $c_I(x) \in \mathbb{R}^{m_I}$ doivent être négatives. Alors, X est fermé si c_E et c_I sont continues. On peut en effet écrire

$$X = c_E^{-1}(\{0\}) \bigcap c_I^{-1}(\mathbb{R}_-^{m_I}),$$

où $c_E^{-1}(\{0\})$ et $c_I^{-1}(\mathbb{R}_-^{m_I})$ sont fermés comme images réciproques par des applications continues des fermées $\{0\}$ et $\mathbb{R}_-^{m_I}$ de $\mathbb{R}^{m_E}$ et $\mathbb{R}^{m_I}$ respectivement.

L'hypothèse « X compact » est souvent restrictive, en particulier, elle n'a pas lieu pour les problèmes sans contrainte. Dans le corollaire suivant, on montre que, *en dimension finie* (le résultat est faux en dimension infinie), on peut remplacer dans l'énoncé du théorème de Weierstrass, l'hypothèse « X compact » par « X fermé et f coercive ».

Définition 1.3 (fonction coercive) Soit $\mathbb{E}$ un espace vectoriel normé. Une fonction $f : \mathbb{E} \to \overline{\mathbb{R}}$ est dite *coercive* sur une partie non bornée X de $\mathbb{E}$ si

$$\lim_{\substack{x \in X \\ \|x\| \to \infty}} f(x) = +\infty \tag{1.4}$$

ou de manière plus précise

$$\forall\, \nu \in \mathbb{R}, \quad \exists\, \rho \geqslant 0 : \quad (x \in X \text{ et } \|x\| \geqslant \rho) \quad \Longrightarrow \quad f(x) \geqslant \nu. \tag{1.5}$$

Si l'on ne spécifie pas la partie X, il est sous-entendu que $X = \mathbb{E}$. □

Insistons bien sur le fait que, lorsque $X \neq \mathbb{E}$, la coercivité sur X est plus faible que le fait de demander que f tende vers l'infini à l'infini (c'est-à-dire pour $\|x\| \to \infty$). Ici on demande que cela ne soit vrai que pour des $x \in X$. Par ailleurs, l'exercice 1.3 propose une expression équivalente de la coercivité, à savoir que les intersections avec X des ensembles de sous-niveau de f sont bornées :

$$\forall\, \nu \in \mathbb{R}, \qquad \{x \in X : f(x) \leqslant \nu\} \text{ est borné.}$$

L'expression de la coercivité d'une forme bilinéaire est examinée à l'exercice 1.4.

Corollaire 1.4 *Soient X est une partie fermée non vide d'un espace vectoriel de dimension finie et $f : X \to \overline{\mathbb{R}}$ est une fonction semi-continue inférieurement et coercive. Alors (P_X) a au moins une solution.*

Démonstration. Soient[1] $x_0 \in X$ (non vide) et

$$X_0 := \{x \in X : f(x) \leqslant f(x_0)\}.$$

Cet ensemble est non vide et compact (il est fermé parce que f est semi-continue inférieurement sur X, borné grâce à (1.4) et est une partie d'un espace vectoriel de dimension finie). Alors, d'après le théorème, le problème

$$\begin{cases} \min f(x) \\ x \in X_0, \end{cases}$$

a une solution x_*. Celle-ci est clairement solution de (P_X), puisque $\forall x \in X \setminus X_0$, on a $f(x) \geqslant f(x_0) \geqslant f(x_*)$. □

Les techniques de démonstration d'existence de solutions de problèmes sont nombreuses. Passons en revue celles qui concernent les problèmes d'optimisation et les systèmes d'équations non linéaires et qui seront abordées dans cet ouvrage.

- Les résultats d'existence de solutions de problèmes d'optimisation présentés ci-dessus sont de *nature topologique* et sont fondés sur le comportement de f à l'infini (pour clarifier ce point, il suffit de remplacer f par $f + \mathcal{I}_X$, où $\mathcal{I}_X$ est la fonction indicatrice de l'ensemble X ou de se rappeler l'hypothèse de coercivité utilisé dans le corollaire 1.4). Cette approche sera systématisée pour les problèmes convexes par l'utilisation de la fonction asymptotique (section 3.3.4), ce qui conduira au résultat d'existence de la proposition 3.29 (voir son point (ii)). Elle s'utilise aussi pour les problèmes non convexes [29], mais nous n'aborderons pas ce sujet ici. On peut rattacher les résultats d'existence de solution d'un problème d'optimisation linéaire (proposition 2.20) ou quadratique (théorème **??**) à cette théorie asymptotique, mais cela demande un peu de gymnastique intellectuelle.
- Une autre possibilité est l'*approche analytique*, fondée sur les conditions d'optimalité qui seront établies au chapitre 4 : si l'on peut montrer l'existence d'un point stationnaire (c'est-à-dire une solution des conditions d'optimalité) par une méthode appropriée (pour les problèmes quadratiques sans contrainte d'inégalité, il s'agit d'un simple système linéaire) et si le problème d'optimisation est convexe, alors ce point stationnaire en est une solution.
- Une troisième approche nous viendra de de l'*analyse convexe* (chapitres 2 et 3) et de la *dualité* (chapitre 14). Elle consiste à montrer que le problème d'optimisation

[1] Il faudrait écrire *Soit* au singulier [284 ; §§ 901(d) et 1045(c)], de la même manière que l'on écrit *vive les vacances* et pas *vivent les vacances* [284 ; §§ 901(e) et 1045(c)]. Nous continuerons pourtant d'écrire avec obstination *soit un objet* et *soient deux objets*, maintenant ainsi une tradition qui ne semble plus se perpétuer que chez les mathématiciens, mais qui a sa logique.

considéré est le dual d'un autre problème (section 14.1.1) dont la fonction valeur (définition 4.71) est sous différentiable en zéro (section 3.6). Cette technique sera par exemple utilisée pour établir le théorème 14.14.

- Le lemme de Farkas (proposition 2.45) et les théorèmes de l'alternatives (exercice 2.39) qui en découlent permettent d'assurer l'existence d'un point satisfaisant des contraintes affines ; voir la discution autour de (2.40). Lorsque des conditions de qualification sont satisfaites, ce lemme permet d'assurer l'existence de multiplicateurs optimaux (section 4.4), qui sont parfois solutions d'un problème d'optimisation dual (chapitre 14).
- Enfin pour montrer l'existence d'un zéro d'un système d'équations non linéaires, nous verrons un résultat (le théorème 10.3 de Kantorovitch) qui est apparenté aux théorèmes d'*existence de point fixe*.

1.3 Problèmes d'optimisation équivalents

Un problème d'optimisation peut se formuler de différentes manières. Certaines formulations permettent de mieux comprendre le problème, d'autres se prêtent à une résolution numérique plus efficace. Les codes d'optimisation tentent d'accepter des formulations les plus générales possibles de manière à pouvoir résoudre le plus grand nombre de problèmes avec l'algorithme implémenté. Par ailleurs, pour de multiples raisons (simplification, mise en évidence de la structure, convention, *etc*), l'étude des algorithmes se fait sur des formulations particulières qui doivent toutefois être suffisamment générales pour pouvoir représenter tous les problèmes d'une classe donnée. La question de savoir si deux problèmes sont équivalents ou si une formulation particulière est représentative d'une classe donnée de problèmes se pose donc souvent. Cette notion d'*équivalence entre problèmes* n'a pas de définition précise, mais on veut souvent dire que les solutions d'une formulation peuvent se déduire aisément des solutions de l'autre, si elles en ont, et que les valeurs optimales des deux formulations ont un lien bien défini entre elles. L'équivalence entre deux problèmes d'optimisation est parfois subtile (voir le chapitre 14 sur la dualité), mais on peut dès à présent donner quelques règles générales élémentaires.

Sachant qu'un problème accepte souvent plusieurs formulations, il est naturel (dans un manuel d'optimisation bien sûr) de se demander s'il n'en existe pas une meilleure que les autres. Cela dépend du critère que l'on se donne, de l'objectif que l'on se fixe. Si l'on s'intéresse à l'analyse du problème, les formulations les plus simples et faisant apparaître au mieux la structure seront préférables. Si l'on s'intéresse à la résolution numérique, il peut être utile de savoir que beaucoup de numériciens pensent qu'il existe une espèce de *loi de conservation des ennuis*, selon laquelle certaines difficultés essentielles ne peuvent pas être supprimées en changeant de formulation (pour autant que celle considérée au départ ne soit pas farfelue), comme la combinatoire, le mauvais conditionnement, la difficulté liée aux inégalités, *etc*. Nous y ferons souvent allusion (déjà dans l'exercice 1.7 de ce chapitre).

Cet ouvrage traite essentiellement de problèmes de minimisation, en particulier parce que les *problèmes de maximisation* leur sont équivalents, dans un sens que nous allons préciser. Ceci implique qu'il suffit d'étudier une seule des deux classes de problèmes. Le choix de la minimisation s'impose alors par le fait qu'il est naturel de

minimiser les fonctions convexes et que ces dernières sont bien étudiées en analyse convexe (il n'y a pas d'analyse concave, car la notion d'ensemble concave n'est pas un bon concept). Cet argument est repris dans le chapitre suivant et résumé par la chaîne d'implications (2.1).

Proposition 1.5 (maximisation) *Soient X un ensemble et $f : X \to \overline{\mathbb{R}}$ une application. Alors les problèmes*

$$\inf_{x\in X} f(x) \qquad \text{et} \qquad \sup_{x\in X} \big(-f(x)\big)$$

ont les mêmes solutions et des valeurs optimales qui s'opposent :

$$\inf_{x\in X} f(x) = -\sup_{x\in X} \big(-f(x)\big). \tag{1.6}$$

Démonstration. Quel que soit $x \in X$, on a $\inf f \leqslant f(x)$ et $-f(x) \leqslant \sup(-f)$, donc $-\inf f \geqslant -f(x)$ et $f(x) \geqslant -\sup(-f)$. On en déduit que $-\inf f \geqslant \sup(-f)$ et $\inf f \geqslant -\sup(-f)$, d'où l'égalité (1.6).

D'après ce qui précède, $\bar{x}$ est solution de $\inf f$ si, et seulement si, $f(\bar{x}) = \inf f$ ou $-f(\bar{x}) = \sup(-f)$; donc si, et seulement si, $\bar{x}$ est solution de $\sup(-f)$. □

On est souvent amener à transformer un problème d'optimisation par un changement de variables, celles que l'on cherche à optimiser. Quand est-ce qu'il y a équivalence entre les deux problèmes

$$(P_X) \quad \inf_{x\in X} f(x) \qquad \text{et} \qquad (P_Y) \quad \inf_{y\in Y} g(y)$$

où X et Y sont des ensembles et $f : X \in \overline{\mathbb{R}}$ et $g : Y \to \overline{\mathbb{R}}$ sont des fonctions. C'est une question très générale. La proposition suivante y donne une réponse lorsque x et y sont liés par un changement de variables, c'est-à-dire par une fonction $\varphi : X \to Y : x \mapsto y$ *surjective*.

Proposition 1.6 (changement de variables) *Dans le cadre défini ci-dessus, on suppose que $Y = \varphi(X)$ et $f = g \circ \varphi$. Alors,*

1) *les problèmes (P_X) et (P_Y) ont les mêmes valeurs optimales,*
2) *si $(x_*, y_*) \in X \times Y$ et $y_* = \varphi(x_*)$, alors x_* est solution de (P_X) si, et seulement si, y_* est solution de (P_Y).*

Démonstration. 1) Soit $\varepsilon > 0$. Alors on peut trouver $x_\varepsilon \in X$ tel que $f(x_\varepsilon) \leqslant \inf_{x\in X} f(x) + \varepsilon$. En utilisant $y_\varepsilon := \varphi(x_\varepsilon) \in Y$, on obtient

$$\inf_{y\in Y} g(y) \leqslant g(y_\varepsilon) = (g \circ \varphi)(x_\varepsilon) = f(x_\varepsilon) \leqslant \inf_{x\in X} f(x) + \varepsilon.$$

Comme $\varepsilon > 0$ est arbitraire, on a $\inf_{y\in Y} g(y) \leqslant \inf_{x\in X} f(x)$.

Inversement, toujours avec un $\varepsilon > 0$ arbitraire, on peut trouver $y_\varepsilon \in Y$ tel que $g(y_\varepsilon) \leqslant \inf_{y\in Y} g(y)+\varepsilon$. Comme $Y = \varphi(X)$, on peut trouver $x_\varepsilon \in X$ tel que $y_\varepsilon = \varphi(x_\varepsilon)$. Alors,

$$\inf_{x\in X} f(x) \leqslant f(x_\varepsilon) = (g\circ\varphi)(x_\varepsilon) = g(y_\varepsilon) \leqslant \inf_{y\in Y} g(y) + \varepsilon.$$

Comme $\varepsilon > 0$ est arbitraire, on a $\inf_{x\in X} f(x) \leqslant \inf_{y\in Y} g(y)$.

2) Soit $(x_*, y_*) \in X \times Y$ et $y_* = \varphi(x_*)$.

$\boxed{\Rightarrow}$ Si x_* est solution de (P_X), on a $g(y_*) = g(\varphi(x_*)) = f(x_*) = \inf_{x\in X} f(x) = \inf_{y\in Y} g(y)$, grâce au point 1. Donc y_* est solution de (P_Y).

$\boxed{\Leftarrow}$ Si y_* est solution de (P_Y), on a $f(x_*) = g(\varphi(x_*)) = g(y_*) = \inf_{y\in Y} g(y) = \inf_{x\in X} f(x)$, grâce au point 1. Donc x_* est solution de (P_X). □

Remarque 1.7 (changement variables en pratique) Souvent, la proposition 1.6 s'utilise lorsqu'on doit résoudre un problème de la forme

$$\inf_{x\in X} g(\varphi(x)),$$

où l'objectif est une composition de fonctions (les notations sont celles de la proposition). On introduit alors la variable $y := \varphi(x)$ et la proposition 1.6 nous dit dans quel sens ce problème est équivalent au problème

$$\inf_{y\in Y} g(y),$$

où $Y := \varphi(X)$. Il y a de nombreuses utilisation de cette technique dans cet ouvrage, en particulier pour résoudre les exercices 1.6 et 4.12(5). □

Venons-en maintenant aux questions liées à la *minimisation emboîtée*. Si X est une réunion d'ensembles, est-il équivalent de minimiser le critère sur chacun de ces ensembles et de minimiser les valeurs ainsi obtenues ? La réponse est affirmative, même s'il s'agit d'une collection non dénombrable d'ensembles, ayant éventuellement des points communs, mais il faut bien comprendre ce que cela veut dire (la remarque 1.9 ci-dessous va le préciser).

Proposition 1.8 (minimisation emboîtée I) *Soient $\{X_i\}_{i\in I}$ une famille quelconque d'ensembles pouvant avoir des points en commun, $X = \cup_{i\in I} X_i$ leur union et $f : X \to \overline{\mathbb{R}}$ une application. Alors*

$$\inf_{x\in X} f(x) = \inf_{i\in I} \left(\inf_{x\in X_i} f(x) \right). \tag{1.7}$$

De plus, $x_* \in X_{i_*}$ est solution du problème à gauche dans (1.7) si, et seulement si, i_* est solution du problème à droite dans (1.7) et x_* est solution du problème $\inf_{x\in X_{i_*}} f(x)$.

Remarques 1.9 1) Il faut comprendre l'expression $\inf_{i\in I} (\inf_{x\in X_i} f(x))$ dans (1.7) comme le problème de minimisation en $i \in I$ de la fonction φ définie par

$$\varphi(i) := \inf_{x\in X_i} f(x). \tag{1.8}$$

Le problème de minimisation dans (1.8) est, quant à lui, appelé le *problème interne* associé à $i \in I$.

2) D'après la proposition, si $x_* \in X_{i_*}$ est solution du problème à gauche dans (1.7), alors le problème $\inf_{x\in X_{i_*}} f(x)$ a une solution (par exemple x_*), mais il se peut que d'autres problèmes internes $\inf_{x\in X_i} f(x)$, avec $i \neq i_*$, n'en ait pas. □

Démonstration. Pour tout $i \in I$, on a $X_i \subseteq X$ et donc

$$\inf_{x\in X} f(x) \leqslant \inf_{x\in X_i} f(x).$$

Comme le membre de gauche ne dépend pas de $i \in I$, on a

$$\inf_{x\in X} f(x) \leqslant \inf_{i\in I} \inf_{x\in X_i} f(x). \tag{1.9a}$$

Inversement, pour tout $x_0 \in X$, il existe un $i_0 \in I$ tel que $x_0 \in X_{i_0}$. On en déduit que

$$f(x_0) \geqslant \inf_{x\in X_{i_0}} f(x) \geqslant \inf_{i\in I} \inf_{x\in X_i} f(x),$$

où la dernière inégalité est évidente. Le membre de droite est donc une borne inférieure de $f(x_0)$, qui est indépendante de $x_0 \in X$. Dès lors,

$$\inf_{x\in X} f(x) \geqslant \inf_{i\in I} \inf_{x\in X_i} f(x). \tag{1.9b}$$

Grâce à (1.9a) et (1.9b), l'identité (1.7) est démontrée.

Par ailleurs, si $x_* \in X_{i_*}$ est solution du problème à gauche dans (1.7), on a $f(x_*) \leqslant f(x)$ pour tout $x \in X$ et donc certainement pour tout $x \in X_{i_*}$, si bien que x_* est solution du problème interne associé à i_* et $f(x_*) = \varphi(i_*)$, où φ est définie par (1.8). On a aussi, quel que soit $i \in I$, $f(x_*) \leqslant f(x)$ pour tout $x \in X_i$, donc $\varphi(i_*) \leqslant \varphi(i)$ si bien que i_* est solution du problème à droite dans (1.7). Inversement, on a successivement

$$\begin{aligned}
f(x_*) &= \inf_{x\in X_{i_*}} f(x) \quad [x_* \text{ est solution du problème interne } i_*] \\
&= \varphi(i_*) \quad [\text{définition de } \varphi] \\
&= \inf_{i\in I} \varphi(i) \quad [i_* \text{ solution du problème à droite dans (1.7)}] \\
&= \inf_{i\in I} \left(\inf_{x\in X_i} f(x) \right) \quad [\text{définition de } \varphi] \\
&= \inf_{x\in X} f(x) \quad [\text{par (1.7)}].
\end{aligned}$$

Donc x_* est solution du problème à gauche dans (1.7). □

Voici deux corollaires bien utiles de la proposition 1.8.

Corollaire 1.10 (minimisation emboîtée II) *Soient X et Y deux ensembles et $f : X \times Y \to \overline{\mathbb{R}}$ une application. Alors*

$$\inf_{(x,y)\in X\times Y} f(x,y) = \inf_{x\in X}\left(\inf_{y\in Y} f(x,y)\right) = \inf_{y\in Y}\left(\inf_{x\in X} f(x,y)\right). \tag{1.10}$$

De plus, (x_*, y_*) est solution du problème à gauche dans (1.10) si, et seulement si, x_* est solution du problème du milieu dans (1.10) et y_* est solution du problème interne $\inf_{y\in Y} f(x_*, y)$ ou encore, si, et seulement si, y_* est solution du problème à droite dans (1.10) et x_* est solution du problème interne $\inf_{x\in X} f(x, y_*)$.

Démonstration. Il suffit d'écrire $X \times Y = \cup_{x\in X}(\{x\} \times Y) = \cup_{y\in Y}(X \times \{y\})$ et d'appliquer la proposition 1.8. □

Remarque 1.11 Le problème à droite dans (1.10) ne veut pas dire que pour minimiser $f(x, y)$, il suffit de minimiser $y \mapsto f(x_0, y)$, pour un $x_0 \in X$ arbitraire, dont la solution serait $\bar{y}$ (si elle existe !), et ensuite de minimiser $x \mapsto f(x, \bar{y})$. Il signifie que minimiser $(x, y) \mapsto f(x, y)$ équivaut à minimiser $x \mapsto \varphi(x)$, où $\varphi(x) := \inf_{y\in Y} f(x, y)$. Pour chaque x, il y a donc un problème de minimisation à résoudre pour déterminer $\varphi(x)$.

En outre, l'identité (1.10) nous apprend que l'on peut inverser l'ordre dans lequel sont pris deux minimisations successives (ou deux maximisations successives) sans modifier la valeur optimale. Il en va tout autrement si une minimisation est suivie d'une maximisation (ou inversement), comme le montrera la section 14.1 sur la dualité min-max. □

Corollaire 1.12 (minimisation emboîtée III) *Soient U et V deux ensembles, X une partie de $U \times V$ et $f : X \to \overline{\mathbb{R}}$. Alors*

$$\inf_{x\in X} f(x) = \inf_{u\in U}\left(\inf_{v\in X_u} f(u,v)\right) = \inf_{v\in V}\left(\inf_{u\in X^v} f(u,v)\right), \tag{1.11}$$

où

$$X_u := \{v \in V : (u, v) \in X\} \qquad \text{et} \qquad X^v := \{u \in U : (u, v) \in X\}.$$

De plus, $x_* = (u_*, v_*)$ est solution du problème à gauche dans (1.11) si, et seulement si, u_* est solution du problème du milieu dans (1.11) et v_* est solution du problème interne $\inf_{v\in X_{u_*}} f(u_*, v)$ ou encore, si, et seulement si, v_* est solution du problème à droite dans (1.11) et u_* est solution du problème interne $\inf_{u\in X^{v_*}} f(u, v_*)$.

Démonstration. On définit $\tilde{f} : U \times V \to \overline{\mathbb{R}}$ par

$$\tilde{f}(x) = \begin{cases} f(x) & \text{si } x \in X \\ +\infty & \text{sinon} \end{cases}$$

et on applique le corollaire 1.10 à $\tilde{f}$. □

Certains algorithmes s'étudient sur des problèmes dont le critère est linéaire (voir, par exemple, le chapitre 16 sur les points intérieurs). Il n'y a pas de restriction dans ce choix si le problème admet déjà des contraintes non linéaires, car on peut toujours *faire passer un terme du critère en contrainte* par l'équivalence mise en évidence dans la proposition suivante.

Proposition 1.13 (passage d'un terme du critère en contrainte) *Soient X un ensemble et f, $g : X \to \mathbb{R} \cup \{+\infty\}$ deux fonctions. Alors*

$$\inf_{x\in X} f(x) + g(x) = \inf_{\substack{(x,\gamma)\in X\times\mathbb{R} \\ g(x)\leqslant\gamma}} f(x) + \gamma. \tag{1.12}$$

De plus, si x_* est solution du problème à gauche dans (1.12) et si $g(x_*)$ est fini, alors $(x_*, g(x_*))$ est solution du problème à droite. Inversement, si (x_*, γ_*) est solution du problème à droite dans (1.12), alors x_* est solution du problème à gauche et $\gamma_* = g(x_*)$.

Si $f = 0$, le problème à droite dans (1.12) consiste de trouver le couple (x, γ) dans l'*épigraphe* de g, qui est l'ensemble

$$\operatorname{epi} g := \{(x, \gamma) \in X \times \mathbb{R} : g(x) \leqslant \gamma\},$$

avec l'ordonnée γ la plus petite possible.

Démonstration. En appliquant le corollaire 1.12, on obtient l'égalité en (1.12) :

$$\inf_{\substack{(x,\gamma)\in X\times\mathbb{R} \\ g(x)\leqslant\gamma}} f(x) + \gamma = \inf_{x\in X} \left(\inf_{\substack{\gamma\in\mathbb{R} \\ g(x)\leqslant\gamma}} f(x) + \gamma \right) = \inf_{x\in X} f(x) + g(x). \tag{1.13}$$

Soit $x_* \in X$ une solution du problème à gauche dans (1.12) telle que $\gamma_* := g(x_*)$ soit fini. Alors $(x_*, \gamma_*) \in \operatorname{epi} g$ (admissibilité pour le problème à droite). De plus, pour tout $(x, \gamma) \in \operatorname{epi} g$, on a $f(x_*) + \gamma_* = f(x_*) + g(x_*) \leqslant f(x) + g(x)$ [car x_* est solution du problème de gauche] $\leqslant f(x) + \gamma$. Donc (x_*, γ_*) est solution du problème de droite.

Inversement, soit (x_*, γ_*) une solution du problème à droite dans (1.12), c'est-à-dire du problème à gauche dans (1.13). D'après le corollaire 1.12, x_* est solution du problème du milieu dans (1.13) (c'est-à-dire solution du problème à gauche dans (1.12)) et γ_* est solution du problème interne avec $x = x_*$ (c'est-à-dire $\gamma_* = g(x_*)$). □

Notes

Auslender et Teboulle [29 ; 2003] dérivent des résultats d'existence de solution de problèmes d'optimisation (convexes ou non) en utilisant une notion de fonction

asymptotique qui étend celle que nous verrons pour les fonctions convexes au chapitre 3. Ils énoncent en particulier des conditions nécessaires et suffisantes d'existence de solution pour des fonctions propres et s.c.i. [29 ; théorème 3.4.1], bien que celles-ci ne soient pas toujours très simples à utiliser. On peut aussi s'intéresser à des résultats d'existence de solution généraux sur des espaces de Banach (de dimension infinie) ou des espaces métriques complets [91, 640].

Malgré l'étendue du corpus de l'optimisation numérique, on peut trouver des synthèses plus ou moins courtes et détaillées donnant une vue d'ensemble de la discipline ; mentionnons [272, 239]. On en apprendra davantage dans les ouvrages généraux d'optimisation numérique, que nous citons par ordre chronologique : le livre de Fiacco et McCormick [207 ; 1968] a connu une renaissance au moment de l'émergence des méthodes de points intérieurs (chapitres 16, 18 et 20) ; Ortega et Rheinboldt [463 ; 1970] ont écrit un ouvrage très classique qui se concentre sur les méthodes itératives de résolution de systèmes d'équations non linéaires, en ayant toutefois une section consacrée à l'optimisation sans contrainte ; Gill, Murray et Wright [251 ; 1981] ; Ciarlet [126 ; 1982] ; McCormick [415 ; 1983] ; Fletcher [212 ; 1987] ; Culioli [148 ; 1994] ; Nazareth [450 ; 1994] ; Bertsekas [54 ; 1995] ; Kelley [348 ; 1995] ; Gauvin [233 ; 1995] ; Hiriart-Urruty [315 ; 1996] ; Polak [482 ; 1997] ; Kelley [349 ; 1999] ; Conn, Gould et Toint [138 ; 2000] sur les régions de confiance ; Biegler et coll. [58 ; 2003] présentent une série de travaux sur l'optimisation de systèmes gouvernés par des équations aux dérivées partielles ; Boyd et Vandenberghe [86 ; 2004] donnent une introduction bien imagée et très abordable de l'optimisation convexe avec de nombreuses applications intéressantes, sans entrer dans la démonstration des résultats les plus fins ; Nesterov [452 ; 2004] ; Nocedal et Wright [460 ; 2006] ; Bonnans, Gilbert, Lemaréchal et Sagastizábal [73 ; 2006] sur l'optimisation sans contrainte, l'optimisation non différentiable, la programmation quadratique successive et les points intérieurs ; Ito et Kunisch [328 ; 2008] présentent la théorie et les algorithmes de l'optimisation en dimension infinie, Bertsekas [56 ; 2015].

Voici pour terminer cette introduction quelques sites de la Toile qu'il pourra être intéressant de visiter. Les sites

http://www-fp.mcs.anl.gov/otc/Guide/SoftwareGuide et
http://plato.asu.edu/guide.html

proposent des guides de codes d'optimisation. Le site de l'Inria

http://www-rocq.inria.fr/estime/modulopt/index.html

contient quelques codes spécialisés qui peuvent s'avérer utiles pour résoudre certains problèmes difficiles ou de très grande taille. À l'adresse

http://plato.la.asu.edu/bench.html,

on trouvera un répertoire de bancs d'essai (collection de problèmes-tests) pour les codes d'optimisation. Sur cette question, l'environnement Libopt qui propose une plate-forme permettant de coupler des solveurs a différentes collections de problèmes-tests pourra s'avérer utile :

http://www-rocq.inria.fr/estime/modulopt/libopt/libopt.html.

En optimisation non linéaire, signalons les collections de problèmes-tests CUTEST et COPS :

http://ccpforge.cse.rl.ac.uk/gf/project/cutest/wiki,
http://www-unix.mcs.anl.gov/~more/cops.

Pour d'autres codes d'analyse numérique et de mathématiques en général, on pourra consulter :

http://gams.nist.gov,
http://gams.nist.gov/serve.cgi,
http://www.netlib.org.

Exercices

1.1. *Inclusion et adhérence d'ensembles admissibles.*

1) Soit X_0 une partie non vide d'un ensemble X et $f : X \to \mathbb{R}$ une fonction. Alors

$$\inf_{x\in X} f(x) \leqslant \inf_{x_0\in X_0} f(x_0).$$

2) Soient X une partie non vide d'un espace topologique et $f : \bar{X} \to \mathbb{R}$ une application semi-continue supérieurement. Alors

$$\inf_{x\in \bar{X}} f(x) = \inf_{x\in X} f(x).$$

1.2. *Minimisation et maximisation d'une somme de fonctions.* Soient X un ensemble et f et $g : X \to \mathbb{R}$ deux applications. Montrez que

(*i*) $\inf_{x\in X} f(x) + \inf_{x\in X} g(x) \leqslant \inf_{x\in X}(f(x) + g(x))$,
(*ii*) $\sup_{x\in X} f(x) + \sup_{x\in X} g(x) \geqslant \sup_{x\in X}(f(x) + g(x))$.

1.3. *Deux notions de coercivité équivalentes.* Soient $\mathbb{E}$ un espace vectoriel normé, X une partie de $\mathbb{E}$ et $f : \mathbb{E} \to \overline{\mathbb{R}}$ une fonction. Montrez que les propriétés suivantes sont équivalentes :

(*i*) f est *coercive* (au sens de la définition 1.3),
(*ii*) $\forall\, \nu \in \mathbb{R}$, l'ensemble $\{x \in X : f(x) \leqslant \nu\}$ est borné.

1.4. *Coercivité d'une forme bilinéaire.* Soient $\mathbb{E}$ un espace vectoriel normé et $a : \mathbb{E}\times\mathbb{E} \to \mathbb{R}$ une forme bilinéaire. Montrez que a est *coercive* si, et seulement si,

$$\exists\, \alpha > 0, \quad \forall\, x \in \mathbb{E} : \qquad a(x,x) \geqslant \alpha\|x\|^2. \tag{1.14}$$

1.5. *Optimisation globale par l'optimisation linéaire sur l'espace des mesures* [375]. Soient X une partie mesurable d'un espace vectoriel $\mathbb{E}$, $\mathcal{M}(X)$ l'ensemble des mesures sur X et $f : X \to \mathbb{R}$ une fonction. Alors l'infimum de f sur X est donné par la valeur optimale d'un problème d'optimisation linéaire sur $\mathcal{M}(X)$, à savoir

$$\inf_{x\in X} f(x) = \inf_{\substack{\mu\in\mathcal{M}(X)\\ \mu(X)=1\\ \mu\geqslant 0}} \int_X f \,\mathrm{d}\mu. \tag{1.15}$$

Remarque. On transforme ainsi le problème potentiellement non convexe de gauche (a priori difficile à résoudre numériquement) en un problème, celui de droite, à la structure très simple (il est linéaire) mais de dimension infinie.

1.6. *Norme de l'inverse d'une matrice par minimisation* ([185] et sans doute beaucoup d'autres articles). Soit $A \in \mathbb{R}^{n\times n}$ une matrice carrée inversible et $\|\cdot\|$ une norme matricielle subordonnée à une norme vectorielle, également notée $\|\cdot\|$. On sait que $\|A\|$ peut se calculer en résolvant le problème de maximisation (B.4). Montrez l'identité suivante

$$\min_{\|x\|=1} \|Ax\| = \|A^{-1}\|^{-1}, \tag{1.16a}$$

qui permet de calculer $\|A^{-1}\|$ en résolvant le problème de minimisation. Montrez également que x_* résout le problème dans le membre de gauche de (1.16a) si, et seulement si, $y_* := \|A^{-1}\|\, Ax_*$ résout le problème dans le membre de gauche de

$$\max_{\|y\|=1} \|A^{-1}y\| = \|A^{-1}\|. \tag{1.16b}$$

1.7. *Problèmes d'optimisation équivalents.* Soient X un ensemble et $f_i : X \to \mathbb{R}$ des applications (pour i dans un ensemble d'indices quelconque I). Montrez que les problèmes ci-dessous ont les mêmes solutions $x_* \in X$:

$$\min_{x\in X} \sup_{i\in I} f_i(x) \qquad \text{et} \qquad \begin{cases} \min \alpha \\ f_i(x) \leqslant \alpha, \quad \forall i \in I \\ x \in X, \quad \alpha \in \mathbb{R}. \end{cases}$$

Remarque. Lorsque les f_i sont différentiables, on a remplacé le problème à gauche, qui est en général non différentiable mais n'a pas de contrainte fonctionnelle, par le problème à droite, qui est différentiable mais présente des contraintes fonctionnelles d'inégalité. On a remplacé la difficulté liée à la non-différentiabilité par celle liée à la présence de contraintes d'inégalité. C'est un exemple où la *loi de conservation des ennuis* se manifeste.

1.8. *Récritures différentiables.* Récrire sous une forme différentiable les problèmes d'optimisation suivants dans lesquels $f : \mathbb{R}^n \to \mathbb{R}$ et $F : \mathbb{R}^n \to \mathbb{R}^m$ sont des fonctions différentiables, X est une partie de $\mathbb{R}^n$ (définie par des fonctions différentiables) et une non-différentiabilité est présente du fait de l'utilisation des normes ℓ_1, notée $\|\cdot\|_1$, ou ℓ_∞, notée $\|\cdot\|_\infty$:

1) $\inf_x \{f(x) + \|F(x)\|_1 : x \in X\}$,
2) $\inf_x \{f(x) + \|F(x)\|_\infty : x \in X\}$.

2 Ensembles convexes

Convexity is a large subject which can hardly be addressed here, but much of the impetus for its growth in recent decades has come from applications in optimization. [...] *In fact the great watershed in optimization isn't between linearity and nonlinearity, but convexity and nonconvexity. Even for problems that aren't themselves of convex type, convexity may enter, for instance, in setting up subproblems as part of an iterative numerical scheme.*

R.T. Rockafellar [526 ; 1993].

We now take for granted that convex analysis is a good subject with worthwhile ideas, yet it was not always that way. There was actually a lot of resistance to it in the early days, from individuals who preferred a geometric presentation to one targeting concepts of analysis. Even on the practical plane, it's fair to say that little respect was paid to convex analysis in numerical optimization until around 1990, say.

R.T. Rockafellar sur le site Wikimization.

Les chapitres 2 et 3 présentent les éléments d'analyse convexe qui nous seront utiles pour étudier les problèmes d'optimisation et les algorithmes qui les résolvent. Le chapitre 2 s'intéresse aux ensembles convexes ; le chapitre 3 aux fonctions convexes. L'*analyse convexe* est une théorie située entre l'algèbre linéaire et l'analyse non linéaire, dans laquelle les objets étudiés, ensembles et fonctions, bien que non linéaires, sont contraints de vérifier une condition particulière qui leur confère des propriétés remarquables. C'est une théorie assez récente ; certains voient la naissance de sa version « moderne », celle renforçant le rôle de l'analyse, dans l'invention du sous-différentiel (section 3.6), de l'application proximale (section 3.7.1) et de l'inf-convolution (section 3.4.4) dans les années 1962-63 [132].

Cette théorie et ses concepts interviennent en optimisation pour de nombreuses raisons. Par exemple, l'écriture des conditions d'optimalité passe par le linéarisé de l'ensemble admissible, qui est un cône ; un objet qui n'appartient pas à l'algèbre linéaire mais à l'analyse convexe. Autre exemple : dans les problèmes d'optimisation convexe (c'est-à-dire avec un critère convexe et un ensemble admissible convexe), tous les points stationnaires sont des minima globaux, ce qui simplifie singulièrement le

problème[1]. Si l'analyse convexe joue un rôle prépondérant en optimisation, la phrase de R.T. Rockafellar donnée en épigraphe parlant de ligne de partage des eaux (*watershed*), nous semble devoir être relativisée ; on sait en effet qu'il existe des problèmes d'optimisation convexe qui sont NP-ardus (l'optimisation copositive [447, 69]) et donc aujourd'hui très difficiles à résoudre lorsque leur dimension est grande.

L'analyse convexe est une théorie très riche et d'une évidente élégance. C'est donc avec regret que nous avons cherché à en dire le moins possible, en ne développant que les concepts et les propriétés qui sont nécessaires à l'étude des quelques problèmes d'optimisation et méthodes de résolution que nous verrons dans cet ouvrage. Le lecteur intéressé pourra approfondir ses connaissances en s'immergeant dans les ouvrages spécialisés cités dans les notes à la fin du chapitre 3. Nous espérons toutefois que l'étude de ces deux chapitres apportera au lecteur une aisance suffisante dans cet univers merveilleux. Si la théorie est parfois difficile, elle contient aussi beaucoup de résultats simples à démontrer pourvu que l'on maîtrise la technique ; leur démonstration est alors proposée en exercice. Mais l'analyse convexe a aussi des affirmations simples à énoncer, qui semblent très naturelles, mais que l'on ne sait pas démontrer. On y prendra garde !

Le fait que l'*Analyse convexe* existe en tant que discipline des mathématiques, et pas l'*Analyse concave*, tient au fait que l'on définit aisément la notion d'ensemble convexe, alors que celle d'*ensemble concave* est moins naturelle, voire pratiquement inexistante. On définit alors les fonctions convexes comme celles ayant un épigraphe convexe (les fonctions concaves ont, elles, un *hypographe* convexe...). Il est normal de minimiser les fonctions convexes, pas de les maximiser, si bien que l'optimisation s'intéressera tout naturellement à la minimisation de fonctions et pas à leur maximisation. La chaîne logique des concepts est donc la suivante :

$$\text{ensemble convexe} \quad \longrightarrow \quad \text{fonction convexe} \quad \longrightarrow \quad \text{minimisation.} \tag{2.1}$$

Mais, on l'a vu (proposition 1.5), les problèmes de maximisation se ramènent aisément à des problèmes de minimisation.

Dans ce chapitre nous introduisons la notion d'ensemble convexe, donnons quelques exemples d'ensembles convexes (en particulier une description assez précise des polyèdres convexes), démontrons leurs principales propriétés (géométriques et topologiques) et décrivons quelques opérations sur les ensembles convexes (projection, séparation, prise du dual). En chemin, nous serons amenés à démontrer un résultat d'existence de solution pour les problèmes d'optimisation linéaire (proposition 2.20) qui nous sera aussi utile au chapitre 17.

Comme c'est le cas dans tout cet ouvrage, l'espace vectoriel sur lequel on travaille, noté $\mathbb{E}$, est supposé défini sur le *corps des réels* $\mathbb{R}$ et *de dimension finie* ; cela ne sera pas toujours précisé dans l'énoncé des résultats. Il n'y a alors pas de restriction à supposer qu'il est euclidien, c'est-à-dire muni d'un produit scalaire que l'on notera $\langle \cdot, \cdot \rangle$.

[1] Les problèmes d'optimisation à données polynomiales présentent aussi des propriétés remarquables, qui contraignent les possibilités d'une autre manière, mais c'est alors la géométrie algébrique plutôt que l'analyse convexe qui y joue un rôle clé. Nous n'aborderons pas ce domaine qui connaît un essor important depuis le début du XXIe siècle [376, 65, 377].

2.1 Définition et premières propriétés

Soient $\mathbb{E}$ un espace vectoriel de dimension finie sur $\mathbb{R}$ et x, $y \in \mathbb{E}$. On appelle *segment* de $\mathbb{E}$, un ensemble noté et défini comme suit

$$[x, y] := \{(1-t)x + ty : t \in [0, 1]\}.$$

Lorsque $x \neq y$, on définit les segments $[x, y[$, $]x, y]$ et $]x, y[$, en remplaçant dans la formule ci-dessus, l'intervalle $[0, 1]$ respectivement par les intervalles $[0, 1[$, $]0, 1]$ et $]0, 1[$. Lorsque $x = y$, les segments $[x, y[$, $]x, y]$ et $]x, y[$ sont vides, par définition.

On dit qu'une partie C de $\mathbb{E}$ est *convexe* si pour tout x, $y \in C$, le segment $[x, y]$ est contenu dans C. On dit aussi que C est « un convexe ». La figure 2.1 illustre cette

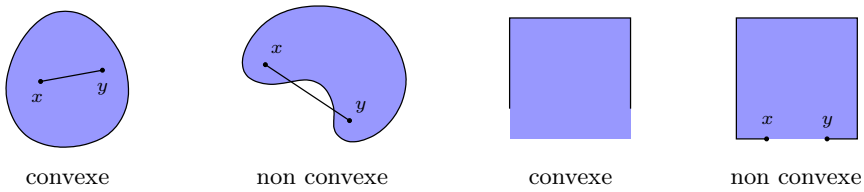


Fig. 2.1. Définition d'un ensemble convexe

notion. L'ensemble de gauche est convexe car il contient tous les segments $[x, y]$ avec des points x et y lui appartenant. Le second ne l'est pas car une partie du segment $[x, y]$ qui y est représenté n'est pas dans l'ensemble. Le cas des deux carrés à droite est plus délicat car la question se joue sur la frontière (la partie de celle-ci appartenant à l'ensemble est marquée d'un trait continu). Le carré de gauche est convexe, bien qu'il ne contienne pas toute sa frontière. Celui de droite ne l'est pas, car le segment $]x, y[$ ne lui appartient pas, alors que x et y sont supposés appartenir à l'ensemble. Cet exemple du carré n'est pas anodin, mais est destiné à faire prendre conscience au lecteur du fait que la validité d'une propriété (comme la convexité, la semi-continuité d'une fonction, sa sous-différentiabilité, *etc*) dépend souvent de ce qui se passe sur la frontière d'un ensemble ; on devra donc toujours traiter avec soin et précision ce qui peut apparaître comme un détail au premier abord.

Exemples et propriétés immédiates

Ci-dessous, $\mathbb{E}$, $\mathbb{E}_1$, $\mathbb{E}_2$ et $\mathbb{F}$ sont des espaces vectoriels ou euclidiens. Les affirmations se démontrent tous sans difficulté, soit en utilisant la définition de la convexité d'un ensemble, soit à partir des exemples et propriétés énoncés précédemment.

1) Un *hyperplan* d'un espace euclidien $(\mathbb{E}, \langle \cdot, \cdot \rangle)$ est un ensemble de la forme

$$H(\xi, \alpha) := \{x \in \mathbb{E} : \langle \xi, x \rangle = \alpha\},$$

où $\xi \in \mathbb{E}$ est *non nul* et $\alpha \in \mathbb{R}$. C'est clairement un convexe de $\mathbb{E}$.

2) Un *demi-espace fermé* de $\mathbb{E}$ est un ensemble de l'une des deux formes suivantes

$$H^-(\xi,\alpha) := \{x \in \mathbb{E} : \langle \xi, x\rangle \leqslant \alpha\}, \tag{2.2a}$$
$$H^+(\xi,\alpha) := \{x \in \mathbb{E} : \langle \xi, x\rangle \geqslant \alpha\}, \tag{2.2b}$$

où $\xi \in \mathbb{E}$ est *non nul* et $\alpha \in \mathbb{R}$. On dit que les demi-espaces sont *ouverts* si l'on utilise des inégalités strictes en (2.2a) et (2.2b). Ce sont aussi clairement des convexes de $\mathbb{E}$.

3) La partie de $\mathbb{R}^n$ définie par $\mathbb{R}^n_+ := \{x \in \mathbb{R}^n : x \geqslant 0\}$ est un convexe appelé *orthant positif*. La notation $x \geqslant 0$ veut dire que $x_i \geqslant 0$ pour tout indice $i \in [1:n]$.

4) La somme (de Minkowski) $C_1 + C_2 := \{x_1 + x_2 : x_1 \in C_1,\ x_2 \in C_2\}$ de deux convexes C_1 et C_2 de $\mathbb{E}$ est un convexe. Le produit $\alpha C := \{\alpha x : x \in C\}$ d'un scalaire $\alpha \in \mathbb{R}$ par un convexe C est un convexe (voir aussi le point 1 de l'exercice 2.1).

5) Si $\{C_i\}_{i\in I}$ est une famille quelconque de convexes de $\mathbb{E}$, alors leur intersection $\cap_{i\in I} C_i$ est un convexe (mais pas leur union !).

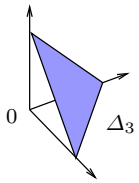
6) Soit $A : \mathbb{E} \to \mathbb{F}$ une application *linéaire*. L'image directe $A(C)$ (resp. l'image réciproque $A^{-1}(C)$) d'un convexe C de $\mathbb{E}$ (resp. de $\mathbb{F}$) par A est un convexe.

7) Soient $C_1 \subseteq \mathbb{E}_1$ et $C_2 \subseteq \mathbb{E}_2$. Alors $C_1 \times C_2$ est convexe dans $\mathbb{E}_1 \times \mathbb{E}_2$ si, et seulement si, C_1 et C_2 sont convexes.

8) On appelle *polyèdre convexe* de $\mathbb{E}$ un ensemble de la forme $\{x \in \mathbb{E} : Ax \leqslant b\}$, où $A : \mathbb{E} \to \mathbb{R}^m$ est une application linéaire, $b \in \mathbb{R}^m$ et l'inégalité $Ax \leqslant b$ se comprend composante par composante : $(Ax)_i \leqslant b_i$, pour tout $i \in [1:m]$. C'est donc l'intersection d'un nombre *fini* de demi-espaces fermés. D'après ce qui précède, c'est un convexe.

9) On appelle *simplexe unitaire* de $\mathbb{R}^n$ l'ensemble défini par

$$\Delta_n := \left\{x \in \mathbb{R}^n : e^\mathsf{T} x = 1,\ x \geqslant 0\right\},$$



où $e = (1,\dots,1) \in \mathbb{R}^n$. C'est un convexe (intersection de deux convexes : un sous-espace affine et l'orthant positif).

10) On note $\mathcal{S}^n$ l'espace vectoriel des matrices d'ordre n symétriques. Il est de dimension $n(n+1)/2$. Les ensembles

$$\mathcal{S}^n_+ := \{A \in \mathcal{S}^n : A \text{ est semi-définie positive}\}$$
$$\mathcal{S}^n_{++} := \{A \in \mathcal{S}^n : A \text{ est définie positive}\}$$

sont convexes. En effet, si A et $B \in \mathcal{S}^n_+$, alors pour tout $t \in [0,1]$ et tout vecteur $v \in \mathbb{R}^n$, on a $v^\mathsf{T}((1-t)A + tB)v = (1-t)v^\mathsf{T}Av + tv^\mathsf{T}Bv \geqslant 0$, ce qui montre que $(1-t)A + tB \in \mathcal{S}^n_+$. On raisonne de même pour $\mathcal{S}^n_{++}$.

11) Si $\mathcal{A}$ est un sous-espace affine de $\mathcal{S}^n$, alors

$$\mathcal{S}^n_+ \cap \mathcal{A}$$

est évidemment un ensemble convexe (intersection de deux convexes). Beaucoup d'ensembles convexes peuvent se ramener à une telle description, au prix de quelques transformations algébriques. Souvent, le sous-espace affine est vu comme le noyau translaté d'une application linéaire $L : \mathcal{S}^n \to \mathbb{R}^m$:

$$\mathcal{A} = \{X \in \mathcal{S}^n : L(X) = b\},$$

où b est fixé dans $\mathbb{R}^m$. On peut aussi voir $\mathcal{A}$ comme l'image d'une application affine $\mathbb{R}^m \to \mathcal{S}^n$:

$$\mathcal{A} = \{A_0 + \textstyle\sum_{i=1}^m \alpha_i A_i : \alpha \in \mathbb{R}^m\},$$

où les matrices A_i sont fixées dans $\mathcal{S}^n$. Dans le premier ou le second cas, $X \in \mathcal{S}^n_+ \cap \mathcal{A}$ si, et seulement si, X vérifie les conditions suivantes

$$L(X) = b, \quad X \succcurlyeq 0 \qquad \text{ou} \qquad X = A_0 + \sum_{i=1}^m \alpha_i A_i, \quad X \succcurlyeq 0.$$

C'est ce qu'on appelle des *inégalités matricielles linéaires* (IML en abrégé, on devrait dire affine plutôt que linéaire).

2.2 Aspects géométriques

En algèbre linéaire, il est naturel de considérer le plus petit sous-espace vectoriel contenant un ensemble donné P de $\mathbb{E}$, ainsi que de son enveloppe affine qui est le plus petit sous-espace affine contenant cet ensemble (section 2.2.1). L'analyse convexe associe à P de nouveaux ensembles : son enveloppe convexe qui est le plus petit convexe contenant P (section 2.2.2), son enveloppe convexe fermée qui est le plus petit convexe fermé contenant P (section 2.5.4), son enveloppe conique qui est le plus petit cône *convexe* contenant P (section 2.2.3) et, lorsque P est convexe, son cône asymptotique (section 2.2.4) et ses faces (section 2.2.5).

2.2.1 Enveloppe affine

Soit P une partie d'un espace vectoriel $\mathbb{E}$. L'intersection de sous-espaces affines étant un sous-espace affine (exercice A.8), on peut parler du plus petit sous-espace affine contenant P, qui est donc l'intersection de tous les sous-espaces affines de $\mathbb{E}$ contenant P. C'est ce que l'on appelle l'*enveloppe affine* de P. On la note

$$\operatorname{aff} P := \bigcap \{A : A \text{ est un sous-espace affine de } \mathbb{E} \text{ contenant } P\}.$$

On appelle *combinaison affine* de $\mathbb{E}$, un élément x de $\mathbb{E}$ de la forme

$$x = \sum_{i=1}^m t_i x_i,$$

où $m \in \mathbb{N}^*$, $t = (t_1, \ldots, t_m) \in \mathbb{R}^m$ vérifie $e^\mathsf{T} t = 1$ et les vecteurs $x_i \in \mathbb{E}$.

Proposition 2.1 1) *Un ensemble est un sous-espace affine si, et seulement si, il contient toutes les combinaisons affines de ses éléments.*
2) *Si $P \subseteq \mathbb{E}$, alors* aff P *est l'ensemble des combinaisons affines des éléments de P :*

$$\operatorname{aff} P = \left\{ \sum_{i=1}^{m} t_i x_i : m \in \mathbb{N}^*,\ t \in \mathbb{R}^m,\ e^{\mathsf{T}} t = 1,\ x_i \in P \right\}. \tag{2.3}$$

Démonstration. 1) Soit A un sous-espace affine. Presque par définition (voir l'exercice A.7), A contient les combinaisons affines formées de deux de ses éléments. On raisonne ensuite par récurrence en supposant qu'un sous-espace affine A contient les combinaisons affines formées de m de ses éléments, avec un certain $m \geqslant 2$. Soient alors $m+1$ éléments $x_i \in A$ et des $t_i \in \mathbb{R}$ vérifiant $\sum_{i=1}^{m+1} t_i = 1$. Il y a au moins un des $t_i \neq 1$. Supposons que ce soit t_1. On écrit :

$$\sum_{i=1}^{m+1} t_i x_i = t_1 x_1 + (1-t_1) \left(\sum_{i=2}^{m+1} \frac{t_i}{1-t_1} x_i \right).$$

Cet élément est dans A car les facteurs de t_1 et de $1-t_1$ sont dans A (par récurrence pour le second). Réciproquement, on savait déjà qu'un ensemble contenant les combinaisons formées de deux de ses éléments est un sous-espace affine.

2) Soit X l'ensemble à droite dans (2.3). On vérifie facilement que cet ensemble contient P (prendre $m = 1$) et est affine, donc aff $P \subseteq X$. Inversement, X est contenu dans l'ensemble défini comme à droite dans (2.3), mais avec des x_i pris dans aff P plutôt que dans P. D'après la première partie de la proposition, ce dernier ensemble est aff P. Donc $X \subseteq$ aff P. □

On dit que les vecteurs $x_0, x_1, \ldots, x_p$ de $\mathbb{E}$ sont *affinement indépendants* si $p = 0$ ou si l'une des conditions équivalentes suivantes est vérifiée :

(A_1) $\sum_{i=0}^{p} \alpha_i x_i = 0$ et $\sum_{i=0}^{p} \alpha_i = 0 \implies$ tous les α_i sont nuls,
(A_2) les vecteurs $\{x_i - x_0 : i \in [1:p]\}$ sont linéairement indépendants,
(A_3) quel que soit $j \in [0:p]$, les vecteurs $\{x_i - x_j : i \in [0:p],\ i \neq j\}$ sont linéairement indépendants.

On parle parfois de la *dimension d'un ensemble convexe* C : c'est la dimension de son enveloppe affine aff(C). Si $\dim C = p$, C contient au moins et au plus $p+1$ vecteurs affinement indépendants.

2.2.2 Enveloppe convexe

Soit P une partie d'un espace vectoriel $\mathbb{E}$. L'intersection de convexes étant convexe, on peut parler du plus petit convexe contenant P, qui est donc l'intersection de tous les convexes contenant P. C'est ce que l'on appelle l'*enveloppe convexe* de P. On la note

$$\operatorname{co} P := \bigcap \{C : C \text{ est un convexe contenant } P\}.$$

On appelle *combinaison convexe* de $\mathbb{E}$, un élément x de $\mathbb{E}$ de la forme

$$x = \sum_{i=1}^{m} t_i x_i,$$

où $m \in \mathbb{N}^*$, $t = (t_1, \dots, t_m) \in \Delta_m$ (simplexe unitaire de $\mathbb{R}^m$, c'est-à-dire $t \geqslant 0$ et $\sum_{i=1}^m t_i = 1$) et les vecteurs $x_i \in \mathbb{E}$.

Proposition 2.2 1) *Un ensemble est convexe si, et seulement si, il contient toutes les combinaisons convexes de ses éléments.*
2) *Si $P \subseteq \mathbb{E}$, alors $\operatorname{co} P$ est l'ensemble des combinaisons convexes des éléments de P :*

$$\operatorname{co} P = \left\{ \sum_{i=1}^{m} t_i x_i : m \in \mathbb{N}^*,\ t \in \Delta_m,\ x_i \in P \right\}. \tag{2.4}$$

Démonstration. 1) Soit C un convexe. Par définition, C contient les combinaisons convexes formées à partir de deux éléments de C ($m = 2$). On raisonne ensuite par récurrence en supposant que C contient les combinaisons convexes formées de m éléments de C ($m \geqslant 2$). Alors pour une combinaison convexe de $m+1$ éléments de C, on écrit (on peut supposer que $t_1 \neq 1$, sinon le résultat est évident) :

$$\sum_{i=1}^{m+1} t_i x_i = t_1 x_1 + (1 - t_1) \left(\sum_{i=2}^{m+1} \frac{t_i}{1 - t_1} x_i \right).$$

Cet élément est dans C car les facteurs de t_1 et de $1 - t_1$ sont dans C (par récurrence pour le second). Réciproquement, si C contient toutes les combinaisons convexes de ses éléments, il contient les combinaisons convexes formées de deux éléments. Donc C est convexe.

2) Il est facile de voir que l'ensemble des combinaisons convexes des éléments de P est un convexe. Il contient donc $\operatorname{co} P$ qui est le plus petit convexe contenant P. Inversement, par la première partie, $\operatorname{co} P$ étant un convexe, il contient toutes les combinaisons convexes des éléments de $\operatorname{co} P$ donc de P. □

À droite dans (2.4), on ne peut pas se contenter de prendre $m = 2$. Par exemple, si P est formé des trois sommets d'un triangle non dégénéré, cet ensemble serait alors la frontière du triangle, qui n'est pas convexe. Cependant, si $P = C_1 \cup \cdots \cup C_m$ est l'union de m convexes C_i, $\operatorname{co} P$ est l'ensemble des combinaisons convexes de m points x_i, chacun des x_i étant pris dans un convexe C_i différent (voir l'exercice 2.5). Le théorème de Carathéodory ci-dessous s'inscrit dans le même esprit, celui de limiter le nombre de termes à prendre dans la somme de (2.4). Il affirme qu'en dimension n, il suffit de prendre $m = n + 1$ dans (2.4). Ce majorant sur le nombre d'éléments à prendre dans la somme à droite dans (2.4) est utile pour passer à la limite dans des combinaisons convexes, comme le montre le corollaire qui suit.

Théorème 2.3 (Carathéodory [106 ; 1907]) *Soit P une partie d'un espace vectoriel $\mathbb{E}$ de dimension n. Alors tout élément de $\operatorname{co} P$ peut s'écrire comme une combinaison convexe de $n+1$ éléments de P.*

DÉMONSTRATION. Supposons que $x \in \operatorname{co} P$ s'écrive comme combinaison convexe de $m > n+1$ éléments x_i de P : $x = \sum_{i=1}^{m} t_i x_i$, avec $t := (t_1, \dots, t_m) \in \Delta_m$. Il suffit de montrer que l'on peut écrire x comme combinaison convexe de $m-1$ des x_i. On peut supposer que tous les $t_i > 0$ (sinon le travail est fait).

En nombre $m > n+1$, les x_i sont affinement dépendants, si bien que l'on peut trouver $\alpha := (\alpha_1, \dots, \alpha_m) \neq 0$, tel que

$$\sum_{i=1}^{m} \alpha_i x_i = 0 \quad \text{et} \quad \sum_{i=1}^{m} \alpha_i = 0.$$

Comme $\alpha \neq 0$, il existe un indice k tel que $\alpha_k > 0$ (cet indice k sera mieux choisi par la suite). On peut donc écrire

$$x_k = - \sum_{\substack{1 \leqslant i \leqslant m \\ i \neq k}} \frac{\alpha_i}{\alpha_k} x_i \quad \text{et} \quad x = \sum_{\substack{1 \leqslant i \leqslant m \\ i \neq k}} \left(t_i - t_k \frac{\alpha_i}{\alpha_k} \right) x_i.$$

On voit que $\sum_{1 \leqslant i \leqslant m,\, i \neq k} (t_i - t_k \alpha_i / \alpha_k) = 1$, si bien que le résultat sera démontré si $t_i - t_k \alpha_i / \alpha_k \geqslant 0$ pour tout $i \neq k$. Ceci s'écrit encore (on se rappelle que les t_i et α_k sont > 0) :

$$\frac{\alpha_k}{t_k} \geqslant \frac{\alpha_i}{t_i}, \quad \text{pour tout } i \neq k.$$

Cette condition spécifie comment choisir l'indice $k \in \arg\max\{\alpha_i / t_i : 1 \leqslant i \leqslant m\}$, qui fournit bien un $\alpha_k > 0$. □

Corollaire 2.4 *L'enveloppe convexe d'une partie compacte d'un espace vectoriel de dimension finie est compacte.*

DÉMONSTRATION. Il est clair que $\operatorname{co} P$ est borné (on utilise la proposition 2.2) ; il reste donc à montrer qu'il est fermé. Soit $\{x_k\} \subseteq \operatorname{co} P$, avec $x_k \to x$ et montrons que $x \in \operatorname{co} P$. D'après le théorème de Carathéodory,

$$x_k = \sum_{i=1}^{n+1} t_{k,i}\, x_{k,i}, \tag{2.5}$$

avec $(t_{k,1}, \dots, t_{k,n+1}) \in \Delta_{n+1}$ et $x_{k,i} \in P$. Comme Δ_{n+1} et P sont compacts, on peut extraire de $\{t_{k,i}\}_k$ et de $\{x_{k,i}\}_k$ des sous-suites convergentes, dont les limites sont dans Δ_{n+1} et P respectivement. En passant à la limite dans (2.5), on voit que $x \in \operatorname{co} P$. □

La démonstration de la proposition suivante est proposée à l'exercice 2.4.

Proposition 2.5 (calcul d'enveloppe convexe) *Soient* $\mathbb{E}$ *et* $\mathbb{F}$ *deux espaces vectoriels.*
1) *Si* $P \subseteq \mathbb{E}$ *et* $a : \mathbb{E} \to \mathbb{F}$ *est affine, alors* $\operatorname{co}(a(P)) = a(\operatorname{co} P)$.
2) *Si* P *et* $Q \subseteq \mathbb{E}$, *alors* $\operatorname{co}(P + Q) = \operatorname{co} P + \operatorname{co} Q$.
3) *Si* $P \subseteq \mathbb{E}$ *et* $Q \subseteq \mathbb{F}$, *alors* $\operatorname{co}(P \times Q) = (\operatorname{co} P) \times (\operatorname{co} Q)$.

2.2.3 Enveloppe conique

On note

$$\mathbb{R}_+ := \{t \in \mathbb{R} : t \geqslant 0\} \qquad \text{et} \qquad \mathbb{R}_{++} := \{t \in \mathbb{R} : t > 0\}.$$

On dit qu'une partie K d'un espace vectoriel $\mathbb{E}$ est un *cône* si $\mathbb{R}_{++}K \subseteq K$, c'est-à-dire si $tx \in K$ chaque fois que $t > 0$ et $x \in K$. On dit qu'un cône K est

- *saillant* si $K \cap (-K) \subseteq \{0\}$, ce qui revient à dire qu'il ne contient pas de droite (sous-espace affine de dimension 1),
- *pointé* si $0 \in K$ (et *épointé* dans le cas contraire).

On a pris soin de ne pas imposer qu'un cône contienne l'origine (on ne demande pas que $\mathbb{R}_+K \subseteq K$), de manière à pouvoir parler de cônes ouverts, tels que $\mathcal{S}^n_{++}$.

Soit P une partie de $\mathbb{E}$. L'intersection de cônes convexes étant un cône convexe, on peut parler du plus petit cône convexe contenant P, qui est donc l'intersection de tous les cônes convexes contenant P. C'est ce que l'on appelle l'*enveloppe conique* de P (on devrait dire son *enveloppe conique convexe*). On la note

$$\operatorname{cone} P := \bigcap \{K : K \text{ est un cône convexe contenant } P\}.$$

On appelle *combinaison conique* de $\mathbb{E}$, un élément x de $\mathbb{E}$ de la forme

$$x = \sum_{i=1}^{m} t_i x_i,$$

où $m \in \mathbb{N}^*$, les $t_i \in \mathbb{R}$ avec $t_i \geqslant 0$ et les vecteurs $x_i \in \mathbb{E}$. La démonstration de la proposition suivante est proposée à l'exercice 2.8.

Proposition 2.6 1) *Un ensemble est un cône convexe si, et seulement si, il contient toutes les combinaisons coniques de ses éléments.*
2) *Si* $P \subseteq \mathbb{E}$, *alors* $\operatorname{cone} P$ *est l'ensemble des combinaisons coniques des éléments de* P *:*

$$\operatorname{cone} P = \left\{ \sum_{i=1}^{m} t_i x_i : m \in \mathbb{N}^*,\ t_i \in \mathbb{R} \text{ avec } t_i \geqslant 0,\ x_i \in P \right\}. \tag{2.6}$$

2.2.4 Cône asymptotique

Soit C un ensemble *convexe fermé non vide* d'un espace vectoriel $\mathbb{E}$ de dimension finie. Le *cône asymptotique* de C est l'ensemble défini par

$$C^\infty := \{d \in \mathbb{E} : C + \mathbb{R}_+ d \subseteq C\} = \{d \in \mathbb{E} : C + d \subseteq C\}.$$

Un élément de C^∞ est appelé une *direction asymptotique*[2]. De façon imagée et à une translation près, le cône asymptotique est l'apparence que prend C lorsqu'on le voit d'infiniment loin. Il apparaît alors comme un cône, réduit éventuellement à zéro (si, et seulement si, il est borné ; c'est ce que nous allons montrer).

La proposition 2.7 ci-dessous donne dans son point (i) une autre expression du cône asymptotique, qui sert parfois à étendre ce concept à des ensembles C non convexes [29], et montre dans son point (ii) que le cône asymptotique peut aussi s'écrire

$$C^\infty(x) := \{d \in \mathbb{E} : x + \mathbb{R}_+ d \subseteq C\} = \bigcap_{t>0} \frac{C - x}{t},$$

quel que soit le point x choisi dans C (le caractère fermé de C est essentiel pour avoir cette propriété). Dans l'expression ci-dessus, $(C - x)/t = \{(y - x)/t : y \in C\}$.

Proposition 2.7 (autres expressions du cône asymptotique) *Soit C un ensemble convexe fermé non vide. Alors C^∞ est un cône convexe fermé contenant zéro. De plus*
(i) $C^\infty = \{d \in \mathbb{E} :$ il existe $\{x_k\} \subseteq C$ et $\{t_k\} \to \infty$ tels que $x_k/t_k \to d\}$,
(ii) $C^\infty = C^\infty(x)$, quel que soit $x \in C$.

Démonstration. Pour tout $t > 0$, $(C - x)/t$ est un convexe fermé. Il en est donc de même de $C^\infty(x)$ et donc de $C^\infty = \cap_{x \in C} C^\infty(x)$. D'autre part, il est clair que C^∞ est un cône ($d \in C^\infty$ et $t > 0$ impliquent que $td \in C^\infty$) et qu'il contient zéro.

Désignons par K l'ensemble dans le membre de droite de (i). Soit $x \in C$. Pour démontrer (i) et (ii), il suffit de montrer que $C^\infty(x) = K$.

Soit $d \in C^\infty(x)$. Avec $\{t_k\} \to \infty$ et $x_k = x + t_k d$, on a $x_k \in C$ et $x_k/t_k \to d$. Donc $d \in K$. Inversement, soient $\{x_k\} \subseteq C$ et $\{t_k\} \to \infty$ tels que $x_k/t_k \to d$. Fixons $t > 0$. Dès que $t_k \geqslant t$,

$$x + \frac{t}{t_k}(x_k - x) \in C$$

et ce point converge vers $x + td \in C$ (fermé). Donc $d \in C^\infty(x)$. □

Si C n'est pas fermé, $C^\infty(x)$ peut dépendre de x. Par exemple, si $C = \{x \in \mathbb{R}^2 : x > 0\} \cup \{(0, 0)\}$, $C^\infty(x)$ est l'orthant positif si $x \neq 0$, mais $C^\infty(0) = C$.

[2] Certains auteurs [517 ; page 61] préfèrent utiliser les appellations *cône de récession* (alors noté 0^+C, en évitant le signe ∞) et *direction de récession* à *cône asymptotique* et *direction asymptotique*, parce que la notion n'a pas de rapport direct avec celle d'asymptote. Le qualificatif *asymptotique* est en réalité utilisé ici comme dans les locutions *comportement asymptotique* et *développement asymptotique*, comme un substitut de l'expression « *à l'infini* ».

D'après le point (i) de la proposition précédente, $C^\infty \subseteq \operatorname{adh}(\mathbb{R}_+ C)$, mais on n'a pas l'égalité en général. Par exemple si $C = \{1\} \subseteq \mathbb{R}$, $C^\infty = \{0\}$ alors que $\operatorname{adh} \mathbb{R}_+ C = \mathbb{R}_+$.

Le corollaire suivant exprime à sa manière qu'un ensemble convexe fermé est borné si, et seulement si, il ne contient pas de *demi-droite*, c'est-à-dire d'ensemble de la forme $x + \mathbb{R}_+ d$, où $x \in \mathbb{E}$ et $d \in \mathbb{E} \setminus \{0\}$.

Corollaire 2.8 (cône asymptotique d'un convexe borné) *Soit C un ensemble convexe fermé non vide. Alors C est borné si, et seulement si, $C^\infty = \{0\}$.*

Démonstration. D'après la proposition 2.7 (i), si C est borné, $C^\infty = \{0\}$. Inversement, si C n'est pas borné, il existe une suite de points $\{x_k\} \subseteq C$ telle que $t_k := \|x_k\| \to \infty$ et $x_k/\|x_k\| \to d \neq 0$. D'après la proposition 2.7 (i), $d \in C^\infty$. □

D'autres propriétés du cône asymptotique sont données dans la proposition ci-dessous, dont la démonstration est proposée à l'exercice 2.9 ; pour d'autres propriétés, voir les exercices 2.20 et 2.21 (cas d'un polyèdre convexe).

Proposition 2.9 (calcul de cône asymptotique)

1) (cône) *K est une cône convexe fermé si, et seulement si, $K^\infty = K$.*
2) (inclusion) *Si C_1 et C_2 sont deux convexes fermés non vides, on a*
$$C_1 \subseteq C_2 \quad \Longrightarrow \quad C_1^\infty \subseteq C_2^\infty.$$
3) (intersection) *Si $\{C_i\}_{i \in I}$ est une famille quelconque d'ensembles convexes fermés, d'intersection non vide, on a*
$$(\cap_{i \in I} C_i)^\infty = \cap_{i \in I} C_i^\infty. \tag{2.7}$$
4) (produit cartésien) *Si C_1 (resp. C_2) est un convexe fermé non vide d'un espace vectoriel $\mathbb{E}_1$ (resp. $\mathbb{E}_2$), alors $C_1 \times C_2$ est un convexe fermé non vide de $\mathbb{E}_1 \times \mathbb{E}_2$ et*
$$(C_1 \times C_2)^\infty = C_1^\infty \times C_2^\infty.$$
5) (préimage linéaire) *Soient $\mathbb{E}$ et $\mathbb{F}$ deux espaces vectoriels de dimension finie, $A : \mathbb{E} \to \mathbb{F}$ une application linéaire et C un convexe fermé non vide de $\mathbb{F}$ tel que $\mathcal{R}(A) \cap C \neq \varnothing$. Alors l'image réciproque $A^{-1}(C)$ de C par A est un convexe fermé de $\mathbb{E}$ et*
$$[A^{-1}(C)]^\infty = A^{-1}(C^\infty).$$
6) (image linéaire) *Soient $\mathbb{E}$ et $\mathbb{F}$ deux espaces vectoriels de dimension finie, $A : \mathbb{E} \to \mathbb{F}$ une application linéaire et C un convexe fermé non vide de $\mathbb{E}$. Alors $A(C)$ est convexe et*
$$\operatorname{adh}(A(C^\infty)) \subseteq (\operatorname{adh} A(C))^\infty.$$

Si, de plus, $\mathcal{N}(A)\cap C^\infty$ est un sous-espace vectoriel, alors $A(C)$ est fermé et

$$A(C^\infty) = A(C)^\infty. \tag{2.8}$$

2.2.5 Faces et points extrêmes

Définitions 2.10 (face, arête) Soit C un convexe. On dit que $F \subseteq C$ est une *face* de C si F est convexe et si tout segment $[x, y]$ de C tel que $]x, y[$ intersecte F est entièrement dans F. On dit qu'une face de C est *propre* si elle est différente de C. Une face de C dont l'enveloppe affine est de dimension un est appelée une *arête*. □

Une partie convexe F d'un convexe C sera donc une face de C si l'on peut écrire

$$\forall\, x, y \in C, \quad \forall\, t \in]0, 1[: \quad (1-t)x + ty \in F \quad \Longrightarrow \quad x, y \in F. \tag{2.9}$$

On peut aussi ne considérer que le cas $t = \frac{1}{2}$:

$$\forall\, x, y \in C : \quad \frac{1}{2}(x + y) \in F \quad \Longrightarrow \quad x, y \in F. \tag{2.10}$$

Une intersection quelconque de faces étant une face, on peut parler de la plus petite face de C contenant une partie $A \subseteq C$, que l'on appelle la *face engendrée* par A. C'est donc l'intersection de toutes les faces de C contenant A. On la note

$$F(A) := \bigcap \{F : F \text{ est une face de } C \text{ contenant } A\}.$$

On notera $F(x)$ la face engendrée par le singleton $\{x\} \subseteq C$.

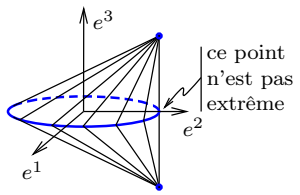
Définition 2.11 (point extrême) Un *point extrême* d'un convexe C est une face de C réduite à un seul point. L'ensemble des points extrêmes de C est noté

$$\text{ext}(C). \qquad \square$$

Un point extrême est donc caractérisé par le fait qu'il ne peut pas s'écrire comme la demi-somme de deux points *distincts* de C ou encore par le fait que $C \setminus \{x\}$ est encore convexe.

Nécessairement $\text{ext}(C) \subseteq \partial C$ (la frontière de C). Par ailleurs, l'ensemble des points extrêmes n'est pas nécessairement un fermé comme le montre l'exemple suivant :

$$\begin{aligned} P &:= \{x \in \mathbb{R}^3 : x_1^2 + x_2^2 = 1,\ x_3 = 0\} \\ &\quad \cup \{(0, 1, 1)\} \cup \{(0, 1, -1)\}, \\ C &:= \text{co}\, P. \end{aligned}$$



En effet, $\text{ext}(C) = P \setminus \{(0, 1, 0)\}$, qui n'est pas fermé.

Tout ensemble convexe fermé non vide n'a pas nécessairement un point extrême (l'espace $\mathbb{E}$ lui-même, qui est convexe, n'en a pas). En réalité, un ensemble convexe a un point extrême si, et seulement si, il ne contient pas de droite, c'est-à-dire d'ensemble de la forme $x+\mathbb{R}d$, où $x \in \mathbb{E}$ et $d \in \mathbb{E}\setminus\{0\}$; nous verrons cela plus loin (proposition 2.35).

Voici quelques propriétés permettant de calculer $\mathrm{ext}(C)$, dont la démonstration est proposée à l'exercice 2.10.

Proposition 2.12 (calcul de points extrêmes)

1) (produit cartésien) *Si C_1 (resp. C_2) est un convexe non vide d'un espace vectoriel $\mathbb{E}_1$ (resp. $\mathbb{E}_2$), alors*

$$\mathrm{ext}(C_1 \times C_2) = (\mathrm{ext}C_1) \times (\mathrm{ext}C_2).$$

2) (image linéaire) *Soient $\mathbb{E}$ et $\mathbb{F}$ deux espaces vectoriels, $A : \mathbb{E} \to \mathbb{F}$ une application linéaire injective, C un convexe non vide de $\mathbb{E}$ et $x \in C$. Alors*

$$x \in \mathrm{ext}(C) \qquad \Longleftrightarrow \qquad Ax \in \mathrm{ext}(A(C)).$$

Sous l'hypothèse d'injectivité de A, cette équivalence est identique à l'identité $A(\mathrm{ext}(C)) = \mathrm{ext}(A(C))$.

On notera qu'aucune des implications du point 2 n'est vraie en l'absence d'injectivité de A. Ainsi, l'implication "$\Rightarrow$" n'a pas lieu si C est la boule unité fermée de $\mathbb{R}^2$ pour la norme ℓ_1, A est le projecteur orthogonal sur la première coordonnée et $x = (0, 1)$ (0 n'est pas un point extrême de $[-1, 1]$). L'implication "$\Leftarrow$" n'a pas lieu si $C = \mathbb{R}$ et $A : \mathbb{R} \to \mathbb{R}$ est l'application nulle (0 est un point extrême du singleton $\{0\}$ mais aucun point de C n'est extrême).

2.3 Aspects topologiques

Soit P une partie d'un espace vectoriel $\mathbb{E}$. En analyse convexe, on rencontre souvent des ensembles convexes dont l'intérieur dans $\mathbb{E}$ est vide : c'est le cas des faces d'un polyèdre convexe (celles différentes du polyèdre lui-même). Il est donc utile d'introduire la notion d'*intérieur relatif* d'un ensemble P (non nécessairement convexe), qui est son intérieur dans son enveloppe affine $\mathrm{aff}\, P$, munie de la topologie induite de celle de $\mathbb{E}$. On le note[3] P° ou $\mathrm{intr}\, P$. On a

$$P^{\circ} \equiv \mathrm{intr}\, P = \{x \in P : \text{il existe } r > 0 \text{ tel que } (B(x, r) \cap \mathrm{aff}\, P) \subseteq P\}.$$

[3] La notation P° nous est propre. Elle est formée du symbole $^{\circ}$ qui rappelle qu'il s'agit d'un intérieur (dans la notation française) et de $^{-}$ qui évoque l'enveloppe affine (plate) dans lequel celui-ci est pris. La notation anglo-saxonne est $\mathrm{ri}\, P$, mais « ri » (*relative interior*) n'est pas très évocateur en français, si bien que nous avons préféré $\mathrm{intr}\, P$ comme expression littérale.

On dit qu'une partie P de $\mathbb{E}$ est un *ouvert relatif* de $\mathbb{E}$ ou est *relativement ouverte* dans $\mathbb{E}$ si $P^{\circ} = P$.

La *frontière relative* d'un ensemble $P \subseteq \mathbb{E}$ est l'ensemble des points de son adhérence $\overline{P}$ qui ne sont pas dans son intérieur relatif P°. On la note

$$\partial_{\mathrm{rel}} P = \overline{P} \setminus P^{\circ}, \tag{2.11}$$

où $\overline{P}$ désigne l'adhérence de P (dans $\mathbb{E}$ ou $\operatorname{aff} P$; c'est la même chose, car $\operatorname{aff} P$ est un fermé ; il n'y a donc pas de notion d'adhérence relative).

Si P est réduit à un point, $\operatorname{aff} P = P$ et donc $P^{\circ} = P$, puis $\partial_{\mathrm{rel}} P = \varnothing$. Par ailleurs, on gardera à l'esprit que l'opération $(\cdot)^{\circ}$ ne préserve pas l'inclusion, même pour des ensembles convexes :

$$C_1,\, C_2 \text{ convexes et } C_1 \subseteq C_2 \qquad \not\Longrightarrow \qquad C_1^{\circ} \subseteq C_2^{\circ}.$$

Par exemple, $C_1 := \{0\} \subseteq C_2 := [0,1] \subseteq \mathbb{R}$, mais $C_1^{\circ} = \{0\} \not\subseteq C_2^{\circ} =]0,1[$. Toutefois, pour des parties P_1 et P_2 ayant la même enveloppe affine, on a bien sûr :

$$\left.\begin{array}{r} P_1 \subseteq P_2 \\ \operatorname{aff} P_1 = \operatorname{aff} P_2 \end{array}\right\} \quad \Longrightarrow \quad P_1^{\circ} \subseteq P_2^{\circ}. \tag{2.12}$$

Proposition 2.13 (intérieur relatif non vide) *Soit C un convexe non vide. Alors son intérieur relatif C° est non vide et* $\operatorname{aff}(C^{\circ}) = \operatorname{aff} C$.

Démonstration. Soit $n := \dim(\operatorname{aff} C)$. On peut supposer que $n \geqslant 1$, car le résultat est évident pour $n = 0$ (C est un singleton). On peut trouver $n+1$ points x_0, x_1, …, x_n de C qui sont affinement indépendants. Alors tout point $x \in \operatorname{aff} C$ peut s'écrire

$$x = x_0 + \sum_{i=1}^{n} \alpha_i (x_i - x_0),$$

avec des coefficients α_1, …, $\alpha_n \in \mathbb{R}$ déterminés de manière unique. L'application

$$\varphi : \alpha = (\alpha_1, \ldots, \alpha_n) \mapsto x = \sum_{i=0}^{n} \alpha_i x_i, \quad \text{avec } \alpha_0 := 1 - \textstyle\sum_{i=1}^{n} \alpha_i$$

est un homéomorphisme de $\mathbb{R}^n \to \operatorname{aff} C$. Dès lors, l'image par φ de l'ouvert

$$\Omega_n := \left\{ \alpha \in \mathbb{R}^n : \sum_{i=1}^{n} \alpha_i < 1,\ \alpha_i > 0 \text{ pour tout } i \right\}$$

est un ouvert dans $\operatorname{aff} C$. Il suffit maintenant de constater que l'ouvert $\varphi(\Omega_n)$ de $\operatorname{aff} C$ est inclus dans C (car $\varphi(\alpha)$ est une combinaison convexe des points x_0, …, x_n de C lorsque $\alpha \in \Omega_n$), pour conclure que l'intérieur relatif de C est non vide.

Les incllusions $\varphi(\Omega_n) \subseteq C^{\circ} \subseteq C$ conduisent à $\operatorname{aff} \varphi(\Omega_n) \subseteq \operatorname{aff} C^{\circ} \subseteq \operatorname{aff} C$. On obtient alors $\operatorname{aff} \varphi(\Omega_n) = \operatorname{aff} C^{\circ}$ en notant que $\operatorname{aff} \varphi(\Omega_n) = \operatorname{aff} C$ parce que $\varphi(\Omega_n)$ est un ouvert dans $\operatorname{aff} C$. □

Le lemme suivant est souvent utile pour traiter des questions d'intériorité relative des ensembles convexes.

Lemme 2.14 (critères d'intériorité relative) *Soit C un convexe non vide. Alors*

$$x \in C^{\oplus} \text{ et } y \in \overline{C} \quad \Longrightarrow \quad [x, y[\subseteq C^{\oplus}.$$

Dès lors, pour un point $x \in \mathbb{E}$, on a

$$x \in C^{\oplus} \quad \Longleftrightarrow \quad \forall\, x_0 \in C,\ \exists\, t > 1 : \quad (1-t)x_0 + tx \in C. \tag{2.13}$$

Démonstration. Considérons la première partie lorsque $x \neq y$ (sinon $[x, y[= \varnothing$ et il n'y a rien à démontrer). On peut aussi supposer que $\operatorname{aff} C = \mathbb{E}$. Soient $t \in [0, 1[$ et B la boule-unité ouverte. Il faut montrer que $z_t = (1-t)x + ty \in C^{\oplus}$ ou que $z_t + \varepsilon B \in C$ pour un $\varepsilon > 0$ assez petit. Comme $y \in \overline{C}$, quel que soit $\varepsilon > 0$, $y \in C + \varepsilon B$. Alors

$$\begin{aligned} z_t + \varepsilon B &= (1-t)x + ty + \varepsilon B \\ &\subseteq (1-t)x + t(C + \varepsilon B) + \varepsilon B \qquad [y \in C + \varepsilon B] \\ &= tC + (1-t)\left(x + \frac{1+t}{1-t}\,\varepsilon B\right) \\ &\subseteq tC + (1-t)C \qquad [\text{pour } \varepsilon > 0 \text{ assez petit}] \\ &= C \qquad [C \text{ convexe}]. \end{aligned} \tag{2.14}$$

En (2.14), on a utilisé le fait que $x \in C^{\oplus}$.

Venons-en à la démonstration de (2.13). Si $x \in C^{\oplus}$, il existe un $\varepsilon > 0$ tel que $B(x, \varepsilon) \cap \operatorname{aff} C \subseteq C$. Alors, pour un $t > 1$ proche de 1, $z := (1-t)y + tx \in B(x, \varepsilon)$, quel que soit $y \in \mathbb{E}$. Comme $z \in \operatorname{aff} C$ lorsque $y \in \operatorname{aff} C$, on en déduit que $z \in C$.

Réciproquement, comme C est non vide, il en est de même de $C^{\oplus}$ et on peut choisir $y \in C^{\oplus}$. De deux choses l'une. Soit $y = x$, auquel cas $x \in C^{\oplus}$ et c'est terminé. Soit $y \neq x$. Dans ce cas, par hypothèse, il existe un $t > 1$ tel que $z := (1-t)y + tx \in C$. Alors $x \in [y, z[$, si bien que par la première partie du lemme, $x \in C^{\oplus}$. □

Nous avons écrit la résultat précédent comme une équivalence facilement mémorisable. En réalité, la preuve a montré que l'on avait en fait les implications plus fortes suivantes :

$$\begin{aligned} x \in C^{\oplus} \quad &\Longrightarrow \quad \forall\, x_0 \in \operatorname{aff} C,\ \exists\, t > 1 : \quad (1-t)x_0 + tx \in C, \\ x \in C^{\oplus} \quad &\Longleftarrow \quad \exists\, x_0 \in C^{\oplus},\ \exists\, t > 1 : \quad (1-t)x_0 + tx \in C. \end{aligned} \tag{2.15}$$

Parfois, on ne connait pas $C^{\oplus}$, mais on cherche à le spécifier ; dans ce cas, l'implication droite-gauche du critère (2.13) est plus utile que (2.15).

On pourra s'entraîner à utiliser ce lemme en démontrant le corollaire suivant et en faisant les exercices 2.12 et 2.13 (ceux-ci utilisent la proposition 2.16 qui est également fondamentale).

On dit que x est un *point absorbant* de C si pour tout $d \in \mathbb{E}$, il existe un $t > 0$ tel que $x + td \in C$. La troisième propriété ci-dessous fait le lien avec le cône des directions admissibles.

Corollaire 2.15 (convexe avec point absorbant) *Soit C un convexe non vide d'un espace vectoriel $\mathbb{E}$. Alors les propriétés suivantes sont équivalentes :*
(i) $x \in C^\circ$,
(ii) x *est un point absorbant de* C,
(iii) $\mathbb{E} = \mathbb{R}_+(C - x)$.

Proposition 2.16 *Soit C un convexe non vide d'un espace vectoriel $\mathbb{E}$. Alors*
1) *son intérieur relatif $C^\oplus$ est convexe,*
2) *son adhérence $\overline{C}$ est un convexe non vide et* $\operatorname{aff}\overline{C} = \operatorname{aff} C$.
3) $\overline{C^\oplus} = \overline{C}$ *et* $(\overline{C})^\oplus = C^\oplus$.
4) *Les ensembles $C^\oplus$, C et $\overline{C}$ ont la même enveloppe affine, le même intérieur relatif, la même adhérence et la même frontière relative.*

Démonstration. 1) Si $x, y \in C^\oplus$, $[x, y] \subseteq C^\oplus$ d'après le lemme 2.14 ; donc $C^\oplus$ est convexe.

2) Soient x, $y \in \overline{C}$ et $t \in [0, 1]$; il faut montrer que $(1-t)x + ty \in \overline{C}$. Il existe alors des suites $\{x_k\} \to x$ et $\{y_k\} \to y$ avec x_k et $y_k \in C$. Le point $(1-t)x_k + ty_k \in C$ (par convexité de C) et converge vers $(1-t)x + ty$, qui appartient donc à $\overline{C}$.

Certainement $\operatorname{aff} C \subseteq \operatorname{aff}\overline{C}$ (car $C \subseteq \overline{C}$). Réciproquement, $\overline{C} \subseteq \operatorname{aff} C$ (car $\operatorname{aff} C$ est un fermé contenant C), donc $\operatorname{aff}\overline{C} \subseteq \operatorname{aff} C$.

La démonstration des autres propriétés est proposée à l'exercice 2.13. □

On trouvera à l'exercice 2.13 d'autres informations sur la topologie des ensembles convexes, qui sont déduites des résultats ci-dessus. Voici pour terminer cette section quelques règles de calcul d'intérieurs relatifs et d'adhérences. Ces règles sont fondamentales. En particulier, les conditions pour avoir l'égalité joueront un rôle essentiel en optimisation convexe où elles prennent le nom de *conditions de qualification*, lesquels permettent d'écrirent des conditions d'optimalité de problèmes d'optimisation (chapitre 4).

Proposition 2.17 (calcul d'intérieurs relatifs et d'adhérences) *Soient $\mathbb{E}$, $\mathbb{E}_1$, $\mathbb{E}_2$ et $\mathbb{F}$ des espaces vectoriels et $A : \mathbb{E} \to \mathbb{F}$ une application linéaire.*
1) (produit cartésien) *Si $C_1 \subseteq \mathbb{E}_1$ et $C_2 \subseteq \mathbb{E}_2$ sont deux convexes, alors*

$$(C_1 \times C_2)^\oplus = C_1^\oplus \times C_2^\oplus \qquad \text{et} \qquad \overline{C_1 \times C_2} = \overline{C_1} \times \overline{C_2}.$$

2) (intersection) *Si $(C_i)_{i\in I}$ est une famille de convexes de $\mathbb{E}$ telle que $\cap_{i\in I} C_i^\oplus \neq \varnothing$, alors*

$$(\cap_{i\in I} C_i)^\oplus \subseteq \cap_{i\in I} C_i^\oplus \qquad \text{et} \qquad \overline{\cap_{i\in I} C_i} = \cap_{i\in I} \overline{C_i}, \tag{2.16}$$

avec égalité à gauche si I est fini.
3) (image linéaire) *Si $C \subseteq \mathbb{E}$ est convexe, alors*

$$(A(C))^\circ = A(C^\circ) \qquad \textit{et} \qquad \overline{A(C)} \supseteq A(\overline{C}), \tag{2.17}$$

avec égalité à droite si $A(\overline{C})$ est fermé.

4) (préimage linéaire) *Si $C \subseteq \mathbb{F}$ est convexe et si l'image réciproque $A^{-1}(C^\circ) \neq \varnothing$, alors*

$$(A^{-1}(C))^\circ = A^{-1}(C^\circ) \qquad \textit{et} \qquad \overline{A^{-1}(C)} = A^{-1}(\overline{C}).$$

5) (multiplication) *Si $C \subseteq \mathbb{E}$ est convexe et $\alpha \in \mathbb{R}$, alors*

$$(\alpha\, C)^\circ = \alpha\, C^\circ \qquad \textit{et} \qquad \overline{\alpha\, C} = \alpha\, \overline{C}.$$

6) (somme) *Si $C_1 \subseteq \mathbb{E}$ et $C_2 \subseteq \mathbb{E}$ sont deux convexes, alors*

$$(C_1 + C_2)^\circ = C_1^\circ + C_2^\circ \qquad \textit{et} \qquad \overline{C_1 + C_2} \supseteq \overline{C_1} + \overline{C_2},$$

avec égalité à droite si $\overline{C_1} + \overline{C_2}$ est fermé.

Démonstration. Nous utiliserons les résultats énoncés à l'exercice 2.13.

1) La proposition 2.1 montre que $\operatorname{aff}(C_1 \times C_2) = (\operatorname{aff} C_1) \times (\operatorname{aff} C_2)$. Par définition de l'intérieur relatif et de la topologie produit, $(x_1, x_2) \in (C_1 \times C_2)^\circ$ si, et seulement si, il existe des ouverts $\theta_i \in \mathbb{E}_i$ contenant x_i $(i = 1, 2)$, tels que

$$\begin{aligned} C_1 \times C_2 &\supseteq (\theta_1 \times \theta_2) \cap \operatorname{aff}(C_1 \times C_2) \\ &= (\theta_1 \times \theta_2) \cap (\operatorname{aff} C_1 \times \operatorname{aff} C_2) \\ &= (\theta_1 \cap \operatorname{aff} C_1) \times (\theta_2 \cap \operatorname{aff} C_2). \end{aligned}$$

Ceci revient à dire que $(\theta_i \cap \operatorname{aff} C_i) \subseteq C_i$ ou encore que $x_i \in C_i^\circ$. La relation $\operatorname{adh}(C_1 \times C_2) = (\operatorname{adh} C_1) \times (\operatorname{adh} C_2)$ est vraie, même si les C_i ne sont pas convexes.

2) Soit $x \in \cap_{i\in I}\overline{C_i}$. Comme il existe un $x_0 \in \cap_{i\in I} C_i^\circ$, $x_t := (1-t)x_0 + tx \in C_i^\circ$, pour tout $t \in [0,1[$ et tout $i \in I$ (lemme 2.14). Alors $x_t \in \cap_{i\in I} C_i^\circ$ pour tout $t \in [0,1[$. À la limite en $t \uparrow 1$, on trouve que $x = x_1$ est dans l'adhérence de $\cap_{i\in I} C_i^\circ$. Enfin, puisqu'une intersection de fermés est fermée, on trouve finalement

$$\cap_{i\in I}\overline{C_i} \subseteq \overline{\cap_{i\in I} C_i^\circ} \subseteq \overline{\cap_{i\in I} C_i} \subseteq \cap_{i\in I}\overline{C_i}.$$

On a donc égalité partout, ce qui démontre l'identité sur les adhérences. On en déduit aussi que $\cap_{i\in I} C_i^\circ$ et $\cap_{i\in I} C_i$ ont la même adhérence et donc le même intérieur relatif (point 4 de l'exercice 2.13), ce qui conduit à l'inclusion sur les intérieurs relatifs :

$$(\cap_{i\in I} C_i)^\circ = (\cap_{i\in I} C_i^\circ)^\circ \subseteq \cap_{i\in I} C_i^\circ.$$

Supposons à présent que I est fini et que $x \in \cap_{i\in I} C_i^\circ$. Soit $y \in \cap_{i\in I} C_i^\circ$. Pour tout $i \in I$, on peut trouver un $t_i > 1$ tel que $y_{t_i} = (1-t)y + tx \in C_i$. Comme I est fini, il existe un $t > 1$ tel que $y_t \in \cap_{i\in I} C_i$. Ceci montre que $x \in (\cap_{i\in I} C_i)^\circ$ (lemme 2.14).

3) L'inclusion $\overline{A(C)} \supseteq A(\overline{C})$ découle de la continuité de A et a lieu même sans la convexité de C et sans la linéarité de A. De même pour le cas où il y a égalité. Pour démontrer l'identité sur les intérieurs relatifs, on observe d'abord que

$$\overline{A(C^\ominus)} \supseteq A(\overline{C^\ominus}) = A(\overline{C}) \supseteq A(C) \supseteq A(C^\ominus).$$

On en déduit que $A(C)$ et $A(C^\ominus)$ ont la même adhérence, si bien qu'ils ont aussi le même intérieur relatif (exercice 2.13) : $(A(C))^\ominus = (A(C^\ominus))^\ominus \subseteq A(C^\ominus)$. Pour montrer l'inclusion inverse, on utilise le lemme 2.14. Soient $y \in A(C^\ominus)$ et $y_0 \in (A(C))^\ominus \subseteq A(C)$. Alors il existe $x \in C^\ominus$ et $x_0 \in C$ tels que $y = Ax$ et $y_0 = Ax_0$. Il existe aussi un $t > 1$ tel que $(1-t)x_0 + tx \in C$. En appliquant A : $(1-t)y_0 + ty \in A(C)$. On en déduit que $y \in (A(C))^\ominus$.

4) On introduit deux ensembles dans $\mathbb{E} \times \mathbb{F}$:

$$X := \mathbb{E} \times C \qquad \text{et} \qquad L := \{(x, Ax) : x \in \mathbb{E}\}.$$

Par hypothèse, il existe un $x \in \mathbb{E}$ tel que $Ax \in C^\ominus$. Comme $X^\ominus = \mathbb{E} \times C^\ominus$, ceci s'écrit $X^\ominus \cap L \neq \varnothing$. On note aussi $\mathrm{P}_\mathbb{E} : \mathbb{E} \times \mathbb{F} \to \mathbb{E} : (x, y) \mapsto x$ le projecteur sur $\mathbb{E}$. Observons que

$$\begin{aligned} A^{-1}(C) &= \{x \in \mathbb{E} : Ax \in C\} = \mathrm{P}_\mathbb{E}(X \cap L) \\ A^{-1}(C^\ominus) &= \{x \in \mathbb{E} : Ax \in C^\ominus\} = \mathrm{P}_\mathbb{E}((\mathbb{E} \times C^\ominus) \cap L) = \mathrm{P}_\mathbb{E}((X \cap L)^\ominus) \\ A^{-1}(\overline{C}) &= \{x \in \mathbb{E} : Ax \in \overline{C}\} = \mathrm{P}_\mathbb{E}((\mathbb{E} \times \overline{C}) \cap L) = \mathrm{P}_\mathbb{E}(\overline{X \cap L}). \end{aligned}$$

Comme $\mathrm{P}_\mathbb{E}$ est une application linéaire, on en déduit que $(A^{-1}(C))^\ominus = A^{-1}(C^\ominus)$ et que $\mathrm{adh}(A^{-1}(C)) \supseteq A^{-1}(\mathrm{adh}\, C)$. L'égalité a lieu dans la dernière relation, car $\mathrm{adh}(A^{-1}(C)) \subseteq A^{-1}(\mathrm{adh}\, C)$, par continuité de A.

5) Les identités sont claires si $\alpha = 0$. Si $\alpha \neq 0$, l'application $x \mapsto \alpha x$ est linéaire bijective et les identités sont des conséquences du point 3 (pour l'adhérence on utilise $\overline{\alpha\, C} \supseteq \alpha\, \overline{C}$ et $\alpha\, \overline{C} = \alpha\, \overline{(\alpha\, C)/\alpha} \supseteq \overline{\alpha\, C}$).

6) On applique le point 3 avec $C = C_1 \times C_2 \subseteq \mathbb{E}^2$ et $A : \mathbb{E}^2 \to \mathbb{E}$ définie par $A(x_1, x_2) = x_1 + x_2$; puis le point 1. □

Les identités de la proposition 2.17 sont fondamentales et les hypothèses qu'elles requièrent se retrouveront sous des formes diverses dans d'autres résultats d'analyse convexe. Pour s'imprégner des raisons de leur présence, voici quelques contre-exemples, tous très simples.

- Point 2. L'identité de gauche est fausse pour les ensembles $[0, 1]$ et $[1, 2]$ dans $\mathbb{R}$: on trouve $\{1\} \not\subseteq \varnothing$. Elle n'a pas lieu avec égalité pour la collection infinie d'intervalles $C_i = [0, 1+1/i]$ avec i entier non nul : on a $(\cap_i C_i)^\ominus = [0, 1]^\ominus =]0, 1[$, tandis que $\cap_i C_i^\ominus =]0, 1]$. L'identité de droite est fausse pour les ensembles $]0, 1[$ et $]1, 2[$ dans $\mathbb{R}$: on trouve $\varnothing \neq \{1\}$.
- Point 3. L'exercice 2.23 donne un exemple d'image de cône convexe fermé par une application linéaire, qui n'est pas fermée. On peut donc ne pas avoir égalité à droite. Cependant on a toujours

$$\overline{A(C)} = \overline{A(\overline{C})}. \tag{2.18}$$

- Point 4. On considère l'application linéaire $A : \mathbb{R} \to \mathbb{R} : x \mapsto 0$. L'identité de gauche n'a pas lieu si $C = [0, 1]$: on trouve $(A^{-1}(C))^\ominus = \mathbb{R}$ et $A^{-1}(C^\ominus) = \varnothing$. Celle de droite n'a pas lieu si $C =]0, 1]$: on trouve $\mathrm{adh}\, A^{-1}(C) = \varnothing$ tandis que $A^{-1}(\mathrm{adh}\, C) = \mathbb{R}$.

- Point 6. La somme des convexes fermés $C_1 = \mathbb{R}_+ \times \{0\} \subseteq \mathbb{R}^2$ et $C_2 = \{(x_1, x_2) : x_2 \geqslant \exp(x_1)\} \subseteq \mathbb{R}^2$ est l'ensemble $\mathbb{R} \times \mathbb{R}_{++}$, qui n'est pas un fermé.

Il ne faut pas manquer de contempler la beauté de l'identité $(C_1+C_2)^\circ = C_1^\circ + C_2^\circ$ (point 6), qui n'a plus lieu si l'on remplace les intérieurs *relatifs* par des intérieurs (par exemple dans le cas où $C_1 = [0,1] \times \{0\} \subseteq \mathbb{R}^2$ et $C_2 = \{0\} \times [0,1] \subseteq \mathbb{R}^2$). Cette sympathique relation devrait contribuer à convaincre de l'utilité du concept d'intériorité relative.

2.4 Polyèdre convexe

2.4.1 Représentations primale et duale

On rappelle qu'un *polyèdre convexe* d'un espace vectoriel $\mathbb{E}$ est un ensemble P de la forme

$$P = \{x \in \mathbb{E} : Ax \leqslant b\}, \tag{2.19}$$

où $A : \mathbb{E} \to \mathbb{R}^m$ est une application linéaire ($m \in \mathbb{N}$; si $m = 0$, $P = \mathbb{E}$), $b \in \mathbb{R}^m$ et l'inégalité $Ax \leqslant b$ se lit composante par composante dans $\mathbb{R}^m$: $(Ax)_i \leqslant b_i$, pour tout $i \in [1\,:m]$. Si l'ensemble se présente avec des égalités linéaires $Cx = d$, on pourra se ramener à la forme (2.19) en les remplaçant par deux inégalités opposées $Cx \leqslant d$ et $-Cx \leqslant -d$. Géométriquement, un polyèdre convexe est donc l'intersection d'un nombre *fini* de demi-espaces fermés de $\mathbb{E}$.

Un *polytope* est un polyèdre convexe borné.

Si $\mathbb{E}$ est de dimensions finie, il n'y a pas de restriction à supposer que $\mathbb{E} = \mathbb{R}^n$ et que A est une matrice $m \times n$ (il suffit de se donner une base de $\mathbb{E}$). Par ailleurs, dans certaines circonstances (par exemple en optimisation linéaire, voir le chapitre 17), il est avantageux de représenter un polyèdre de $\mathbb{R}^n$ sous la forme dite *standard* suivante :

$$P = \{x \in \mathbb{R}^n : Ax = b,\ x \geqslant 0\}. \tag{2.20}$$

Il n'y a aucune perte de généralité dans cette représentation. Tout polyèdre de la forme (2.19) se représente sous la forme (2.20) en introduisant des *variables d'écart* $s \in \mathbb{R}^m$ et en décomposant $x = u - v$, avec $u, v \in \mathbb{R}^n_+$:

$$\{(u, v, s) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m : (A\ -A\ I)(u^\mathsf{T}\ v^\mathsf{T}\ s^\mathsf{T})^\mathsf{T} = b,\ (u, v, s) \geqslant 0\}.$$

Il faut toutefois noter qu'il faut alors travailler dans un espace de dimension plus grande. D'autre part, un polyèdre de la forme (2.20) s'écrit comme en (2.19) en remplaçant $Ax = b$ par les deux inégalités $Ax \leqslant b$ et $-Ax \leqslant -b$.

Les représentations (2.19) et (2.20) d'un polyèdre sont dites *duales*, car elles font intervenir des applications linéaires (éléments du dual de $\mathbb{E}$). Dans la *représentation primale* d'un polyèdre convexe, on écrit celui-ci comme une somme de combinaisons convexe et conique d'éléments de $\mathbb{E}$. Si on se donne $x^1, \ldots, x^p \in \mathbb{E}$ et $y^1, \ldots, y^q \in \mathbb{E}$ (on peut prendre les $y^j = 0$), l'ensemble

$$P = \mathrm{co}\{x^1, \ldots, x^p\} + \mathrm{cone}\{y^1, \ldots, y^q\} \tag{2.21}$$

est un polyèdre convexe. C'est ce qu'affirme la proposition 2.23, que nous ne démontrerons qu'à la fin de cette section. Le choix de la représentation dépendra du type de résultat que l'on veut obtenir (voir les exercices 2.20 et 2.21).

Des caractérisations primale et duale de la bornitude ou de la réduction à un point d'un polyèdre décrit par (2.19) sont données à la section 2.5.5.

2.4.2 Image linéaire

Cette section s'intéresse à la description de l'image d'un polyèdre convexe par une application linéaire. Nous allons montrer que si le polyèdre convexe est écrit sous forme duale (2.19), son image par une application linéaire est également un un polyèdre convexe écrit sous forme duale (2.19). Cette affirmation serait plus simple à démontrer si l'on savait déjà qu'un ensemble de la forme (2.21) avec des $x^1, \dots, x^p \in \mathbb{E}$ et des $y^1, \dots, y^q \in \mathbb{E}$, est un polyèdre convexe (cette démonstration est proposée à l'exercice 2.21), mais nous utiliserons précisément le résultat de la proposition 2.18 ci-dessous pour établir ce fait !

La démonstration que nous proposons repose sur l'*élimination de Fourier* [219 ; 1827] : trouver $x \in \mathbb{R}^n$ tel que

$$Ax \leqslant b. \tag{2.22}$$

Elle s'apparente à l'*élimination gaussienne* dans le sens où à chaque étape elle transforme le système d'inégalités affines courant en un système d'inégalités affines *équivalent*, mais avec une inconnue de moins. Cependant, contrairement à l'élimination gaussienne, qui, à chaque étape, *enlève* une équation, l'élimination de Fourier *ajoute* de nombreuses inégalités, si bien que cette dernière n'est guère exploitable en pratique car le nombre d'opérations, peut croître très rapidement au cours des éliminations.

Voyons comment la méthode de Fourier élimine x_n ; elle procède de la même manière pour les autres variables. On note a_{ij} l'élément (i, j) de A. Le vecteur $x = (x_1, \dots, x_n)$ est solution du système d'inégalités (2.22) si, et seulement si,

- pour tout i_1 tel que $a_{i_1 n} > 0$, on a $x_n \leqslant \dfrac{1}{a_{i_1 n}} \left(b_{i_1} - \displaystyle\sum_{j=1}^{n-1} a_{i_1 j} x_j \right)$,
- pour tout i_2 tel que $a_{i_2 n} < 0$, on a $x_n \geqslant \dfrac{1}{a_{i_2 n}} \left(b_{i_2} - \displaystyle\sum_{j=1}^{n-1} a_{i_2 j} x_j \right)$,
- pour tout i_3 tel que $a_{i_3 n} = 0$, on a $0 \leqslant b_{i_3} - \displaystyle\sum_{j=1}^{n-1} a_{i_3 j} x_j$.

On peut à présent éliminer x_n : x est solution de (2.22) si, et seulement si, pour tout i_1 tel que $a_{i_1 n} > 0$, pour tout i_2 tel que $a_{i_2 n} < 0$ et pour tout i_3 tel que $a_{i_3 n} = 0$, $\tilde{x} := (x_1, \dots, x_{n-1})$ vérifie

$$\frac{1}{a_{i_2 n}}\left(b_{i_2}-\sum_{j=1}^{n-1} a_{i_2 j}x_j\right) \leqslant \frac{1}{a_{i_1 n}}\left(b_{i_1}-\sum_{j=1}^{n-1} a_{i_1 j}x_j\right) \tag{2.23a}$$

$$0 \leqslant b_{i_3}-\sum_{j=1}^{n-1} a_{i_3 j}x_j. \tag{2.23b}$$

et x_n est choisi dans l'intervalle défini par le jeu d'inégalités dans (2.23a) (la borne inférieure de l'intervalle est le maximum des quantités à gauche des inégalités pour les i_2 tels que $a_{i_2 n} < 0$ et la borne droite de l'intervalle est le minimum des quantités à droite des inégalités pour les i_1 tels que $a_{i_1 n} > 0$; la borne gauche est bien inférieure à la borne droite, car le premier jeu d'inégalités dans (2.23a) doit être vérifié pour tous les indices i_1 et i_2 désignés précédemment). On obtient ainsi un nouveau système d'inégalités affines ne portant plus que sur $\tilde{x}$. On peut ensuite éliminer x_{n-1} jusqu'à x_2, pour finalement obtenir un dernier jeu d'inégalités ne portant que sur x_1. Le nombre d'inégalités croît rapidement puisqu'après la première élimination, on peut en avoir jusqu'à $(\lfloor m/2 \rfloor)^2$, après la seconde jusqu'à $(\lfloor (\lfloor m/2 \rfloor)^2/2 \rfloor)^2$, *etc.*

L'opération permettant de passer de x vérifiant (2.22) à $\tilde{x}$ vérifiant (2.23a) est une projection cartésienne, puisque $\tilde{x} = (x_1, \ldots, x_{n-1})$ vérifie (2.23a) si, et seulement si, il existe un x_n tel que $x = (x_1, \ldots, x_{n-1}, x_n)$ vérifie (2.22). Dès lors, si l'on introduit le projecteur

$$\Phi_n : \mathbb{R}^n \to \mathbb{R}^{n-1} : (x_1, \ldots, x_n) \mapsto (x_1, \ldots, x_{n-1}),$$

l'ensemble défini par le jeu d'inégalités (2.23a) n'est autre que $\Phi_n(P)$, où P est le polyèdre convexe défini par le jeu d'inégalités (2.22). On vient donc de démontrer que $\Phi_n(P)$ est un polyèdre convexe. La proposition suivante utilise ce fait pour aller un peu plus loin.

Proposition 2.18 (image d'un polyèdre convexe par une application linéaire) *Soient $\mathbb{E}$ et $\mathbb{F}$ deux espaces vectoriels, $T : \mathbb{E} \to \mathbb{F}$ une application linéaire et P un polyèdre convexe de $\mathbb{E}$. Alors $T(P)$ est un polyèdre convexe de $\mathbb{F}$.*

Démonstration. On peut supposer que $\mathbb{E}$ et $\mathbb{F}$ sont les espaces vectoriels $\mathbb{R}^n$ et $\mathbb{R}^p$. Nous donnons ci-dessous une démonstration dans le cas où P est de la forme $\{x \in \mathbb{R}^n : Ax \leqslant b\}$, avec une matrice A de type $m \times n$ et $b \in \mathbb{R}^m$. Le fait que le résultat est aussi valable pour des polyèdres convexes sous représentation primale (2.21) sera une conséquence de la proposition 2.23, qui n'utilisera que le résultat démontré ci-dessous.

Observons d'abord que le projecteur $\Pi_{p,p+n} : (y, x) \in \mathbb{R}^p \times \mathbb{R}^n \mapsto y \in \mathbb{R}^p$ peut s'écrire

$$\Pi_{p,p+n} = \Phi_{p+1} \circ \cdots \circ \Phi_{p+n}.$$

où les Φ_k réalisent l'élimination de Fourier de la k-ième variable. Ensuite, il suffit d'observer que

$$\begin{aligned} T(P) &= \{Tx : Ax \leqslant b\} \\ &= \{y : y = Tx \text{ et } Ax \leqslant b\} \\ &= \Pi_{p,p+n}\{(y, x) : y = Tx \text{ et } Ax \leqslant b\} \\ &= (\Phi_{p+1} \circ \cdots \circ \Phi_{p+n})\{(y, x) : y = Tx \text{ et } Ax \leqslant b\}. \end{aligned}$$

Comme l'ensemble $\{(y,x) : y = Tx \text{ et } Ax \leqslant b\}$ est un polyèdre convexe et comme chaque élimination de Fourier transforme un polyèdre convexe en un autre, on voit que $T(P)$ est un polyèdre convexe. □

Remarque 2.19 La proposition précédente a un corollaire que nous utiliserons quelque fois : pour une matrice A et des vecteurs x de dimensions appropriées,

$$\{Ax : x \geqslant 0\} \text{ est un cône convexe fermé.} \tag{2.24}$$

Le fait que ce soit un cône convexe peut se vérifier facilement en utilisant les définitions. Le caractère fermé résulte du fait que l'orthant positif est un polyèdre convexe ; donc, par la proposition précédente, il en est de même de $\{Ax : x \geqslant 0\}$ qui, comme polyèdre convexe, est fermé. On notera, qu'en général, l'image par une application linéaire d'un cône convexe fermé quelconque n'est pas nécessairement fermée (exercice 2.23). D'autre part, il faut se garder de confondre (2.24) avec le fait, trivial lui, que l'ensemble $\{x : Ax \geqslant 0\}$ est un cône convexe fermé non vide (il est fermé parce qu'il est l'image réciproque de l'orthant positif par l'application continue A). L'affirmation (2.24) peut se voir avec un point de vue plus géométrique : l'enveloppe conique d'un nombre *fini* de vecteurs (les colonnes de A) est un fermé. De ce point de vue, la seconde partie de la proposition 2.23 est un peu plus riche. □

2.4.3 Optimisation linéaire

Voici une conséquence importante de la proposition 2.18, qui concerne l'existence de solution d'un problème d'optimisation linéaire, c'est-à-dire d'un problème qui consiste à minimiser une fonction linéaire sur un polyèdre convexe P d'un espace euclidien $(\mathbb{E}, \langle\cdot,\cdot\rangle)$. La fonction linéaire peut s'écrire $x \in \mathbb{E} \mapsto \langle c, x\rangle \in \mathbb{R}$, où $c \in \mathbb{E}$. Dès lors, un *problème d'optimisation linéaire* est de la forme

$$(P_L) \qquad \inf_{x \in P} \langle c, x\rangle.$$

Proposition 2.20 (existence de solution de (P_L)) *Le problème (P_L) ci-dessus a une solution si, et seulement si, sa valeur optimale est finie.*

Démonstration. [⇒] Si $\bar{x}$ est solution de (P_L), la valeur optimale du problème s'écrit $\langle c, \bar{x}\rangle$ et est donc finie (c'est un élément de $\mathbb{R}$).

[⇐] Par la proposition 2.18, l'ensemble $I := \{\langle c, x\rangle : x \in P\}$ est un polyèdre convexe de $\mathbb{R}$, donc un intervalle fermé.

Si la valeur optimale $\mathrm{val}(P_L)$ de (P_L) est finie, $P \neq \varnothing$ (voir (1.1)) et on peut trouver une suite minimisante du problème, c'est-à-dire une suite $\{x_k\} \subseteq P$ telle que $\langle c, x_k\rangle \to \mathrm{val}(P_L)$ (par définition de l'infimum). Clairement, $\langle c, x_k\rangle \in I$. Comme I est fermé et $\langle c, x_k\rangle \to \mathrm{val}(P_L) \in \mathbb{R}$, on a nécessairement que $\mathrm{val}(P_L) \in I$, ce qui veut dire qu'il existe un point $\bar{x} \in P$ tel que $val(P_L) = \langle c, \bar{x}\rangle$. Ce point $\bar{x}$ est manifestement une solution de (P_L). □

Une autre manière, sans doute plus expressive, d'exprimer le résultat de la proposition 2.20 est de dire que *le problème* (P_L) *a une solution si, et seulement si, il est réalisable* (sa valeur optimale n'est pas $+\infty$, voir (1.1)) *et borné* (sa valeur optimale n'est pas $-\infty$).

2.4.4 Faces et sommets

On rappelle qu'une arête d'un convexe est une face de dimension 1 et qu'un point extrême est une face de dimension 0. Un point extrême d'un polyèdre convexe P est aussi appelé un *sommet*. On note $\mathrm{ext}(P)$ l'ensemble des sommets de P. La figure 2.2 illustre les différents types de faces d'un polyèdre convexe dans $\mathbb{R}^2$.

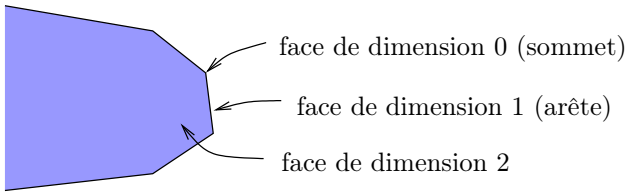


Fig. 2.2. Faces d'un polyèdre convexe

Si P est donné dans la représentation primale (2.21),

$$\mathrm{ext}(P) \subseteq \{x^1, \ldots, x^p\},$$

sans que l'on ait nécessairement l'égalité (exercice 2.21). La représentation duale (2.20) fournit une description plus précise des sommets d'un polyèdre. Pour $x \in \mathbb{R}^n$, on note

$$I^+(x) := \{i : x_i > 0\} \quad \text{et} \quad I^0(x) := \{i : x_i = 0\}.$$

Proposition 2.21 (faces et sommets d'un polyèdre convexe) *Considérons un polyèdre P représenté par (2.20) et un point $x \in P$, dont on note $F(x)$ la face qu'il engendre. Alors*

$$\dim F(x) = \dim \mathcal{N}\left(A_{:I^+(x)}\right),$$

où $A_{:I^+(x)}$ est la matrice formée des colonnes A^j de A avec indices $j \in I^+(x)$. En particulier, $x \in P$ est un sommet de P si, et seulement si, les colonnes $\{A^j : x_j > 0\}$ de A sont linéairement indépendantes.

Démonstration. Simplifions les écritures en notant $B := I^+(x)$ et $N := I^0(x)$. On note alors $A_{:B}$ (resp. $A_{:N}$) la matrice extraite de A formée de ses colonnes avec indices dans B (resp. N). On fait de même pour les vecteurs de $\mathbb{R}^n$, si bien que $x_B > 0$ et $x_N = 0$. Le résultat découle des équivalences suivantes, dans lesquelles $k \geqslant 1$:

$\dim \mathcal{N}(A_{:B}) \geqslant k$,

$\iff$ on peut trouver des directions $d^1, \ldots, d^k$ linéairement indépendantes telles que, pour tout i, $Ad^i = 0$ et $d^i_N = 0$,

$\iff$ on peut trouver des directions $d^1, \ldots, d^k$ linéairement indépendantes telles que, pour tout $\alpha = (\alpha_1, \ldots, \alpha_k) \in \mathbb{R}^k$ voisin de zéro, on a $x + \sum_{i=1}^k \alpha_i d^i \in P$,

$\boxed{\Rightarrow}$ on a $A(x + \sum_i \alpha_i d^i) = b$, $(x + \sum_i \alpha_i d^i)_N = 0$ et $(x + \sum_i \alpha_i d^i)_B = x_B + \sum_i \alpha_i d^i_B \geqslant 0$ pour α petit (car $x_B > 0$),

$\boxed{\Leftarrow}$ en prenant $\alpha = (0, \ldots, 0, \alpha_i, 0, \ldots, 0)$ avec $|\alpha_i|$ petit non nul, la relation $A(x + \sum_i \alpha_i d^i) = b$ implique que $Ad^i = 0$; d'autre part, la relation $(x + \sum_i \alpha_i d^i)_N \geqslant 0$ implique que $\alpha_i d^i_N \geqslant 0$ pour $|\alpha_i|$ petit, ou encore que $d^i_N = 0$,

$\iff \dim F(x) \geqslant k$.

On en déduit que $\dim \mathcal{N}(A_{:B}) = \dim F(x)$. □

Corollaire 2.22 *Un polyèdre non vide représenté par (2.20) a au plus $\binom{n}{r}$ sommets, où r est le rang de A.*

Démonstration. Soit

$$\mathcal{B} := \{B : A_{:B} \text{ est injective et } \mathcal{R}(A_{:B}) = \mathcal{R}(A)\}.$$

Observons que les $A_{:B}$ avec $B \in \mathcal{B}$ sont des sous-matrices extraites de A, de type $m \times r$, si bien que $|\mathcal{B}| \leqslant \binom{n}{r}$. D'autre part, si l'on introduit l'application $\varphi : \mathcal{B} \to \mathbb{R}^n$ qui à $B \in \mathcal{B}$ fait correspondre l'unique $x \in \mathbb{R}^n$ tel que $Ax = b$ et $x_i = 0$ pour $i \notin B$, on voit que tout sommet est de la forme $\varphi(B)$, avec $B \in \mathcal{B}$ (proposition 2.21 et théorème B.1 de la base incomplète). Le nombre de sommets est donc $\leqslant |\varphi(\mathcal{B})| \leqslant |\mathcal{B}| \leqslant \binom{n}{r}$. □

Le nombre de sommets d'un polyèdre convexe représenté sous forme standard peut augmenter exponentiellement avec n. Par exemple, le polyèdre convexe de $\mathbb{R}^{2n}$ écrit sous forme standard (on note e un vecteur dont toutes les composantes valent 1)

$$\{(x, y) \in \mathbb{R}^{2n} : x + y = e,\ x \geqslant 0,\ y \geqslant 0\} \text{ a } 2^n \text{ sommets.}$$

En effet, ce polyèdre P est une récriture sous forme standard du polyèdre $P_0 = \{x \in \mathbb{R}^n : 0 \leqslant x \leqslant e\}$. On voit facilement que (x, y) est un sommet de P si, et seulement si, x est un sommet de P_0 et $y = e - x$, si bien que P et P_0 ont le même nombre de sommets. Comme $P_0 = \frac{1}{2}(B_\infty + 1)$ est un translaté-contracté de la boule unité de $\mathbb{R}^n$ pour la norme ℓ_∞, il a 2^n sommets (exercice 2.17) ; il en est donc de même de P.

Un *cône polyédrique convexe* est un cône qui est aussi un polyèdre convexe. Comme polyèdre convexe, il peut s'écrire sous la forme standard (2.20) et pour que celui-ci soit un cône, il faut que $b = 0$ (prendre un point de la forme tx avec $t \downarrow 0$). Sous sa forme standard, un cône polyédrique convexe s'écrit donc comme suit :

$$K = \{x \in \mathbb{R}^n : Ax = 0,\ x \geqslant 0\}. \tag{2.25}$$

Ce cône est d'ailleurs le cône asymptotique du polyèdre convexe (2.20) : $K = P^\infty$ (exercice 2.20). Évidemment 0 est un sommet de K et on démontre facilement qu'il n'y en a pas d'autre. Quant aux arêtes de K, nécessairement de la forme $\mathbb{R}_+ x$ avec $x \neq 0$, elles sont repérées par le fait que $\dim \mathcal{N}(A_{:I^+(x)}) = 1$ (proposition 2.21).

2.4.5 Équivalence des représentations

Proposition 2.23 (résolution d'un polyèdre convexe) *Un polyèdre convexe P écrit sous la forme standard (2.20) peut aussi s'écrire*

$$P = \mathrm{co}\{x^1, \ldots, x^p\} + \mathrm{cone}\{y^1, \ldots, y^q\}, \tag{2.26}$$

où les vecteurs $x^1, \ldots, x^p$ $(p \geqslant 1)$ sont les sommets de P et les demi-droites $\mathbb{R}_+ y^1, \ldots, \mathbb{R}_+ y^q$ $(q \geqslant 1)$ associées à des y^j non nuls sont les arêtes de P^∞ ; en particulier, un polyèdre convexe écrit sous la forme standard (2.20) a au moins un sommet. Inversement, tout ensemble de la forme (2.26) est un polyèdre convexe.

Démonstration. 1) Montrons qu'un polynôme convexe P écrit sous forme standard (2.20) peut s'écrire sous la forme (2.26).

[$\supseteq$] Soient des $x^i \in P$ (c'est-à-dire $Ax^i = b$ et $x^i \geqslant 0$), des $y^j \in P^\infty$ (c'est-à-dire $Ay^j = 0$ et $y^j \geqslant 0$), des $\alpha_i \geqslant 0$ vérifiant $\sum_i \alpha_i = 1$ et des $\beta_j \geqslant 0$. Alors $x = \sum_i \alpha_i x^i + \sum_j \beta_j y^j$ est clairement dans P (c'est-à-dire $Ax = b$ et $x \geqslant 0$).

[$\subseteq$] Commençons par montrer que $P \subseteq \mathrm{co}\{x^1, \ldots, x^p\} + P^\infty$, où les x^i sont les sommets de P, en nombre $p \geqslant 1$. Soit $x \in P$. On procède par récurrence sur $\dim \mathcal{N}(A_{:I^+(x)})$.

- Si $\dim \mathcal{N}(A_{:I^+(x)}) = 0$, x est un sommet (proposition 2.21) et le résultat est clairement démontré.
- Supposons que le résultat soit démontré pour les $x \in P$ tels que $\dim \mathcal{N}(A_{:I^+(x)}) = 0, \ldots, k-1$, avec $k \geqslant 1$, et démontrons le lorsque x vérifie $\dim \mathcal{N}(A_{:I^+(x)}) = k$. Comme $k \geqslant 1$, on peut trouver une direction d telle que

$$d \neq 0, \qquad d_{I^0(x)} = 0 \qquad \text{et} \qquad Ad = 0.$$

  On envisage trois cas, selon le signe des composantes de d, illustrés à la figure 2.3.

  - Cas 1 : $d \leqslant 0$. Prenons le plus grand $\alpha' > 0$ tel que $x' := x + \alpha' d \geqslant 0$. Alors $x' \in P$ et $I^+(x') \subsetneqq I^+(x)$. Remarquons que

$$\mathcal{N}(A_{:I^+(x')}) \times 0_{I^+(x)\setminus I^+(x')} \subseteq \mathcal{N}(A_{I^+(x)})$$
$$d_{I^+(x)} \in \mathcal{N}(A_{:I^+(x)}), \text{ mais } d_{I^+(x)} \notin \mathcal{N}(A_{:I^+(x')}) \times 0_{I^+(x)\setminus I^+(x')}.$$

    La dernière non-appartenance vient du fait que d a nécessairement une composante non nulle avec indice $i \in I^+(x) \setminus I^+(x')$ (un indice $i \in I^+(x)$ tel que $x'_i = 0$, alors que $x_i > 0$). Dès lors $\dim \mathcal{N}(A_{:I^+(x')}) = \dim \mathcal{N}(A_{:I^+(x')}) \times$

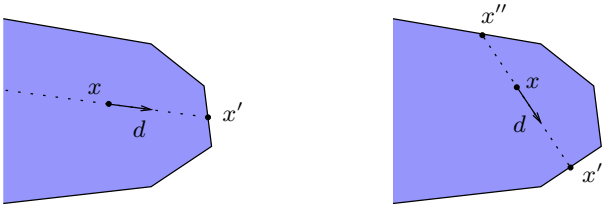


Fig. 2.3. Deux cas de la démonstration de la proposition 2.23

$0_{I^+(x)\setminus I^+(x')} \leqslant \dim \mathcal{N}(A_{:I^+(x)}) - 1$. Par récurrence, on a $x' \in \mathrm{co}\{x^1, \dots, x^p\} + P^\infty$. Comme d'autre part, $-d \in P^\infty$, le résultat est démontré pour $x = x' + \alpha'(-d)$ qui est alors aussi dans $\mathrm{co}\{x^1, \dots, x^p\} + P^\infty$.

- Cas 2 : $d \geqslant 0$. On s'y prend de la même manière, avec $\alpha'' > 0$ le plus grand possible tel que $x'' := x - \alpha'' d \geqslant 0$.
- Cas 3 : $d \not\leqslant 0$ et $d \not\geqslant 0$. On définit $x' := x + \alpha' d$ et $x'' := x - \alpha'' d$, avec $\alpha' > 0$ et $\alpha'' > 0$ les plus grands possibles tels que $x' \geqslant 0$ et $x'' \geqslant 0$. On a

$$x = \frac{\alpha''}{\alpha' + \alpha''} x' + \frac{\alpha'}{\alpha' + \alpha''} x''.$$

Comme ci-dessus, x' et $x'' \in \mathrm{co}\{x^1, \dots, x^p\} + P^\infty$, et donc il en est de même de x qui est une combinaison convexe de x' et de x''.

Il reste à montrer que $P^\infty \subseteq \mathrm{cone}\{y^1, \dots, y^q\}$, où les y^j non nuls engendrent les arêtes de P^∞. Ce résultat est clair si $P^\infty = \{0\}$. Dans le cas contraire, on considère le polyèdre convexe $P_1 := \{y : Ay = 0, e^\mathsf{T} y = 1, y \geqslant 0\}$, qui est obtenu en normalisant les éléments de P^∞ (e est un vecteur dont les éléments valent tous 1) et qui est non vide. Clairement $P_1^\infty = \{0\}$ et donc P_1 est borné (corollaire 2.8). Par ce que l'on vient de démontrer, $P_1 = \mathrm{co}\{y^1, \dots, y^q\}$, où les y^j sont les sommets de P_1. Alors $P^\infty = \mathbb{R}_+ P_1 = \mathrm{cone}\{y^1, \dots, y^q\}$. Il reste à montrer que les y^j engendrent les arêtes de P^∞ ou encore, d'après la proposition 2.21, que $\dim \mathcal{N}(A_{:I^+(y^j)}) = 1$. D'une part $\dim \mathcal{N}(A_{:I^+(y^j)}) \geqslant 1$ car $(y^j)_{I^+(y^j)}$ est non nul et dans $\mathcal{N}(A_{:I^+(y^j)})$ (du fait que $e^\mathsf{T} y^j = 1$ et $Ay^j = 0$). D'autre part, comme y^j est un sommet de P_1,

$$\mathcal{N}\begin{pmatrix} A_{:I^+(y^j)} \\ e^\mathsf{T}_{I^+(y^j)} \end{pmatrix} = \{0\}.$$

Ceci implique que $\dim \mathcal{N}(A_{:I^+(y^j)}) \leqslant 1$ (exercice B.2).

2) Montrons la réciproque. Un ensemble P de la forme (2.26) peut aussi s'écrire $T(Q)$, où

$$Q = \left\{ (\alpha, \beta) \in \mathbb{R}^p \times \mathbb{R}^q : \alpha \geqslant 0,\ \sum_{i=1}^p \alpha_i = 1,\ \beta \geqslant 0 \right\}$$

et $T : \mathbb{R}^p \times \mathbb{R}^q : (\alpha, \beta) \mapsto \sum_i^p \alpha_i x^i + \sum_j \beta_j y^j$. Clairement, l'ensemble Q est un polyèdre convexe (écrit sous forme standard) et T est une application linéaire. Par la proposition 2.18 (démontrée pour les polyèdres convexes écrits sous forme standard), $P = T(Q)$ est un polyèdre convexe. □

Le passage d'une représentation (primale ou duale) à l'autre est une opération difficile. Un algorithme a été proposé par Motzkin et al. [445] et redécouvert par Chernikova [119, 120, 121]. La complexité pire cas de ces transformations est exponentielle en la dimension [32, 88], ce qui les rend inutilisables dès que les dimensions ne sont pas très petites.

2.5 Opérations

2.5.1 Projection

Soit $\mathbb{E}$ un espace vectoriel euclidien, dont le produit scalaire est noté $\langle \cdot, \cdot \rangle$ et la norme associée $\| \cdot \|$. On sait bien ce qu'est le projeté d'un point x de $\mathbb{E}$ sur un sous-espace vectoriel $\mathbb{F}$ de $\mathbb{E}$: c'est le point $\bar{x} \in \mathbb{F}$ tel que $x - \bar{x}$ est orthogonal à tout vecteur de $\mathbb{F}$; mais c'est aussi le point $\bar{x} \in \mathbb{F}$ le plus proche de x. Cette dernière caractérisation peut être utilisée pour définir le projeté d'un point sur certaines parties de $\mathbb{E}$ qui ne sont pas des sous-espaces vectoriels.

On appelle *projection orthogonale* sur une partie C de $\mathbb{E}$, l'opération qui à un point $x \in \mathbb{E}$ associe toute solution éventuelle des problèmes équivalents (ils ont les mêmes solutions) suivants :

$$\mathrm{d}_C(x) := \inf_{y \in C} \|y - x\| \qquad \text{et} \qquad \mathrm{e}_C(x) := \inf_{y \in C} \frac{1}{2}\|y - x\|^2. \tag{2.27}$$

La valeur optimale $\mathrm{d}_C(x)$ du premier problème est la *distance de x à C* et $\mathrm{e}_C(x) = \frac{1}{2}[\mathrm{d}_C(x)]^2$. Toute solution de ces problèmes est appelée *projeté* de x sur C. Clairement, x est le projeté de x si $x \in C$. La proposition 2.25 ci-dessous nous montrera que ces problèmes ont une solution et une seule si C est un convexe fermé non vide le résultat ne serait pas vrai pour une norme arbitraire, qui n'est pas associée à un produit scalaire ; le qualificatif *orthogonal* est là pour rappeler cette propriété de la norme utilisée dans la projection. On note alors $\mathrm{P}_C(x)$ cette solution et l'opérateur $\mathrm{P}_C : \mathbb{E} \to C$ est alors appelé le *projecteur orthogonal* sur C.

Donnons en premier lieu des propriétés qui caractérisent un projeté orthogonal. La première, (2.28a), est la plus souvent utilisée ; elle le sera déjà dans la démonstration du résultat d'existence et d'unicité de la proposition 2.25.

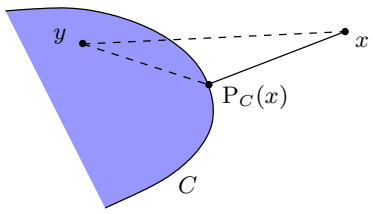


Fig. 2.4. Illustration de la proposition 2.24

Proposition 2.24 (caractérisation d'un projeté orthogonal) *Soit C un convexe non vide de $\mathbb{E}$. Un point $\bar{x} \in C$ est un projeté de $x \in \mathbb{E}$ sur C si, et seulement si, l'une des conditions équivalentes suivantes est vérifiée :*

$$\forall y \in C, \quad \langle y - \bar{x}, \bar{x} - x \rangle \geqslant 0, \tag{2.28a}$$
$$\forall y \in C, \quad \langle y - \bar{x}, y - x \rangle \geqslant 0, \tag{2.28b}$$
$$\forall y \in C, \quad \langle y - x, \bar{x} - x \rangle \geqslant \|\bar{x} - x\|^2. \tag{2.28c}$$

Démonstration. [$\bar{x}$ *est caractérisé par* (2.28a)] Soit $y \in C$. Par convexité de C, $z_t = \bar{x} + t(y - \bar{x}) \in C$ pour tout $t \in [0, 1]$. Comme $\bar{x}$ est solution de (2.27), on a

$$\|\bar{x} - x\|^2 \leqslant \|z_t - x\|^2 = \|\bar{x} - x\|^2 + 2t\langle \bar{x} - x, y - \bar{x} \rangle + t^2 \|y - \bar{x}\|^2.$$

Après retranchement de $\|\bar{x} - x\|^2$, division par $t > 0$ et passage à la limite lorsque $t \downarrow 0$, on obtient (2.28a). Inversement, en écrivant $y - x = (y - \bar{x}) + (\bar{x} - x)$ et en utilisant (2.28a), on obtient

$$\|y - x\|^2 = \|y - \bar{x}\|^2 + 2\langle y - \bar{x}, \bar{x} - x \rangle + \|\bar{x} - x\|^2 \geqslant \|\bar{x} - x\|^2.$$

Donc $\bar{x}$ est solution de (2.27).

[(2.28a) $\Leftrightarrow$ (2.28b)] Si (2.28a) est vérifiée, on a pour $y \in C$:

$$\langle y - \bar{x}, y - x \rangle = \|y - \bar{x}\|^2 + \langle y - \bar{x}, \bar{x} - x \rangle \geqslant 0,$$

c'est-à-dire (2.28b). Inversement, si $y \in C$ et $t \in]0, 1]$, on a $z_t := \bar{x} + t(y - \bar{x}) \in C$. Donc, d'après (2.28b) :

$$\begin{aligned} 0 &\leqslant \langle z_t - \bar{x}, z_t - x \rangle \\ &= t\langle y - \bar{x}, (\bar{x} - x) + t(y - \bar{x}) \rangle \\ &= t\langle y - \bar{x}, \bar{x} - x \rangle + t^2 \|y - \bar{x}\|^2. \end{aligned}$$

En divisant par $t > 0$, puis en faisant tendre t vers 0, on trouve (2.28a).

[(2.28a) $\Leftrightarrow$ (2.28c)] L'équivalence s'obtient en observant que $\langle y - x, \bar{x} - x \rangle = \langle y - \bar{x}, \bar{x} - x \rangle + \|\bar{x} - x\|^2$. □

Avec l'apport du chapitre 4, on comprendra que (2.28a) n'est autre que la condition nécessaire et suffisante d'optimalité (4.13). Ce point de vue permettrait de simplifier la démonstration.

Proposition 2.25 (existence et unicité du projeté orthogonal) *Soient C une partie convexe fermée non vide d'un espace euclidien $\mathbb{E}$ et x un point de $\mathbb{E}$. Alors il existe un unique élément $\bar{x} \in C$ tel que*

$$\forall y \in C, \quad \|\bar{x} - x\| \leqslant \|y - x\|.$$

DÉMONSTRATION. Il s'agit de montrer que le problème (2.27) a une solution et une seule.

Existence : elle découle du corollaire 1.4 ; C est un fermé non vide et le critère de (2.27) est coercif.

Unicité : soient $\bar{x}_1$ et $\bar{x}_2$ deux projetés de x sur C ; par (2.28a), on a

$$\langle \bar{x}_2 - \bar{x}_1, \bar{x}_1 - x \rangle \geqslant 0 \quad \text{et} \quad \langle \bar{x}_1 - \bar{x}_2, \bar{x}_2 - x \rangle \geqslant 0\,;$$

en sommant ces deux inégalités, on trouve que $\bar{x}_1 = \bar{x}_2$. □

Proposition 2.26 (propriétés de la projection) *Le projecteur orthogonal* P_C *sur un convexe fermé non vide* C *est fortement monotone dans le sens suivant*

$$\forall\, x_1, x_2 \in \mathbb{E} : \quad \langle \mathrm{P}_C(x_2) - \mathrm{P}_C(x_1), x_2 - x_1 \rangle \geqslant \| \mathrm{P}_C(x_2) - \mathrm{P}_C(x_1) \|^2 \tag{2.29a}$$

et contractant dans le sens suivant

$$\forall\, x_1, x_2 \in \mathbb{E} : \quad \| \mathrm{P}_C(x_1) - \mathrm{P}_C(x_2) \| \leqslant \|x_1 - x_2\|. \tag{2.29b}$$

DÉMONSTRATION. 1) D'après la caractérisation (2.28a) du projeté, on a

$$\begin{aligned} \langle \mathrm{P}_C(x_2) - \mathrm{P}_C(x_1), \mathrm{P}_C(x_1) - x_1 \rangle &\geqslant 0, \\ \langle \mathrm{P}_C(x_1) - \mathrm{P}_C(x_2), \mathrm{P}_C(x_2) - x_2 \rangle &\geqslant 0. \end{aligned}$$

En sommant ces inégalités, on trouve

$$\langle \mathrm{P}_C(x_2) - \mathrm{P}_C(x_1), x_2 - x_1 - (\mathrm{P}_C(x_2) - \mathrm{P}_C(x_1)) \rangle \geqslant 0.$$

On en déduit (2.29a).

2) On a ensuite

$$\begin{aligned} \| \mathrm{P}_C(x_2) - \mathrm{P}_C(x_1) \|^2 &\leqslant \langle \mathrm{P}_C(x_2) - \mathrm{P}_C(x_1), x_2 - x_1 \rangle \qquad [(2.29a)] \\ &\leqslant \| \mathrm{P}_C(x_2) - \mathrm{P}_C(x_1) \| \, \|x_2 - x_1\| \qquad [\text{Cauchy-Schwarz}]. \end{aligned}$$

On en déduit (2.29b). □

Si la projection P_C est lipschitzienne par (2.29), elle n'est généralement pas différentiable. Elle a toutefois des dérivées directionnelles en un point de C, mais pas nécessairement en un point n'appartenant pas à C (voir Kruskal [366 ; 1969] pour un contre-exemple dans $\mathbb{R}^3$ et Shapiro [547 ; 1994] pour un contre-exemple simple dans $\mathbb{R}^2$).

2.5.2 Cône normal et hyperplan d'appui

Normale

Définition 2.27 Soient $(\mathbb{E}, \langle\cdot,\cdot\rangle)$ un espace euclidien et $C \subseteq \mathbb{E}$ un ensemble convexe *fermé*. Le *cône normal* à C en un point $x \in C$ est l'ensemble noté et défini par

$$\mathrm{N}_x C \equiv \mathrm{N}_C(x) := \{\nu \in \mathbb{E} : \langle \nu, y - x\rangle \leqslant 0,\ \forall y \in C\}.$$

Par convention, $\mathrm{N}_C(x) = \varnothing$ si $x \notin C$. Les éléments de $\mathrm{N}_C(x)$ sont appelés des *normales* à C en x. □

Proposition 2.28 *Soient C un convexe fermé non vide d'un espace euclidien $\mathbb{E}$. Alors $\mathrm{N}_c C$ est un cône convexe fermé non vide.*

Démonstration. Le cône normal peut en effet s'écrire comme une intersection de demi-espaces fermés (des convexes) :

$$\mathrm{N}_x C = \bigcap_{y \in C} \{\nu \in \mathbb{E} : \langle \nu, y - x\rangle \leqslant 0\}.$$

□

Il est clair que l'orthogonal $\mathbb{E}_0^\perp$ du sous-espace vectoriel $\mathbb{E}_0$ parallèle à $\operatorname{aff} C$ est dans $\mathrm{N}_C(x)$; ce dernier contient donc un élément non nul si $\operatorname{aff} C \neq \mathbb{E}$. Par ailleurs, il est également clair que $0 \in \mathrm{N}_C(x) \cap \mathbb{E}_0$. La proposition suivante en dit un peu plus : si $x \in \partial_{\mathrm{rel}} C$, la frontière relative de C, $\mathrm{N}_C(x) \cap \mathbb{E}_0$ n'est pas réduit à l'élément nul. Il y a une hypothèse implicite dans le fait que $\partial_{\mathrm{rel}} C \neq \varnothing$, à savoir que C n'est ni réduit à un point (auquel cas, $\mathbb{E}_0 = \{0\}$, C n'a pas de frontière, le cône normal en ce point est $\mathbb{E}$, mais $\mathrm{N}_C(x) \cap \mathbb{E}_0$ est réduit à $\{0\}$), ni $\mathbb{E}$ tout entier (auquel cas, le cône normal en tout point est réduit à $\{0\}$).

Proposition 2.29 (existence d'une normale non nulle) *Soient C un convexe fermé non vide d'un espace euclidien $\mathbb{E}$, $\mathbb{E}_0$ le sous-espace vectoriel parallèle à $\operatorname{aff} C$ et $x \in \partial_{\mathrm{rel}} C$. Alors $\mathrm{N}_C(x) \cap \mathbb{E}_0$ contient un élément non nul.*

Démonstration. On peut supposer que $\operatorname{aff} C = \mathbb{E}$. Si x est sur la frontière de C, alors $C \neq \mathbb{E}$ et il existe une suite $\{x_k\}_{k\geqslant 1} \subseteq \mathbb{E} \setminus C$ qui converge vers x. Soit $y_k = \mathrm{P}_C(x_k)$. Comme $y_k \neq x_k$, on peut définir la vecteur unitaire $\nu_k := (x_k - y_k)/\|x_k - y_k\|$. Par caractérisation (2.28a) du projeté, on a

$$\forall\, y \in C : \quad \langle y - y_k, \nu_k\rangle \leqslant 0.$$

La suite $\{y_k\} \to x$, par la propriété de contraction (2.29) de la projection : $\|y_k - x\| \leqslant \|x_k - x\| \to 0$. Par ailleurs, la suite $\{\nu_k\}$ étant bornée, on peut en extraire une sous-suite convergente, encore notée $\{\nu_k\} \to \nu$, avec ν de norme un. En passant à la limite

dans l'inégalité ci-dessus, on obtient quel que soit $y \in C$: $\langle y - x, \nu \rangle \leqslant 0$. Ceci montre que ν est une normale non nulle. □

La démonstration des quelques règles de calcul de cônes normaux données ci-dessous est proposée à l'exercice 2.41 (la démonstration de l'égalité dans (2.30) est difficile à démontrer) et celle de la proposition 2.31 est proposée à l'exercice 2.42.

Proposition 2.30 (calcul de cône normal)
1) (intersection) *Si C_1 et C_2 sont deux ensembles convexes fermés et $x \in C_1 \cap C_2$, alors*
$$\mathrm{N}_x(C_1 \cap C_2) \supseteq \mathrm{N}_x C_1 + \mathrm{N}_x C_2, \tag{2.30}$$
avec égalité si $0 \in (C_1 - C_2)^\oplus$ ou si $C_1^\oplus \cap C_2^\oplus \neq \varnothing$.
2) (produit cartésien) *Si C_1 (resp. C_2) est un convexe fermé non vide d'un espace vectoriel $\mathbb{E}_1$ (resp. $\mathbb{E}_2$), alors $C_1 \times C_2 := \{(x_1, x_2) : x_1 \in C_1,\ x_2 \in C_2\}$ est un convexe fermé non vide de $\mathbb{E}_1 \times \mathbb{E}_2$ et en $(x_1, x_2) \in C_1 \times C_2$, on a*
$$\mathrm{N}_{(x_1,x_2)}(C_1 \times C_2) = \mathrm{N}_{x_1} C_1 \times \mathrm{N}_{x_2} C_2.$$

Proposition 2.31 (semi-continuité supérieure) *Soit $S := \{x \in \mathbb{E} : \|x\| = 1\}$ la sphère unité de $\mathbb{E}$ et $\mathrm{N}_C : \mathbb{E} \multimap \mathbb{E} : x \mapsto \mathrm{N}_C(x)$ la multifonction cône normal. Alors la multifonction $\mathrm{N}_C \cap S$ est semi-continue supérieurement.*

Hyperplan d'appui

Une notion étroitement reliée à celle de normale est celle d'hyperplan d'appui. Le résultat d'existence de la proposition 2.33 se déduit d'ailleurs directement de celui sur l'existence d'une normale non nulle (proposition 2.29).

Définition 2.32 Soient $(\mathbb{E}, \langle \cdot, \cdot \rangle)$ un espace euclidien et $C \subseteq \mathbb{E}$ un ensemble convexe. Un *hyperplan d'appui* de C est un hyperplan contenant C dans un de ses demi-espaces fermés et contenant un point de la frontière de C. C'est donc un ensemble qui doit pouvoir s'écrire $H(\xi, \alpha) := \{x \in \mathbb{E} : \langle \xi, x \rangle = \alpha\}$ avec $\xi \in \mathbb{E} \setminus \{0\}$ et $\alpha \in \mathbb{R}$ et vérifie $C \subseteq H^-(\xi, \alpha)$ et $\overline{C} \cap H(\xi, \alpha) \neq \varnothing$.

Un *hyperplan d'appui de C en $x \in \overline{C}$* est un hyperplan d'appui de C qui contient x. □

Si $\operatorname{aff} C \neq \mathbb{E}$, alors $\operatorname{aff} C$ est un hyperplan d'appui trivial, mais rarement intéressant. La proposition suivante donne dans ce cas un hyperplan d'appui souvent plus utile, pourvu que sa frontière relative $\partial_{\mathrm{rel}} C$ soit non vide.

Proposition 2.33 (existence d'un hyperplan d'appui) *Soient C un convexe non vide d'un espace euclidien $\mathbb{E}$, $\mathbb{E}_0$ le sous-espace vectoriel parallèle à* aff C *et $x \in \partial_{\mathrm{rel}} C$. Alors il existe un hyperplan d'appui $H(\xi,\alpha)$ de C en x tel que $\xi \in \mathbb{E}_0$.*

Démonstration. Comme $\operatorname{aff}\overline{C} = \operatorname{aff} C$ et $(\overline{C})^{\oplus} = C^{\oplus}$ (proposition 2.16), $\mathbb{E}_0$ est aussi l'espace vectorel parallèle au convexe fermé non vide $\overline{C}$ et $x \in \partial_{\mathrm{rel}}\overline{C} = \partial_{\mathrm{rel}} C$. Soit $\xi \in \mathrm{N}_{\overline{C}}(x) \cap \mathbb{E}_0$ la normale au convexe fermé non vide $\overline{C}$ en x, dont l'existence est assurée par la proposition 2.29. Alors $\langle \xi, y - x\rangle \leqslant 0$ pour tout $y \in C$, ce qui montre que $C \subseteq H^-(\xi,\alpha)$ avec $\alpha = \langle \xi, x\rangle$. De plus $x \in H(\xi,\alpha) \cap \partial_{\mathrm{rel}} C \subseteq H(\xi,\alpha) \cap \partial C$. Donc $H(\xi,\alpha)$ est un hyperplan d'appui de C en x, avec $\xi \in \mathbb{E}_0$. □

Existence de point extrême

Tout ensemble convexe fermé non vide n'a pas nécessairement un point extrême. On peut démontrer une condition nécessaire et suffisante pour qu'il en soit ainsi (proposition 2.35) en se fondant sur le lemme suivant.

Lemme 2.34 *Soient C un ensemble convexe et H un hyperplan d'appui de C. Alors tout point extrême de $C \cap H$ est aussi un point extrême de C.*

Démonstration. Soit z un point extrême de $C \cap H$ et supposons qu'il s'écrive $z = (1-t)x + ty$ avec $t \in]0,1[$ et x, $y \in C$. Il s'agit de montrer que $x = y = z$.

Il suffit en réalité de montrer que x, $y \in H$, car alors x, $y \in C \cap H$ et $z \in \mathrm{ext}(C \cap H)$ impliquent que $x = y = z$. Soit $H \equiv H(\xi,\alpha)$ une représentation de H telle que $C \subseteq H^-(\xi,\alpha)$. On a $\langle \xi, x\rangle \leqslant \alpha$ ou $\langle \xi, y\rangle \leqslant \alpha$, parce que x, $y \in C \subseteq H^-(\xi,\alpha)$. Mais, on doit avoir $\langle \xi, x\rangle = \alpha$ et $\langle \xi, y\rangle = \alpha$, sinon $\langle \xi, x\rangle = (1-t)\langle \xi, x\rangle + t\langle \xi, y\rangle < \alpha$ $(t \in]0,1[)$ et z n'appartiendrait pas à H. On a démontré que x, $y \in H$, comme souhaité. □

On appelle *droite*, un sous-espace affine de dimension un. C'est donc un ensemble de la forme $x + \mathbb{R}d$, où $x \in \mathbb{E}$ et $d \in \mathbb{E} \setminus \{0\}$.

Proposition 2.35 (CNS d'existence de point extrême) *Un ensemble convexe fermé non vide a un point extrême si, et seulement si, il ne contient pas de droite.*

Démonstration. Soit C le convexe fermé non vide considéré.

[⇒] On raisonne par l'absurde en supposant que C contient la droite $x_0 + \mathbb{R}d$ avec $d \neq 0$. Nécessairement $x_0 \in C$ et d et $-d \in C^{\infty}$. Alors pour tout $x \in C$, on a $x + \mathbb{R}d \in C$ (proposition 2.7(ii)). Comme $x = \frac{1}{2}((x+d) + (x-d))$, avec deux points $(x+d)$ et $(\hat{x}-d)$ distincts de C, on en déduit que $x \notin \mathrm{ext}(C)$. Donc C n'a pas de point extrême.

[⇐] On raisonne par récurrence sur la dimension de aff C (ou dimension de C), quel que soit la dimension de $\mathbb{E}$. La propriété est vraie si $\dim(\operatorname{aff} C) = 1$ car alors $C := \{x + td\colon t \in I\}$ où $x \in \mathbb{E}$, $d \in \mathbb{E} \setminus \{0\}$ et I est un intervalle fermé différent de $\mathbb{R}$, donc de la forme $[a, b]$ ou $[a, +\infty[$ ou $]-\infty, a]$. Clairement, $x + ad$ est un point extrême de $\mathbb{E}$.

Supposons maintenant que l'implication soit vraie lorsque $\dim(\operatorname{aff} C) \leqslant m < \dim \mathbb{E}$ et démontrons la lorsque $\dim(\operatorname{aff} C) = m+1$. Si C ne contient pas de droite, on peut trouver un point $x \in \partial_{\mathrm{rel}} C$ (on prend un $x_0 \in C$, une direction $d \neq 0$ et $x_0 + \mathbb{R}d$ rencontre $\partial_{\mathrm{rel}} C$). Par la proposition 2.33, on peut trouver un hyperplan d'appui H de C en x. Alors $C \cap H$ est un convexe fermé de dimension $\leqslant m$, qui ne contient pas de droite. Par récurrence, $C \cap H$ a un point extrême et, par le lemme 2.34, ce point extrême de $C \cap H$ est aussi un point extrême de C. □

2.5.3 Séparation

Dans cette section, on suppose que l'espace vectoriel $\mathbb{E}$ est de dimension finie, muni d'un produit scalaire, noté $\langle \cdot, \cdot \rangle$.

Une notion essentielle d'analyse convexe est celle de la séparation des ensembles convexes. La séparation des deux convexes se fait géométriquement dans $\mathbb{E}$ en utilisant un hyperplan $H(\xi, \alpha)$. On dit que cet hyperplan *sépare* deux convexes C_1 et C_2, ou que ceux-ci sont *séparables* par cet hyperplan, si l'on a

$$\forall x_1 \in C_1, \quad \forall x_2 \in C_2 : \quad \langle \xi, x_1 \rangle \leqslant \alpha \leqslant \langle \xi, x_2 \rangle,$$

ce qui est équivalent à l'existence d'un vecteur ξ *non nul* dans $\mathbb{E}$ tel que

$$\sup_{x_1 \in C_1} \langle \xi, x_1 \rangle \leqslant \inf_{x_2 \in C_2} \langle \xi, x_2 \rangle. \tag{2.31}$$

On dit que cet hyperplan *sépare strictement* ces deux convexes, ou que ceux-ci sont *strictement séparables* par cet hyperplan, s'il existe deux scalaires α_1 et α_2 tels que $\alpha_1 < \alpha_2$ et

$$\forall x_1 \in C_1, \quad \forall x_2 \in C_2 : \quad \langle \xi, x_1 \rangle \leqslant \alpha_1 < \alpha_2 \leqslant \langle \xi, x_2 \rangle,$$

ce qui est équivalent à l'existence d'un vecteur ξ (nécessairement non nul) dans $\mathbb{E}$ tel que

$$\sup_{x_1 \in C_1} \langle \xi, x_1 \rangle < \inf_{x_2 \in C_2} \langle \xi, x_2 \rangle. \tag{2.32}$$

La figure 2.5 illustre ces notions de séparation. On y a utilisé le produit scalaire

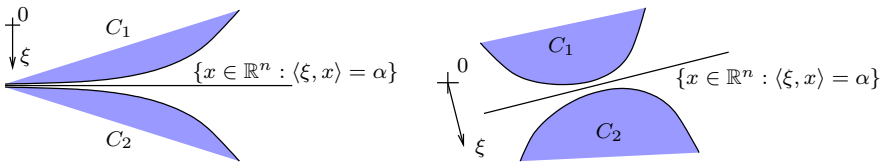


Fig. 2.5. séparation de deux convexes : non stricte à gauche et stricte à droite

euclidien sur $\mathbb{R}^2$. L'hyperplan $\{x \in \mathbb{R}^2 : \langle \xi, x\rangle = \alpha\}$ sépare les convexes C_1 et C_2, non strictement à gauche et strictement à droite. Cet hyperplan est déterminé par un vecteur ξ qui lui est orthogonal et un scalaire α.

Nous donnons ci-dessous deux résultats de séparation. Le premier (théorème 2.36) énonce des conditions équivalentes à la séparation stricte de C_1 et C_2 en termes de $C_1 - C_2$ et de la distance de C_1 à C_2. Son corollaire 2.37 énonce un certain nombre de cas où cette séparation stricte peut être réalisée. Signalons le cas où les convexes disjoints sont l'un fermé et l'autre compact, qui est utilisé lorsqu'on veut montrer qu'un ensemble convexe fermé est inclus dans un autre ensemble convexe fermé (en raisonnant par l'absurde). Le second (théorème 2.38) exprime que l'on peut séparer (non strictement cette fois) deux ensembles convexes quelconques (en dimension finie).

Théorème 2.36 (séparation stricte de convexes) *Soient C_1 et C_2 deux convexes non vides d'un espace euclidien $\mathbb{E}$. Alors les propriétés suivantes sont équivalentes :*
(i) on peut séparer C_1 et C_2 strictement,
(ii) $0 \notin \operatorname{adh}(C_1 - C_2)$,
(iii) $\inf\{\|x_1 - x_2\| : x_1 \in C_1,\ x_2 \in C_2\} > 0$.

Démonstration. [$(i) \Rightarrow (ii)$] Il existe donc $\xi \in \mathbb{E}$ tel que l'on ait (2.32). Alors

$$0 < \inf_{\substack{x_1 \in C_1 \\ x_2 \in C_2}} \langle \xi, x_1 - x_2\rangle = \inf_{x \in C_1 - C_2} \langle \xi, x\rangle = \inf_{x \in \operatorname{adh}(C_1 - C_2)} \langle \xi, x\rangle.$$

Donc $0 \notin \operatorname{adh}(C_1 - C_2)$.

[$(ii) \Rightarrow (iii)$] Le problème dans (iii) s'écrit $\inf\{\|x\| : x \in C_1 - C_2\} = \inf\{\|x\| : x \in \operatorname{adh}(C_1 - C_2)\} > 0$ grâce à (ii).

[$(iii) \Rightarrow (i)$] Comme le convexe fermé non vide $C := \operatorname{adh}(C_1 - C_2)$ ne contient pas zéro, le projeté ξ de zéro sur C est non nulle et, d'après (2.28a), vérifie pour tout $x \in C$: $0 \leqslant \langle \xi, x - \xi\rangle$. Alors pour tout $x_i \in C_i$, $0 < \|\xi\|^2 \leqslant \langle \xi, x_1 - x_2\rangle$ ou encore

$$\forall x_1 \in C_1, \quad \forall x_2 \in C_2 : \quad \langle \xi, x_2\rangle + \|\xi\|^2 \leqslant \langle \xi, x_1\rangle.$$

On en déduit (i). □

Corollaire 2.37 (séparation stricte de convexes) *On peut séparer strictement deux convexes fermés non vides disjoints C_1 et C_2 d'un espace euclidien $\mathbb{E}$ dans chacune des situations suivantes :*
1) *$C_1 - C_2$ est fermé,*
2) *$C_1^\infty \cap C_2^\infty = \{0\}$,*
3) *C_1 ou C_2 est compact,*
4) *C_1 et C_2 sont polyédriques.*

Démonstration. 1) Il suffit d'utiliser l'implication $(ii) \Rightarrow (i)$ de la proposition 2.36, après avoir constaté que $0 \notin C_1 - C_2$, car C_1 et C_2 sont disjoints.

2) D'après le point 1, il suffit de montrer que $C := C_1 - C_2$ est fermé. Soit $x_k^1 - x_k^2 \to x$, avec $\{x_k^i\} \subseteq C_i$, $i = 1, 2$. Les suites $\{x_k^i\}$ sont bornées (en effet, si $\{x_k^1\}$ n'est pas bornée, alors, avec $t_k := \|x_k^1\|$, on aurait une sous-suite de x_k^1/t_k qui convergerait vers un $d \in C_1^\infty$ non nul et pour la même sous-suite $x_k^2/t_k \to d$; donc $d \in C_1^\infty \cap C_2^\infty = \{0\}$, ce qui contredirait le fait que $d \neq 0$). En extrayant des sous-suites convergentes des $\{x_k^i\}$, on voit que $x \in C$. Donc C est fermé.

3) Si C_1 est compact, $C_1^\infty = \{0\}$ (corollaire 2.8) et le résultat se déduit du point 2.

4) Si les C_i sont polyédriques, $C_1 - C_2$ est polyédrique (exercice 2.20 ou 2.21), donc fermé. On applique alors le point 1. □

On ne peut pas se libérer de la condition $C_1^\infty \cap C_2^\infty = \{0\}$ du point 2 du corollaire précédent sans en modifier la conclusion. Par exemple, les deux convexes fermés $C_1 = \{(x, y) \in \mathbb{R}^2 : y \geqslant e^x\}$ et $C_2 = \{(x, y) \in \mathbb{R}^2 : y \leqslant 0\}$ sont disjoints dans $\mathbb{R}^2$, mais ne peuvent pas être séparés strictement (observons que dans ce cas $C_1^\infty \cap C_2^\infty = \mathbb{R}_- \times \{0\}$).

Le point 3 du corollaire précédent s'utilise souvent pour montrer qu'un point x appartient à un convexe fermé non vide C. On raisonne par l'absurde en suppossant que ce n'est pas le cas. Alors $\{x\}$ et C sont deux convexes fermés disjoints, dont l'un est fermé et l'autre (le singleton $\{x\}$) est compact. On les sépare et on démontre que cela conduit à une contradiction. Cet argument est utilisé dans la démontration des propositions 2.40 et 2.45.

Le point 3 du corollaire précédent reste vrai en dimension infinie. On peut l'obtenir à partir d'un résultat de prolongement d'application linéaire continue et c'est l'existence d'un projeté qui s'en déduit [91 ; 1983].

En dimension infinie, il y a un autre résultat permettant de séparer deux convexes, non strictement cette fois. Il faut pour cela que l'un des deux soit d'intérieur non vide. En dimension finie, cette hypothèse n'est pas nécessaire : on peut toujours séparer deux convexes quelconques disjoints par un hyperplan. C'est ce qu'affirme le théorème suivant.

Théorème 2.38 (séparation de convexes) *On peut séparer deux convexes non vides disjoints d'un espace euclidien.*

Démonstration. On considère $C := C_2 - C_1$, qui est convexe (car C_1 et C_2 sont convexes), non vide (car C_1 et C_2 sont non vides) et ne contient pas zéro (car C_1 et C_2 sont disjoints). Montrons que l'on peut trouver un vecteur non nul $\xi \in \mathbb{E}$ tel que

$$\forall\, (x_1, x_2) \in C_1 \times C_2 : \quad \langle x_2 - x_1, \xi \rangle \geqslant 0, \tag{2.33}$$

relation qui implique immédiatement que l'on peut séparer C_1 et C_2.

- Si $0 \notin \overline{C}$, on prend pour ξ le projecté de zéro sur C, qui vérifie par (2.28a) : $\langle (x_2 - x_1) - \xi, \xi - 0 \rangle \geqslant 0$, quels que soient $x_i \in C_i$. On en déduit (2.33).
- Dans le cas contraire, $0 \notin \overline{C} \setminus C \subseteq \partial_{\mathrm{rel}} C$ et on prend pour ξ l'opposé d'une normale non nulle à C en zéro (proposition 2.29), qui satisfait $\langle (x_2 - x_1) - 0, -\xi \rangle \leqslant 0$, quels que soient $x_i \in C_i$. On en déduit (2.33). □

2.5.4 Enveloppe convexe fermée

L'enveloppe convexe d'un ouvert relatif est un ouvert relatif (exercice 2.32) ; l'enveloppe convexe d'un compact est un compact (corollaire 2.4) ; mais l'enveloppe convexe d'un fermé n'est pas nécessairement fermée. C'est le cas par exemple de la partie fermée de $\mathbb{R}^2$ union de l'origine et de $\{x \in \mathbb{R}^2 : x_1 x_2 \geqslant 1\}$, dont l'enveloppe convexe $\{(0,0)\} \cup \{x \in \mathbb{R}^2 : x_1 > 0,\ x_2 > 0\}$ n'est pas fermée (voir la figure 2.6). La notion suivante a donc tout son sens.

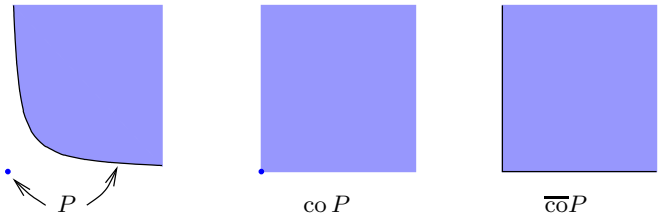


Fig. 2.6. Un ensemble P dont l'enveloppe convexe $\text{co}\, P$ diffère de l'enveloppe convexe fermée $\overline{\text{co}}P$

Soit P une partie de $\mathbb{E}$. L'intersection de convexes fermés étant un convexe fermé, on peut parler du plus petit convexe fermé contenant P, qui est donc l'intersection de tous les convexes fermés contenant P. C'est ce que l'on appelle l'*enveloppe convexe fermée* de P. On la note

$$\overline{\text{co}}P := \bigcap \{C : C \text{ est un convexe fermé contenant } P\}. \tag{2.34}$$

Évidemment, si C est un convexe fermé, $\overline{\text{co}}C = C$. La proposition suivante donne, en particulier, une autre manière de définir l'enveloppe convexe fermée : c'est aussi l'adhérence de l'enveloppe convexe, mais nous avons préféré la définition (2.34) du fait de l'idée de minimalité qu'elle contient.

Proposition 2.39 *On a*

$$P_1 \subseteq P_2 \quad \Longrightarrow \quad \overline{\text{co}}P_1 \subseteq \overline{\text{co}}P_2,$$

$$P \subseteq \text{co}\, P \subseteq \text{co}\, \overline{P} \subseteq \overline{\text{co}}P = \overline{\text{co}}\overline{P} = \overline{\text{co}\, P}.$$

Démonstration. 1) Évidemment, $P_1 \subseteq P_2 \subseteq \overline{\text{co}}P_2$. Donc, $\overline{\text{co}}P_2$ est un convexe fermé contenant P_1, si bien que $\overline{\text{co}}P_1 \subseteq \overline{\text{co}}P_2$.

2) Les deux premières inclusions sont claires. En notant que $\overline{P}$ est contenu dans tout convexe fermé contenant P, on a $\overline{P} \subseteq \overline{\text{co}}P$ (par définition de $\overline{\text{co}}P$) et donc aussi $\text{co}\, \overline{P} \subseteq \overline{\text{co}}P$ (car $\overline{\text{co}}P$ est convexe). L'égalité $\overline{\text{co}}P = \overline{\text{co}}\overline{P}$ se déduit des observations suivantes : $P \subseteq \overline{P}$ implique $\overline{\text{co}}P \subseteq \overline{\text{co}}\overline{P}$ (première partie de la proposition) ; inversement $\overline{P} \subseteq \overline{\text{co}}P$ ($\overline{\text{co}}P$ est fermé), donc $\overline{\text{co}}\overline{P} \subseteq \overline{\text{co}}P$. Pour montrer $\overline{\text{co}}P = \overline{\text{co}\, P}$, on observe

d'abord que $\overline{\text{co}}P \subseteq \overline{\text{co}\,P}$, parce que ce dernier ensemble est un convexe fermé contenant P (proposition 2.16). Inversement, $\text{co}\,P \subseteq \overline{\text{co}}P$ et comme ce dernier ensemble est fermé, on a $\overline{\text{co}\,P} \subseteq \overline{\text{co}}P$. □

Voici une autre manière de décrire l'enveloppe convexe fermée. La proposition suivante exprime le fait qu'à droite dans (2.34), on peut ne sélectionner que les demi-espaces fermés.

Proposition 2.40 (description externe d'un convexe fermé) *L'enveloppe convexe fermée d'une partie $P \subseteq \mathbb{E}$ est l'intersection de tous les demi-espaces fermés contenant P.*

Démonstration. Soit C l'intersection de tous les demi-espaces fermés contenant P. Comme C est un convexe fermé, on a certainement $\overline{\text{co}}(P) \subseteq C$. Inversement, si $x_0 \in C \setminus \overline{\text{co}}(P)$, on peut séparer strictement le convexe compact $\{x_0\}$ et le convexe fermé $\overline{\text{co}}(P)$ (point 3 du corollaire 2.37) : il existe $\xi \in \mathbb{E}$ et $\alpha \in \mathbb{R}$ tels que

$$\forall x \in \overline{\text{co}}(P), \quad \langle \xi, x\rangle \leqslant \alpha < \langle \xi, x_0\rangle.$$

On en déduit deux conséquances contradictoires :

- par les inégalités de gauche, $P \subseteq H^-(\xi,\alpha)$, c.-à-d., $H^-(\xi,\alpha)$ est un demi-espace fermé contenant P ; alors $x_0 \in H^-(\xi,\alpha)$ (car x_0 est dans C et donc dans tous les demi-espaces fermés contenant P) ;
- par l'inégalité de droite, $x_0 \notin H^-(\xi,\alpha)$.

On a démontré que $C \subseteq \overline{\text{co}}(P)$. □

Corollaire 2.41 *Soit C un ensemble convexe. L'intersection de tous les demi-espaces fermés contenant C est $\overline{C}$.*

Corollaire 2.42 *C est un convexe fermé si, et seulement si, il est l'intersection de tous les demi-espaces fermés contenant C.*

2.5.5 Cône dual

Définition

Soient $\mathbb{E}$ un espace vectoriel euclidien, dont le produit scalaire est noté $\langle\cdot,\cdot\rangle$, et P une partie de $\mathbb{E}$. On appelle *cône dual* de P l'ensemble P^+ défini par

$$P^+ := \{x \in \mathbb{E} : \langle x, y\rangle \geqslant 0,\ \forall y \in P\}.$$

La notion de cône dual est illustrée à la figure 2.7, dans laquelle on a utilisé le produit

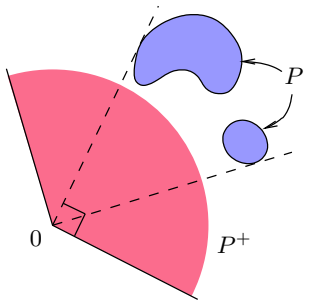


Fig. 2.7. Cône dual P^+ d'un ensemble P de $\mathbb{R}^2$ pour le produit scalaire euclidien (les demi-droites formant le bord du cône P^+ sont orthogonales aux demi-droites en trait discontinu qui délimitent le secteur angulaire contenant l'ensemble P)

scalaire euclidien de $\mathbb{R}^2$. Le *cône bidual* P^{++} de P est l'ensemble $(P^+)^+$, le dual du dual. On définit aussi le *cône dual négatif* P^- de P comme l'opposé de son cône dual, c'est-à-dire

$$P^- := \{d \in \mathbb{E} : \langle d, x\rangle \leqslant 0,\ \forall x \in P\} = -P^+.$$

Enfin, un cône K de $\mathbb{E}$ est dit *autodual* si $K^+ = K$; on en trouvera des exemples à l'exercice 2.35.

Proposition 2.43 (premières propriétés) *Soient $\mathbb{E}$ un espace euclidien, P, P_1 et P_2 des parties non vides de $\mathbb{E}$ et $(P_i)_{i\in I}$ une famille quelconque de parties P_i non vides de $\mathbb{E}$. Alors*

1) *P^+ est un cône convexe fermé non vide,*
2) *$P_1 \subseteq P_2 \Longrightarrow P_1^+ \supseteq P_2^+$ et $P_1^{++} \subseteq P_2^{++}$,*
3) *$P^+ = (\mathbb{R}_+ P)^+$, $P^+ = (\operatorname{co} P)^+$, $P^+ = (\operatorname{adh} P)^+$,*
4) *$P^{++} = \overline{\operatorname{co}}(\mathbb{R}_+ P)$, en particulier $P \subseteq P^{++}$,*
5) *$P^{++} = P$ si, et seulement si, P est un cône convexe fermé,*
6) *$(P_1 + P_2)^+ \supseteq P_1^+ \cap P_2^+$, avec égalité si $0 \in \operatorname{adh}(P_1) \cap \operatorname{adh}(P_2)$,*
7) *$(\cup_{i\in I} P_i)^+ = \cap_{i\in I} P_i^+$.*
8) *$(Q_1 \times Q_2)^+ \supseteq Q_1^+ \times Q_2^+$, si $\varnothing \neq Q_1 \subseteq \mathbb{E}_1$ et $\varnothing \neq Q_2 \subseteq \mathbb{E}_2$, $(\mathbb{E}_1, \langle\cdot,\cdot\rangle_1)$ et $(\mathbb{E}_2, \langle\cdot,\cdot\rangle_2)$ sont deux espaces euclidiens et $\mathbb{E}_1 \times \mathbb{E}_2$ est muni du produit scalaire $\langle (x_1, x_2), (y_1, y_2)\rangle = \langle x_1, y_1\rangle_1 + \langle x_2, y_2\rangle_2$, avec égalité si $0 \in \operatorname{adh}(Q_1) \cap \operatorname{adh}(Q_2)$.*

Démonstration. 1) P^+ est clairement non vide (il contient 0). D'autre part, on peut écrire P^+ comme une intersection de cônes convexes fermés :

$$P^+ = \bigcap_{y\in P} \{x \in \mathbb{E} : \langle x, y\rangle \geqslant 0\}.$$

2) Évident.

3) Par le point 2, $P^+ \supseteq (\operatorname{adh} P)^+$. Inversement, si $d \in P^+$ et $x \in \operatorname{adh} P$, il existe une suite $\{x_k\} \subseteq P$ telle que $x_k \to x$. En passant à la limite dans $\langle d, x_k\rangle \geqslant 0$, on trouve $\langle d, x\rangle \geqslant 0$. Donc $d \in (\operatorname{adh} P)^+$. Les autres identités se démontrent aisément.

4) D'après la proposition 2.40, $\overline{\text{co}}(\mathbb{R}_+P)$ est l'intersection de tous les demi-espaces fermés $H^-(\xi,\alpha)$ contenant $\mathbb{R}_+P$. De par la structure de $\mathbb{R}_+P$, $H^-(\xi,\alpha) \supseteq \mathbb{R}_+P$ implique que $H^-(\xi,\alpha) \supseteq H^-(\xi,0) \supseteq \mathbb{R}_+P$, si bien que $\overline{\text{co}}(\mathbb{R}_+P)$ est l'intersection de tous les demi-espaces fermés $H^-(\xi,0) = \{-\xi\}^+$ contenant $\mathbb{R}_+P$ ou P. Mais $\{-\xi\}^+$ contient P revient à dire que $-\xi \in P^+$. Dès lors $\overline{\text{co}}(\mathbb{R}_+P) = \cap_{-\xi\in P^+}\{-\xi\}^+ = \{x \in \mathbb{E} : \langle d,x\rangle \geqslant 0, \text{ pour tout } d \in P^+\} = P^{++}$.

5) Conséquence directe du point 4.

6) [$\supseteq$] Soient $d \in P_1^+ \cap P_2^+$, $x_1 \in P_1$ et $x_2 \in P_2$. Alors $\langle d, x_1+x_2\rangle \geqslant 0$, car les $\langle d,x_i\rangle \geqslant 0$. Donc $d \in (P_1+P_2)^+$. [$\subseteq$] Soient $d \in (P_1+P_2)^+$, $x \in P_1$ et $\{x'_k\} \subseteq P_2$ avec $x'_k \to 0$ (0 adhère à P_2). Alors $x+x'_k \in P_1+P_2$ et donc $\langle d, x+x'_k\rangle \geqslant 0$. À la limite lorsque $k\to\infty$, on trouve $\langle d,x\rangle \geqslant 0$, ce qui montre que $d \in P_1^+$.

7) $d \in \cap_{i\in I}P_i^+ \iff \forall i \in I\colon d \in P_i^+ \iff \forall i \in I,\ \forall x \in P_i\colon \langle d,x\rangle \geqslant 0 \iff \forall x \in \cup_{i\in I}P_i\colon \langle d,x\rangle \geqslant 0 \iff d \in (\cup_{i\in I}P_i)^+$.

8) L'inclusion $\supseteq$ est immédiate. On a l'égalité lorsque $0 \in \overline{Q_1}\cap\overline{Q_2}$, car si $(d_1,d_2) \in \mathbb{E}_1\times\mathbb{E}_2$ vérifie $\langle (d_1,d_2),(x_1,x_2)\rangle \geqslant 0$ pour tout $x_i \in \mathbb{E}_i$, on obtient $d_1 \in Q_1^+$ en prenant $\{x_2^k\}_k \subseteq Q_2$ avec $x_2^k \to 0$ (de même pour $d_2 \in Q_2^+$). □

La démonstration du corollaire suivant est proposée à l'exercice 2.36.

Corollaire 2.44 (dual d'une somme et d'une intersection de cônes)
1) *Si K_1 et K_2 sont des cônes non vides, alors*

$$(K_1+K_2)^+ = K_1^+ \cap K_2^+.$$

2) *Si K_1 et K_2 sont des cônes convexes fermés non vides, alors*

$$(K_1\cap K_2)^+ = \overline{K_1^+ + K_2^+}. \tag{2.35}$$

Le lemme de Farkas et ses conséquences

La notion de cône dual généralise celle de sous-espace vectoriel orthogonal, puisque si P est un sous-espace vectoriel, $P^+ = P^\perp$. On connaît bien la relation $\mathcal{N}(A^\mathsf{T})^\perp = \mathcal{R}(A)$, rappelée en (A.6), qui nous apprend ce qu'est le cône dual d'un ensemble défini par des équations linéaires homogènes. Une question naturelle est alors de se demander ce qu'est le cône dual d'un ensemble donné par des *inégalités* linéaires homogènes. La réponse à cette question sera un corollaire du résultat plus général suivant. Dans celui-ci, on note $A^*\colon \mathbb{F}\to\mathbb{E}$ l'application linéaire adjointe de l'application linéaire $A\colon \mathbb{E}\to\mathbb{F}$ et $A(K) := \{Ax : x \in K\}$ l'image du cône K par A. On notera que le cône K de la proposition ne doit pas être fermé.

Proposition 2.45 (Farkas généralisé) *Soient $\mathbb{E}$ et $\mathbb{F}$ deux espaces euclidiens, $A\colon \mathbb{E}\to\mathbb{F}$ une application linéaire et K un cône convexe non vide de $\mathbb{E}$. Alors*

$$\{y \in \mathbb{F} : A^*y \in K^+\}^+ = \overline{A(K)}. \tag{2.36}$$

DÉMONSTRATION. On note $D := \{y \in \mathbb{F} : A^*y \in K^+\}$.

[$\supseteq$] Comme D^+ est fermé, il suffit de montrer que $A(K) \subseteq D^+$. Soit $x \in K$ et montrons que $Ax \in D^+$. Pour cela on prend $y \in D$ et on constate que

$$\langle Ax, y\rangle = \langle x, A^*y\rangle \geqslant 0,$$

car $x \in K$ et $A^*y \in K^+$.

[$\subseteq$ (par l'absurde)] Supposons qu'il existe un vecteur $b \in D^+ \setminus \overline{A(K)}$. On peut alors séparer strictement le convexe compact $\{b\}$ et le convexe fermé $\overline{A(K)}$ (point 3 du corollaire 2.37) : il existe $y_0 \in \mathbb{F}$ et $\alpha \in \mathbb{R}$, tels que

$$\forall x \in K : \quad \langle y_0, b\rangle < \alpha \leqslant \langle y_0, Ax\rangle. \tag{2.37}$$

Exploitons cela pour mettre en évidence une contradiction. En prenant $x \to 0$, on trouve

$$\langle y_0, b\rangle < 0. \tag{2.38}$$

Par ailleurs, on peut prendre x de la forme tx avec $t > 0$ et $x \in K$ dans (2.37). En divisant l'inégalité de droite par t et en passant à la limite lorsque $t \to \infty$, on trouve $\langle y_0, Ax\rangle \geqslant 0$ ou $\langle A^*y_0, x\rangle \geqslant 0$ ou $A^*y_0 \in K^+$ (car x est arbitraire dans K). Dès lors $y_0 \in D$ et, comme $b \in D^+$, on obtient $\langle y_0, b\rangle \geqslant 0$, ce qui contredit (2.38). □

Comme un cône dual est fermé (point 1 de la proposition 2.43), on pourra ôter l'adhérence dans (2.36) si, et seulement si, $A(K)$ est fermé.

L'identité (2.36) peut s'interpréter géométriquement comme suit (voir la figure 2.8). Elle signifie qu'un vecteur $b \notin \overline{A(K)}$ si, et seulement si, $b \notin \{y \in \mathbb{F} : A^*y \in K^+\}^+$, ce qui revient à dire qu'il existe un vecteur $y_0 \in \mathbb{F}$ tel que $\langle y_0, b\rangle < 0$ et $A^*y_0 \in K^+$ (ou $\langle y_0, Ax\rangle \geqslant 0$ pour tout $x \in K$). Cette propriété exprime donc le fait que l'hyperplan $\{y \in \mathbb{F} : \langle y_0, y\rangle = \langle y_0, b\rangle/2\}$ sépare strictement le singleton $\{b\}$ de l'adhérence du cône $A(K)$. C'est d'ailleurs cet argument de séparation de ces ensembles qui a été utilisé dans la démonstration de la proposition 2.45.

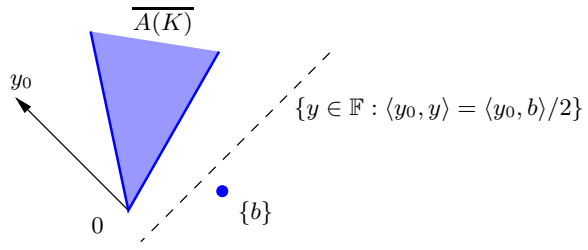


Fig. 2.8. Illustration du lemme de Farkas généralisé (proposition 2.45)

Observons que $\{y \in \mathbb{F} : A^*y \in K^+\} = (A^*)^{-1}(K^+)$ est l'image réciproque par l'application linéaire (continue) A^* du cône convexe fermé K^+ ; il s'agit donc d'un cône convexe fermé. D'après le point 5 de la proposition 2.43, il est égal à son bidual. Dès lors, par le lemme de Farkas généralisé :

$$(A(K))^+ = \{y \in \mathbb{F} : A^*y \in K^+\}, \tag{2.39}$$

sans que l'on ait besoin de prendre d'adhérence à gauche (point 3 de la proposition 2.43).

L'identité (2.36) permet de donner une condition nécessaire pour que le système linéaire $Ax = b$ ait une solution x dans K. Il faut en effet que $b \in A(K) \subseteq \overline{A(K)}$ et donc que

$$\text{pour tout } y \text{ tel que } A^*y \in K^+ \text{ on ait } \langle y, b\rangle \geqslant 0. \tag{2.40}$$

Si $A(K)$ est fermé, cette condition sur A et b est aussi suffisante. Si $A(K)$ n'est pas fermé, on peut trouver des conditions nécessaires et suffisantes pour que $b \in A(K)$, qui renforcent (2.40), voir [374]. Lorsque K est l'orthant positif de $\mathbb{R}^n$, $A(K)$ est fermé (remarque 2.19) et le résultat, exprimé sous une forme différente, est alors connu sous le nom de *théorème de l'alternative*. Diverses variantes sont considérées à l'exercice 2.39.

Voici un corollaire de la proposition 2.45, plus proche de la contribution originale de Farkas.

Corollaire 2.46 (Farkas) *Soient A_1 et A_2 deux matrices ayant le même nombre de lignes. Alors*

$$\{y : A_1^{\mathsf{T}}y = 0,\ A_2^{\mathsf{T}}y \geqslant 0\}^+ = \{A_1x_1 + A_2x_2 : x_1 \text{ quelconque}, x_2 \geqslant 0\},$$

où $\{\cdot\}^+$ désigne le cône dual pour le produit scalaire euclidien.

Démonstration. Supposons que A_i $(i = 1, 2)$ soit de type $m \times n_i$. On applique la proposition 2.45 avec $\mathbb{E} = \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$, $\mathbb{F} = \mathbb{R}^m$, tous deux munis du produit scalaire euclidien, $A = (A_1\ A_2)$, $K = \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}_+$. On calcule aisément

$$A^* = \begin{pmatrix} A_1^{\mathsf{T}} \\ A_2^{\mathsf{T}} \end{pmatrix} \qquad \text{et} \qquad K^+ = \{0\} \times \mathbb{R}^{n_2}_+.$$

La proposition 2.45 conduit alors au résultat puisque $\{A_1x_1 + A_2x_2 : x_2 \geqslant 0\} = \{A_1x_1' - A_1x_1'' + A_2x_2 : x_1' \geqslant 0,\ x_1'' \geqslant 0,\ x_2 \geqslant 0\}$ est un fermé (remarque 2.19). □

Terminons cette section par d'autres propriétés des cônes duaux et quelques règles de calcul.

Corollaire 2.47 (polyédricité d'un cône dual) *Soit P un polyèdre convexe d'un espace euclidien $\mathbb{E}$. Alors P^+ est un polyèdre convexe.*

DÉMONSTRATION. Par le point 3 de la proposition 2.43, $P^+ = K^+$, où $K = \operatorname{cone} P$. Par le point 4 de l'exercice 2.21, K est un cône polyédrique, qui est donc de la forme $K = \{x \in \mathbb{E} : Ax \geqslant 0\}$, où $A : \mathbb{E} \to \mathbb{R}^m$ est linéaire. Par la proposition 2.45, son dual s'écrit $K^+ = A^*(\mathbb{R}^m_+)$. Comme image par A^* du cône polyédrique $\mathbb{R}^m_+$, K^+ est un cône polyédrique (proposition 2.18). □

Lemme 2.48 (somme fermée de cônes duaux)
1) *$P_1^+ + \cdots + P_m^+$ est un polyèdre convexe (donc un fermé), si les P_i sont des polyèdres convexes,*
2) *$K_1^+ + \cdots + K_m^+$ est fermé, si les K_i sont des cônes convexes et si $K_1^\ominus \cap \cdots \cap K_m^\ominus \neq \varnothing$.*

DÉMONSTRATION. 1) Si les P_i sont polyédriques, les P_i^+ sont aussi polyédriques (corollaire 2.47) et donc aussi leur somme $P_1^+ + \cdots + P_m^+$ (exercice 2.20 ou 2.21), qui est donc fermée.

2) Supposons à présent qu'il existe un point $\check{x} \in K_1^\ominus \cap \cdots \cap K_m^\ominus$ et que, pour $i \in [1:m]$, on ait des suites $\{d_k^i\} \subseteq K_i^+$ telles que $d_k^1 + \cdots + d_k^m \to d$. Il s'agit de montrer que $d \in K_1^+ + \cdots + K_m^+$.

On note P_i est le projecteur orthogonal sur le sous-espace vectoriel $\mathbb{E}_i := \operatorname{vect} K_i$ et $\mathrm{Q}_i := I - \mathrm{P}_i$ le projecteur orthogonal sur $\mathbb{E}_i^\perp$. Montrons d'abord que les suites $\{\mathrm{P}_i\, d_k^i\}$ sont bornées. Par hypothèse, il existe un $\varepsilon > 0$ tel que $\bar{B}(\check{x}, \varepsilon) \cap (\operatorname{vect} K_i) \subseteq K_i$. Alors $\langle d_k^i, \check{x} - \varepsilon\, \mathrm{P}_i\, d_k^i / \|\mathrm{P}_i\, d_k^i\| \rangle \geqslant 0$, d'où l'on déduit que $\langle d_k^i, \check{x} \rangle \geqslant \varepsilon \|\mathrm{P}_i\, d_k^i\|$ et $\langle \sum_i d_k^i, \check{x} \rangle \geqslant \varepsilon \sum_i \|\mathrm{P}_i\, d_k^i\|$. Comme $\sum_i d_k^i \to d$, on voit que les $\{\mathrm{P}_i\, d_k^i\}$ sont bornées.

On peut alors trouver une sous-suite d'indices $\mathcal{K} \subseteq \mathbb{N}$ telle que chaque $\{\mathrm{P}_i\, d_k^i\}_{k \in \mathcal{K}}$ converge vers un $\bar{d}^i \in \mathbb{E}_i$. Comme pour tout $x \in K_i$, on a $0 \leqslant \langle d_k^i, x \rangle = \langle \mathrm{P}_i\, d_k^i, x \rangle \to \langle \bar{d}^i, x \rangle$, il vient que $\bar{d}^i \in K_i^+$.

On déduit de ce qui précède que $\{\sum_i \mathrm{Q}_i\, d_k^i\}_{k \in \mathcal{K}}$ converge vers $d - \sum_i \bar{d}^i$ dans le sous-espace vectoriel $\sum_i \mathbb{E}_i^\perp$, qui est un fermé. Dès lors $d - \sum_i \bar{d}^i \in \sum_i \mathbb{E}_i^\perp$ et on peut trouver des $\hat{d}^i \in \mathbb{E}_i^\perp$ tel que $d - \sum_i \bar{d}^i = \sum_i \hat{d}^i$. Observons maintenant que pour tout $x \in K_i$, on a $\langle \hat{d}_k^i, x \rangle = 0$.

On a donc montré que $\bar{d}^i + \hat{d}^i \in K_i^+$ et donc que $d = \sum_i (\bar{d}^i + \hat{d}^i) \in \sum_i K_i^+$. □

Corollaire 2.49 (dual d'une intersection) *Si K_i, $i \in [1:m]$, sont des cônes convexes fermés d'un espace euclidien $\mathbb{E}$, alors*

$$(K_1 \cap \cdots \cap K_m)^+ = \overline{K_1^+ + \cdots + K_m^+}.$$

On peut enlever l'adhérence si les K_i sont polyédriques ou si $K_1^\ominus \cap \cdots \cap K_m^\ominus \neq \varnothing$.

DÉMONSTRATION. Soient $A : \mathbb{E}^m \to \mathbb{E} : (x_1, \ldots, x_m) \mapsto x_1 + \cdots + x_m$ et $K = K_1^+ \times \cdots \times K_m^+$. On vérifie aisément que $A^* : \mathbb{E} \to \mathbb{E}^m : x \mapsto (x, \ldots, x)$, que K est un cône convexe fermé et que $K^+ = K_1 \times \cdots \times K_m$. L'identité se déduit alors du lemme

de Farkas généralisé (proposition 2.45). Par le lemme 2.48, $K_1^+ + \cdots + K_m^+$ est fermé si les K_i sont polyédriques ou si $K_1^\oplus \cap \cdots \cap K_m^\oplus \neq \varnothing$. □

On ne peut pas se passer de l'adhérence dans le résultat précédent, même si $K_1 \cap K_2^\oplus \cap \cdots \cap K_m^\oplus \neq \varnothing$. Par exemple, dans $\mathbb{R}^3$, si $m = 2$, $K_1 = \mathbb{R}^3_\nabla$ et $K_2 = \{x \in \mathbb{R}^3 : x_2 = x_3\}$, on a $K_1^+ = \mathbb{R}^3_\nabla$ (autodualité du cornet), $K_2^+ = \{d \in \mathbb{R}^3 : d_1 = 0,\ d_2 + d_3 = 0\}$, $K_1 \cap K_2^\oplus = K_1 \cap K_2 = \{x \in \mathbb{R}^3 : x_1 = 0,\ x_2 = x_3 \geqslant 0\} \neq \varnothing$, alors que $K_2^+ + K_3^+ = \{d : d_2 + d_3 > 0\} \cup \{d \in \mathbb{R}^3 : d_1 = 0,\ d_2 + d_3 = 0\}$ n'est pas fermé.

Polyèdre borné et singleton

La caractérisation de la bornitude du polyèdre P donné par (2.19) s'obtient aisément par l'examen de son cône asymptotique P^∞. Les caractérisations obtenues montrent que la propriété de bornitude ne dépend pas de b (mais pour certains b, P pourra être vide, ce qui ne contredit pas sa bornitude).

Proposition 2.50 (polyèdre borné) *Soient $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ et $P := \{x \in \mathbb{R}^n : Ax \leqslant b\}$. Alors les propriétés suivantes sont équivalentes :*
(i) P est borné,
(ii) tout vecteur $d \in \mathbb{R}^n$ vérifiant $Ad \leqslant 0$ est nul,
(iii) $A^\mathsf{T}(\mathbb{R}^m_+) = \mathbb{R}^n$.

Démonstration. [$(i) \Leftrightarrow (ii)$] D'après la proposition 2.8, P est borné si, et seulement si, $P^\infty = \{0\}$. L'équivalence se déduit alors du fait que $P^\infty = \{d : Ad \leqslant 0\}$.

[$(ii) \Leftrightarrow (iii)$] $(ii) \Leftrightarrow \{d : Ad \leqslant 0\} = \{0\} \Leftrightarrow \mathbb{R}^n = \{d : Ad \leqslant 0\}^+ = A^\mathsf{T}(\mathbb{R}^m_-)$ (lemme de Farkas [proposition 2.45], $(\mathbb{R}^m_-)^+ = \mathbb{R}^m_-$ [point 1 de l'exercice 2.35] et $A^\mathsf{T}(\mathbb{R}^m_-)$ fermé [(2.24)]) $\Leftrightarrow (iii)$. □

Lorsqu'on considère le système linéaire $Ax = b$, avec $A \in \mathbb{R}^{n\times n}$ et $b \in \mathbb{R}^n$, il est naturel de se demander si celui-ci a une solution x unique. On sait qu'il en sera ainsi, quel que soit $b \in \mathbb{R}^n$, si, et seulement si, A est inversible. On cherche ici des conditions similaires pour que le système d'inégalités affines $Ax \leqslant b$ ait une unique solution x, autrement dit, pour que le polyèdre convexe P défini par (2.19) soit un singleton. Pour ce cas, une condition nécessaire et suffisante ne peut s'exprimer qu'en terme de la matrice A, car si b devient « grand », le polyèdre P est « grand » aussi. Pour pouvoir utiliser les conditions nécessaires et suffisantes du résultat suivant, il faut aussi connaître un point du polyèdre convexe.

Proposition 2.51 (polyèdre singleton) *Soient $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $x \in \mathbb{R}^n$ vérifiant $Ax \leqslant b$, $I := \{i \in [1:m] : (Ax-b)_i = 0\}$, $m_I := |I|$ et A_I la sous-matrice de A formée de ses lignes avec indices dans I. Alors les propriétés suivantes sont équivalentes :*
(i) x est l'unique solution de $Ax \leqslant b$,

(ii) tout vecteur $d \in \mathbb{R}^n$ vérifiant $A_I d \leqslant 0$ est nul,
(iii) $A_I^{\mathsf{T}}(\mathbb{R}_+^{m_I}) = \mathbb{R}^n$.

DÉMONSTRATION. $[(i) \Rightarrow (ii)]$ Avec un vecteur d vérifiant les conditions données, on peut trouver $t > 0$ tel que $x' = x + td$ vérifie $Ax' \leqslant b$. En effet, si $(Ax - b)_i < 0$, on a $(Ax')_i = (Ax)_i + t(Ad)_i \leqslant b_i$ si $t > 0$ est assez petit car $(Ax)_i < b_i$ (et il y a un nombre fini de tels indices i) ; si $(Ax - b)_i = 0$, on a $(Ax')_i = (Ax)_i + t(Ad)_i = b_i + t(Ad)_i \leqslant b_i$, car $(Ad)_i \leqslant 0$ par hypothèse. L'unicité de x implique que $x' = x$, donc $d = 0$.

$[(ii) \Rightarrow (i)]$ Soit x' tel que $Ax' \leqslant b$. On pose $d := x' - x$. Si $(Ax - b)_i = 0$, on a $(Ad)_i = (Ax')_i - (Ax)_i = (Ax')_i - b_i \leqslant 0$. Par (ii), $d = 0$ et donc $x' = x$.

$[(ii) \Leftrightarrow (iii)]$ Voir la démonstration de l'équivalence $(ii) \Leftrightarrow (iii)$ de la proposition 2.50. □

La condition (ii) (resp. (iii)) de la proposition 2.51 est plus forte (moins souvent vérifiée) que la condition (ii) (resp. (iii)) de la proposition 2.50, ce qui est consistant avec le fait qu'un singleton est borné !

2.5.6 Cône tangent

Soit $\mathbb{E}$ est un espace vectoriel de dimension finie et C est un convexe non vide de $\mathbb{E}$, que l'on ne suppose pas nécessairement fermé.

Définitions 2.52 On dit que $d \in \mathbb{E}$ est une *direction admissible* de C en $x \in \mathbb{E}$ si $x + td \in C$ pour tout $t > 0$ petit. L'ensemble des directions directions admissibles forme un cône, noté

$$\mathrm{T}_x^a C := \{d \in \mathbb{E} : x + td \in C \text{ pour tout } t > 0 \text{ petit}\},$$

que l'on appelle le *cône des directions admissibles*. □

La conicité de $\mathrm{T}_x^a C$ ne fait aucun doute. Par ailleurs, la définition montre que

$$x \notin \overline{C} \quad \Longrightarrow \quad \mathrm{T}_x^a C = \varnothing. \tag{2.41}$$

Enfin, par la convexité de C, on voit que

$$\begin{aligned} x \in C \quad \Longrightarrow \quad \mathrm{T}_x^a C &= \{d \in \mathbb{E} : x + td \in C \text{ pour un } t > 0\}, && (2.42)\\ &= \bigcup_{t>0} \frac{C - x}{t}, && (2.43)\\ &= \mathbb{R}_{++}(C - x), && (2.44)\\ &= \mathbb{R}_{+}(C - x). && (2.45) \end{aligned}$$

Le cas où $x \in \overline{C} \setminus C$ est moins aisé à décrire, si bien que certains auteurs préfèrent supposer d'emblée que C est fermé [317]. On notera que $\mathrm{T}_x^a C$ n'est pas nécessairement fermé, même si C est fermé et $x \in C$; ainsi pour le disque $C := \{x \in \mathbb{R}^2 : x_1^2 + (x_2 - 1)^2 \leqslant 1\}$ et le point $x = 0$ puisque l'on a $\mathrm{T}_x^a C = \{x \in \mathbb{R}^2 : x_2 > 0\} \cup \{0\}$. La définition suivante a donc bien un sens.

Définition 2.53 On appelle *cône tangent* à C en $x \in \mathbb{E}$ l'adhérence de $\mathrm{T}_x^a C$. On le note

$$\mathrm{T}_x C := \overline{\mathrm{T}_x^a C}. \qquad \square$$

On déduit de (2.41), (2.44) et (2.45) que

$$\begin{aligned} x \notin \overline{C} \quad &\Longrightarrow \quad \mathrm{T}_x C = \varnothing, \\ x \in C \quad &\Longrightarrow \quad \mathrm{T}_x C = \overline{\mathbb{R}_{++}(C-x)} = \overline{\mathbb{R}_{+}(C-x)}. \end{aligned} \tag{2.46}$$

Le cas où $x \in \overline{C} \setminus C$ ne posera pas de difficulté d'expression du cône tangent, grâce aux formulations en termes de suites données dans la proposition suivante.

Proposition 2.54 (autres expressions du cône tangent) *Si C est une partie convexe d'un espace vectoriel de dimension finie $\mathbb{E}$ et $x \in \overline{C}$, alors*

$$\begin{aligned} \mathrm{T}_x C &= \left\{ d \in \mathbb{E} : \exists \{x_k\} \subseteq C,\ \exists \{t_k\} \downarrow 0 \text{ telles que } \frac{x_k - x}{t_k} \to d \right\} \\ &= \left\{ d \in \mathbb{E} : \exists \{x_k\} \subseteq C^{\ominus},\ \exists \{t_k\} \downarrow 0 \text{ telles que } \frac{x_k - x}{t_k} \to d \right\}. \end{aligned}$$

Démonstration. Désignons par $\mathrm{T}'_x C$ la première nouvelle expression du cône tangent et par $\mathrm{T}''_x C$ la seconde.

[$\mathrm{T}_x C \subseteq \mathrm{T}'_x C$] Soit $d \in \mathrm{T}_x C$. Par la définition 2.53, il existe une suite $\{d_k\} \subseteq \mathrm{T}_x^a C$ telle que $d_k \to d$. Par conséquent, pour tout k, on peut trouver un $t_k > 0$ tel que $x_k := x + t_k d_k \in C$. Comme les t_k peuvent être pris arbitrairement petits, on peut s'arranger pour que $t_k \to 0$. Alors $(x_k - x)/t_k = d_k \to d$ montre que $d \in \mathrm{T}'_x C$.

[$\mathrm{T}'_x C \subseteq \mathrm{T}''_x C$] Soit $d \in \mathrm{T}'_x C$, si bien qu'il existe des suites $\{x_k\} \subseteq C$ et $\{t_k\} \subseteq \mathbb{R}_{++}$ telles que $t_k \to 0$ et $(x_k - x)/t_k \to d$. Soit $\check{x} \in C^{\ominus}$ qui est non vide (point 1 de la proposition 2.16). On peut supposer que $t_k \leqslant 1$. Alors $x'_k = (1 - t_k^2)x_k + t_k^2 \check{x} \in C^{\ominus}$ par le lemme 2.14. Comme $(x'_k - x)/t_k = (1 - t_k^2)(x_k - x)/t_k + t_k(\check{x} - x) \to d$, on a montré que $d \in \mathrm{T}''_x C$ comme limite d'éléments de $\mathrm{T}_x^a C$.

[$\mathrm{T}''_x C \subseteq \mathrm{T}_x C$] Soit $d \in \mathrm{T}''_x C$, si bien qu'il existe des suites $\{x_k\} \subseteq C^{\ominus}$ et $\{t_k\} \subseteq \mathbb{R}_{++}$ telles que $t_k \to 0$ et $d_k := (x_k - x)/t_k \to d$. Par le lemme 2.14, $x + t d_k = (1 - (t/t_k))x + (t/t_k)x_k \in C^{\ominus} \subseteq C$ pour tout $t \in]0, t_k]$, si bien que $d_k \in \mathrm{T}_x^a C$. Donc $d \in \mathrm{T}_x C$. $\square$

Comme C et $\overline{C}$ ont les mêmes intérieurs relatifs (point 3 de l'exercice 2.13), on déduit de la seconde expression du cône tangent dans la proprosition précédente que, quel que soit $x \in \mathbb{E}$:

$$\boxed{\mathrm{T}_x \overline{C} = \mathrm{T}_x C.} \tag{2.47}$$

Si $x \notin \overline{C}$ les deux cônes ci-dessus sont vides.

La démonstration de la proposition suivante est proposée à l'exercice 2.34.

Proposition 2.55 (propriétés des cônes $\mathrm{T}^a_x C$ et $\mathrm{T}_x C$) *Si C est une partie convexe d'un espace vectoriel de dimension finie $\mathbb{E}$ et $x \in C$, alors*

1) $\mathrm{T}^a_x C$ *est un cône convexe pointé et* $\mathrm{T}_x C$ *est un cône convexe fermé pointé,*
2) $\operatorname{aff}(\mathrm{T}_x C) = \operatorname{aff}(\mathrm{T}^a_x C) = (\operatorname{aff} C) - x$,
3) $(\mathrm{T}_x C)^\ominus = (\mathrm{T}^a_x C)^\ominus = R_{++}(C^\ominus - x)$,
4) $\mathrm{T}_x C = (\mathrm{N}_x C)^-$ *et* $\mathrm{N}_x C = (\mathrm{T}_x C)^-$.

Proposition 2.56 (calcul de cône tangent) *Soient $\mathbb{E}_1$ et $\mathbb{E}_2$ des espaces vectoriels.*

1) *Si $C_1 \subseteq \mathbb{E}_1$ et $C_2 \subseteq \mathbb{E}_2$ sont deux convexes, alors*

$$\mathrm{T}_{(x_1,x_2)}(C_1 \times C_2) = \mathrm{T}_{x_1} C_1 \times \mathrm{T}_{x_2} C_2.$$

2) *Si $\{C_i\}_{i\in I}$ est une famille de parties convexes fermées de $\mathbb{E}$ et $x \in \cap_{i\in I} C_i$, alors*

$$\mathrm{T}^a_x(\cap_{i\in I} C_i) \subseteq \cap_{i\in I}(\mathrm{T}^a_x C_i), \text{ avec égalité si } I \text{ est fini,}$$
$$\mathrm{T}_x(\cap_{i\in I} C_i) \subseteq \cap_{i\in I}(\mathrm{T}_x C_i), \text{ avec égalité si } I \text{ est fini et si } \cap_{i\in I} C_i^\ominus \neq \varnothing.$$

Démonstration. 1) On a successivement

$$\begin{aligned}
\mathrm{T}_{(x_1,x_2)}(C_1 \times C_2) &= \operatorname{adh}(\mathbb{R}_+(C_1 \times C_2 - (x_1, x_2)) \\
&= \operatorname{adh}(\mathbb{R}_+((C_1 - x_1) \times (C_2 - x_2))) \\
&= \operatorname{adh}(\mathbb{R}_+(C_1 - x_1) \times \mathbb{R}_+(C_2 - x_2)) \qquad (2.48)\\
&= \operatorname{adh}(\mathbb{R}_+(C_1 - x_1)) \times \operatorname{adh}(\mathbb{R}_+(C_2 - x_2)) \qquad (2.49)\\
&= \mathrm{T}_{x_1} C_1 \times \mathrm{T}_{x_2} C_2.
\end{aligned}$$

L'inclusion $\subseteq$ en (2.48) ne pose pas de difficulté et l'inclusion réciproque repose sur l'identité

$$\Big(t_1(x_1' - x_1), t_2(x_2' - x_2)\Big) = t_2\Big((1 - t_1/t_2)x_1 + (t_1/t_2)x_1' - x_1, x_2' - x_2\Big),$$

qui permet de conclure lorsque $0 \leqslant t_1 \leqslant t_2$ (non restrictif) et les x_i' sont arbitraires dans les C_i. Pour (2.49), on utilise le fait que $\operatorname{adh}(P_1 \times P_2) = (\operatorname{adh} P_1) \times (\operatorname{adh} P_2)$, quels que soient les ensembles $P_i \subseteq \mathbb{E}_i$.

2) On a $\mathrm{T}^a_x(\cap_{i\in I} C_i) = \mathbb{R}_+((\cap_{i\in I} C_i) - x) = \mathbb{R}_+(\cap_{i\in I}(C_i - x)) \subseteq \cap_{i\in I}(\mathbb{R}_+(C_i - x))$ $= \cap_{i\in I}(\mathrm{T}^a_x C_i)$.

Si $I = [1\,{:}\,m]$ est fini, la dernière inclusion devient une égalité. En effet, si un point $y \in \cap_i(\mathbb{R}_+(C_i - x))$, il peut s'écrire $y = t_i(x_i - x)$ avec $t_i \geqslant 0$ et $x_i \in C_i$, quel que soit $i \in [1\,{:}\,m]$. On peut supposer que $t_m = \max_i t_i$. Si $t_m = 0$, alors $y = 0$ qui est clairement dans $\mathbb{R}_+((\cap_{i\in I} C_i) - x)$. Si $t_m > 0$, on introduit $\bar{x} := x + y/t_m$. On a

$\bar{x} = x + (t_i/t_m)(x_i - x) = (1 - t_i/t_m)x + (t_i/t_m)x_i \in C_i$, par convexité de C_i. Donc $\bar{x} \in \cap_i C_i$. Alors $y = t_m(\bar{x} - x)$ est dans $\mathbb{R}_+((\cap_{i\in I} C_i) - x)$.

La seconde inclusion s'obtient en prenant l'adhérence des deux membres de la première inclusion, car $\operatorname{adh}(\cap_i P_i) \subseteq \cap_i(\operatorname{adh} P_i)$.

Si $I = [1:m]$ est fini, la première identité du point 2 nous apprend que

$$\mathrm{T}^a_x(\cap_{i\in I} C_i) = \cap_{i\in I}(\mathrm{T}^a_x C_i) \qquad \text{et} \qquad \mathrm{T}^a_x(\cap_{i\in I} C_i^{\ominus}) = \cap_{i\in I}(\mathrm{T}^a_x(C_i^{\ominus})). \tag{2.50}$$

Maintenant, si $\cap_{i\in I} C_i^{\ominus} \neq \varnothing$, on peut y trouver un point x_0. Clairement, $x_0 - x \in \mathrm{T}^a_x(\cap_{i\in I} C_i^{\ominus})$, si bien que $\cap_{i\in I}(\mathrm{T}^a_x(C_i^{\ominus})) \neq \varnothing$ par $(2.50)_2$.

En prenant l'adhérence des deux membres de cette identité, on obtient $\mathrm{T}_x(\cap_{i\in I} C_i) = \cap_{i\in I}(\mathrm{T}_x C_i)$, grâce au point 2 de la proposition 2.17 et à l'hypothèse de qualification $\cap_{i\in I} C_i^{\ominus} \neq \varnothing$. □

Notes

Hermann Minkowski (1864-1909) fut l'un des premiers à avoir étudié systématiquement les ensembles convexes [421 ; 1896]. Au départ, ce fut dans le but de résoudre des problèmes en théorie des nombres, alors que notre présentation est davantage de nature qualitative, orientée vers l'analyse. On lui doit par exemple la proposition 2.23 sur la « résolution » d'un polyèdre convexe.

L'*élimination de Fourier* [219 ; 1827] a été redécouverte plus d'un siècle plus tard par Motzkin [444 ; 1936]. Celle-ci a été reconsidérée à maintes reprises en vue de son amélioration [381, 380, 324, 339, 113], mais l'obtention d'une description explicite du projeté d'un polytope est un problème NP-ardu [588 ; 2008].

La proposition 2.18 a été étendue à des cas beaucoup plus généraux. Ainsi, on peut montrer que l'image par une application linéaire d'un ensemble de la forme $c^{-1}(0)$, où $c : \mathbb{R}^n \to \mathbb{R}^m$ a ses composantes convexes et « *als* » (*asymptotically level stable*), est fermée [29 ; corollaire 3.7.1]. Cette classe de fonctions c_i inclut celles qui sont convexes et quadratiques par morceaux (leur domaine est une union finie de polyèdres convexes, sur chacun desquels la fonction est quadratique [29 ; proposition 3.3.3]).

Le lemme de Farkas a été énoncé la première fois dans [202 ; 1902] .

Exercices

2.1. *Ensembles convexes.* On note $\mathbb{E}$ un espace vectoriel ou euclidien au besoin.

1) Soient Λ un intervalle non vide de $\mathbb{R}_+$ et C un convexe de $\mathbb{E}$. Alors $\Lambda C := \{\lambda x : \lambda \in \Lambda,\ x \in C\}$ est convexe.
2) Si $C \subseteq \mathbb{E}$ est un convexe, alors $C + C = 2C$.
3) Si $K \subseteq \mathbb{E}$ est un cône, alors K est convexe si, et seulement si, $K + K = K$.
4) *Diagramme de Voronoï.* Soient p un point de $\mathbb{E}$ et Q une partie arbitraire et non vide de $\mathbb{E}$. On note d_P la distance à un ensemble P de $\mathbb{E}$. Alors $V_{p,Q} := \{x \in \mathbb{E} : \mathrm{d}_{\{p\}}(x) \leqslant \mathrm{d}_Q(x)\}$ est convexe.

2.2. *Calcul d'enveloppe affine.* Soient P et Q des parties d'un espace vectoriel $\mathbb{E}$ et C une partie convexe.

1) $P \subseteq Q \implies \operatorname{aff} P \subseteq \operatorname{aff} Q$.
2) $\operatorname{aff}(P \cap Q) \subseteq (\operatorname{aff} P) \cap (\operatorname{aff} Q)$.
3) $\operatorname{aff}(P + Q) = (\operatorname{aff} P) + (\operatorname{aff} Q)$.
4) $\operatorname{aff} \overline{P} = \operatorname{aff} P$.
5) $0 \in \overline{P} \implies \operatorname{aff} P = \operatorname{aff}(\mathbb{R}_{++}P) = \operatorname{aff}(\mathbb{R}_{+}P) = \operatorname{vect} P$.
6) $\operatorname{aff}(\operatorname{cone} P) = \operatorname{vect} P$.
7) C convexe $\implies \operatorname{aff} C = \{tx + (1-t)y : t \in \mathbb{R},\ x, y \in C\}$ (comparez avec (2.3)).

Si P et Q sont des parties des espaces vectoriels $\mathbb{E}$ et $\mathbb{F}$ respectivement, alors

8) $\operatorname{aff}(P \times Q) = (\operatorname{aff} P) \times (\operatorname{aff} Q)$.

2.3. *Maximum d'une combinaison convexe de valeurs.* Soient $\{v_i\}_{i\in[1\,:\,p]} \subseteq \mathbb{R} \cup \{+\infty\}$ un nombre fini p de valeurs et Δ_p le simplexe unité de $\mathbb{R}^p$. Alors

$$\max_{\alpha\in\Delta_p} \left(\sum_{\substack{i\in[1\,:\,p]\\ \alpha_i\neq 0}} \alpha_i v_i \right) = \max_{i\in[1\,:\,p]} v_i.$$

2.4. *Calcul d'enveloppe convexe.* Démontrez la proposition 2.5.

2.5. *Enveloppe convexe d'un nombre fini de convexes.* Si $P = C_1 \cup \cdots \cup C_m$, où les C_i sont convexes, alors $\operatorname{co} P = \{\sum_{i=1}^m t_i x_i : (t_1, \ldots, t_m) \in \Delta_m,\ x_i \in C_i \text{ pour tout } i\}$ (comparez avec la proposition 2.2).

2.6. *Exemples d'enveloppes convexes.*

1) Soient $X = \{x_1, \ldots, x_n\}$ une partie d'un espace vectoriel $\mathbb{E}$ formée de n points et $k \in [1\,:n]$. Alors $\operatorname{co}\{x_{i_1} + \cdots + x_{i_k} : 1 \leqslant i_1 \leqslant \cdots \leqslant i_k \leqslant n\} = k\,(\operatorname{co} X)$.
2) Soient $X = \{x_1, \ldots, x_n\}$ une partie d'un espace vectoriel $\mathbb{E}$ formée de n points et $k \in [1\,:n]$. Alors $\operatorname{co}\{x_{i_1} + \cdots + x_{i_k} : 1 \leqslant i_1 < \cdots < i_k \leqslant n\} = \{\sum_{i=1}^n t_i x_i : t_i \in [0,1] \text{ pour tout } i \text{ et } \sum_{i=1}^n t_i = k\}$.

2.7. *Enveloppe conique d'une somme.* Soient P et Q deux ensembles d'un même espace vectoriel. Montrez que

$$\operatorname{cone}(P + Q) \subseteq \operatorname{cone} P + \operatorname{cone} Q \subseteq \operatorname{cone}(\mathbb{R}_+P + \mathbb{R}_+Q).$$

En déduire que si P et Q sont des cônes convexes, alors $P+Q$ est un cône convexe.

2.8. *Enveloppe conique.* Démontrez la proposition 2.6.

2.9. *Calcul de cône asymptotique.* Démontrez la proposition 2.9. De plus, donnez un exemple dans lequel (2.7) n'a pas lieu parce que l'intersection des C_i est vide. Donnez également un exemple dans lequel (2.8) n'a pas lieu parce que $\mathcal{N}(A) \cap C^\infty$ n'est pas un sous-espace vectoriel.

2.10. *Calcul de points extrêmes.* Démontrez la proposition 2.12.

2.11. *Face exposée d'un convexe.* Soit C un convexe d'un espace euclidien $\mathbb{E}$. On dit qu'une partie E de C est *exposée* s'il existe $\xi \in \mathbb{E}$ tel que $E = \arg\min\{\langle \xi, x\rangle : x \in C\}$. Montrez que toute partie exposée d'un convexe est une face, mais que la réciproque n'est pas nécessairement vraie (elle est vraie dans le cas d'un polyèdre convexe, voir l'exercice 2.20).

2.12. *Autour du lemme 2.14 d'intériorité relative.* Soit $\mathbb{E}$ un espace vectoriel.

1) Démontrez le corollaire 2.15.
2) Si C_1 et C_2 sont deux convexes de $\mathbb{E}$ tels que $C_1 \cap C_2^{\oplus} \neq \varnothing$, alors $(C_1 \cap \operatorname{aff} C_2)^{\oplus} \cap C_2^{\oplus} \neq \varnothing$.
3) *Ensemble convexe dense.* Si C est un convexe dense de $\mathbb{E}$, alors $C = \mathbb{E}$.

2.13. *Topologie des ensembles convexes.* Soit P une partie d'un espace vectoriel $\mathbb{E}$. On note $F(x)$ la face d'un convexe C engendrée par un point $x \in C$.

1) $\operatorname{aff} P^{\circ} \subseteq \operatorname{aff} P$, avec égalité si $P^{\circ} \neq \varnothing$ (c'est le cas si P est convexe).
2) A affine et $C^{\circ} \subseteq A \subseteq C$ impliquent que $C = A$.
3) $\overline{P^{\circ}} \subseteq \overline{P}$ et $P^{\circ} \subseteq (\overline{P})^{\circ}$, avec des égalités si P est convexe.
4) Si C est convexe, alors C°, C et $\overline{C}$ ont la même enveloppe affine, le même intérieur relatif, la même adhérence et la même frontière relative.
5) Deux convexes ont le même intérieur relatif si, et seulement si, ils ont la même adhérence.
6) Pour tout $x \in C$, $x \in F(x)^{\circ}$.

2.14. *Propriété topologique des cônes.* Soit K un cône. Alors son adhérence $\overline{K}$ et son intérieur relatif K° sont également des cônes.

2.15. *Conditions de qualification équivalentes.* Si C_1 et C_2 sont deux convexes, alors

$$0 \in (C_1 - C_2)^{\circ} \qquad \Longleftrightarrow \qquad C_1^{\circ} \cap C_2^{\circ} \neq \varnothing.$$

2.16. *Face d'un convexe.* Soient C un convexe et F une face non vide de C.

1) Si F_0 est une face de F, alors F_0 est une face de C.
2) $\operatorname{ext}(F) \subseteq \operatorname{ext}(C)$.
3) Si C est fermé, alors $\operatorname{ext}(F) \neq \varnothing \Leftrightarrow \operatorname{ext}(C) \neq \varnothing$.

2.17. *Points extrêmes des boules unités ℓ_1 et ℓ_∞.* Dans $\mathbb{R}^n$, la boule-unité pour la norme ℓ_1 a $2n$ points extrêmes et la boule-unité pour la norme ℓ_∞ en a 2^n.

2.18. *Théorème de Birkhoff* [60, 364]. On dit qu'une matrice G est *doublement stochastique* si ses éléments sont positifs ($G_{ij} \geqslant 0$ pour tout i et j) et si la somme des éléments de chaque ligne et de chaque colonne vaut 1 ($\sum_i G_{ij} = 1$ pour tout j et $\sum_j G_{ij} = 1$ pour tout i). On note $\mathbb{G}^n$ l'ensemble des matrices d'ordre n doublement stochastiques. Calculez $\operatorname{ext}(\mathbb{G}^n)$.

2.19. *Cônes convexes.* Soit K un cône d'un espace vectoriel $\mathbb{E}$.

1) K est convexe ssi pour tout $\forall x, y \in K$ et $\forall \alpha, \beta \in \mathbb{R}_+$, on a $\alpha x + \beta y \in K$.
2) Si K est convexe, $x \in K^{\circ}$, $y \in K$, $\alpha > 0$ et $\beta \geqslant 0$, alors $\alpha x + \beta y \in K^{\circ}$.
3) Si $K = \operatorname{cone}\{x_1, \ldots, x_p\}$, alors $\operatorname{aff} K = \operatorname{vect}\{x_1, \ldots, x_p\}$ et $K^{\circ} = \{\sum_{i=1}^{p} \alpha_i x_i : \alpha_i > 0 \text{ pour tout } i\}$.

2.20. *Polyèdre convexe sous représentation duale.* Soit $P := \{x \in \mathbb{R}^n : Ax = a,\ Bx \leqslant b\}$ un polyèdre convexe de $\mathbb{R}^n$ (on suppose les dimensions consistantes).

1) L'ensemble F est une face de P si, et seulement si, il existe un ensemble d'indices I tel que $F = \{x \in P : (Bx - b)_I = 0\}$.
2) Toute face de P est une partie exposée (on sait que la réciproque est vraie pour un convexe quelconque : toute partie exposée d'un convexe est une face, voir l'exercice 2.11).
3) Le point $x \in P$ est un sommet de P si, et seulement si, la matrice $(A^{\mathsf{T}}\ \ B_I^{\mathsf{T}})^{\mathsf{T}}$ formée de la matrice A et des lignes de B d'indices dans $I := \{i : (Bx - b)_i = 0\}$ est injective.
4) $P^{\infty} = \{d \in \mathbb{R}^n : Ad = 0,\ Bd \leqslant 0\}$.
5) Si I est l'ensemble des indices i tels que $\{x \in P : B_i x < b_i\} \neq \varnothing$ et I^c son complémentaire, alors $\operatorname{aff} P = \{x \in \mathbb{R}^n : Ax = a,\ B_{I^c} x = b_{I^c}\}$ et $P^{\circ} = \{x \in P : B_I x < b_I\}$ ($= P$ si $I = \varnothing$).
6) La somme de deux polyèdres convexes est un polyèdre convexe (donnez la représentation duale de cette somme).

7) L'image réciproque d'un polyèdre convexe par une application linéaire est un polyèdre convexe.
8) Le produit cartésien de deux polyèdres convexes est un polyèdre convexe.

2.21. *Polyèdre convexe sous représentation primale.* Soit P un polyèdre convexe d'un espace vectoriel $\mathbb{E}$, donné sous la forme $P = \mathrm{co}\{x_1, \dots, x_p\} + \mathrm{cone}\{y_1, \dots, y_q\}$, avec des x_i et des $y_j \in \mathbb{E}$.

1) $\mathrm{ext}(P) \subseteq \{x_1, \dots, x_p\}$ sans que l'on ait nécessairement l'égalité.
2) $P^\infty = \mathrm{cone}\{y_1, \dots, y_q\}$.
3) $P^\oplus = \{\sum_i \alpha_i x_i + \sum_j \beta_j y_j : \sum_i \alpha_i = 1,\ \alpha > 0,\ \beta > 0\}$.
4) $\mathrm{cone}\, P = \mathrm{cone}\{x_1, \dots, x_p, y_1, \dots, y_q\}$, qui est donc un cône polyédrique.
5) La somme de deux polyèdres convexes est un polyèdre convexe (donnez la représentation primale de cette somme).
6) L'image d'un polyèdre convexe par une application linéaire est un polyèdre convexe.

2.22. *Projeté d'un polyèdre convexe.* On considère le polyèdre convexe $P := \{(x_1, x_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} : A_1 x_1 + A_2 x_2 \leqslant b\}$, où les matrices A_i sont de type $m \times n_i$ et $b \in \mathbb{R}^m$. Montrez que le projeté de P sur $\mathbb{R}^{n_1}$, qui est l'ensemble $\{x_1 \in \mathbb{R}^{n_1} : \text{il existe un } x_2 \in \mathbb{R}^{n_2} \text{ tel que } (x_1, x_2) \in P\}$, est le polyèdre convexe $P_1 := \{x_1 \in \mathbb{R}^{n_1} : y^\mathsf{T} A_1 x_1 \leqslant y^\mathsf{T} b$, pour tout $y \in \mathbb{R}^m_+ \cap \mathcal{N}(A_2^\mathsf{T})\}$.

2.23. *Image d'un cône convexe fermé par une application linéaire.* On considère le cône convexe fermé $K := \{x \in \mathbb{R}^3 : x_1^2 + x_2^2 \leqslant x_3^2,\ x_3 \geqslant 0\}$ (voir l'exercice 2.35) et l'application linéaire $A : x \in \mathbb{R}^3 \to \mathbb{R}^2 : x \mapsto (x_1, x_2 + x_3)$. Montrez que le cône $\{Ax : x \in K\}$ n'est pas fermé dans $\mathbb{R}^2$.

Conséquence. L'image par une application linéaire d'un cône convexe fermé n'est pas nécessairement fermée.

2.24. *Somme d'un cône convexe fermé et d'un sous-espace vectoriel.* Montrez que la somme d'un cône convexe fermé et d'un sous-espace vectoriel n'est pas nécessairement fermée.

2.25. *Propriétés variationnelles de la projection.* Soit $\mathbb{E}$ un espace vectoriel euclidien (produit scalaire $\langle \cdot, \cdot \rangle$ et norme associée $\| \cdot \|$). Soient C une partie convexe fermée non vide de $\mathbb{E}$ et x un point de $\mathbb{E}$. On note $\bar{x}$ le projeté de x sur C. Montrez que

$$\forall\, y \in C : \quad \langle y - x, \bar{x} - x \rangle \geqslant 0 \quad \text{et} \quad \|y - \bar{x}\| \leqslant \|y - x\|.$$

Montrez par un exemple qu'aucune de ces propriétés ne caractérise le projeté $\bar{x}$.

2.26. *Projection sur un cône convexe fermé.* polyedre convexe@polyèdre convexe!projection d'un – Soit $\mathbb{E}$ un espace vectoriel euclidien (produit scalaire $\langle \cdot, \cdot \rangle$). Soient K un cône convexe fermé non vide de $\mathbb{E}$ et $x \in \mathbb{E}$. Montrez que $\bar{x} \in K$ est le projeté de x sur K si, et seulement si,

$$\langle \bar{x} - x, \bar{x} \rangle = 0 \qquad \text{et} \qquad \forall y \in K : \quad \langle \bar{x} - x, y \rangle \geqslant 0. \tag{2.51}$$

2.27. *Décomposition de Moreau* [441]. Soient $\mathbb{E}$ un espace euclidien (produit scalaire $\langle \cdot, \cdot \rangle$), K un cône convexe fermé de $\mathbb{E}$ et K^- son cône dual négatif. On note P_K et P_{K^-} les projecteurs orthogonaux sur K et K^- respectivement. Montrez que, pour x, y et z donnés dans $\mathbb{E}$, les propriétés suivantes sont équivalentes :

(i) $z = x + y$, $x \in K$, $y \in K^-$ et $\langle x, y \rangle = 0$,
(ii) $x = \mathrm{P}_K(z)$ et $y = \mathrm{P}_{K^-}(z)$.

En déduire que

$$\mathrm{P}_K(z) = 0 \qquad \Longleftrightarrow \qquad z \in K^-. \tag{2.52}$$

La décomposition de z en $x+y$ comme en (i) est appelée la *décomposition de Moreau*.

2.28. *Projection sur $\mathcal{S}_+^n$.* Soit $A = V\Lambda V^\mathsf{T}$ la factorisation spectrale d'une matrice $A \in \mathcal{S}^n$ ($\Lambda = \mathrm{Diag}(\lambda_1, \dots, \lambda_n)$ est la matrice diagonale des valeurs propres λ_i de A). Montrez que le projeté de A sur le cône $\mathcal{S}_+^n$ des matrices semi-définies positives est la matrice $V\Lambda^+V^\mathsf{T}$, où $\Lambda^+ = \mathrm{Diag}(\lambda_1^+, \dots, \lambda_n^+)$ et $\lambda_i^+ = \max(0, \lambda_i)$.

2.29. *Projection sur une somme de sous-espaces affines.* Soient A_1 et A_2 deux sous-espaces affines d'un espace euclidien $\mathbb{E}$. Si $\bar{x}$ est le projeté d'un point $x \in \mathbb{E}$ sur $A_1 + A_2$, alors $x - \bar{x}$ est orthogonal aux sous-espaces vectoriels parallèles aux sous-espaces affines.

2.30. *Projection en deux temps.* Soit $\mathbb{E}$ un espace euclidien, $\mathcal{A}$ un sous-espace affine de $\mathbb{E}$ et C un convexe fermé non vide de $\mathbb{E}$ *contenu dans* $\mathcal{A}$. On note P_C (resp. $\mathrm{P}_\mathcal{A}$) le projecteur orthogonal sur C (resp. sur $\mathcal{A}$). Montrez que $\mathrm{P}_C = \mathrm{P}_C \circ \mathrm{P}_\mathcal{A}$.

2.31. *Réalisation de la distance entre deux ensembles.* Soient P_1 et P_2 deux ensembles non vides d'un espace euclidien $\mathbb{E}$ tels que $P_1 - P_2$ soit un convexe fermé. Alors il existe des points $\bar{x}_1 \in P_1$ et $\bar{x}_2 \in P_1$ réalisant la distance entre P_1 et P_2, c'est-à-dire tels que

$$\|\bar{x}_1 - \bar{x}_2\| = \inf_{\substack{x_1 \in P_1 \\ x_2 \in P_2}} \|x_1 - x_2\|.$$

2.32. *L'enveloppe convexe ouverte.* L'enveloppe convexe d'une partie relativement ouverte est relativement ouverte.

2.33. *Minimisation d'une fonction linéaire sur une enveloppe convexe fermée.* Soient $\mathbb{E}$ un espace euclidien, $c \in \mathbb{E}$ et $X \subseteq \mathbb{E}$. Alors

$$\inf_{x \in X} \langle c, x\rangle = \inf_{x \in \overline{\mathrm{co}}\, X} \langle c, x\rangle \tag{2.53}$$

et

$$\overline{\mathrm{co}}\left(\arg\min_{x \in X} \langle c, x\rangle\right) \subseteq \arg\min_{x \in \overline{\mathrm{co}}\, X} \langle c, x\rangle, \tag{2.54}$$

avec égalité en (2.54) si X est compact, mais pas nécessairement autrement.

2.34. *Propriétés des cônes $\mathrm{T}_x^a C$ et $\mathrm{T}_x C$.* Démontrez la proposition 2.55.

2.35. *Exemples de cônes duaux.*

1) $\mathbb{R}^n_+$ (orthant positif de $\mathbb{R}^n$) est autodual pour le produit scalaire euclidien de $\mathbb{R}^n$.
2) Le *cornet* ou *cône du second ordre* ou *cône de Lorentz*

$$\mathbb{R}^{n+1}_\triangledown := \{(x, z) \in \mathbb{R}^n \times \mathbb{R} : \|x\|_2 \leqslant z\}$$

est autodual pour le produit scalaire euclidien de $\mathbb{R}^{n+1}$.
3) L'ensemble des matrices d'ordre n symétriques semi-définies positives

$$\mathcal{S}_+^n := \{A \in \mathcal{S}^n : x^\mathsf{T} A x \geqslant 0 \text{ pour tout } x \in \mathbb{R}^n\}$$

est autodual pour le produit scalaire $\langle A, B\rangle = \mathrm{tr}\, AB$ de $\mathcal{S}^n$.
4) Le *simplexe ordonné* de $\mathbb{R}^n$, défini par

$$\mathbb{R}^n_\leqslant := \{x \in \mathbb{R}^n : x_1 \leqslant \cdots \leqslant x_n\},$$

a pour cône dual

$$(\mathbb{R}^n_\leqslant)^+ = \{d \in \mathbb{R}^n : \textstyle\sum_{i=1}^j d_i \leqslant 0, \text{ pour } j = 1, \dots, n-1, \text{ et } \sum_{i=1}^n d_i = 0\}.$$

5) L'ensemble des matrices *symétriques copositives* $\mathcal{C}^n$ et l'ensemble des matrices *complètement positives* $\mathcal{C}^{n+}$, définis par

$$\mathcal{C}^n := \{A \in \mathcal{S}^n : x^\mathsf{T} A x \geqslant 0 \text{ pour tout } x \geqslant 0\},$$
$$\mathcal{C}^{n+} := \{A \in \mathcal{S}^n : A = BB^\mathsf{T},\ B \geqslant 0\},$$

où $B \geqslant 0$ signifie que B est *positive* (c'est-à-dire que tous ses éléments B_{ij} sont positifs), sont des cônes convexes fermés non vides, duaux l'un de l'autre, et on a

$$\mathcal{C}^{n+} \subseteq \mathcal{S}^n_+ \subseteq \mathcal{C}^n.$$

2.36. *Dual d'une somme et d'une intersection de cônes.* Démontrez le corollaire 2.44.

2.37. *Intérieur et intérieur relatif du cône dual.* Soient $\mathbb{E}$ un espace euclidien (produit scalaire et norme associés notés $\langle\cdot,\cdot\rangle$ et $\|\cdot\|$ respectivement), $P \subseteq \mathbb{E}$ et $\mathrm{P}_{(\mathrm{aff}(P^+))}$ le projecteur orthogonal sur le sous-espace vectoriel $\mathrm{aff}(P^+)$, qui est l'enveloppe affine du cône dual P^+ de P. Montrez que

$$d \in \mathrm{int}(P^+) \iff \exists\, \varepsilon > 0,\ \forall\, x \in P, \text{ on a } \langle d, x\rangle \geqslant \varepsilon \|x\|. \tag{2.55}$$
$$d \in \mathrm{intr}(P^+) \iff \exists\, \varepsilon > 0,\ \forall\, x \in P, \text{ on a } \langle d, x\rangle \geqslant \varepsilon \|\mathrm{P}_{(\mathrm{aff}\, P^+)}\, x\|. \tag{2.56}$$

2.38. *Autour du lemme de Farkas.*

1) *Ajout d'un cône convexe à l'image.* Soient $\mathbb{E}$ et $\mathbb{F}$ deux espaces euclidiens, $A : \mathbb{E} \to \mathbb{F}$ une application linéaire, K un cône convexe non vide de $\mathbb{E}$ et L un cône convexe non vide de $\mathbb{F}$. Alors $\{y \in L^+ : A^* y \in K^+\}^+ = \overline{A(K) + L}$.
2) *Effet d'une translation.* Quel que soit x_0, on a $Ax_0 + \{y : A^\mathsf{T} y \geqslant 0\}^+ = \{Ax : x \geqslant x_0\}$.

2.39. *Théorèmes de l'alternative.* Voir aussi l'exercice 17.9.

1) *Théorème de l'alternative de Farkas* [202 ; 1902]. Soient $A \in \mathbb{R}^{m\times n}$ et $b \in \mathbb{R}^m$. Alors, des deux affirmations suivantes, une et une seule est vraie :
   (*i*) $\exists\, x \in \mathbb{R}^n_+ : Ax = b$,
   (*ii*) $\exists\, y \in \mathbb{R}^m : A^\mathsf{T} y \geqslant 0$ et $b^\mathsf{T} y < 0$.
2) *Théorème de l'alternative de Motzkin homogène* (1936). Soient $A \in \mathbb{R}^{m_A\times n}$, $B \in \mathbb{R}^{m_B\times n}$ et $C \in \mathbb{R}^{m_C\times n}$ des matrices ayant un même nombre de colonnes. Alors, des deux affirmations suivantes, une et une seule est vraie :
   (*i*) $\exists\, x \in \mathbb{R}^n : Ax = 0$, $Bx \leqslant 0$ et $Cx < 0$,
   (*ii*) $\exists\, (\alpha, \beta, \gamma) \in \mathbb{R}^{m_A} \times \mathbb{R}^{m_B}_+ \times \mathbb{R}^{m_C}_+ : A^\mathsf{T}\alpha + B^\mathsf{T}\beta + C^\mathsf{T}\gamma = 0$ et $\gamma \neq 0$.

   Remarque. Cette alternative peut s'utiliser sans les matrices A et B (on obtient alors l'*alternative de Gordan* [270 ; 1873]), mais pas sans la matrice C (dans ce dernier cas, (*i*) et (*ii*) sont tous les deux trivialement vrais). Elle permet d'avoir des conditions duales exprimant la *compatibilité d'égalités et d'inégalités linéaires* (strictes et non strictes) homogènes.
3) *Théorème de l'alternative de Motzkin non-homogène* [295 ; théorème 3.17]. Soient $A \in \mathbb{R}^{m_A\times n}$, $B \in \mathbb{R}^{m_B\times n}$ et $C \in \mathbb{R}^{m_C\times n}$ des matrices ayant un même nombre de colonnes et $a \in \mathbb{R}^{m_A}$, $b \in \mathbb{R}^{m_B}$ et $c \in \mathbb{R}^{m_C}$ des vecteurs. Alors, des deux affirmations suivantes, une et une seule est vraie :
   (*i*) $\exists\, x \in \mathbb{R}^n : Ax = a$, $Bx \leqslant b$ et $Cx < c$,
   (*ii*) $\exists\, (\alpha, \beta, \gamma, \gamma_0) \in \mathbb{R}^{m_A} \times \mathbb{R}^{m_B}_+ \times \mathbb{R}^{m_C}_+ \times \mathbb{R}_+ : A^\mathsf{T}\alpha + B^\mathsf{T}\beta + C^\mathsf{T}\gamma = 0$, $a^\mathsf{T}\alpha + b^\mathsf{T}\beta + c^\mathsf{T}\gamma + \gamma_0 = 0$ et $(\gamma_0, \gamma) \neq 0$.
4) *Théorème de l'alternative de Ville* [609 ; 1938]. Soit $A \in \mathbb{R}^{m\times n}$. Alors, des deux affirmations suivantes, une et une seule est vraie :

(*i*) $\exists\, x \in \mathbb{R}^n_+ \setminus \{0\} : Ax \leqslant 0$,
(*ii*) $\exists\, y \in \mathbb{R}^m_+ : A^\mathsf{T} y > 0$.

5) *Variations.*
   a) Il existe un $x \in \mathbb{R}^n$ tel que $Ax \leqslant b$ si, et seulement si, $b^\mathsf{T} y \geqslant 0$ pour tout $y \in \mathbb{R}^m_+$ tel que $A^\mathsf{T} y = 0$.
   b) Il existe un $x \in \mathbb{R}^n$ tel que $a \leqslant Ax \leqslant b$ si, et seulement si, $a^\mathsf{T} y \leqslant b^\mathsf{T} z$ pour tout $(y, z) \in \mathbb{R}^m_+ \times \mathbb{R}^m_+$ tel que $A^\mathsf{T} y = A^\mathsf{T} z$.

2.40. *Sous-système d'inégalités affines incompatibles* [125]. Soit A une matrice de type $m \times n$. Si le système d'inégalités affines $Ax \leqslant b$ n'a pas de solution, alors il existe un sous-ensemble d'indices $I \subseteq \{1, \dots, m\}$ tel que $|I| \leqslant n + 1$ et $A_I x \leqslant b_I$ n'a pas non plus de solution (on a noté A_I la sous-matrice de A formée des lignes d'indices dans I, de même pour b_I).

2.41. *Calcul de cônes normaux.* Démontrez la proposition 2.30.

2.42. *Semi-continuité supérieure du cône normal unitaire.* Démontrez la proposition 2.31.

2.43. *Cône tangent à un polyèdre convexe.* Soient $P := \{x \in \mathbb{R}^n : Ax \leqslant b\}$ un polyèdre convexe de $\mathbb{R}^n$ (A est de type $m \times n$, $b \in \mathbb{R}^m$ et $\mathbb{R}^n$ est muni du produit scalaire euclidien) et $x \in P$. Alors le cône tangent et le cône des directions admissibles sont identiques :

$$\mathrm{T}_x P = \mathrm{T}^a_x P.$$

Par ailleurs, si on note $I := \{i \in [1:m] : (Ax - b)_i = 0\}$, on a

$$\mathrm{T}_x P = \{d \in \mathbb{R}^n : (Ad)_I \leqslant 0\} \qquad \text{et} \qquad \mathrm{N}_x P = \mathrm{cone}\{A_i^\mathsf{T} : i \in I\},$$

où les A_i désignent les lignes de A, ce qui montre en particulier que $\mathrm{T}_x P$ et $\mathrm{N}_x P$ sont des cônes polyédriques.

2.44. *Cônes tangent et normal à $\mathcal{S}^n_+$.* Montrez que les cônes tangent et normal à $\mathcal{S}^n_+$ en $S \in \mathcal{S}^n_+$ s'écrivent

$$\mathrm{T}_S \mathcal{S}^n_+ = \{D \in \mathcal{S}^n : v^\mathsf{T} D v \geqslant 0, \text{ pour tout } v \in \mathcal{N}(S)\} \tag{2.57}$$

$$\mathrm{N}_S \mathcal{S}^n_+ = \{N \in \mathcal{S}^n_- : \langle S, N \rangle = 0\}. \tag{2.58}$$

3 Fonctions convexes

The fundamental idea to be understood is that the convex functions on $\mathbb{R}^n$ can be identified with certain convex subsets of $\mathbb{R}^{n+1}$ (their epigraphs), while the convex sets in $\mathbb{R}^n$ can be identified with certain convex functions on $\mathbb{R}^n$ (their indicators). These identifications make it easy to pass back and forth between a geometric approach and an analytic approach.

R.T. Rockafellar (1970). Convex Analysis [517].

Moreau and I independently in those days at first, but soon in close exchanges with each other, made the crucial changes in outlook which, I believe, created convex analysis out of convexity. For instance, he and I passed from the basic objects in Fenchel's work, which were pairs consisting of a convex set and a finite convex function on that set, to extended-real-valued functions implicitly having effective domains, for which we moreover introduced set-valued subgradient mappings.

R.T. Rockafellar sur le site Wikimization.

3.1 Définition

Soit $\mathbb{E}$ un espace vectoriel sur $\mathbb{R}$. On note $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, +\infty\}$ la droite achevée. En analyse convexe et en optimisation, il est parfois intéressant de pouvoir considérer des fonctions pouvant prendre des valeurs infinies. En analyse convexe, plutôt que de définir une fonction convexe comme un couple formé d'un ensemble convexe C et d'une fonction f définie sur C ayant une propriété bien particulière (une approche que l'on rencontre encore parfois), il est préférable de dire que cette fonction prend des valeurs infinies en dehors de C. On évite ainsi de devoir décrire C avant de définir f, en particulier lorsque f est construite par une des opérations non triviales que nous verrons dans lesquels il est compliqué de déterminer l'ensemble C. C'est ce à quoi fait allusion la seconde épigraphe de ce chapitre. De même, il peut être utile de représenter un problème d'optimisation avec contrainte $\min\{f(x) : x \in X\}$, où X est une partie de $\mathbb{E}$, par le problème sans contrainte équivalent $\min\{\tilde{f}(x) : x \in \mathbb{E}\}$, où $\tilde{f}$ prend les mêmes valeurs que f sur X et vaut $+\infty$ sur le complémentaire de X.

Cette astuce conduit à un cadre théorique permettant de traiter en même temps les problèmes avec et sans contraintes ; nous l'utiliserons souvent.

On rappelle que le *domaine effectif* (ou simplement *domaine*) d'une fonction $f : \mathbb{E} \to \overline{\mathbb{R}}$ est l'ensemble des points où elle ne prend pas la valeur $+\infty$. On le note

$$\operatorname{dom} f := \{x \in \mathbb{E} : f(x) < +\infty\}.$$

On accepte que f prenne la valeur $-\infty$ sur son domaine pour que celui-ci soit convexe lorsque f est convexe (proposition 3.2, ci-dessous), mais nous considérerons le plus souvent des fonctions ne prenant pas la valeur $-\infty$. On rappelle que l'*épigraphe* de f est la partie de l'espace produit $\mathbb{E} \times \mathbb{R}$ qui est au-dessus de son graphe :

$$\operatorname{epi} f := \{(x, \alpha) \in \mathbb{E} \times \mathbb{R} : f(x) \leqslant \alpha\}.$$

Quant à l'*épigraphe stricte*, il est obtenu en prenant l'inégalité stricte ci-dessus. On le note

$$\operatorname{epi}_s f := \{(x, \alpha) \in \mathbb{E} \times \mathbb{R} : f(x) < \alpha\}. \tag{3.1}$$

Comme le montre le point 3 de l'exercice 3.1, cet épigraphe stricte est généralement différent de l'intérieur relatif de l'épigraphe.

Définitions 3.1 On dit qu'une fonction $f : \mathbb{E} \to \overline{\mathbb{R}}$ est *convexe* si son épigraphe (ou son épigraphe stricte) est convexe dans $\mathbb{E} \times \mathbb{R}$. On dit que $f : \mathbb{E} \to \overline{\mathbb{R}}$ est *concave* si $-f$ est convexe. □

Le fait qu'il soit équivalent de prendre l'épigraphe ou l'épigraphe stricte dans cette définition est le sujet du point 2 de l'exercice 3.1.

Proposition 3.2 (charactérisation de la convexité d'une fonction) *Le domaine d'une fonction convexe est convexe. Une fonction $f : \mathbb{E} \to \overline{\mathbb{R}}$ est convexe si, et seulement si, pour tout $x, y \in \operatorname{dom} f$ et tout $t \in]0, 1[$, on a*

$$f\Big((1-t)x + ty\Big) \leqslant (1-t)f(x) + tf(y). \tag{3.2}$$

Avant de donner la démonstration de la proposition, observons que le membre de droite de (3.2) n'est pas ambigu, car les coefficients $(1 - t)$ et t sont *strictement* positifs. Donc les produits ne sont pas indéterminés (comme le produit de $-\infty$ par zéro), même si $f(x)$ ou $f(y)$ valent $-\infty$. Par ailleurs, $f(x)$ ou $f(y)$ ne peuvent pas prendre la valeur $+\infty$ (x et $y \in \operatorname{dom} f$) et donc on n'a jamais l'indétermination $\infty - \infty$.

Démonstration. Supposons que f soit convexe. Soient x, $y \in \operatorname{dom} f$, $t \in]0, 1[$ et α, $\beta \in \mathbb{R}$ tels que $\alpha > f(x)$ et $\beta > f(y)$ (de tels α et β existent car x et $y \in \operatorname{dom} f$). Comme (x, α) et (y, β) sont dans l'épigraphe de f, qui est convexe, il en est de même de $(1-t)(x, \alpha) + t(y, \beta) = ((1-t)x + ty, (1-t)\alpha + t\beta)$, ce qui se traduit par $f((1-t)x + ty) \leqslant (1-t)\alpha + t\beta$. Comme $\alpha > f(x)$ et $\beta > f(y)$ sont arbitraires, on obtient (3.2). En particulier, $((1-t)x + ty) \in \operatorname{dom} f$; donc $\operatorname{dom} f$ est convexe.

Supposons à présent que $f : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ vérifie (3.2) pour tout $x, y \in \operatorname{dom} f$ et tout $t \in]0, 1[$. Soient (x, α) et $(y, \beta) \in \operatorname{epi} f$; donc x et $y \in \operatorname{dom} f$. D'après (3.2),

on a pour $t \in]0,1[$: $f((1-t)x+ty) \leqslant (1-t)\alpha + t\beta$. Donc $(1-t)(x,\alpha) + t(y,\beta) = ((1-t)x+ty, (1-t)\alpha+t\beta) \in \operatorname{epi} f$, si bien que $\operatorname{epi} f$ est convexe. □

La signification géométrique de l'*inégalité de convexité* (3.2) est claire (voir la figure 3.1) : pour être convexe, il faut que, sur tout segment $[x,y] \subseteq \operatorname{dom} f$, f reste en-dessous de la fonction affine valant $f(x)$ en x et $f(y)$ en y.

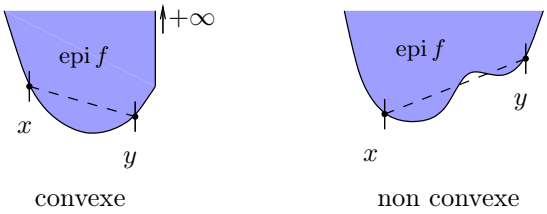


Fig. 3.1. Définition d'une fonction convexe

Définition 3.3 On dit que $f : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ est *strictement convexe* si pour tout $x, y \in \operatorname{dom} f$ avec $x \neq y$ et pour tout $t \in]0,1[$ on a

$$f\Big((1-t)x+ty\Big) < (1-t)f(x) + tf(y).$$

Sur un espace normé $\mathbb{E}$, on dit que $f : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ est *fortement convexe* de *module* $\alpha > 0$, si pour tout $x, y \in \operatorname{dom} f$ et tout $t \in [0,1]$, on a

$$f\Big((1-t)x+ty\Big) \leqslant (1-t)f(x) + tf(y) - \frac{\alpha}{2}\, t(1-t)\|x-y\|^2.$$

Si la norme dérive d'un produit scalaire, cette propriété qui revient à affirmer que $f - (\alpha/2)\|\cdot\|^2$ est convexe. □

Une fonction fortement convexe est donc strictement convexe avec une inégalité de convexité renforcée par un terme quadratique lui donnant une « courbure » au moins égale à α.

Une fonction identiquement égale à $+\infty$ est convexe (son épigraphe est vide), mais présente peu d'intérêt. Par ailleurs, une fonction convexe prenant la valeur $-\infty$ est très particulière (exercice 3.3).

Définition 3.4 (fonction propre) Une fonction définie sur un ensemble à valeurs dans $\overline{\mathbb{R}}$ est dite *propre* si elle ne prend pas la valeur $-\infty$ et n'est pas identiquement égale à $+\infty$. Dans le cas contraire, elle est dite *impropre*. □

On note

$$\operatorname{Conv}(\mathbb{E})$$

l'ensemble des fonctions de $\mathbb{E} \to \overline{\mathbb{R}}$ qui sont convexes et propres. Ce n'est pas un espace vectoriel (la différence de deux fonctions convexes n'est généralement pas convexe !), ni même un cône convexe (à moins que l'on ne suppose les fonctions à valeurs

réelles, car deux fonctions de Conv($\mathbb{E}$) peuvent avoir des domaines qui ne s'intersectent pas, auquel cas leur somme est impropre). Si $f \in \mathrm{Conv}(\mathbb{E})$, son épigraphe n'est pas vide ; donc l'intérieur relatif de celui-ci non plus (proposition 2.16). Le point 3 de l'exercice 3.1 en donne une expression explicite :

$$(\mathrm{epi}\, f)^{\circ} = \{(x,\alpha) : x \in (\mathrm{dom}\, f)^{\circ},\ f(x) < \alpha\}. \tag{3.3}$$

Il s'agit donc d'une partie de l'épigraphe stricte de f.

On rappelle qu'une une fonction $f : \mathbb{E} \to \overline{\mathbb{R}}$ est dite *fermée* si son épigraphe est fermé (il revient au même de dire que f est s.c.i.). La figure 3.2 représente à gauche

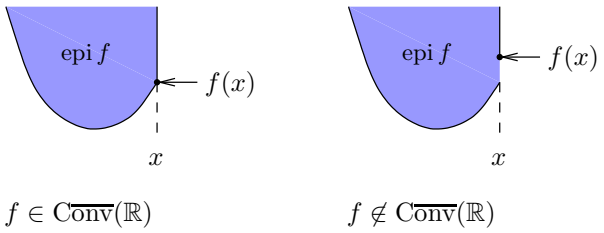


Fig. 3.2. Fonctions convexes fermée (à gauche) et non fermée (à droite)

une fonction convexe fermée et à droite une fonction convexe non fermée, ne différant de la première que par sa valeur en x. On note

$$\mathrm{C\overline{onv}}(\mathbb{E})$$

la partie de Conv($\mathbb{E}$) formée des fonctions fermées.

Considérons le problème d'optimisation suivant :

$$(P_X) \quad \begin{cases} \min\ f(x) \\ x \in X, \end{cases}$$

dans lequel on cherche à minimiser une fonction f définie sur un espace topologique X à valeurs dans $\mathbb{R} \cup \{+\infty\}$ (voir la section 1.1). Le résultat suivant donne des conditions simples assurant l'unicité de la solution de (P_X).

Théorème 3.5 (unicité de solution) *Si X est une partie convexe d'un espace vectoriel $\mathbb{E}$ et si f est strictement convexe sur X, alors (P_X) a au plus une solution.*

Démonstration. On raisonne par l'absurde. Soient $x_1 \in X$ et $x_2 \in X$ deux solutions distinctes, vérifiant donc

$$f(x_1) = f(x_2) = \min f.$$

Comme X est convexe et f est strictement convexe sur X, on a

$$f\left(\frac{x_1 + x_2}{2}\right) < \frac{1}{2} f(x_1) + \frac{1}{2} f(x_2) = \min f.$$

On aurait donc un point $x = \frac{x_1+x_2}{2}$ appartenant à X (car cet ensemble est convexe) et tel que $f(x) < f(x_1)$. Ceci contredirait l'optimalité de x_1 et x_2. □

3.2 Exemples

3.2.1 Indicatrice

On appelle *fonction indicatrice* d'une partie $P \subseteq \mathbb{E}$, ou simplement *indicatrice* de P, la fonction $\mathcal{I}_P : \mathbb{E} \to \overline{\mathbb{R}}$ définie par

$$\mathcal{I}_P(x) = \begin{cases} 0 & \text{si } x \in P \\ +\infty & \text{sinon.} \end{cases}$$

La démonstration du résultat suivant est proposée au point 1 de l'exercice 3.28.

Proposition 3.6 (fonction indicatrice) *Si P est une partie non vide de $\mathbb{E}$, alors $\mathcal{I}_P \in \mathrm{Conv}(\mathbb{E})$ (resp. $\mathcal{I}_P \in \mathrm{C}\overline{\mathrm{onv}}(\mathbb{E})$) si, et seulement si, P est convexe (resp. convexe fermé).*

3.2.2 Fonction affine et minorante affine

Une fonction $a : \mathbb{E} \to \mathbb{R}$ est dite *affine* si elle vérifie pour tout $x, y \in \mathbb{E}$ et tout $t \in \mathbb{R}$:

$$a((1-t)x + ty) = (1-t)a(x) + ta(y).$$

Ceci revient à dire que la fonction $x \in \mathbb{E} \mapsto a(x) - a(0)$ est linéaire. Il s'agit évidemment d'une fonction convexe, puisque (3.2) est vérifiée avec égalité sans condition sur $t \in \mathbb{R}$.

En dimension finie, il n'y a pas de restriction à supposer que $\mathbb{E}$ est muni d'un produit scalaire $\langle \cdot, \cdot \rangle$. Une fonction affine est alors donnée par un élément $x^* \in \mathbb{E}$ et une scalaire $\alpha \in \mathbb{R}$ et s'écrit

$$a : x \in \mathbb{E} \mapsto a(x) = \langle x^*, x \rangle - \alpha.$$

L'élément x^*, qui représente l'application linéaire $x \mapsto a(x) - a(0)$ dans $\mathbb{E}$, est appelé la *pente* de a. On voit que, si l'on munit l'espace produit $\mathbb{E} \times \mathbb{R}$ du produit scalaire $\langle (x, \alpha), (y, \beta) \rangle = \langle x, y \rangle + \alpha\beta$, l'épigraphe de a s'écrit

$$\begin{aligned} \operatorname{epi} a &= \{(x,t) \in \mathbb{E} \times \mathbb{R} : \langle x^*, x \rangle - \alpha \leqslant t\} \\ &= \{(x,t) \in \mathbb{E} \times \mathbb{R} : \langle (x^*, -1), (x,t) \rangle \leqslant \alpha\}, \end{aligned}$$

qui n'est autre que le demi-espace fermé $H^-((x^*, -1), \alpha)$ de $\mathbb{E} \times \mathbb{R}$ (notation (2.2a)).

On appelle *minorante affine* d'une fonction $f : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$, une fonction affine a qui minore f sur $\mathbb{E}$: pour tout $x \in \mathbb{E}$, $f(x) \geqslant a(x)$. L'épigraphe d'une minorante affine de f est donc un demi-espace fermé de $\mathbb{E} \times \mathbb{R}$ qui contient l'épigraphe de f. Forcément, si f a une minorante affine, elle ne peut pas prendre la valeur $-\infty$.

On dit qu'une minorante affine a de f est *exacte* en x_0 si $a(x_0) = f(x_0)$. Dans ce cas, elle s'écrit

$$a(x) = f(x_0) + \langle x^*, x - x_0 \rangle.$$

Une fonction convexe et propre a une minorante affine et celle-ci peut être choisie exacte en un point donné de $(\operatorname{dom} f)^{\diamond}$, c'est ce qu'affirme la proposition suivante.

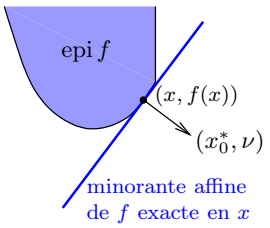
Proposition 3.7 (existence de minorante affine) *Si $f \in \operatorname{Conv}(\mathbb{E})$ et $x \in (\operatorname{dom} f)^{\diamond}$, alors il existe x^* dans le sous-espace vectoriel parallèle à $\operatorname{aff}(\operatorname{dom} f)$ tel que*

$$\forall\, y \in \mathbb{E} : \quad f(y) \geqslant f(x) + \langle x^*, y - x \rangle.$$

Démonstration. Observons d'abord que $(x, f(x))$ est sur la frontière relative de l'ensemble convexe fermé $C := \operatorname{adh}(\operatorname{epi} f)$. En effet, $(x, f(x))$ est sur la frontière relative de l'épigraphe de f (un convexe non nécessairement fermé), puisque $(x, f(x)) \in \operatorname{epi} f$ et que $(x, \alpha) \in (\operatorname{aff}(\operatorname{epi} f)) \setminus (\operatorname{epi} f)$ pour tout $\alpha < f(x)$. Donc $(x, f(x)) \notin (\operatorname{epi} f)^{\diamond}$. Mais ce dernier ensemble est aussi $C^{\diamond}$ (point 3 de la proposition 2.16). Comme $(x, f(x))$ appartient à C mais pas à $C^{\diamond}$, il est sur la frontière relative de C.

La proposition 2.29 nous assure alors qu'il existe une normale non nulle (x_0^*, ν) à C en $(x, f(x))$ qui est dans l'espace vectoriel parallèle à $\operatorname{aff}(\operatorname{epi} f) = \operatorname{aff}(\operatorname{dom} f) \times \mathbb{R}$ (point 1 de l'exercice 3.1) qui est $\mathbb{E}_0 \times \mathbb{R}$, où $\mathbb{E}_0$ est l'espace vectoriel parallèle à $\operatorname{aff}(\operatorname{dom} f)$. Pour tout $(y, \alpha) \in \operatorname{epi} f$, on a donc $\langle (y, \alpha) - (x, f(x)), (x_0^*, \nu) \rangle \leqslant 0$ ou, si l'on munit $\mathbb{E} \times \mathbb{R}$ du produit scalaire produit (A.10) :

$$\forall\, (y, \alpha) \in \operatorname{epi} f : \quad \langle x_0^*, y - x \rangle + \nu(\alpha - f(x)) \leqslant 0.$$



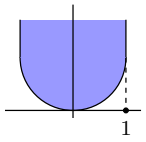
Exploitons cette inégalité :

- on peut prendre $\alpha \uparrow \infty$ dans cette inégalité, car $(\alpha, f(x))$ reste alors dans l'épigraphe de f, ce qui montre que $\nu \leqslant 0$;
- on ne peut avoir $\nu = 0$ (normale horizontale), car $y = x + t x_0^* \in \operatorname{dom} f$ pour $t > 0$ petit (car $x \in (\operatorname{dom} f)^{\diamond}$ et $x_0^* \in \mathbb{E}_0$), si bien que l'inégalité ci-dessus impliquerait $t\|x_0^*\|^2 \leqslant 0$ et donc x_0^* serait aussi nul, contredisant le fait que $(x_0^*, \nu) \neq 0$.

Alors, en définissant $x^* = -x_0^*/\nu$ et en prenant $\alpha = f(y)$ dans l'inégalité exposée ci-dessus, on obtient l'inégalité de l'énoncé. □

Comme souvent, on a obtenu des résultats confortables en des points *qui ne sont pas sur la frontière relative des ensembles convexes considérés*. Dans le cas présent, si x est sur la frontière relative du domaine de f, le résultat de la proposition précédente ne tient plus nécessairement. Ceci est illustré par la fonction d'une variable $x \in \mathbb{R}$ suivante, qui a un graphe en demi-cercle :

$$f(x) = \begin{cases} 1 - \sqrt{1-x^2} & \text{si } x \in [-1,1] \\ +\infty & \text{sinon.} \end{cases}$$



Cette fonction convexe a évidemment une minorante affine (par exemple la fonction nulle), mais n'a pas de minorante affine exacte en $x = 1$ (la tangente y est verticale et celle-ci n'est pas le graphe d'une fonction affine ; $\nu = 0$ dans la démonstration précédente).

3.2.3 Fonction convexe polyédrique

On dit qu'une fonction $f : \mathbb{E} \to \overline{\mathbb{R}}$ est (*convexe*) *polyédrique* si son épigraphe est un polyèdre convexe. Une fonction polyédrique est donc nécessairement convexe et fermée (elle peut ne pas être propre cependant).

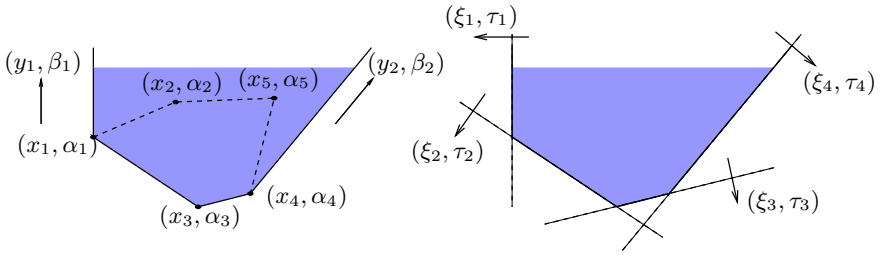


Fig. 3.3. Représentation primale (à gauche) et duale (à droite) d'une fonction polyédrique

Selon la représentation choisie du polyèdre convexe de $\mathbb{E} \times \mathbb{R}$ qu'est l'épigraphe d'un fonction polyédrique f (voir la section 2.4), on obtient une représentation primale ou duale de f. Dans la représentation primale (figure 3.3, à gauche), epi f s'écrit

$$\operatorname{epi} f = \operatorname{co}\{(x_1, \alpha_1), \ldots, (x_p, \alpha_p)\} + \operatorname{cone}\{(y_1, \beta_q), \ldots, (y_q, \beta_q)\},$$

où les x_i et les y_j sont donnés dans $\mathbb{E}$ et les α_i et β_j sont donnés dans $\mathbb{R}$. Alors

$$f(x) = \inf \left\{ \sum_{1 \leqslant i \leqslant p} t_i \alpha_i + \sum_{1 \leqslant j \leqslant q} s_j \beta_j : \sum_{1 \leqslant i \leqslant p} t_i x_i + \sum_{1 \leqslant j \leqslant q} s_j y_j = x, \right.$$
$$\left. \sum_{1 \leqslant i \leqslant p} t_i = 1,\ t_i \geqslant 0,\ s_j \geqslant 0 \right\}. \qquad (3.4)$$

Dans la représentation duale (figure 3.3, à droite), epi f est vu comme une intersection d'un nombre fini de demi-espaces fermés de $\mathbb{E} \times \mathbb{R}$, qui sont des ensembles de la forme

$$\{(x, \alpha) \in \mathbb{E} \times \mathbb{R} : \langle \xi_i, x \rangle + \tau_i \alpha \leqslant t_i\},$$

où $\xi_i \in \mathbb{E}$, $\tau_i \in \mathbb{R}$ et $t_i \in \mathbb{R}$. Puisque le polyèdre convexe ainsi obtenu doit être un épigraphe de fonction, α peut y être pris aussi grand que l'on veut, si bien que les $\tau_i \leqslant 0$. Si $\tau_i < 0$, le demi-espace fermé est l'épigraphe de la fonction affine $a_i : x \mapsto -\langle \xi_i/\tau_i, x\rangle + t_i/\tau_i$. Si $\tau_i = 0$, le demi-espace fermé est vertical et la condition $\langle \xi_i, x\rangle \leqslant t_i$ impose à x d'appartenir à un demi-espace fermé de $\mathbb{E}$: le domaine de f est donc un polyèdre convexe P de $\mathbb{E}$. De ce point de vue, f s'écrit

$$f = \max_{1\leqslant i\leqslant k} a_i + \mathcal{I}_P. \tag{3.5}$$

Cette représentation permet de voir que la somme de deux fonctions convexes polyédriques est une fonction convexe polyédrique (exercice 3.15).

3.2.4 Fonction sous-linéaire

Définition 3.8 On dit qu'une fonction $\sigma : \mathbb{E} \to \mathbb{R}\cup\{+\infty\}$ est *sous-linéaire* si elle est convexe et positivement homogène de degré 1 (c'est-à-dire : pour tout $t > 0$ et tout $x \in \mathbb{E}$, on a $\sigma(tx) = t\sigma(x)$). □

Un exemple de fonction sous-linéaire est l'application $d \mapsto f'(x; d)$, dérivée directionnelle d'une fonction convexe en un point x. Nous verrons cela plus loin.

L'appellation « sous-linéaire » vient de la propriété (iii) de la proposition ci-dessous. Elle est semblable à la condition de convexité (3.2), mais doit avoir lieu ici, même si $t_1 + t_2 \neq 1$.

Proposition 3.9 (sous-linéarité) *Soit $\sigma : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ une fonction avec domaine non vide. Les propriétés suivantes sont équivalentes :*

(i) *σ est sous-linéaire ;*

(ii) *$\operatorname{epi}\sigma$ est un cône convexe de $\mathbb{E} \times \mathbb{R}$;*

(iii) *$\forall\,(x_1, x_2) \in (\operatorname{dom}\sigma)^2$ et $\forall\,(t_1, t_2) \in \mathbb{R}^2_+$, on a $\sigma(t_1x_1 + t_2x_2) \leqslant t_1\sigma(x_1) + t_2\sigma(x_2)$.*

Démonstration. [$(i) \iff (ii)$] Par définition même, la convexité de σ est équivalente à la convexité de son épigraphe. Si σ est sous-linéaire et si $(x, r) \in \operatorname{epi}\sigma$, on a $\sigma(x) \leqslant r$, puis $\sigma(tx) = t\sigma(x) \leqslant tr$ pour $t > 0$, donc $(tx, tr) \in \operatorname{epi}\sigma$, qui montre que $\operatorname{epi}\sigma$ est un cône. Inversement, si $\operatorname{epi}\sigma$ est un cône, $(tx, t\sigma(x)) \in \operatorname{epi}\sigma$ pour tout $t > 0$, c'est-à-dire $\sigma(tx) \leqslant t\sigma(x)$ pour tout $t > 0$, donc aussi $\sigma(x) \leqslant \frac{1}{t}\sigma(tx)$; on en déduit $\sigma(tx) = t\sigma(x)$ et l'homogénéité positive de σ.

[$(i) \iff (iii)$] Supposons que σ soit sous-linéaire. Soient $(x_1, x_2) \in (\operatorname{dom}\sigma)^2$ et $(t_1, t_2) \in \mathbb{R}^2_+$. Alors pour $t = t_1 + t_2 \neq 0$, on a grâce à l'homogénéité positive et la convexité de σ :

$$\sigma(t_1x_1 + t_2x_2) = t\sigma(\tfrac{t_1}{t}x_1 + \tfrac{t_2}{t}x_2) \leqslant t_1\sigma(x_1) + t_2\sigma(x_2).$$

Inversement la condition (iii) implique la convexité de σ (proposition 3.2) ainsi que, pour tout $t > 0$, $\sigma(tx) = \sigma(\frac{t}{2}x + \frac{t}{2}x) = t\sigma(x)$. □

En prenant $x_1 = -x_2 = x$ et $t_1 = t_2 = 1$ dans le point (iii) de la proposition précédente, on voit qu'une fonction sous-linéaire σ vérifie

$$\forall\, x \in \operatorname{dom}\sigma : \qquad \sigma(x) + \sigma(-x) \geqslant 0.$$

Pour que σ devienne linéaire sur un sous-espace vectoriel, il suffit que l'on ait égalité dans cette relation pour des vecteurs engendrant ce sous-espace.

Proposition 3.10 (sous-linéarité et linéarité) *Soit $\sigma : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ une fonction sous-linéaire telle que, pour des points $x_1, \ldots, x_m \in \operatorname{dom}\sigma$, on ait*

$$\sigma(x_i) + \sigma(-x_i) = 0. \tag{3.6}$$

Alors σ est linéaire sur le sous-espace vectoriel engendré par $x_1, \ldots, x_m$.

Démonstration. Par (3.6), les $-x_i \in \operatorname{dom}\sigma$. Par ailleurs si $x = \sum_{i=1}^m t_i x_i$ avec des $t_i \in \mathbb{R}$, on a

$$\begin{aligned}
\sigma(x) &= \sum_{i=1}^m \sigma\Big(|t_i|\,(\operatorname{sgn} t_i)\, x_i\Big) \\
&\leqslant \sum_{i=1}^m |t_i|\,\sigma\Big((\operatorname{sgn} t_i)\, x_i\Big) \qquad \text{[sous-linéarité]} \\
&= \sum_{i=1}^m t_i\,\sigma(x_i) \qquad \text{[(3.6)]}.
\end{aligned}$$

Donc aussi

$$-\sigma(x) \leqslant \sigma(-x) \leqslant \sum_{i=1}^m t_i\,\sigma(-x_i) = -\sum_{i=1}^m t_i\,\sigma(x_i).$$

On en déduit $\sigma(x) = \sum_{i=1}^m t_i\,\sigma(x_i)$, c'est-à-dire la linéarité de σ. □

3.2.5 Fonction d'appui

Soit $\mathbb{E}$ un espace euclidien (produit scalaire $\langle\cdot,\cdot\rangle$). La *fonction d'appui* d'une partie non vide $P \subseteq \mathbb{E}$ est la fonction $\sigma_P : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ définie par

$$\sigma_P(d) := \sup_{x \in P} \langle d, x\rangle. \tag{3.7}$$

Elle donne donc la valeur optimale du problème de maximisation de la fonction linéaire $x \in \mathbb{E} \mapsto \langle d, x\rangle$ sur P.

Proposition 3.11 *Soient P, P_1 et P_2 des parties non vides de $\mathbb{E}$. Alors*
1) *σ_P est sous-linéaire et fermée,*
2) *$\sigma_{P_1} \leqslant \sigma_{P_2} \iff \overline{\text{co}}P_1 \subseteq \overline{\text{co}}P_2$,*

3) $\sigma_{P_1} = \sigma_{P_2} \iff \overline{\text{co}}P_1 = \overline{\text{co}}P_2$,
4) $\sigma_P = \sigma_{\text{co}\,P} = \sigma_{\overline{P}} = \sigma_{\overline{\text{co}}P}$,
5) $\text{dom}\,\sigma_P \subseteq ((\overline{\text{co}}P)^\infty)^-$, $\text{dom}\,\sigma_P = (P^\infty)^-$ *si* P *est un polyèdre convexe.*

DÉMONSTRATION. 1) Il s'agit de l'enveloppe supérieure de fonctions linéaires ; celles-ci étant convexes et fermées, la fonction d'appui est convexe et fermée (proposition 3.34). L'homogénéité positive est immédiate.

2) $\boxed{\Rightarrow}$ Il suffit de montrer que $P_1 \subseteq \overline{\text{co}}P_2$. Si ce n'est pas le cas, il existe un point $\hat{x} \in P_1 \setminus \overline{\text{co}}P_2$. On sépare strictement le compact $\{\hat{x}\}$ du fermé $\overline{\text{co}}P_2$: il existe $d \in \mathbb{E}$ et $\alpha \in \mathbb{R}$ tels que

$$\forall\, x \in \overline{\text{co}}P_2, \quad \langle d, x\rangle \leqslant \alpha < \langle d, \hat{x}\rangle.$$

On en déduirait que $\sigma_{P_2}(d) < \sigma_{P_1}(d)$, une contradiction.

$\boxed{\Leftarrow}$ Réciproquement, supposons qu'il existe $d \in \mathbb{E}$ tel que $\sigma_{P_2}(d) \leqslant \alpha < \sigma_{P_1}(d)$. Alors P_2 est contenu dans le demi-espace fermé $H^-(d,\alpha) := \{x \in \mathbb{E} : \langle d, x\rangle \leqslant \alpha\}$, mais celui-ci ne contient pas un point de $\hat{x} \in P_1$. Alors $\overline{\text{co}}P_2 \subseteq H^-(d,\alpha)$ et on aurait $P_1 \not\subseteq \overline{\text{co}}P_2$.

3) C'est une conséquence immédiate du point 2.

4) C'est une conséquence du point 3 puisque $\overline{\text{co}}P = \overline{\text{co}}(\text{co}\,P) = \overline{\text{co}}\,\overline{P} = \overline{\text{co}}(\overline{\text{co}}P)$.

5) On note $C := \overline{\text{co}}P$. Si $d \notin (C^\infty)^-$, il existe une direction $p \in C^\infty$ telle que $2\alpha := \langle d, p\rangle > 0$. Le fait que $p \in C^\infty$ implique à son tour qu'il existe des suites $\{x_k\} \subseteq C$ et $\{t_k\} \to \infty$ telles que $x_k/t_k \to p$ (point 1 de la proposition 2.7). Dès lors, pour k assez grand, on a $\langle d, x_k\rangle \geqslant t_k\alpha$, ce qui implique que $\sigma_C(d) \geqslant \sup_k \langle d, x_k\rangle = +\infty$. Comme $\sigma_C = \sigma_P$ (point 4), l'inclusion $\text{dom}\,\sigma_P \subseteq (C^\infty)^-$ est démontrée.

Si P est un polyèdre convexe, il peut s'écrire $P = \{x \in \mathbb{E} : Ax \leqslant b\}$ pour une application linéaire $A : \mathbb{E} \to \mathbb{R}^m$ et $b \in \mathbb{R}^m$. On sait qu'alors $P^\infty = \{p \in \mathbb{E} : Ap \leqslant 0\}$ (point 4 de l'exercice 2.20). Soit $d \in (P^\infty)^-$. Par le lemme de Farkas (proposition 2.45), on peut écrire $d = A^*v$ avec $v \geqslant 0$ dans $\mathbb{R}^m$. Alors pour tout $x \in P$, on a $\langle d, x\rangle = v^\mathsf{T}(Ax) \leqslant v^\mathsf{T}b$ (var $v \geqslant 0$ et $Ax \leqslant b$), ce qui montre que $\sigma_P(d) \leqslant v^\mathsf{T}b < +\infty$, c'est-à-dire que $d \in \text{dom}\,\sigma_P$. □

On remarquera que l'on peut avoir $\text{dom}\,\sigma_C \neq (C^\infty)^-$, si C est un convexe fermé non vide, *non polyédrique*. C'est le cas si $C = \{x \in \mathbb{R}^2 : x_2 \geqslant x_1^2\}$, puisqu'alors $C^\infty = \mathbb{R}_+e^2$, $e^1 \in (C^\infty)^-$, mais que $\sigma_C(e^1) \geqslant \sup_k (e^1)^\mathsf{T}(k^2, k^2) = \sup_k k = \infty$.

Proposition 3.12 (calcul de fonctions d'appui) 1) *Pour* $i \in \{1, \ldots, m\}$, *on suppose donnés des parties non vides* P_i *de* $\mathbb{E}$ *et des scalaires* $\alpha_i \geqslant 0$. *Alors*

$$\sigma_{(\sum_{i=1}^m \alpha_i P_i)} = \textstyle\sum_{i=1}^m \alpha_i\, \sigma_{P_i}.$$

2) *Soient* $\mathbb{F}$ *un espace euclidien,* $A : \mathbb{E} \to \mathbb{F}$ *une fonction linéaire,* A^* *son adjointe et* P *une partie non vide de* $\mathbb{E}$. *Alors, pour tout* $h \in \mathbb{F}$ *:*

$$\sigma_{A(P)}(h) = \sigma_P(A^*h).$$

DÉMONSTRATION. 1) Soit $d \in \mathbb{E}$. On a

$$\sigma_{(\sum_{i=1}^m \alpha_i P_i)}(d) = \sup_{x_i \in P_i} \langle d, \sum_{i=1}^m \alpha_i x_i \rangle = \sup_{x_i \in P_i} \Big(\sum_{i=1}^m \alpha_i \langle d, x_i \rangle \Big).$$

Ci-dessus, le supremum est pris sur chaque terme séparément, si bien que l'on peut sortir la somme du sup, ce qui conduit au résultat.

2) On a $\sigma_P(A^*h) = \sup\{\langle A^*h, x\rangle : x \in P\} = \sup\{\langle h, Ax\rangle : x \in P\} = \sigma_{A(P)}(h)$. □

3.3 Régularité

3.3.1 Continuité lipschitzienne

Les fonctions convexes sont régulières sur leur domaine ; plus précisément elles sont lipschitziennes sur tout convexe compact contenu dans l'intérieur relatif de leur domaine. C'est ce que nous allons montrer à la proposition 3.14 ci-dessous. Dès lors, les accidents ne peuvent arriver que sur la frontière (relative) du domaine. Ce sont ces accidents-frontière qui peuvent rendre la fonction non fermée ou non sous-différentiable. Il faudra donc à l'avenir se concentrer sur le comportement de f sur la frontière relative de son domaine.

Lemme 3.13 *Si $f \in \mathrm{Conv}(\mathbb{E})$ est bornée (supérieurement et inférieurement) sur un voisinage de $\mathbb{E}$, alors f est localement lipschitzienne sur l'intérieur de ce voisinage (tout point de ce voisinage est contenu dans un voisinage sur lequel f est lipschitzienne).*

DÉMONSTRATION. Par hypothèse, il existe un voisinage V et une constante $C > 0$ tels que $|f(x)| \leqslant C$ lorsque $x \in V$. Fixons $x \in \mathrm{int}(V)$ et $\delta > 0$ tels que $\bar{B}(x, \delta) \subseteq V$ et montrons que f est lipschitzienne de module $\frac{4C}{\delta}$ sur $U := \bar{B}(x, \frac{\delta}{2})$.

Soient $y, y' \in U$, $y \neq y'$, et posons $\Delta := \|y - y'\| \in]0, \delta]$. Avec

$$y'' := y' + \frac{\delta}{2\Delta}(y' - y) \in \bar{B}(x, \delta)$$

on a

$$y' = \frac{2\Delta}{2\Delta + \delta} y'' + \frac{\delta}{2\Delta + \delta} y.$$

Par convexité de f et sa bornitude sur $\bar{B}(x, \delta)$, on obtient alors

$$f(y') \leqslant \frac{2\Delta}{2\Delta+\delta} f(y'') + \frac{\delta}{2\Delta+\delta} f(y) = f(y) + \frac{2\Delta}{2\Delta+\delta}(f(y'') - f(y)) \leqslant f(y) + \frac{4C}{\delta}\|y - y'\|.$$

On obtient le résultat en inversant le rôle de y et y' dans l'inégalité ci-dessus. □

Les voisinages U sur lesquels f est lipschitzienne sont éventuellement plus petits que le voisinage V sur lequel elle est bornée et la constante de Lipschitz sur U pourra

être d'autant plus grande que U entoure un point proche du bord de V. C'est le cas par exemple pour la fonction convexe définie par

$$\begin{cases} -\log x & \text{si } x > 0 \\ +\infty & \text{si } x \leqslant 0. \end{cases}$$

Cette fonction n'est pas lipschitzienne sur $]0,\infty[$, mais l'est sur $]\varepsilon,\infty[$, quel que soit $\varepsilon > 0$ (le module de Lipschitz peut y être pris égal à $1/\varepsilon$).

Proposition 3.14 (continuité lipschitzienne) *Si $f \in \mathrm{Conv}(\mathbb{E})$, alors f est lipschitzienne sur tout convexe compact inclus dans $(\mathrm{dom}\, f)^\circ$.*

Démonstration. Sans perte de généralité, on peut supposer que le domaine de f est d'intérieur non vide (dans le cas contraire, on travaille sur $\mathrm{aff}(\mathrm{dom}\, f)$, comme dans la démonstration de la proposition 3.7). Soient $n = \dim \mathbb{E}$ et K un convexe compact non vide de $(\mathrm{dom}\, f)^\circ$.

Montrons, en utilisant le lemme 3.13, que *pour tout $x_0 \in K$, il existe un $\varepsilon_0 > 0$ et une constante $L > 0$ tels que $B(x_0,\varepsilon_0) \subseteq (\mathrm{dom}\, f)^\circ$ et f est lipschitzienne de module L sur $B(x_0,\varepsilon_0)$*. Soit $x_0 \in K$. Comme dans la démonstration de la proposition 2.16, on peut construire un simplexe

$$\Delta = \mathrm{co}\{z_0, z_1, \dots, z_n\} \subseteq \mathrm{dom}\, f, \qquad \text{avec } x_0 \in \Delta^\circ.$$

Dès lors, il existe $\varepsilon_0' > 0$ tel que $B(x_0,\varepsilon_0') \subseteq \Delta$. Alors f est majorée par une constante M sur cette boule, car tout $x \in \Delta$ s'écrit $x = \sum_{i=0}^n \alpha_i z_i$ avec $(\alpha_0,\dots,\alpha_n) \in \Delta_{n+1}$, si bien que, par convexité de f, on a

$$f(x) \leqslant \sum_{i=0}^n \alpha_i f(z_i) =: M.$$

Par ailleurs, f est aussi minorée inférieurement sur $B(x_0,\varepsilon_0')$, car elle a une minorante affine (proposition 3.7). Par le lemme 3.13, il existe un $\varepsilon_0 \in]0,\varepsilon_0']$ tel que f est lipschitzienne sur $B(x_0,\varepsilon_0)$.

En utilisant la propriété de Heine-Borel pour le compact K, on peut déterminer des points $x_1,\dots,x_p \in K$, tels qu'avec les ε_i déterminés comme ci-dessus, les boules $B(x_0,\varepsilon_i)$ recouvrent K et f est L_i-lipschitzienne sur $B(x_0,\varepsilon_i)$. En prenant $L = \max(L_1,\dots,L_p)$, on voit que f est L-lipschitzienne sur K. En effet, un segment $[x,y] \subseteq K$ sera divisé par ces boules en au plus p sous-segments ($[x,y]$ intersecte chaque boule en un unique sous-segment) sur chacun desquels f est L-lipschitzienne ; en ordonnant ces sous-segments de $[x,y]$, on trouve que $|f(y) - f(x)| \leqslant L\|y - x\|$. □

3.3.2 Différentiabilité

En tout point où elle prend une valeur finie, une fonction convexe admet des dérivées directionnelles suivant toutes directions (on se rappelle que l'on n'a pas besoin de topologie sur $\mathbb{E}$ pour que cette notion de différentiabilité ait un sens). C'est ce que

l'on montre avec la proposition suivante. Dans cette affirmation, la notion de dérivée directionnelle est prise dans un sens élargi : on accepte les valeurs $-\infty$ ou $+\infty$. Ceci revient à dire que la limite dans

$$f'(x;d) = \lim_{t \downarrow 0} \frac{f(x+td) - f(x)}{t},$$

qui est normalement prise dans $\mathbb{R}$ pour les fonctions réelles, est prise ici dans $\overline{\mathbb{R}}$.

Proposition 3.15 (différentiabilité directionnelle) *Soient $\mathbb{E}$ un espace vectoriel, $f : \mathbb{E} \to \overline{\mathbb{R}}$ une fonction convexe, $x \in \mathbb{E}$ un point tel que $f(x)$ est fini et $d \in \mathbb{E}$. Alors*

(i) la fonction

$$t \in \mathbb{R}_{++} \mapsto \frac{f(x+td) - f(x)}{t} \in \overline{\mathbb{R}}$$

est croissante ;

(ii) $f'(x;d)$ existe dans $\overline{\mathbb{R}}$ (elle vaut éventuellement $-\infty$ ou $+\infty$) ;

(iii) $f'(x;d)$ vaut $+\infty$ si, et seulement si, $x + td \notin \operatorname{dom} f$ pour tout $t > 0$;

(iv) on a

$$f'(x;d) \geqslant -f'(x;-d)\,; \tag{3.8}$$

en particulier, si l'une des deux dérivées directionnelles $f'(x;d)$ ou $f'(x;-d)$ vaut $-\infty$ l'autre vaut $+\infty$.

Démonstration. [(i)] Soient $0 < t_1 < t_2$. Si $x + t_2 d \notin \operatorname{dom} f$, la croissance du quotient dans (i) est claire. Si $x + t_2 d \in \operatorname{dom} f$, alors $x + t_1 d \in \operatorname{dom} f$ (convexité de $\operatorname{dom} f$ et $x \in \operatorname{dom} f$) et par la convexité de f, on a

$$\begin{aligned} f(x+t_1 d) &= f\left(\left(1 - \frac{t_1}{t_2}\right) x + \frac{t_1}{t_2}(x + t_2 d)\right) \\ &\leqslant \left(1 - \frac{t_1}{t_2}\right) f(x) + \frac{t_1}{t_2} f(x + t_2 d). \end{aligned}$$

Comme $f(x) \in \mathbb{R}$, on en déduit

$$\frac{f(x+t_1 d) - f(x)}{t_1} \leqslant \frac{f(x+t_2 d) - f(x)}{t_2}.$$

[(ii) et (iii)] Donc, lorsque $t \downarrow 0$ de façon monotone, la fonction ξ définie en (i) décroît. Si $x + td \notin \operatorname{dom} f$ pour tout $t > 0$, alors $\xi(t) = +\infty$ et $f'(x;d) = +\infty$. Dans le cas contraire, la limite $\xi(0^+)$ existe comme limite d'une suite décroissante (et vaut éventuellement $-\infty$).

[(iv)] Si $f'(x;d)$ ou $f'(x;-d) = +\infty$, il n'y a rien à démontrer. Dans le cas contraire, $x - td$ et $x + td \in \operatorname{dom} f$ pour $t > 0$ assez petit. Pour ces $t > 0$, la convexité de f implique que

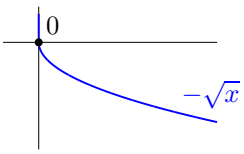
$$f(x) = f\left(\frac{x+td}{2} + \frac{x-td}{2}\right) \leqslant \frac{1}{2} f(x+td) + \frac{1}{2} f(x-td).$$

On en déduit (iv) par passage à la limite lorsque $t \downarrow 0$, après avoir retranché $f(x)$ aux deux membres. □

Voici deux remarques sur ce résultat.

- D'après le point (iii), on n'aura $f'(x;d) = +\infty$ que si $f(x+td) = +\infty$ pour tout $t > 0$. Mais on peut très bien avoir $f'(x;d) = -\infty$, alors que $f(x+td) > -\infty$ pour tout $t > 0$. C'est le cas en $x = 0$ pour la fonction convexe définie par

$$f(x) = \begin{cases} -\sqrt{x} & \text{si } x \geqslant 0 \\ +\infty & \text{sinon} \end{cases} \tag{3.9}$$



  et la direction $d = 1$.

- Le point (iv) montre que l'on peut comparer $f'(x;d)$ et $f'(x;-d)$ lorsque la fonction f est convexe. Si f n'est pas dérivable en x, en général $f'(x;d) \neq -f'(x;-d)$. Par exemple, si $f(x) = |x|$, $x \in \mathbb{R}$, on a $f'(0;1) = f'(0;-1) = 1$.

  Par ailleurs, la fonction (3.9) donne un exemple d'application de la formule (3.8) avec des valeurs infinies : $f'(0;1) = -\infty$ implique que $f'(0;-1) = +\infty$.

Pour x fixé tel que $f(x) \in \mathbb{R}$, la proposition suivante étudie les propriétés de l'*application dérivée directionnelle*

$$\delta_x : d \in \mathbb{E} \mapsto \delta_x(d) = f'(x;d) \in \overline{\mathbb{R}}. \tag{3.10}$$

Proposition 3.16 (application dérivée directionnelle) *Soient $\mathbb{E}$ un espace normé, $f : \mathbb{E} \to \overline{\mathbb{R}}$ une fonction convexe et $x \in \mathbb{E}$ un point tel que $f(x)$ soit fini. Alors l'application dérivée directionnelle (3.10) vérifie les propriétés suivantes :*

1) *δ_x est sous-linéaire,*
2) *si $x \in (\operatorname{dom} f)^{\circ}$, alors*
   a) $\operatorname{dom} \delta_x$ est le sous-espace vectoriel $\mathbb{E}_0$ de $\mathbb{E}$ parallèle à $\operatorname{aff}(\operatorname{dom} f)$,
   b) $\delta_x \in \operatorname{Conv}(\mathbb{E})$,
   c) δ_x est lipschitzienne sur $\mathbb{E}_0$.

Démonstration. 1) Montrons d'abord la convexité de δ_x. Soient d_1, $d_2 \in \operatorname{dom} \delta_x$ et $\alpha \in [0,1]$. Alors $x + td_1$ et $x + td_2 \in \operatorname{dom} f$ pour $t > 0$ assez petit (proposition 3.15, point (iii)). Par convexité de f, on a

$$\begin{aligned}
\delta_x((1-\alpha)d_1 + \alpha d_2) &= \lim_{t\downarrow 0} \frac{1}{t}\Big\{ f\Big(\underbrace{x + t[(1-\alpha)d_1 + \alpha d_2]}_{\substack{=(1-\alpha)[x+td_1]+\alpha[x+td_2] \\ \in \operatorname{dom} f \text{ pour } t>0 \text{ petit}}} \Big) - f(x) \Big\} \\
&\leqslant \lim_{t\downarrow 0} \frac{1}{t}\Big\{ (1-\alpha) f(x+td_1) + \alpha f(x+td_2) - f(x) \Big\} \\
&= \lim_{t\downarrow 0} \frac{1}{t}\Big\{ (1-\alpha)\Big(f(x+td_1) - f(x) \Big) + \alpha \Big(f(x+td_2) - f(x) \Big) \Big\} \\
&= (1-\alpha)\delta_x(d_1) + \alpha \delta_x(d_2).
\end{aligned}$$

Montrons à présent l'homogénéité positive de degré 1 de δ_x. Soient $d \in \mathbb{E}$ et $\alpha > 0$. On a

$$\delta_x(\alpha d) = \lim_{t \downarrow 0} \alpha \frac{f(x + t\alpha d) - f(x)}{\alpha t} = \alpha \lim_{\tau \downarrow 0} \frac{f(x + \tau d) - f(x)}{\tau} = \alpha \delta_x(d).$$

2) On suppose à présent que $x \in (\operatorname{dom} f)^\circ$ (donc f est propre, voir l'exercice 3.3) et on note $\mathbb{E}_0$ le sous-espace vectoriel parallèle à $\operatorname{aff}(\operatorname{dom} f)$.

2.a) Si $d \notin \mathbb{E}_0$, $x + td \notin \operatorname{dom} f$ pour tout $t > 0$, si bien que $\delta_x(d) = +\infty$ et $d \notin \operatorname{dom} \delta_x$. Par ailleurs, si $d \in \mathbb{E}_0$, $x + td \in \operatorname{dom} f$ pour tout $t > 0$ petit (car $x \in (\operatorname{dom} f)^\circ$, si bien que $\delta_x(d) < +\infty$ et $d \in \operatorname{dom} \delta_x$.

2.b) La fonction δ_x est convexe (point 1) et propre (car son domaine est un sous-espace vectoriel et que $\delta_x(0) = 0$, donc elle ne peut pas prendre la valeur $-\infty$ sur son domaine). Elle est donc localement lipschitzienne sur $\mathbb{E}_0$ (proposition 3.14), donc certainement s.c.i. sur $\mathbb{E}_0$.

2.c) D'après la proposition 3.14, f est localement lipschitzienne sur $(\operatorname{dom} f)^\circ$. Comme $x \in (\operatorname{dom} f)^\circ$, il existe $r > 0$ tel que f soit lipschitzienne sur $B(x, r) \cap \operatorname{aff}(\operatorname{dom} f)$, disons de module $L > 0$, Alors pour $d \in \mathbb{E}_0$ et $t > 0$ assez petit, $|f(x + td) - f(x)| \leqslant Lt\|d\|$. En divisant par $t > 0$ en passant à la limite lorsque $t \downarrow 0$, on trouve

$$\forall\, d \in \mathbb{E}_0, \quad |\delta_x(d)| \leqslant L\|d\|. \tag{3.11}$$

Soient d_1 et $d_2 \in \mathbb{E}_0$. Dès que $t \geqslant 1$, la monotonie du quotient différentiel de la fonction convexe δ_x (proposition 3.15) et l'homogénéité positive de cette même fonction assurent que

$$\begin{aligned} \delta_x(d_2) - \delta_x(d_1) &\leqslant \frac{\delta_x(d_1 + t(d_2 - d_1)) - \delta_x(d_1)}{t} \\ &= \delta_x\left(\frac{1}{t} d_1 + d_2 - d_1\right) - \delta_x\left(\frac{1}{t} d_1\right). \end{aligned}$$

Comme δ_x est continue sur $\mathbb{E}_0$, lorsque $t \to \infty$, le membre de droite converge vers $\delta_x(d_2 - d_1)$ qui, par (3.11), est majoré par $L\|d_2 - d_1\|$. En inversant le rôle de d_1 et d_2, on obtient $|\delta_x(d_2) - \delta_x(d_1)| \leqslant L\|d_2 - d_1\|$. □

Intéressons-nous à présent à des propriétés de différentiabilité plus forte que la différentiabilité directionnelle, qui est toujours vérifiée par les fonctions convexes, comme nous l'avons vu à la proposition 3.15. La proposition 3.18 ci-dessous montre que les trois notions de différentiabilité suivantes coïncident pour une fonction convexe $f : \mathbb{E} \to \overline{\mathbb{R}}$ en un point $x \in \mathbb{E}$ où elle prend une valeur finie (la situation est plus complexe en dimension infinie [480 ; 1993]).

- On dit que f a une *dérivée partielle* en x suivant un vecteur $d \in \mathbb{E}$ si la fonction $t \in \mathbb{R} \mapsto f(x + td)$ est différentiable en $t = 0$.
- On dit que f est *Gâteaux-différentiable* en x si la dérivée directionnelle $f'(x; d)$ existe et est finie pour tout $d \in \mathbb{E}$ et si $d \in \mathbb{E} \mapsto f'(x; d)$ est linéaire.
- On dit que f est *Fréchet-différentiable* en x s'il existe un vecteur $D \in \mathbb{E}$ tel que

$$\lim_{\substack{d \to 0 \\ d \neq 0}} \frac{f(x + d) - f(x) - \langle D, d \rangle}{\|d\|} = 0. \tag{3.12}$$

Dans ce cas, le vecteur D est appelé le *gradient* de f en x et est noté $\nabla f(x)$. D'après la définition, si f est Fréchet-différentiable en x, f prend des valeurs finies dans un voisinage de x.

Commençons par un lemme qui montre que la dérivée directionnelle en un point (non nécessairement linéaire par rapport à la direction de dérivation) est une approximation au premier ordre de f en ce point. On pourra rapprocher (3.13) de (3.12). Cette propriété est essentiellement due à la lipschitzianité de f et de $f'(x;\cdot)$.

Lemme 3.17 (développement au premier ordre) *Soient $\mathbb{E}$ un espace vectoriel de dimension finie, $f \in \mathrm{Conv}(\mathbb{E})$ et $x \in (\mathrm{dom}\, f)^\circ$. Alors*

$$\lim_{\substack{d\to 0\\ d\neq 0}} \frac{f(x+d) - f(x) - f'(x;d)}{\|d\|} = 0. \tag{3.13}$$

Démonstration. Si (3.13) n'a pas lieu, il existe un $\varepsilon > 0$ et une suite de directions non nulles $\{d_k\}$ convergeant vers zéro, tels que

$$\forall\, k \geqslant 1 : \qquad |f(x+d_k) - f(x) - f'(x;d_k)| \geqslant \varepsilon \|d_k\|.$$

En extrayant une sous-suite au besoin, on peut supposer qu'avec $t_k := \|d_k\|$, la suite bornée $\{d_k/t_k\}$ converge vers un $d \in \mathbb{E}$. Alors

$$\begin{aligned} \varepsilon \;\leqslant\; & \frac{1}{t_k} |f(x+d_k) - f(x+t_k d)| \\ & + \frac{1}{t_k} |f(x+t_k d) - f(x) - f'(x;t_k d)| \\ & + \frac{1}{t_k} |f'(x;t_k d) - f'(x;d_k)|. \end{aligned}$$

Comme $x \in (\mathrm{dom}\, f)^\circ$, f est lipschitzienne dans un voisinage de x (proposition 3.14) et $f'(x;\cdot)$ est lipschitzienne sur $\mathbb{E}$ (proposition 3.16), disons de même constante $L > 0$. Dès lors, les premier et troisième termes du membre de droite ci-dessus sont majorés par $L\|d_k/t_k - d\| \to 0$. Quant au second terme, il tend aussi vers zéro, par la définition et la positive homogénéité de la dérivée directionnelle. On aboutit donc bien à une contradiction puisque $\varepsilon > 0$. □

Proposition 3.18 (différentiabilité) *Soient $\mathbb{E}$ un espace vectoriel de dimension n, $f \in \mathrm{Conv}(\mathbb{E})$ et $x \in (\mathrm{dom} f)^\circ$. Alors les propriétés suivantes sont équivalentes :*

(*i*) *f a des dérivées partielles en x suivant n directions linéairement indépendantes,*

(ii) *f est Gâteaux-différentiable en x,*
(iii) *f est Fréchet-différentiable en x.*

DÉMONSTRATION. [$(i) \Rightarrow (ii)$] Comme $f'(x;\cdot)$ est sous-linéaire (point 1 de la proposition 3.16) et linéaire suivant n directions linéairement indépendantes (hypothèse), elle est linéaire sur l'espace vectoriel engendré par ces directions (proposition 3.10), c'est-à-dire sur $\mathbb{E}$.

[$(ii) \Rightarrow (iii)$] La linéarité supposée de $f'(x;\cdot)$ et (3.13) montrent que f est Fréchet différentiable en x.

[$(iii) \Rightarrow (i)$] Vrai quelle que soit la fonction f. □

3.3.3 Reconnaître une fonction convexe par ses dérivées

Les résultats de cette section donnent des critères (souvent même des caractérisations, c'est-à-dire des conditions nécessaires et suffisantes) de convexité d'une fonction au moyen de ses dérivées. Ces critères sont souvent le moyen le plus rapide de vérifier qu'une fonction régulière est convexe.

Pour étudier la convexité d'une fonction on est amené à examiner ses valeurs le long des segments $[x, y]$, pour tout $x, y \in \mathbb{E}$. C'est la définition qui l'impose. Chaque fois que l'on se fixe les points x et y, on se ramène à l'étude d'une fonction réelle d'une variable réelle. Les premières caractérisations (propositions 3.19, 3.21 et 3.24) n'utilisent que ces fonctions « réduites » aux segments $[x, y]$. Au contraire, les corollaires 3.20 et 3.22 et la proposition 3.26 utilisent les valeurs de la fonction sur des voisinages. On a alors besoin d'une topologie sur $\mathbb{E}$ et d'avoir une fonction f dont le domaine contienne un *ouvert*. Si ce n'est pas le cas, le domaine de f est toutefois d'intérieur relatif non vide. Lorsqu'on peut représenter ce sous-espace comme l'image d'une application linéaire continue bijective $L : \mathbb{E}_0 \to \operatorname{aff}(\operatorname{dom} f)$, il est possible d'étudier la convexité de f à partir de celle de $f \circ L$ sur un ouvert de l'espace normé $\mathbb{E}_0$.

Utilisation des dérivées premières

CONVEXITÉ

La proposition suivante montre que l'on peut caractériser la convexité d'une fonction f par le fait que celle-ci est au-dessus des fonctions affines obtenues par linéarisation de f (point (ii)) ou encore en exprimant que ses dérivées directionnelles sont monotones (point (iii)). Ceci est illustré à la figure 3.4. La fonction f qui y est représentée est convexe. On voit qu'elle est au-dessus de la fonction affine $y \mapsto f(x) + f'(x; y-x)$ et que $f'\left(y; \frac{y-x}{\|y-x\|}\right) \geqslant f'\left(x; \frac{y-x}{\|y-x\|}\right)$. On rappelle qu'une fonction convexe admet des dérivées directionnelles (pouvant valoir $-\infty$ ou $+\infty$) en tout point de son domaine (proposition 3.15).

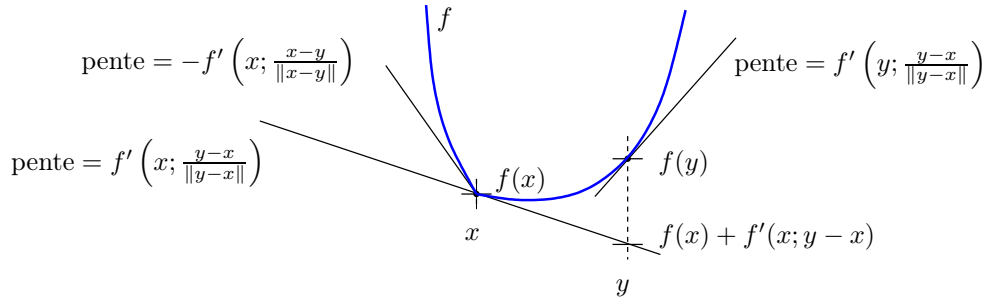


Fig. 3.4. Caractérisations de la convexité

Proposition 3.19 (convexité et dérivées premières) *Soient $\mathbb{E}$ un espace normé et $f : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ une fonction non identiquement égale à $+\infty$, ayant un domaine convexe et admettant des dérivées directionnelles en tout point de son domaine (celles-ci pouvant valoir $-\infty$ ou $+\infty$). Alors, les propriétés suivantes sont équivalentes :*

(i) f est convexe ;
(ii) $\forall\, x, y \in \operatorname{dom} f$, $x \neq y$: f est continue sur $]x, y[$ et $f(y) \geqslant f(x) + f'(x; y-x)$;
(iii) $\forall\, x, y \in \operatorname{dom} f$, $x \neq y$: f est continue sur $]x, y[$ et $f'(y; y-x) \geqslant f'(x; y-x)$.

Ces propriétés impliquent que

(iv) $\forall\, x, y \in \operatorname{dom} f$: $f'(x; y-x) + f'(y; x-y) \leqslant 0$.

Démonstration. $[(i) \Rightarrow (ii)]$ Soient $x, y \in \operatorname{dom} f$, $x \neq y$. Si $f'(x; y-x) = -\infty$, l'inégalité dans (ii) est trivialement vérifiée. Sinon $f'(x; y-x)$ a une valeur finie (par la proposition 3.15 : f est convexe et $x, y \in \operatorname{dom} f$, donc $f'(x; y-x) < +\infty$) et l'inégalité dans (ii) s'obtient grâce à la convexité de f par

$$\begin{aligned} f'(x; y-x) &= \lim_{t \downarrow 0} \frac{1}{t}\Big(f(x + t(y-x)) - f(x)\Big) \\ &= \lim_{t \downarrow 0} \frac{1}{t}\Big(f((1-t)x + ty) - f(x)\Big) \\ &\leqslant \lim_{t \downarrow 0} \frac{1}{t}\Big((1-t)f(x) + tf(y) - f(x)\Big) \\ &= f(y) - f(x). \end{aligned}$$

D'autre part, l'application $\xi : [0,1] \to \mathbb{R} : t \mapsto f((1-t)x + ty)$ est convexe et bornée supérieurement par $\max(f(x), f(y))$. Elle est aussi bornée inférieurement. Pour montrer cela, prenons par exemple le point milieu $z = \frac{x+y}{2}$. Alors $f'(z; x-z)$ et $f'(z; y-z)$ sont toutes les deux finies (elles ne valent pas $+\infty$ car f est convexe et $x, y, z \in \operatorname{dom} f$ et ne peuvent alors pas prendre la valeur $-\infty$ par la proposition 3.15 (iv)). Alors ξ est bornée inférieurement par $\inf(f(z), f(z) + f'(z; x-z), f(z) + f'(z; y-z))$. On en déduit que ξ est continue sur $]0,1[$ par le lemme 3.13.

$[(ii) \Rightarrow (i)]$ Si f n'est pas convexe, il existe $x, y \in \operatorname{dom} f$, $x \neq y$, et $\hat{t} \in]0,1[$ tel que $\hat{z} := (1-\hat{t})x + \hat{t}y$ vérifie

$$f(\hat{z}) > (1-\hat{t})f(x) + \hat{t}f(y).$$

Soit $t_1 := \inf\{\bar{t} \in [0,\hat{t}] : f((1-t)x+ty) > (1-t)f(x)+tf(y), \forall\, t \in]\bar{t},\hat{t}]\}$. Par continuité de f sur $]x,y[$, un tel t_1 existe. D'autre part, soit parce que $t_1 = 0$, soit par continuité de f sur $]x,y[$, on a avec $z_1 := (1-t_1)x + t_1 y$

$$f(z_1) = (1-t_1)f(x) + t_1 f(y).$$

Remarquons que f est continue sur $[z_1,\hat{z}]$. En effet, f étant continue sur $]x,y[$ par hypothèse, il reste à montrer la continuité de f en z_1 lorsque $t_1 = 0$. Mais si $t_1 = 0$, $\liminf_{t\downarrow 0} f((1-t)x + ty) \geqslant f(x)$. Or $\limsup_{t\to 0^+} f((1-t)x + ty)$ ne peut être $> f(x)$ sinon $f'(x; y-x)$, qui est supposé exister, vaudrait $+\infty$, ce qui contredirait l'inégalité de (ii) ($f(x)$ et $f(y)$ sont finis). Donc f est continue sur $[z_1,\hat{z}]$ et on peut appliquer la version généralisée du théorème de Rolle (corollaire C.11) : il existe $t_2 \in]t_1,\hat{t}[$ tel que $z_2 := (1-t_2)x + t_2 y$ vérifie

$$f'\left(z_2; \frac{y-z_2}{\|y-z_2\|}\right) \geqslant \frac{f(\hat{z}) - f(z_1)}{\|\hat{z}-z_1\|}.$$

On peut alors conclure comme suit :

$$\begin{aligned} f(z_2) + f'(z_2; y-z_2) &> (1-t_2)f(x) + t_2 f(y) \\ &\quad + \frac{1-t_2}{\hat{t}-t_1}\Big((1-\hat{t})f(x) + \hat{t}f(y) - (1-t_1)f(x) - t_1 f(y)\Big) \\ &= (1-t_2)f(x) + t_2 f(y) + (1-t_2)(f(y) - f(x)) \\ &= f(y), \end{aligned}$$

ce qui contredit l'inégalité de (ii).

$[(ii) \Rightarrow (iv)]$ Soient $x, y \in \operatorname{dom} f$. L'inégalité est triviale si $x = y$. Sinon, on utilise l'inégalité de (ii) en intervertissant le rôle de x et de y

$$\begin{aligned} f(y) &\geqslant f(x) + f'(x; y-x) \\ f(x) &\geqslant f(y) + f'(y; x-y). \end{aligned}$$

En sommant, on obtient (iv).

$[(i),(ii) \Rightarrow (iii)]$ Soient $x, y \in \operatorname{dom} f$, $x \neq y$. Si $f'(y; y-x) = +\infty$, (iii) est trivialement vérifiée. Sinon, d'après la proposition 3.15 (iv), $f'(y; y-x)$ est fini (dans le cas contraire, on devrait avoir $f'(y; x-y) = +\infty$, ce qui contredirait le fait que $x, y \in \operatorname{dom} f$) et $f'(y; x-y) \geqslant -f'(y; y-x)$. En utilisant cette inégalité et en intervertissant les rôles de x et y dans l'inégalité de (ii), on obtient

$$\begin{aligned} f(y) &\geqslant f(x) + f'(x; y-x) \\ f(x) &\geqslant f(y) - f'(y; y-x). \end{aligned}$$

En sommant, on obtient (iii).

$[(iii) \Rightarrow (i)]$ Si f n'est pas convexe, il existe $x, y \in \operatorname{dom} f$, $x \neq y$, et $\hat{t} \in]0,1[$ tels que $\hat{z} := (1-\hat{t})x + \hat{t}y$ vérifie

$$f(\hat{z}) > (1-\hat{t})f(x) + \hat{t}f(y).$$

On peut trouver $\hat{t}_1 \in [0, \hat{t}]$ tel que $\hat{z}_1 := (1-\hat{t}_1)x + \hat{t}_1 y$ vérifie

$$f(\hat{z}_1) = (1-\hat{t}_1)f(x) + \hat{t}_1 f(y)$$

et tel que f soit continue sur $[\hat{z}_1, \hat{z}]$. On peut s'y prendre, par exemple, comme dans la détermination de t_1 dans la démonstration de l'implication $(ii) \Rightarrow (i)$; ici, si $\hat{t}_1 = 0$, on ne peut pas avoir $\limsup_{t \downarrow 0} f((1-t)x + ty) > f(x)$ car cela impliquerait que $f'(x; y-x) = +\infty$, ce qui est en contradiction avec l'hypothèse de monotonie des dérivées directionnelles et le fait qu'il y a des points z entre x et y où $f'(z; y-x)$ est finie. Pour les mêmes raisons, on peut trouver $\hat{t}_2 \in [\hat{t}, 1]$ tel que $\hat{z}_2 := (1-\hat{t}_2)x + \hat{t}_2 y$ vérifie

$$f(\hat{z}_2) = (1-\hat{t}_2)f(x) + \hat{t}_2 f(y)$$

et tel que f soit continue sur $[\hat{z}, \hat{z}_2]$. On peut alors appliquer la variante du théorème de Rolle donnée au corollaire C.11 : il existe $t_1 \in]\hat{t}_1, \hat{t}[$ et $t_2 \in]\hat{t}, \hat{t}_2[$ tels que $z_1 := (1-t_1)x + t_1 y$ et $z_2 := (1-t_2)x + t_2 y$ vérifient

$$f'(z_1; y-x) \geqslant \frac{f(\hat{z}) - f(\hat{z}_1)}{\hat{t} - \hat{t}_1} \quad \text{et} \quad f'(z_2; y-x) \leqslant \frac{f(\hat{z}_2) - f(\hat{z})}{\hat{t}_2 - \hat{t}}.$$

On en déduit que $f'(z_2; y-x) < f(y) - f(x) < f'(z_1; y-x)$, ce qui contredit l'hypothèse de monotonie des dérivées directionnelles dans (iii). □

Sans la continuité de f sur les segments $]x, y[$, l'inégalité de (ii) peut très bien être vérifiée par une fonction non convexe. En voici un exemple sur $\mathbb{R}$: $f(x) = 0$, si $x \neq 0$, et $f(0) = 1$ (on a $f'(x; \pm 1) = 0$, si $x \neq 0$, et $f'(0; \pm 1) = -\infty$).

La démonstration de la proposition précédente est longue parce que nous avons fait des hypothèses minimales sur la régularité de f. Si l'on suppose que le domaine de f est un ouvert convexe d'un espace normé $\mathbb{E}$ et que f est dérivable sur cet ouvert, on a un résultat plus simple à mémoriser et à démontrer. Nous le donnons sous forme de corollaire.

Corollaire 3.20 *Soient $\mathbb{E}$ un espace normé, Ω un ouvert convexe de $\mathbb{E}$ et $f : \Omega \to \mathbb{R}$ une fonction dérivable. Alors, les propriétés suivantes sont équivalentes :*

(i) f est convexe sur Ω ;
(ii) $\forall\, x, y \in \Omega : f(y) \geqslant f(x) + f'(x) \cdot (y-x)$;
(iii) $\forall\, x, y \in \Omega : (f'(y) - f'(x)) \cdot (y-x) \geqslant 0$.

Convexité stricte

On a un résultat analogue à celui de la proposition 3.19 pour les fonctions strictement convexes.

Proposition 3.21 (stricte convexité et dérivées premières I) *Soient $\mathbb{E}$ un espace normé et $f : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ une fonction non identiquement égale à $+\infty$, ayant un domaine convexe et admettant des dérivées directionnelles en tout point de son domaine (celles-ci pouvant valoir $-\infty$ ou $+\infty$). Alors, les propriétés suivantes sont équivalentes :*
(i) f est strictement convexe ;
(ii) $\forall\, x, y \in \operatorname{dom} f$, $x \neq y$: f est continue sur $]x, y[$ et $f(y) > f(x) + f'(x; y-x)$;
(iii) $\forall\, x, y \in \operatorname{dom} f$, $x \neq y$: f est continue sur $]x, y[$ et $f'(y; y-x) > f'(x; y-x)$.

Démonstration. $[(i) \Rightarrow (ii)]$ Supposons que f soit strictement convexe et soient $x, y \in \operatorname{dom} f$, $x \neq y$. La continuité de f sur $]x, y[$ découle de la proposition 3.19. Pour démontrer l'inégalité stricte de (ii), on ne peut plus utiliser le même raisonnement que dans la démonstration de la proposition 3.19, car le passage à la limite détruirait l'inégalité stricte que l'on a pour les fonctions strictement convexes. On s'y prend alors comme suit. Pour $t \in]0, 1]$, on a

$$f'(x; y-x) = \frac{1}{t} f'(x; t(y-x)) \leqslant \frac{1}{t}\Big(f(x + t(y-x)) - f(x)\Big) < f(y) - f(x),$$

où pour la première inégalité on a utilisé la proposition 3.19 (ii) et pour la seconde on a utilisé la stricte convexité de f.

$[(ii) \Rightarrow (iii)]$ Si (iii) n'a pas lieu, il existe $x, y \in \operatorname{dom} f$, $x \neq y$, tels que $f'(y; y - x) \leqslant f'(x; y - x)$. Par la proposition 3.19, on a $f'(z; y - x) = f'(x; y - x)$ pour tout $z \in [x, y]$ et cette valeur est finie. On en déduit (par exemple par le corollaire C.10, en y prenant $\varphi(t) = \pm(f((1-t)x + ty) - tf'(x; y - x))$ pour $t \in [0, 1]$) que $f(y) = f(x) + f'(x; y - x)$, ce qui contredit (ii).

$[(iii) \Rightarrow (i)]$ Si f n'est pas strictement convexe, il existe $x, y \in \operatorname{dom} f$, $x \neq y$, et $\hat{t} \in]0, 1[$ tels que $\hat{z} := (1-\hat{t})x + \hat{t}y$ vérifie

$$f(\hat{z}) \geqslant (1-\hat{t}) f(x) + \hat{t} f(y).$$

Comme f est convexe (proposition 3.19) ceci ne peut avoir lieu que si f est affine entre x et y. Mais alors, la stricte monotonie des dérivées directionnelles dans (iii) ne serait pas vérifiée. □

Énonçons également sous forme de corollaire, le résultat que donne la proposition 3.21 lorsque f est dérivable.

Corollaire 3.22 *Soient $\mathbb{E}$ un espace normé, Ω un ouvert convexe de $\mathbb{E}$ et $f : \Omega \to \mathbb{R}$ une fonction dérivable. Alors, les propriétés suivantes sont équivalentes :*
(i) f est strictement convexe sur Ω ;
(ii) $\forall\, x, y \in \Omega$, $x \neq y$: $f(y) > f(x) + f'(x) \cdot (y - x)$;
(iii) $\forall\, x, y \in \Omega$, $x \neq y$: $(f'(y) - f'(x)) \cdot (y - x) > 0$.

Pour une fonction strictement convexe, contnûment différentiable, définie sur un espace de dimension finie, les inégalités de la proposition 3.21 et de son corollaire 3.22

peuvent être renforcées [252 ; lemmes 1.1 et 1.2, pages 61 et 63]. Les inégalités (3.14) et (3.15) qui sont établies dans la proposition suivante seront à nouveau renforcées pour les fonction fortement convexes (proposition 3.24).

Proposition 3.23 (stricte convexité et dérivées premières II) *Soient $\mathbb{E}$ un espace vectoriel de dimension finie, $f : \mathbb{E} \to \mathbb{R}$ une fonction de classe C^1 et $t \in]0,1[$. Alors les propriétés suivantes sont équivalentes :*

(i) f est strictement convexe,

(ii) pour tout $\beta > 0$, il existe une fonction $g_\beta : [0, 2\beta] \to \mathbb{R}_+$ continue, strictement croissante, vérifiant $g_\beta(0) = 0$ et

$$\forall\, x, y \in \beta\bar{B} : \quad f(y) - f(x) \geqslant f'(x) \cdot (y - x) + (1 - t) g_\beta(t\|y - x\|), \tag{3.14}$$

(iii) pour tout $\beta > 0$, il existe une fonction $g_\beta : [0, 2\beta] \to \mathbb{R}_+$ continue, strictement croissante, vérifiant $g_\beta(0) = 0$ et

$$\forall\, x, y \in \beta\bar{B} : \quad (f'(y) - f'(x)) \cdot (y - x) \geqslant g_\beta(\|y - x\|). \tag{3.15}$$

Démonstration. [$(i) \Rightarrow (iii)$] Soit $\beta > 0$. Au vu de la propriété désirée (3.15), il est naturel d'introduire la fonction $g^0_\beta : [0, 2\beta] \to \mathbb{R}_+$ définie en $\alpha \in [0, 2\beta]$ par

$$g^0_\beta(\alpha) := \inf_{\substack{\|y-x\|=\alpha \\ x,y \in \beta\bar{B}}} \Big(f'(y) - f'(x)\Big) \cdot (y - x).$$

Cette fonction a toutes les propriétés requises par le point (ii), hormis la continuité. Montrons cela.

1. $g^0_\beta(0) = 0$, clairement.
2. $g^0_\beta > 0$ *sur* $]0, 2\beta]$, grâce à la stricte convexité de f et au point (iii) de la proposition 3.21.
3. g^0_β *est strictement croissante sur* $[0, 2\beta]$. En effet, soient $0 < \alpha_1 < \alpha_2 \leqslant 2\beta$. Remarquons d'abord que l'infimum dans la définition de g^0_β est atteint, car $\{(x, y) \in \beta\bar{B} \times \beta\bar{B} : \|y - x\| = \alpha\}$ est compact ($\mathbb{E}$ est de dimension finie) et $(x, y) \mapsto (f'(y) - f'(x)) \cdot (y - x)$ est continue (f est continûment différentiable), puis on utilise alors le théorème de Weierstrass. Il existe donc des points $(x_2, y_2) \in \beta\bar{B} \times \beta\bar{B}$ tels que $\|y_2 - x_2\| = \alpha_2$ et

$$g^0_\beta(\alpha_2) = \Big(f'(y_2) - f'(x_2)\Big) \cdot (y_2 - x_2).$$

En introduisant $y := x_2 + (\alpha_1/\alpha_2)(y_2 - x_2)$, on a grâce à la stricte convexité de f et $0 < \alpha_1 < \alpha_2$:

$$\Big(f'(y_2) - f'(x_2)\Big) \cdot (y_2 - x_2) > \Big(f'(y) - f'(x_2)\Big) \cdot (y_2 - x_2).$$

En combinant ces deux relations, on obtient

$$\begin{aligned} g^0_\beta(\alpha_2) &> \Big(f'(y)-f'(x_2)\Big)\cdot(y_2-x_2) \\ &= \frac{\alpha_2}{\alpha_1}\Big(f'(y)-f'(x_2)\Big)\cdot(y-x_2) \qquad [\text{définition de } y] \\ &> \Big(f'(y)-f'(x_2)\Big)\cdot(y-x_2) \qquad [\alpha_2>\alpha_1 \text{ et } y\neq x_2] \\ &\geqslant g^0_\beta(\alpha_1) \qquad [\|y-x_2\|=\alpha_1]. \end{aligned}$$

D'où la croissance stricte de g^0_β.

4. Par définition même de g^0_β, (3.15) a lieu avec $g_\beta = g^0_\beta$.

Pour ontenir une fonction continue g_β, on utilise une technique de dérivation intégration. Soient x et $y \in \beta\bar{B}$. Comme ci-dessus, en intégrant la dérivée de $s \mapsto f(x+s(y-x))$, on obtient

$$f(y)-f(x) = \int_0^1 f'(x+s(y-x))\cdot(y-x)\,\mathrm{d}s.$$

Mais $x+s(y-x) \in \beta\bar{B}$ (convexité de $\beta\bar{B}$) et (3.15) a lieu avec $g_\beta = g^0_\beta$, si bien que

$$\Big(f'(x+s(y-x))-f'(x)\Big)\cdot(y-x) \geqslant \frac{1}{s}\,g^0_\beta(s\|y-x\|).$$

qui, avec l'identité précédente donne

$$f(y)-f(x) \geqslant f'(x)\cdot(y-x) + \int_0^1 \frac{1}{s}\,g^0_\beta(s\|y-x\|)\,\mathrm{d}s.$$

En échangeant x et y et en additionnant, on obtient

$$\Big(f'(y)-f'(x)\Big)\cdot(y-x) \geqslant \int_0^1 \frac{2}{s}\,g^0_\beta(s\|y-x\|)\,\mathrm{d}s = \int_0^{\|y-x\|} \frac{2}{r}\,g^0_\beta(r)\,\mathrm{d}r.$$

On obtient alors (3.15) en définissant $g_\beta : [0,2\beta] \to \mathbb{R}+$ en $s \in [0,2\beta]$ par

$$g_\beta(s) = \int_0^s \frac{2}{r}\,g^0_\beta(r)\,\mathrm{d}r.$$

Par ailleurs, il est clair que g_β est continue, strictement croissante et vérifie $g_\beta(0)=0$.

$[(iii)\Rightarrow(ii)]$ Il s'agit de montrer que (3.15) implique (3.14). Comme ci-dessus, en intégrant la dérivée de $s \mapsto f(x+s(y-x))$, on obtient

$$\begin{aligned} f(y)-f(x) &= \int_0^1 f'(x+s(y-x))\cdot(y-x)\,\mathrm{d}s \\ &\geqslant f'(x)\cdot(y-x) + \int_0^1 \frac{1}{s}\,g_\beta(s\|y-x\|)\,\mathrm{d}s \qquad [(3.15)] \\ &\geqslant f'(x)\cdot(y-x) + \int_t^1 \frac{1}{s}\,g_\beta(s\|y-x\|)\,\mathrm{d}s \qquad [g_\beta \geqslant 0] \\ &\geqslant f'(x)\cdot(y-x) + (1-t)\,g_\beta(t\|y-x\|), \end{aligned}$$

car pour $s \in [t,1]$, $g_\beta(s\|y-x\|) \geqslant g_\beta(t\|y-x\|)$ (croissance de g_β) et $s \leqslant 1$, si bien que $g_\beta(s\|y-x\|)/s \geqslant g_\beta(t\|y-x\|)$.

[$(ii) \Rightarrow (i)$] Soient x et y deux points distincts de $\mathbb{E}$. On prend $\beta := \|y-x\|$, qui est strictement positif. Par (3.14) avec $t = \frac{1}{2}$, on a $f(y) - f(x) \geqslant f'(x)\cdot(y-x) + \frac{1}{2}g_\beta(\frac{1}{2}\|y-x\|) > f'(x)\cdot(y-x)$, car $g_\beta(\frac{1}{2}\|y-x\|) > 0$. Donc f est strictement convexe. □

Convexité forte

On a aussi des caractérisations de la forte convexité au moyen d'inégalités renforcées. La démonstration du résultat est proposée à l'exercice 3.11.

Proposition 3.24 (forte convexité et dérivées premières) *Soient $\mathbb{E}$ un espace vectoriel euclidien (produit scalaire $\langle\cdot,\cdot\rangle$ et norme associée $\|\cdot\|$), $\alpha > 0$ et $f : \mathbb{E} \to \mathbb{R}\cup\{+\infty\}$ une fonction non identiquement égale à $+\infty$, ayant un domaine convexe et admettant des dérivées directionnelles en tout point de son domaine (celles-ci pouvant valoir $-\infty$ ou $+\infty$). Alors, les propriétés suivantes sont équivalentes :*

(i) f est fortement convexe de module α ;

(ii) pour tout $x, y \in \operatorname{dom} f$, $x \neq y$: f est continue sur $]x,y[$ et $f(y) \geqslant f(x) + f'(x;y-x) + \frac{\alpha}{2}\|y-x\|^2$;

(iii) pour tout $x, y \in \operatorname{dom} f$, $x \neq y$: f est continue sur $]x,y[$ et $f'(y;y-x) \geqslant f'(x;y-x) + \alpha\|y-x\|^2$.

Ce résultat s'énonce plus simplement lorsque la fonction est différentiable.

Corollaire 3.25 *Soient $\mathbb{E}$ un espace euclidien, Ω un ouvert convexe de $\mathbb{E}$ et $f : \Omega \to \mathbb{R}$ une fonction dérivable. Alors, les propriétés suivantes sont équivalentes :*

(i) f est fortement convexe sur Ω ;

(ii) $\forall\, x, y \in \Omega$: $f(y) \geqslant f(x) + f'(x)\cdot(y-x) + \frac{\alpha}{2}\|y-x\|^2$;

(iii) $\forall\, x, y \in \Omega$: $(f'(y) - f'(x))\cdot(y-x) \geqslant \alpha\|y-x\|^2$.

Utilisation des dérivées secondes

On peut également donner une caractérisation de convexité d'une fonction (mais pas de sa stricte convexité) au moyen de ses dérivées secondes.

Proposition 3.26 (convexité et dérivées secondes) *Soient $\mathbb{E}$ un espace normé, Ω un ouvert convexe de $\mathbb{E}$ et $f : \Omega \to \mathbb{R}$ une fonction deux fois dérivable. Alors f est convexe sur Ω si, et seulement si,*

$$\forall\, x \in \Omega, \quad \forall\, d \in \mathbb{E} : \quad f''(x) \cdot d^2 \geqslant 0. \tag{3.16}$$

D'autre part, si

$$\forall\, x \in \Omega, \quad \forall\, d \in \mathbb{E} \setminus \{0\} : \quad f''(x) \cdot d^2 > 0, \tag{3.17}$$

alors f est strictement convexe sur Ω.

Démonstration. Si f est convexe sur Ω, on a pour tout $x \in \Omega$ et tout $d \in \mathbb{E}$ (en utilisant les propositions C.17 et 3.19) :

$$f''(x) \cdot d^2 = \lim_{t \downarrow 0} \frac{1}{t}(f'(x + td) \cdot d - f'(x) \cdot d) \geqslant 0.$$

Inversement, supposons que (3.16) ait lieu. Si $x, y \in \Omega$, le théorème C.19 donne

$$f(y) = f(x) + f'(x) \cdot (y - x) + \frac{1}{2} f''(x + \theta(y - x)) \cdot (y - x)^2,$$

où $\theta \in]0, 1[$. Comme le dernier terme est positif ou nul, on voit par la proposition 3.19 que f est convexe. Le cas où (3.17) a lieu se démontre de la même manière. □

Remarquons que la condition (3.17) n'est pas nécessairement vérifiée pour une fonction strictement convexe comme le montre la fonction $x \in \mathbb{R} \mapsto f(x) = x^4$. Bien que celle-ci soit strictement convexe, on a $f''(0) = 0$. Par contre, pour une fonction quadratique, (3.17) est une condition nécessaire et suffisante de convexité stricte (point 2 de l'exercice 3.8).

Lorsque $\mathbb{E}$ est un espace de Hilbert, on peut traduire le résultat de la proposition 3.26 en utilisant la hessienne de f : (i) f est convexe si, et seulement si, sa hessienne $\nabla^2 f(x)$ est semi-définie positive et (ii) si $\nabla^2 f(x)$ est défini positif, alors f est strictement convexe.

3.3.4 Fonction asymptotique

L'épigraphe d'une fonction $f \in \mathrm{C}\overline{\mathrm{onv}}(\mathbb{E})$ est un convexe fermé non vide. On peut donc considérer son cône asymptotique $(\mathrm{epi}\, f)^\infty$. Celui-ci a des propriétés intéressantes.

Proposition 3.27 (définition de f^∞) *Soit $f \in \mathrm{C}\overline{\mathrm{onv}}(\mathbb{E})$. Alors*

(i) $(\mathrm{epi}\, f)^\infty$ est l'épigraphe d'une fonction $f^\infty : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$: $(\mathrm{epi}\, f)^\infty = \mathrm{epi}\, f^\infty$,

(ii) $\forall\, x \in \mathrm{dom}\, f$, $\forall\, d \in \mathbb{E}$:

$$f^\infty(d) = \lim_{t \to +\infty} \frac{f(x + td) - f(x)}{t}, \tag{3.18}$$

(*iii*) $\operatorname{dom} f^\infty \subseteq (\operatorname{dom} f)^\infty$,
(*iv*) $f^\infty \in \overline{\operatorname{Conv}}(\mathbb{E})$ *et est sous-linéaire.*

DÉMONSTRATION. Commençons par une caractérisation de l'appartenance à $(\operatorname{epi} f)^\infty$. Soit $x \in \operatorname{dom} f$ (qui est non vide) ; donc $f(x)$ est fini et $(x, f(x)) \in \operatorname{epi} f$. Alors

$$\begin{aligned}(d,\delta) \in (\operatorname{epi} f)^\infty &\iff (x,f(x)) + t(d,\delta) \in \operatorname{epi} f, \quad \forall\, t \geqslant 0 \\ &\iff f(x+td) \leqslant f(x) + t\delta, \quad \forall\, t \geqslant 0. \end{aligned} \tag{3.19}$$

[(*i*)] Si $(d,\delta) \in (\operatorname{epi} f)^\infty$ et $\delta' \geqslant \delta$, $(d,\delta') \in (\operatorname{epi} f)^\infty$, grâce à (3.19). Fixons à présent $d \in \mathbb{E}$ et posons

$$\delta_0 := \inf\{\delta : (d,\delta) \in (\operatorname{epi} f)^\infty\}.$$

Il faut montrer que $(d,\delta_0) \in (\operatorname{epi} f)^\infty$. D'après (3.19)

$$\delta_0 = \sup_{t>0} \frac{f(x+td) - f(x)}{t}.$$

On en déduit que $f(x+td) \leqslant f(x) + t\delta_0$, $\forall\, t \geqslant 0$, et donc $(d,\delta_0) \in (\operatorname{epi} f)^\infty$ (par (3.19)).

[(*ii*)] Soit f^∞ la fonction dont $(\operatorname{epi} f)^\infty$ est l'épigraphe. On a $f^\infty(d) = \delta_0$, ou encore (3.18), car le quotient ci-dessus est croissant avec t (de fait de la convexité de f).

[(*iii*)] D'après (3.19), si $d \in \operatorname{dom} f^\infty$, $(d, f^\infty(d)) \in (\operatorname{epi} f)^\infty$ et $f(x+td) \leqslant \alpha + tf^\infty(d)$, qui est fini quel que soit $t \geqslant 0$. Donc $d \in (\operatorname{dom} f)^\infty$.

[(*iv*)] D'une part, f^∞ est convexe fermée et non identiquement égale à $+\infty$, parce que son épigraphe est le convexe fermé non vide $(\operatorname{epi} f)^\infty$. D'autre part, f^∞ ne prend pas la valeur $-\infty$ car sa valeur est obtenue comme limite d'une suite croissante (voir (3.18)). Enfin f est sous-linéaire, car son épigraphe est un cône convexe (proposition 3.9). □

Définition 3.28 On appelle *fonction asymptotique* (ou *fonction de récession*) de $f \in \overline{\operatorname{Conv}}(\mathbb{E})$ la fonction $f^\infty \in \overline{\operatorname{Conv}}(\mathbb{E})$ introduite dans la proposition 3.27. □

Quelques remarques sur la proposition 3.27.

- D'après la proposition 3.15, on sait que, lorsque $x \in \operatorname{dom} f$, le quotient de la formule (3.18) croît avec t ; il converge vers $f^\infty(d)$ si $t \to +\infty$ et vers $f'(x;d)$ si $t \downarrow 0$.
- Comme $f(x) \in \mathbb{R}$, la formule du point (3.18) s'écrit aussi

$$f^\infty(d) = \lim_{t\to+\infty} \frac{f(x+td)}{t}, \tag{3.20}$$

avec un quotient $f(x+td)/t$ qui, contrairement au quotient différentiel dans (3.18), n'est pas nécessairement monotone en t.

- L'utilisation de la formule (3.20) est souvent le moyen le plus rapide de calculer la valeur de la fonction asymptotique en une direction d. Insistons sur le fait que la limite dans (3.20) ne dépend pas du point x choisi dans le domaine de f.
- La formule (3.20) montre que si la fonction $t \mapsto f(x+td)$ a une asymptote à droite, $f^\infty(d)$ en est sa pente ; sinon $f^\infty(d) = +\infty$.
- Si $x \in \operatorname{dom} f$ et $f(x+td) = +\infty$ pour un $t > 0$, alors $d \notin (\operatorname{dom} f)^\infty$ et donc $d \notin \operatorname{dom} f^\infty$ par (iii), c'est-à-dire $f^\infty(d) = +\infty$.
- On n'a pas nécessairement égalité au point (iii). En effet, si $f : \mathbb{R} \to \mathbb{R}$ est l'exponentielle, on a $f^\infty(1) = +\infty$. Donc $1 \notin \operatorname{dom} f^\infty$, alors que $1 \in (\operatorname{dom} f)^\infty = (\mathbb{R})^\infty = \mathbb{R}$.

La notion de fonction asymptotique est illustrée à la figure 3.5.

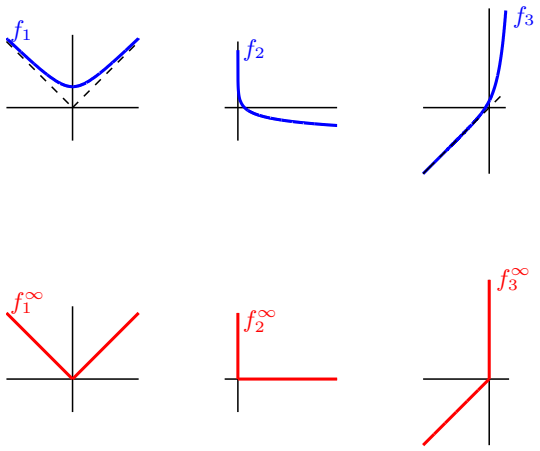


Fig. 3.5. Exemples de fonctions convexes (en haut) et de leur fonction asymptotique (en bas) ; de gauche à droite : $f_1(x) = (x^2+10)^{1/2}$, $f_2(x) = -\log x$ et $f_3(x) = x + e^x$

On note l'*ensemble de sous-niveau* $\nu \in \mathbb{R}$ d'une fonction $f : \mathbb{E} \to \overline{\mathbb{R}}$ de la manière suivante :

$$N_\nu(f) := \{x \in \mathbb{E} : f(x) \leqslant \nu\}.$$

C'est un ensemble convexe, lorsque f est convexe. La notion de fonction asymptotique est très utile, du fait du résultat suivant, qui montre que pour les fonctions de $\Gamma_0(\mathbb{E})$ ces ensembles de sous-niveau ont tous le même cône asymptotique (s'ils sont non vides). En particulier, si l'un d'eux est borné non vide, ils sont tous bornés (éventuellement vides). Un de ces ensembles de sous-niveau est l'ensemble de ses minimiseurs :

$$\arg\min f := \{x \in \mathbb{E} : f(x) \leqslant f(x'),\ \forall\, x' \in \mathbb{E}\} = N_{\inf f}(f).$$

La fonction asymptotique permet alors de donner des conditions nécessaires et suffisantes pour que l'ensemble des minimiseurs soit non vide et compact. D'autres conditions, ne faisant pas intervenir f^∞ et valables pour les fonctions de $\operatorname{Conv}(\mathbb{E})$, sont données à l'exercice 3.39.

Proposition 3.29 (ensembles de sous-niveau d'une fonction convexe) *Soit $f \in \Gamma_0(\mathbb{E})$. Alors, pour tout $\nu \in \mathbb{R}$ tel que $N_\nu(f) \neq \varnothing$, on a*

$$(N_\nu(f))^\infty = \{d \in \mathbb{E} : f^\infty(d) \leqslant 0\}. \tag{3.21}$$

En particulier, les propriétés suivantes sont équivalentes :
(i) $\exists\, \nu \in \mathbb{R}$ *tel que* $N_\nu(f)$ *est non vide et compact,*
(ii) $\forall\, \nu \in \mathbb{R}$, $N_\nu(f)$ *est compact,*
(iii) $\arg\min f$ *est non vide et compact,*
(iv) $\forall\, d \in \mathbb{E} \setminus \{0\}$, $f^\infty(d) > 0$.

Démonstration. Soient $\nu \in \mathbb{R}$ et $x \in N_\nu(f)$ supposé non vide. Alors

$$d \in (N_\nu(f))^\infty \quad \Longleftrightarrow \quad f(x+td) \leqslant \nu, \quad \forall\, t \geqslant 0.$$

Donc si $d \in (N_\nu(f))^\infty$, $f^\infty(d) \leqslant 0$ par passage à la limite ci-dessus (après avoir divisé par $t > 0$). Inversement, si $f^\infty(d) \leqslant 0$, $f(x+td) \leqslant f(x)$, pour tout $t > 0$ (le quotient dans (3.18) est croissant et borné par 0). Alors $x \in N_\nu(f)$ implique que $x+td \in N_\nu(f)$, pour tout $t > 0$; donc $d \in (N_\nu(f))^\infty$.

L'équivalence des propriétés (i)-(iv) se déduit facilement de (3.21). □

En pratique, pour montrer que f a un ensemble non vide et borné de minimiseurs (point (iii)), on utilise le point (iv) : quelle que soit la direction non nulle d, $f^\infty(d) > 0$. Comme souvent en analyse convexe, on obtient une propriété globale (la bornitude de l'ensemble des minimiseurs) à partir de propriétés unidirectionnelles (la stricte positivité de la fonction asymptotique dans toutes les directions non nulles).

Dans le reste de cette section, nous énonçons quelques règles de calcul de fonctions asymptotiques. La démonstration de la première règle est proposée à l'exercice 3.5.

Proposition 3.30 (pré-composition par une fonction affine) *Soient $\mathbb{E}$ et $\mathbb{F}$ deux espaces vectoriels, $A : \mathbb{E} \to \mathbb{F}$ une fonction linéaire, $b \in \mathbb{F}$ et $a : x \in \mathbb{E} \to Ax + b \in \mathbb{F}$ l'application affine associée et $f \in \Gamma_0(\mathbb{E})$ telle que $(\operatorname{dom} f) \cap \mathcal{R}(a) \neq \varnothing$. Alors, pour tout $d \in \mathbb{F}$,*

$$(f \circ a)^\infty(d) = f^\infty(Ad).$$

Voici maintenant un résultat permettant de calculer, dans certains cas, la fonction asymptotique d'une post-composition d'une fonction convexe par une fonction convexe croissante [27]. La formule (3.22) rappelle celle de la dérivation en chaîne.

Proposition 3.31 (post-composition par une fonction convexe croissante) *Supposons données deux fonctions $f \in \overline{\mathrm{Conv}}(\mathbb{E})$ et $g \in \overline{\mathrm{Conv}}(\mathbb{R})$ telles que $(\mathrm{dom}\, g) \cap f(\mathbb{E}) \neq \varnothing$. On suppose que g est croissante et vérifie $g^\infty(1) > 0$. Alors $(g \circ f) \in \overline{\mathrm{Conv}}(\mathbb{E})$ et pour tout $d \in \mathbb{E}$, on a*

$$(g \circ f)^\infty(d) = g^\infty(f^\infty(d)). \tag{3.22}$$

Dans ce résultat, on a adopté les conventions suivantes : $g(f(x)) = +\infty$ si $x \notin \mathrm{dom}\, f$ et $g^\infty(f^\infty(d)) = +\infty$ si $d \notin \mathrm{dom}\, f^\infty$.

Démonstration. D'après [317 ; proposition IV.2.1.8], $(g \circ f) \in \overline{\mathrm{Conv}}(\mathbb{E})$. Montrons que

$$\delta < f^\infty(d) \quad \Longrightarrow \quad g^\infty(\delta) \leqslant (g \circ f)^\infty(d). \tag{3.23}$$

Si $\delta < f^\infty(d)$, alors pour t assez grand et $x \in \mathrm{dom}\, f$ tel que $f(x) \in \mathrm{dom}\, g$, on a $\delta < (f(x+td) - f(x))/t$ ou encore $f(x) + t\delta < f(x+td)$. Comme g est croissante et $f(x) \in \mathrm{dom}\, g$, on en déduit

$$\frac{g(f(x)+t\delta) - g(f(x))}{t} \leqslant \frac{(g \circ f)(x+td) - (g \circ f)(x)}{t}, \quad \text{pour } t \text{ grand.}$$

Comme $f(x) \in \mathrm{dom}\, g$, la limite dans cette relation conduit à (3.23).

Si $d \notin \mathrm{dom}\, f^\infty$, on peut prendre $\delta \to \infty$ dans (3.23), et comme $\lim_{\delta \to \infty} g^\infty(\delta) = \infty$ (car $g^\infty(1) > 0$), on en déduit que $(g \circ f)^\infty(d) = \infty$.

Si $d \in \mathrm{dom}\, f^\infty$, on prend $\delta \downarrow f^\infty(d)$ dans (3.23). Comme g^∞ est fermée, on obtient

$$g^\infty(f^\infty(d)) \leqslant \liminf_{\delta \downarrow f^\infty(d)} g^\infty(\delta) \leqslant (g \circ f)^\infty(d).$$

Pour montrer l'inégalité inverse, on prend $x \in \mathrm{dom}\, f$ tel que $f(x) \in \mathrm{dom}\, g$. Alors, quel que soit $t > 0$, $f(x+td) \leqslant f(x) + tf^\infty(d)$ et comme g est croissante :

$$(g \circ f)(x+td) \leqslant g(f(x) + tf^\infty(d)).$$

On divise par $t > 0$ et on passe à la limite pour $t \to \infty$. Ceci donne $(g \circ f)^\infty(d) \leqslant g^\infty(f^\infty(d))$. □

3.4 Opérations

J'ai seul la clef de cette parade sauvage.

A. Rimbaud, Parade, *Illuminations* (1873-1875).

3.4.1 Composition

La *fonction composée* de deux fonctions f et g, dont l'ensemble d'arrivée de f est l'ensemble de définition de g, est la fonction notée $g \circ f$ et définie en x par

$$(g \circ f)(x) = g(f(x)).$$

Si f et g sont convexes, la fonction composée $g \circ f$ n'est généralement pas convexe. Relevons toutefois deux exceptions heureuses que l'on rencontre fréquemment, celle où f est affine et celle où g est convexe croissante. Nous les examinons l'une après l'autre.

Proposition 3.32 (pré-composition par une fonction affine) *Soient $\mathbb{E}$ et $\mathbb{F}$ deux espaces vectoriels, $a : \mathbb{E} \to \mathbb{F}$ une application affine et $g : \mathbb{F} \to \overline{\mathbb{R}}$ une fonction telle que $\mathcal{R}(a) \cap \operatorname{dom} g \neq \varnothing$.*
1) *Si $g \in \operatorname{Conv}(\mathbb{F})$, alors $g \circ a \in \operatorname{Conv}(\mathbb{E})$.*
2) *Si $g \in \operatorname{C\overline{onv}}(\mathbb{F})$, alors $g \circ a \in \operatorname{C\overline{onv}}(\mathbb{E})$.*
3) *Si g est convexe polyédrique, alors $g \circ a$ est convexe polyédrique.*

Démonstration. 1) Clairement $g \circ a \in \operatorname{Conv}(\mathbb{E})$ lorsque $g \in \operatorname{Conv}(\mathbb{F})$ et que $\mathcal{R}(a) \cap \operatorname{dom} g \neq \varnothing$.

2) Introduisons l'application affine

$$\tilde{a} : \mathbb{E} \times \mathbb{R} \to \mathbb{F} \times \mathbb{R} : (x, \alpha) \mapsto (a(x), \alpha).$$

Alors, $(x, \alpha) \in \operatorname{epi}(g \circ a)$ si, et seulement si, $(a(x), \alpha) \in \operatorname{epi} g$, ce qui s'écrit

$$\operatorname{epi}(g \circ a) = \tilde{a}^{-1}(\operatorname{epi} g).$$

Si $g \in \operatorname{C\overline{onv}}(\mathbb{F})$, $\operatorname{epi} g$ est fermé, donc aussi $\tilde{a}^{-1}(\operatorname{epi} g) = \operatorname{epi}(g \circ a)$, ce qui montre que $g \circ a \in \operatorname{C\overline{onv}}(\mathbb{E})$.

3) Si g est convexe polyédrique, $\operatorname{epi} g$ est un polyèdre convexe, donc aussi $\tilde{a}^{-1}(\operatorname{epi} g) = \operatorname{epi}(g \circ a)$ (exercice 2.20), ce qui montre que $g \circ a$ est convexe polyédrique. □

Le second cas où la convexité est préservée par composition se situe dans le cadre suivant. Soient $\mathbb{E}$ un espace vectoriel, $F : \mathbb{E} \to (\mathbb{R} \cup \{+\infty\})^m$ une première fonction et $g : \mathbb{R}^m \to \overline{\mathbb{R}}$ une seconde fonction. Avec F pouvant prendre la valeur $+\infty$, la fonction composée $(g \circ F) : \mathbb{E} \to \overline{\mathbb{R}}$ est définie en $x \in \mathbb{E}$ par

$$(g \circ F)(x) = \begin{cases} g(F(x)) & \text{si } F(x) \in \mathbb{R}^m \\ +\infty & \text{sinon.} \end{cases}$$

On munit $\mathbb{R}^m$ de l'*ordre* habituel (c.-à-d., $[x \leqslant y] \Leftrightarrow [x_i \leqslant y_i \text{ pour tout } i = 1, \dots, m]$) et on dit que g est *croissante* si $g(x) \leqslant g(y)$ chaque fois que $x \leqslant y$ dans $\mathbb{R}^m$.

Proposition 3.33 (post-composition par une fonction convexe croissante) *Dans le cadre défini ci-dessus, si F est convexe composante par composante et si g est convexe et croissante, alors $g \circ F$ est convexe.*

Démonstration. Soient x, $y \in \operatorname{dom}(g \circ F)$ et $t \in]0, 1[$. Alors $F(x)$ et $F(y) \in \mathbb{R}^m$ et la convexité composante par composante de F conduit à

$$F((1-t)x+ty) \leqslant (1-t)F(x)+tF(y) \qquad \text{(inégalité dans } \mathbb{R}^m\text{)}.$$

Ensuite

$$\begin{array}{rcll}(g\circ F)\big((1-t)x+ty\big) & \leqslant & g\big((1-t)F(x)+tF(y)\big) & \text{[croissance de } g\text{]}\\ & \leqslant & (1-t)(g\circ F)(x)+t(g\circ F)(y) & \text{[convexité de } g\text{]}.\end{array}$$

C'est l'inégalité recherchée. □

3.4.2 Enveloppes supérieure et inférieure

L'enveloppe supérieure d'une famille de fonctions est la fonction dont la valeur en un point est le supremum (ou borne supérieure) des valeurs prises par ces fonctions en ce point. Soyons plus précis.

Soient $\mathbb{E}$ un premier ensemble utilisé comme espace de définition de fonctions et I un second ensemble utilisé comme ensemble (quelconque) d'indices. Pour tout $i \in I$, on suppose donnée une fonction $f_i : \mathbb{E} \to \overline{\mathbb{R}}$. On peut alors introduire la fonction

$$f = \sup_{i\in I} f_i : \mathbb{E} \to \overline{\mathbb{R}},$$

définie pour $x \in \mathbb{E}$ par

$$f(x) = \sup_{i\in I} \Big(f_i(x)\Big).$$

Cette fonction est appelée *l'enveloppe supérieure* des fonctions f_i. On définit de manière similaire, l'*enveloppe inférieure* des fonctions f_i :

$$\inf_{i\in I} f_i : \mathbb{E} \to \overline{\mathbb{R}} : x \mapsto \Big(\inf_{i\in I} f_i\Big)(x) := \inf_{i\in I} \Big(f_i(x)\Big).$$

De manière plus abstraite, on peut dire que l'enveloppe supérieure de la famille $\{f_i\}_{i\in I}$ n'est autre que le supremum (ou borne supérieure) de cette famille dans l'ensemble ordonné réticulé des fonctions définies sur $\mathbb{E}$ à valeurs dans $\overline{\mathbb{R}}$ (ou treilli, c.-à-d., l'ordre $[f \leqslant g] \Leftrightarrow$ [pour tout $x \in \mathbb{E}$, $f(x) \leqslant g(x)$] est tel que tout couple de fonctions a une borne supérieure et une borne inférieure). Ceci justifie la notation $\sup_{i\in I} f_i$ [83 ; 1971, p. IV.21].

Proposition 3.34 (épigraphe d'une enveloppe supérieure) *Soit $\{f_i\}_{i\in I}$ une famille quelconque de fonctions. Alors*

$$\operatorname{epi}\left(\sup_{i\in I} f_i\right) = \bigcap_{i\in I} (\operatorname{epi} f_i)\,.$$

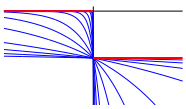
On en déduit que :
1) $\sup_{i\in I} f_i$ *est convexe si* $\mathbb{E}$ *est un espace vectoriel et les* f_i *sont convexes,*
2) $\sup_{i\in I} f_i$ *est fermée si* $\mathbb{E}$ *est un espace topologique et les* f_i *sont fermées.*

DÉMONSTRATION. En effet $(x,\alpha) \in \text{epi}(\sup_{i\in I} f_i) \iff \alpha \geqslant f_i(x)$, pour tout $i \in I$ $\iff (x,\alpha) \in \text{epi}\, f_i$, pour tout $i \in I \iff (x,\alpha) \in \cap_{i\in I}(\text{epi}\, f_i)$. On utilise ensuite le fait qu'une intersection de convexes [resp. fermés] est convexe [resp. fermée]. □

Il se peut cependant que l'enveloppe supérieure de fonctions convexes soit identiquement égale à $+\infty$, même si les $f_i \in \text{Conv}(\mathbb{E})$ (par exemple l'enveloppe supérieure des fonctions constantes).

Le point 2 de la proposition précédente met en évidence l'importance du concept de fonction *fermée* (ou s.c.i.) en optimisation : cette propriété est stable pour l'enveloppe supérieure, une opération souvent rencontrée dans cette discipline. On notera qu'à l'inverse la continuité des fonctions n'est pas conservée par l'enveloppe supérieure, comme le montre l'exemple où les fonctions $f_i : \mathbb{R} \to \mathbb{R}$, pour $i \in \mathbb{R}_+$, sont définies en $x \in \mathbb{R}$ par

$$f_i(x) = -e^{ix}.$$



L'enveloppe supérieure de ces fonctions continues est nulle sur $\mathbb{R}_-^*$ et vaut -1 sur $\mathbb{R}_+$; elle n'est donc pas continue. Bien sûr, une fonction continue étant s.c.i., l'enveloppe supérieure de fonctions continues est s.c.i.. Dans l'exemple ci-dessus, la valeur de l'enveloppe supérieure en zéro est -1, pas zéro ou tout autre valeur > -1.

Le fait que l'enveloppe supérieure de fonctions convexes soit convexe se montre aussi aisément en utilisant l'inégalité de convexité (3.2), grâce au fait que l'enveloppe supérieure d'une somme est inférieure à la somme des enveloppes supérieures. Seule la convexité des fonctions $x \mapsto f_i(x)$, à $i \in I$ fixé, joue dans l'obtention de ce résultat (il n'y a d'ailleurs pas de structure sur l'ensemble I). La situation est bien différente pour l'enveloppe inférieure, qui ne bénéficie pas des deux propriétés numérotées de la proposition précédente. On peut toutefois avoir la convexité de l'enveloppe inférieure si l'on a une structure vectorielle sur I et si $(x,i) \in \mathbb{E} \times I \mapsto f_i(x)$ est convexe. Cela conduit à la notion de fonction marginale de la section suivante.

3.4.3 Fonction marginale

Soient $\mathbb{E}$ et $\mathbb{F}$ deux espaces vectoriels et $\varphi : \mathbb{E} \times \mathbb{F} \to \overline{\mathbb{R}}$ une fonction. On associe à cette dernière la *fonction marginale* $f : \mathbb{E} \to \overline{\mathbb{R}}$ définie par :

$$f(x) = \inf_{y\in\mathbb{F}} \varphi(x,y). \tag{3.24}$$

Proposition 3.35 1) *f est convexe, si φ est convexe.*
2) *$f \in \text{Conv}(\mathbb{E})$, si $\varphi \in \text{Conv}(\mathbb{E}\times\mathbb{F})$ et si f ne prend pas la valeur $-\infty$.*

DÉMONSTRATION. 1) L'épigraphe stricte $\text{epi}_s\, f$ est la projection sur $\mathbb{E}\times\mathbb{R}$ de $\text{epi}_s\, \varphi \subseteq (\mathbb{E}\times\mathbb{F})\times\mathbb{R}$. Comme epi_s est convexe (convexité de φ) et que l'image par une application linéaire d'un convexe est convexe (section 2.1), on en déduit que $\text{epi}_s\, f$ est convexe ; donc f est convexe.

2) Il reste à montrer que $f \not\equiv +\infty$. Mais si φ est propre, il existe un (x, y) tel que $\varphi(x, y) < \infty$, donc $f(x) < \infty$. □

La fonction marginale est une enveloppe *inférieure* de fonctions convexes $x \mapsto \varphi(x, y)$, paramétrées par $y \in \mathbb{F}$. On pourrait donc, à juste titre, s'étonner de sa convexité. C'est évidemment la convexité *conjointe* sur $\mathbb{E} \times \mathbb{F}$ qui permet d'avoir cette propriété. En comparaison, dans la proposition 3.34, on ne fait aucune hypothèse sur la structure de l'ensemble I, ni sur la manière dont $f_i(x)$ dépend de i.

3.4.4 Inf-convolution

Soient f et g deux fonctions $\mathbb{E} \to \mathbb{R} \cup \{+\infty\}$, définies sur un espace vectoriel $\mathbb{E}$. Leur *inf-convolution* est la fonction $(f \uplus g) : \mathbb{E} \to \overline{\mathbb{R}}$ définie par[1]

$$(f \uplus g)(x) = \inf_{y \in \mathbb{E}} \Big(f(y) + g(x-y) \Big) = \inf\{ f(y) + g(z) : y + z = x \}. \tag{3.25}$$

Voilà une opération bien singulière ! Observons d'abord qu'il s'agit de la fonction marginale associée à la fonction définie sur $\mathbb{E}^2$ par $\varphi(x, y) = f(y) + g(x-y)$, si bien qu'elle hérite des propriétés des fonctions marginales. D'autre part, si la définition de l'inf-convolution donnée ci-dessus peut paraître obscure, son expression en termes d'épigraphe est particulièrement simple (la démonstration de la proposition ci-dessous est proposée à l'exercice 3.6).

Proposition 3.36 *Soient f et g deux fonctions $\mathbb{E} \to \mathbb{R} \cup \{+\infty\}$. Alors*

$$\operatorname{epi}_s(f \uplus g) = \operatorname{epi}_s f + \operatorname{epi}_s g. \tag{3.26}$$

En particulier :
1) $\operatorname{dom}(f \uplus g) = \operatorname{dom} f + \operatorname{dom} g$,
2) *commutativité :* $f \uplus g = g \uplus f$,
3) *associativité :* $(f \uplus g) \uplus h = f \uplus (g \uplus h)$,
4) $f \uplus g$ *est convexe si* f *et* g *sont convexes,*
5) $f \uplus g \in \operatorname{Conv}(\mathbb{E})$ *si* f *et* $g \in \operatorname{Conv}(\mathbb{E})$ *et ont une minorante affine commune.*

L'identité (3.26) portant sur les épigraphes strictes n'a pas lieu si l'on utilise les épigraphes, bien que l'on ait toujours

$$\operatorname{epi}(f \uplus g) \supseteq \operatorname{epi} f + \operatorname{epi} g.$$

Le problème vient du fait que la somme de deux fermés n'est pas nécessairement un fermé, si bien que $\operatorname{epi} f + \operatorname{epi} g$ n'est pas nécessairement un épigraphe ; il peut ne

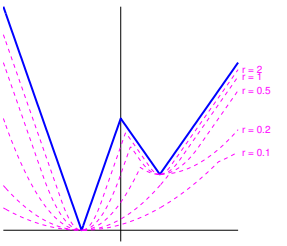
[1] La notation $f \uplus g$ nous est propre, mais les autres notations ne sont guère stabilisées : on trouve $f \square g$ chez Rockafellar [517], $f \dot{+} g$ chez Hiriart-Urruty et Lemaréchal [317] et $f \odot g$ chez Borwein et Lewis [78]. Avec un peu de bonne volonté, le lecteur verra dans le symbole $\uplus$ le fait que l'épigraphe (symbole $\cup$) stricte de $f \uplus g$ est obtenu en sommant (symbole $+$) ceux de f et g (proposition 3.36).

pas contenir sa « frontière inférieure », comme dans l'exemple suivant : si $f = \mathcal{I}_{[0,1[}$ et $g = \mathrm{Id} + \mathcal{I}_{[0,+\infty[}$, alors $(x,\alpha) \in \operatorname{epi} f + \operatorname{epi} g$ si, et seulement si, $x \geqslant 0$, $\alpha \geqslant 0$ et $\alpha > x - 1$, si bien que $\operatorname{epi} f + \operatorname{epi} g$ n'est pas l'épigraphe d'une fonction. Cependant, en prenant $\alpha := f(y) + g(z)$ dans la définition (3.25) de $f \boxplus g$, on a $(x,\alpha) = (y, f(y)) + (z.g(z))$, si bien que

$$(f \boxplus g)(x) = \inf\{\alpha : (x,\alpha) \in \operatorname{epi} f + \operatorname{epi} g\}. \tag{3.27}$$

L'inf-convolution d'une fonction f (éventuellement non convexe) avec la fonction quadratique $q = \frac{1}{2}\|\cdot\|^2$ a un effet régularisant sur certains points de non-différentiabilité de f. La figure ci-joint considère le cas de la fonction non convexe et non différentiable

$$f : x \in \mathbb{R} \mapsto |x+1| - \frac{3}{2}\,x^+ + (x-1)^+ \in \mathbb{R},$$



qui y est représentée par la courbe en trait plein. Les courbes en tirets sont les inf-convolutions

$$x \mapsto \left(f \boxplus \frac{r}{2}\,q\right)(x) = \inf_{y\in\mathbb{R}} \; f(y) + \frac{r}{2}\,\|y - x\|^2$$

pour les valeurs de $r = 0.1$, $r = 0.2$, $r = 0.5$, $r = 1$ et $r = 2$. On voit en prenant $y = x$ dans l'infimum ci-dessus, que $(f \boxplus \frac{r}{2}\,q)(x) \in [\inf f, f(x)]$, lorsque $r \geqslant 0$; on verra aussi au point 3 de la proposition 13.2 que $(f \boxplus \frac{r}{2}\,q)(x)$ croît avec $r > 0$; ces propriétés peuvent s'observer dans la figure. Cet effet régularisant sera examiné dans le cas d'une fonction convexe f à la section 3.7.2.

3.4.5 Inf-image sous une application linéaire

Soient $f : \mathbb{E} \to \overline{\mathbb{R}}$ une fonction et $A : \mathbb{E} \to \mathbb{F}$ une application linéaire. L'*inf-image de f sous A* est la fonction $(f \veebar A) : \mathbb{F} \to \overline{\mathbb{R}}$ définie par[2]

$$(f \veebar A)(y) = \inf\{f(x) : x \in \mathbb{E},\ Ax = y\}. \tag{3.28}$$

Il s'agit donc de la valeur minimale de f sur les sous-espaces affines $\{x \in \mathbb{E} : Ax = y\}$ paramétrés par y. On appellera aussi cette fonction la *fonction valeur* associée au problème d'optimisation $\inf\{f(x) : Ax = b\}$, où b est donné dans $\mathbb{F}$. C'est un des intérêts de cette opération.

En fait, la fonction marginale et l'inf-convolution peuvent être vues comme des l'inf-images :

- la fonction marginale de $\varphi : \mathbb{E} \times \mathbb{F} \to \overline{\mathbb{R}}$ n'est autre que $\varphi \veebar A$, où A est la projection de $\mathbb{E} \times \mathbb{F}$ sur $\mathbb{E}$,

[2] L'appellation « inf-image de f sous A » et la notation $f \veebar A$ nous sont propres. Rockafellar [517] et Hiriart-Urruty et Lemaréchal [317] appellent simplement cette fonction l'*image de f sous A*, donc sans faire mention de l'opération de minimalité qui intervient dans la définition, ce qui nous a paru dommage. D'autre part, leur notation « Af » nous semble être trop proche d'une composition (chez Rockafellar [517], « fA » est la composition $f \circ A$, alors que « Af » désigne l'inf-image de f sous A). La singularité de cette opération nous a semblé mériter une notation particulière suggérant l'opération de minimalité (symbole $\vee$) sur des sous-espaces affines (symbole $_$).

- si f_1 et $f_2 : \mathbb{E} \to \overline{\mathbb{R}}$, alors
$$f_1 \uplus f_2 = f \veebar A, \tag{3.29}$$
avec $f : \mathbb{E} \times \mathbb{E} \to \overline{\mathbb{R}}$: et $A : \mathbb{E} \times \mathbb{E} \to \mathbb{E}$ définies par
$$f(x_1, x_2) = f_1(x_1) + f_2(x_2) \qquad \text{et} \qquad A(x_1, x_2) = x_1 + x_2. \tag{3.30}$$

Par ailleurs, l'inf-image apparaît dans la conjuguée de la pré-composition d'une fonction convexe avec une application linéaire (proposition 3.52).

Des conditions assurant la convexité de $f \veebar A$ sont données dans la proposition suivante. La polyédricité de $f \veebar A$ est étudiée à la proposition 3.38. Des conditions pour que $f \veebar A$ soit propre et fermée seront données à la proposition 3.50.

Proposition 3.37 *Soient $f : \mathbb{E} \to \overline{\mathbb{R}}$ une fonction convexe et $A : \mathbb{E} \to \mathbb{F}$ une application linéaire. Alors $f \veebar A$ est convexe.*

Démonstration. Soit $B : \mathbb{E} \times \mathbb{R} \to \mathbb{F} \times \mathbb{R}$ l'application linéaire définie par $B(x, \alpha) = (Ax, \alpha)$. Montrons que
$$B(\operatorname{epi}_s f) = \operatorname{epi}_s(f \veebar A).$$
En effet, $(y, \alpha) \in B(\operatorname{epi}_s f)$ si, et seulement si, il existe un $x \in \mathbb{E}$ tel que $y = Ax$ et $f(x) < \alpha$, ce qui revient à dire que $(f \veebar A)(y) < \alpha$ ou encore que $(y, \alpha) \in \operatorname{epi}_s(f \veebar A)$. Alors la convexité de $\operatorname{epi}_s f$ entraîne celle de $\operatorname{epi}_s(f \veebar A)$ et donc la convexité de $f \veebar A$. □

Proposition 3.38 (inf-image d'une fonction convexe polyédrique sous une application linéaire) *Soient $f : \mathbb{E} \to \overline{\mathbb{R}}$ une fonction convexe polyédrique et $A : \mathbb{E} \to \mathbb{F}$ une application linéaire. Alors $f \veebar A$ est convexe polyédrique et l'infimum dans (3.28) est atteint lorsqu'il est fini.*

Démonstration. On utilise l'application linéaire $B : \mathbb{E} \times \mathbb{R} \to \mathbb{F} \times \mathbb{R}$ de la démonstration de la proposition 3.37, définie par $B(x, \alpha) = (Ax, \alpha)$. Montrons que
$$B(\operatorname{epi} f) = \operatorname{epi}(f \veebar A).$$
En effet,

$(y, \alpha) \in B(\operatorname{epi} f)$
$\iff$ il existe un $x \in \mathbb{E}$ tel que $y = Ax$ et $f(x) \leqslant \alpha$,
$\iff$ $(f \veebar A)(y) \leqslant \alpha$,

$\boxed{\Rightarrow}$ clair,

$\boxed{\Leftarrow}$ le problème $\inf\{f(x) : Ax = y\}$ définissant $(f \veebar A)(y)$ s'écrit aussi $\inf\{t : Ax = y,\ f(x) \leqslant t\}$; ce dernier consiste à minimiser une fonction linéaire sur un polyèdre convexe non vide (polyédricité de f et

$(f \veebar A)(y) < +\infty)$; soit il est non borné, soit il a une solution (proposition 2.20) ; dans chaque cas l'implication est claire ; le même raisonnement montre que l'infimum dans la définition de l'inf-image est atteint lorsqu'il est fini,

$\iff (y, \alpha) \in \operatorname{epi}(f \veebar A)$.

La polyédricité de $f \veebar A$ résulte maintenant du fait que $\operatorname{epi} f$ est convexe polyédrique, donc aussi $B(\operatorname{epi} f) = \operatorname{epi}(f \veebar A)$ (proposition 2.18). □

3.4.6 Adhérence

On peut construire d'autres fonctions à partir de l'épigraphe d'une fonction $f : \mathbb{E} \to \overline{\mathbb{R}}$ non nécessairement convexe, par exemple en prenant son adhérence. Comme le montre la proposition suivante, par cette opération on obtient l'épigraphe d'une fonction notée

$$\bar{f} : \mathbb{E} \to \overline{\mathbb{R}}.$$

Cette fonction est appelée l'*adhérence* de f.

Proposition 3.39 (définition de l'adhérence) *Soit $f : \mathbb{E} \to \overline{\mathbb{R}}$ une fonction (non nécessairement convexe). Alors* $\operatorname{adh}(\operatorname{epi} f)$ *est l'épigraphe d'une fonction $\bar{f} : \mathbb{E} \to \overline{\mathbb{R}}$:*

$$\operatorname{epi} \bar{f} = \overline{\operatorname{epi} f}.$$

Démonstration. Il faut d'abord montrer que si $(x, \alpha) \in \operatorname{adh}(\operatorname{epi} f)$ et $\beta > \alpha$, alors $(x, \beta) \in \operatorname{adh}(\operatorname{epi} f)$. Si $(x, \alpha) \in \operatorname{adh}(\operatorname{epi} f)$, il existe une suite de couples $(x_k, \alpha_k) \in \operatorname{epi} f$ convergeant vers (x, α). Alors, pour k assez grand, $\alpha_k \leqslant \beta$ et donc $(x_k, \beta) \in \operatorname{epi} f$. En passant à la limite, on trouve que $(x, \beta) \in \operatorname{adh}(\operatorname{epi} f)$.

Il faut ensuite montrer que si $(x, \alpha_k) \in \operatorname{adh}(\operatorname{epi} f)$ et $\alpha_k \downarrow \alpha$, on a $(x, \alpha) \in \operatorname{adh}(\operatorname{epi} f)$. Ceci se déduit directement du fait que $\operatorname{adh}(\operatorname{epi} f)$ est fermé. □

Proposition 3.40 (propriétés de l'adhérence) *Soit $f : \mathbb{E} \to \overline{\mathbb{R}}$ une fonction (non nécessairement convexe) et $\bar{f}$ son adhérence. Alors*

(i) $\bar{f}$ est fermée et $\bar{f} \leqslant f$,

(ii) pour tout $x \in \operatorname{dom} \bar{f}$,

$$\bar{f}(x) = \liminf_{y \to x} f(y) := \lim_{\varepsilon \downarrow 0} \inf_{y \in B(x,\varepsilon)} f(y), \tag{3.31}$$

(iii) en $x \in \operatorname{dom} f$, $f(x) = \bar{f}(x)$ si, et seulement si, f est s.c.i. en x.

Démonstration. [(i)] L'épigraphe de $\bar{f}$ étant fermé, $\bar{f}$ est fermée. Par ailleurs $\bar{f} \leqslant f$ car $\operatorname{epi} \bar{f} \supseteq \operatorname{epi} f$.

[(*ii*)] Comme $\bar f$ est fermée et $\bar f \leqslant f$, $\bar f(x) = \liminf_{y\to x} \bar f(y) \leqslant \liminf_{y\to x} f(y)$. Inversement, il existe une suite de couples $(x_k, \alpha_k) \in \operatorname{epi} f$ convergeant vers $(x, \bar f(x))$. Comme $f(x_k) \leqslant \alpha_k$, $\bar f(x) = \lim_k \alpha_k \geqslant \liminf_k f(x_k) \geqslant \liminf_{y\to x} f(y)$.

[(*iii*)] Si $f(x) = \bar f(x)$, on voit par (3.31) que f est s.c.i. en x. Inversement, si f est s.c.i. en x, $f(x) \leqslant \liminf_{y\to x} f(y) = \bar f(x)$ [par (3.31)] $\leqslant f(x)$ [car $\bar f \leqslant f$], donc $f(x) = \bar f(x)$. □

Proposition 3.41 (enveloppe supérieure de minorantes affines) *Soit $f \in \operatorname{Conv}(\mathbb{E})$. Alors $\bar f$ est l'enveloppe supérieure des minorantes affines de f.*

Démonstration. Soit $\{a_i\}_{i\in I}$ la famille des minorantes affines de f. On cherche à montrer que $\bar f = \sup_{i\in I} a_i$ ou en termes d'épigraphe :

$$\operatorname{epi} \bar f = \bigcap_{i\in I} \operatorname{epi} a_i.$$

L'épigraphe de f étant un convexe, le corollaire 2.41 nous apprend que $\operatorname{epi} \bar f = \operatorname{adh}(\operatorname{epi} f)$ est l'intersection de tous les demi-espaces fermés de $\mathbb{E} \times \mathbb{R}$ contenant $\operatorname{epi} f$. Comme l'épigraphe d'un a_i est un demi-espace fermé contenant $\operatorname{epi} f$, on a déjà l'inclusion $\operatorname{epi} \bar f \subseteq \cap_{i\in I} \operatorname{epi} a_i$. Pour montrer l'inclusion inverse, il s'agit d'écrire les demi-espaces fermés contenant $\operatorname{epi} f$ au moyen des épigraphes des a_i.

Un demi-espace fermé de $\mathbb{E} \times \mathbb{R}$ est défini par un triplet $(\xi, \tau, t) \in \mathbb{E} \times \mathbb{R} \times \mathbb{R}$; on le note $H^-(\xi, \tau, t) := \{(x, \alpha) : \langle \xi, x\rangle + \tau\alpha \leqslant t\}$. S'il contient $\operatorname{epi} f$ il faut que l'on ait

$$\forall\, (x, \alpha) \in \operatorname{epi} f \ : \quad \langle \xi, x\rangle + \tau\alpha \leqslant t.$$

On a nécessairement $\tau \leqslant 0$ (on peut prendre α aussi grand que l'on veut dans l'inégalité ci-dessus). Si $\tau < 0$, $H^-(\xi, \tau, t)$ est l'épigraphe de la minorante affine de f suivante $x \mapsto \langle -\xi/\tau, x\rangle + t/\tau$ (prendre $\alpha = f(x)$ ci-dessus).

Que faire des demi-espaces fermés de la forme $H^-(\xi, 0, t)$? Par hypothèse, $f \in \operatorname{Conv}(\mathbb{E})$; elle a donc une minorante affine (proposition 3.7), dont l'épigraphe est un certain $H^-(\xi_0, \tau_0, t_0)$, avec $\tau_0 < 0$. Si l'on établit que

$$H^-(\xi, 0, t) \cap H^-(\xi_0, \tau_0, t_0) = \bigcap_{\rho\geqslant 0} H^-(\xi_0 + \rho\xi, \tau_0, t_0 + \rho t),$$

on aura terminé la démonstration, puisque l'on aura exprimé l'intersection des demi-espaces fermés contenant $\operatorname{epi} f$ comme l'intersection des $\operatorname{epi} a_i$. Cette identité résulte en fait de l'équivalence triviale

$$\begin{aligned} &\langle \xi, x\rangle \leqslant t \quad \text{et} \quad \langle \xi_0, x\rangle + \tau_0\alpha \leqslant t_0 \\ \iff \quad &\forall\, \rho \geqslant 0 : \quad \langle \xi_0 + \rho\xi, x\rangle + \tau_0\alpha \leqslant t_0 + \rho t. \end{aligned}$$

□

3.5 Conjugaison

Soit $f : \mathbb{E} \to \overline{\mathbb{R}}$ une fonction, non nécessairement convexe. Nous allons dans cette section lui associer une autre fonction $f^* : \mathbb{E} \to \overline{\mathbb{R}}$, appelée conjuguée de Fenchel de f. Elle est construite explicitement à partir de f. Cette construction peut paraître abstraite et arbitraire au premier abord, mais elle est utile à plus d'un titre :

- nous verrons à la section 3.5.2 qu'en réitérant le processus de conjugaison, on dispose d'un moyen analytique de *convexifier* f,
- la conjuguée de Fenchel peut aussi être utilisée comme outil intermédiaire dans le *calcul de sous-gradient*, un concept généralisant la notion de gradient pour les fonctions convexes non différentiables, que nous verrons à la section 3.6,
- enfin la conjuguée de Fenchel permet aussi d'introduire un *problème d'optimisation dual* d'un autre problème d'optimisation (section 14.2).

On supposera dans cette section que $\mathbb{E}$ est un espace euclidien, dont le produit scalaire est noté $\langle \cdot, \cdot \rangle$.

3.5.1 Conjuguée

Comme nous le signalions ci-dessus, la notion de fonction conjuguée intervient dans la définition du sous-différentiel d'une fonction convexe f et c'est par cette notion de sous-différentiel que nous allons motiver l'introduction de la fonction conjuguée.

On verra que le sous-différentiel $\partial f(x)$ de f en x est l'ensemble des pentes x^* des minorantes affines de f exactes en x. Il définit une multifonction

$$\partial f : x \in \mathbb{E} \mapsto \partial f(x) \subseteq \mathbb{E}.$$

Il n'est pas toujours aisé d'expliciter cette application car, pour $x \in \mathbb{E}$, il n'est pas nécessairement simple de déterminer toutes les minorantes affines de f exactes en x. Que l'on songe par exemple à la fonction valeur-propre-maximale $\lambda_{\max}$ qui à une matrice $A \in \mathcal{S}^n$ fait correspondre sa valeur propre maximale $\lambda_{\max}(A)$. Cette fonction est convexe (exercice 3.34), mais quelles sont les minorantes affines de $\lambda_{\max}$ qui sont exactes en une matrice $A \in \mathcal{S}^n$ donnée ? Il est parfois plus facile de déterminer la multifonction réciproque de ∂f, à savoir

$$\partial f^{-1} : x^* \in \mathbb{E} \mapsto \{x \in \mathbb{E} : x^* \in \partial f(x)\}.$$

De ce point de vue, pour une pente $x^* \in \mathbb{E}$ donnée, on cherche les points $x \in \mathbb{E}$ tels que f a une minorante affine de pente x^* qui est exacte en x. Pour déterminer ces points x, on s'y prend de la manière suivante ; c'est dans ce calcul qu'apparaît la fonction conjuguée.

Une fonction affine de pente x^* est une application de la forme

$$a_{x^*,\alpha} : \mathbb{E} \to \mathbb{R} : x \mapsto a_{x^*,\alpha}(x) = \langle x^*, x \rangle - \alpha,$$

où $\alpha \in \mathbb{R}$ est l'opposé de sa valeur en zéro. Pour que cette fonction affine minore f et soit exacte en certains points, on cherche à prendre α le plus petit possible (donc $-\alpha$ le plus grand possible) tout en conservant l'inégalité de minoration $a_{x^*,\alpha} \leqslant f$. Les

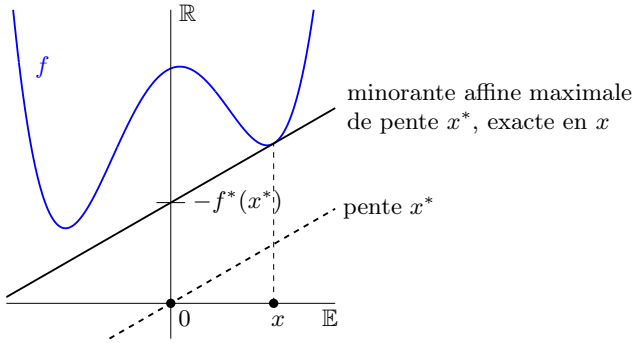


Fig. 3.6. Interprétation de $f^*(x^*)$

points où f et cette minorante affine maximale $a_{x^*,\alpha}$ prennent les mêmes valeurs sont les points $x \in \partial f^{-1}(x^*)$ recherchés. La démarche est illustrée à la figure 3.6. De manière plus explicite, on cherche le plus petit $\alpha \in \mathbb{R}$ tel que

$$a_{x^*,\alpha} \leqslant f \qquad \text{ou} \qquad \Big(\forall\, x \in \mathbb{E} : \quad \langle x^*, x\rangle - \alpha \leqslant f(x)\Big) \tag{3.32}$$
$$\text{ou} \qquad \Big(\forall\, x \in \mathbb{E} : \quad \langle x^*, x\rangle - f(x) \leqslant \alpha\Big).$$

On voit clairement que la plus petite valeur de α est donnée par

$$\boxed{f^*(x^*) := \sup_{x\in\mathbb{E}} \Big(\langle x^*, x\rangle - f(x)\Big).} \tag{3.33}$$

C'est la valeur de la conjuguée de f en x^*.

Définition 3.42 Soit $f : \mathbb{E} \to \overline{\mathbb{R}}$ une fonction (non nécessairement convexe). Sa *fonction conjuguée* $f^* : \mathbb{E} \to \overline{\mathbb{R}}$ est la fonction prenant en $x^* \in \mathbb{E}$ la valeur donnée par (3.33). L'application $f \mapsto f^*$ est appelée *transformation de Legendre-Fenchel*. □

Revenons au problème que nous nous posions au début de ce paragraphe : si x est solution du problème en (3.33), alors

$$\forall\, y \in \mathbb{E} : \quad \langle x^*, x\rangle - f(x) \geqslant \langle x^*, y\rangle - f(y) \tag{3.34}$$

ou encore (on vérifiera l'équivalence même lorsque f peut prendre des valeurs infinies)

$$\forall\, y \in \mathbb{E} : \quad f(y) \geqslant f(x) + \langle x^*, y - x\rangle, \tag{3.35}$$

ce qui montre que $y \mapsto f(x) + \langle x^*, y - x\rangle$ est une minorante affine de f, exacte en x. Dès lors, les sous-différentiels de f aux points-solutions du problème en (3.33) contiennent x^*.

L'expression (3.33) de la conjuguée et (3.32) montrent que l'on a une relation très simple entre les minorantes affines de f et les éléments de l'épigraphe de f^* :

$$\boxed{a_{x^*,\alpha} \text{ est une minorante affine de } f \quad \Longleftrightarrow \quad (x^*, \alpha) \in \operatorname{epi} f^*.} \tag{3.36}$$

Comme une fonction est entièrement spécifiée par son épigraphe, on peut dire que la conjuguée f^* de f est la fonction qui décrit toutes les minorantes affines de f.

La convexité de f^* mise en évidence par la proposition ci-dessous est remarquable ; on se rappelle en effet que f n'est, elle, pas nécessairement convexe. On note $f \not\equiv -\infty$ (resp. $f \not\equiv +\infty$) pour signifier que f n'est pas identiquement égale à $-\infty$ (resp. $+\infty$) et on note $f > -\infty$ (resp. $f < +\infty$) pour exprimer le fait que f ne prend pas la valeur $-\infty$ (resp. $+\infty$).

Proposition 3.43 (conjuguée convexe fermée) *La fonction conjuguée f^* d'une fonction $f : \mathbb{E} \to \overline{\mathbb{R}}$ est convexe et fermée. Par ailleurs, on a les équivalences suivantes*

$$f \not\equiv +\infty \iff f^* > -\infty \iff f^* \not\equiv -\infty, \tag{3.37}$$

$$f \text{ a une minorante affine} \iff f^* \not\equiv +\infty, \tag{3.38}$$

$$f \text{ propre avec minorante affine} \iff f^* \text{ propre} \iff f^* \in \overline{\mathrm{Conv}}(\mathbb{E}). \tag{3.39}$$

Démonstration. Certains arguments seront donnés plusieurs fois, de manière à garder une démonstrstion structurée.

[f^* est convexe fermée] Il y a trois cas possibles.

- Si f prend la valeur $-\infty$, alors $f^* \equiv +\infty$ (par la définition de f^*), une fonction convexe et fermée (l'épigraphe de f^* est vide, donc convexe et fermé).
- Si $f \equiv +\infty$, alors $f^* \equiv -\infty$ (par la définition de f^*), une fonction convexe et fermée (l'épigraphe de f^* est égal à $\mathbb{E} \times \mathbb{R}$, donc convexe et fermé).
- Si f est propre, f^* s'écrit comme le supremum de la famille indicée par $x \in \operatorname{dom} f$ (non vide) des fonctions affines (donc convexes et fermées) $x^* \mapsto \langle x^*, x\rangle - f(x)$. On en déduit que f^* est elle-même convexe et fermée (proposition 3.34).

[(3.37)] Si $f \not\equiv +\infty$, il existe un $x_0 \in \mathbb{E}$ tel que $f(x_0) < +\infty$; dans ce cas, quel que soit $x^* \in \mathbb{E}$, $f^*(x^*) \geqslant \langle x^*, x_0\rangle - f(x_0) > -\infty$, si bien que $f^* > -\infty$. Évidemment, $f^* > -\infty$ implique que $f^* \not\equiv -\infty$. Enfin, si $f \equiv +\infty$, alors $f^* \equiv -\infty$.

[(3.38)] Soit $x \mapsto \langle x_0^*, x\rangle + \alpha_0$ une minorante affine de f, avec $x_0^* \in \mathbb{E}$ et $\alpha_0 \in \mathbb{R}$; alors, pour tout $x \in \mathbb{E}$, $f(x) \geqslant \langle x_0^*, x\rangle + \alpha_0$; on en déduit que $-\alpha_0 \geqslant \sup\{\langle x_0^*, x\rangle - f(x) : x \in \mathbb{E}\} = f^*(x_0^*)$ et donc que $f^* \not\equiv +\infty$. La réciproque est obtenue par le raisonnement inverse : si $f^* \not\equiv +\infty$, il existe $x_0^* \in \mathbb{E}$ et $\alpha_0 \in \mathbb{R}$ tels que $f^*(x_0^*) \leqslant -\alpha_0$; alors, pour tout $x \in \mathbb{E}$, $\langle x_0^*, x\rangle - f(x) \leqslant -\alpha_0$, si bien que $x \mapsto \langle x_0^*, x\rangle + \alpha_0$ est une minorante affine de f.

[(3.39)] Par (3.37) et (3.38), f est propre et a une minorante affine si, et seulement si, f^* est propre. Enfin, il revient au même de dire que f^* est propre et que $f^* \in \overline{\mathrm{Conv}}(\mathbb{E})$, parce que f^* est toujours convexe et fermée. □

3.5.2 Biconjuguée

On peut bien sûr appliquer la transformation de Legendre-Fenchel à la fonction conjuguée $f^* : \mathbb{E} \to \overline{\mathbb{R}}$; on obtient ainsi la biconjuguée de f, notée f^{**}.

Définition 3.44 Soit $f : \mathbb{E} \to \overline{\mathbb{R}}$ une fonction (non nécessairement convexe) et $f^* : \mathbb{E} \to \overline{\mathbb{R}}$ sa conjuguée. La *fonction biconjuguée* de f est la fonction $f^{**} : \mathbb{E} \to \overline{\mathbb{R}}$ définie en $x \in \mathbb{E}$ par

$$\boxed{f^{**}(x) = \sup_{x^* \in \mathbb{E}} \Big(\langle x^*, x \rangle - f^*(x^*) \Big).}$$

□

Proposition 3.45 (biconjuguée convexe fermée) *La fonction biconjuguée f^{**} d'une fonction $f : \mathbb{E} \to \overline{\mathbb{R}}$ est convexe et fermée. De plus*

$$f \text{ est propre et a une minorante affine} \iff f^{**} \in \mathrm{C}\overline{\mathrm{onv}}(\mathbb{E}). \tag{3.40}$$

Démonstration. La biconjuguée est convexe et fermée en tant que conjuguée (proposition 3.43).

Si f est propre et a une minorante affine, alors f^* est propre par (3.39). Mais f^* a aussi une minorante affine puisque, avec $x_0 \in \operatorname{dom} f \neq \varnothing$, on a

$$\forall\, x^* \in \mathbb{E} : \qquad f^*(x^*) \geqslant \langle x^*, x_0 \rangle - f(x_0).$$

Dès lors, $f^{**} \in \mathrm{C}\overline{\mathrm{onv}}(\mathbb{E})$ par (3.39) appliqué à f^* au lieu de f.

Réciproquement, si $f^{**} \in \mathrm{C}\overline{\mathrm{onv}}(\mathbb{E})$, alors f^* est propre par (3.39) appliqué à f^*. On en déduit que f est propre et a une minorante affine, à nouveau par (3.39). □

Si l'argument x^* de f^* est une forme linéaire (identifiée à un élément de $\mathbb{E}$ au moyen du produit scalaire $\langle \cdot, \cdot \rangle$), l'argument x de f^{**} est dans l'espace de départ $\mathbb{E}$. On peut alors se demander s'il y a un lien entre f et f^{**}. La proposition suivante examine cette question.

Proposition 3.46 (enveloppe convexe fermée) *Soit $f : \mathbb{E} \to \overline{\mathbb{R}}$ une fonction propre ayant une minorante affine. Alors*

1) *f^{**} est l'enveloppe supérieure des minorantes affines de f,*
2) *$f \in \mathrm{Conv}(\mathbb{E}) \implies f^{**} = \bar{f}$ (l'adhérence de f),*
3) *$f \in \mathrm{C}\overline{\mathrm{onv}}(\mathbb{E}) \iff f^{**} = f$,*
4) *la transformation de Legendre-Fenchel $f \mapsto f^*$ est une bijection sur l'ensemble $\mathrm{C}\overline{\mathrm{onv}}(\mathbb{E})$.*

Démonstration. 1) Soit $x \in \mathbb{E}$. La valeur en x de l'enveloppe supérieure des minorantes affines de f s'écrit

$$\begin{aligned}
\sup_{\substack{x^*\in\mathbb{E},\ \alpha\in\mathbb{R}\\ \langle x^*,y\rangle-\alpha\leqslant f(y),\ \forall\, y\in\mathbb{E}}} \Big(\langle x^*,x\rangle-\alpha\Big) &= \sup_{x^*\in\mathbb{E}} \sup_{\substack{\alpha\in\mathbb{R}\\ \langle x^*,y\rangle-\alpha\leqslant f(y),\ \forall\, y\in\mathbb{E}}} \Big(\langle x^*,x\rangle-\alpha\Big)\\
&= \sup_{x^*\in\mathbb{E}} \sup_{\alpha=f^*(x^*)} \Big(\langle x^*,x\rangle-\alpha\Big)\\
&= \sup_{x^*\in\mathbb{E}} \Big(\langle x^*,x\rangle-f^*(x^*)\Big)\\
&= f^{**}(x).
\end{aligned}$$

2) Si $f\in\mathrm{Conv}(\mathbb{E})$, $\bar f$ est l'enveloppe supérieure des minorante affines de f (proposition 3.41), c'est-à-dire f^{**} d'après le point 1.

3) Si $f\in\mathrm{C}\overline{\mathrm{onv}}(\mathbb{E})$, $\mathrm{epi}\, f=\mathrm{adh}(\mathrm{epi}\, f)$ [car $\mathrm{epi}\, f$ est fermé] $=\mathrm{epi}\,\bar f$ [par définition de $\bar f$] $=\mathrm{epi}\, f^{**}$ [par le point 2, donc $f=f^{**}$. Inversement, l'égalité $f^{**}=f$ implique que f est convexe et fermée (car $f^{**}\in\mathrm{C}\overline{\mathrm{onv}}(\mathbb{E})$).

4) D'après la proposition 3.43, la transformation de conjugaison $\mathcal{C}: f\mapsto f^*$ est bien définie sur $\mathrm{C}\overline{\mathrm{onv}}(\mathbb{E})$ et est de toute façon à valeurs dans $\mathrm{C}\overline{\mathrm{onv}}(\mathbb{E})$. Elle est injective, car si f_1 et $f_2\in\mathrm{C}\overline{\mathrm{onv}}(\mathbb{E})$ sont telles que $f_1^*=f_2^*$, on a $f_1^{**}=f_2^{**}$ et donc $f_1=f_2$ d'après le point 3. Elle est surjective, car si $g\in\mathrm{C}\overline{\mathrm{onv}}(\mathbb{E})$, on a $\mathcal{C}(g^*)=g^{**}=g$ d'après le point 3 ; donc g est l'image de g^* par $\mathcal{C}$. □

Ce résultat permet de comparer les valeurs de f^{**} et de f.

Corollaire 3.47 (comparaison de f^{**} et f) *Quelle que soit $f:\mathbb{E}\to\overline{\mathbb{R}}$, on a*

$$f^{**}\leqslant f.$$

Si $f\in\mathrm{Conv}(\mathbb{E})$ et $x\in\mathrm{dom}\, f$, alors

$$f^{**}(x)=f(x)\qquad\Longleftrightarrow\qquad f \text{ est s.c.i. en } x.$$

Démonstration. Si $f\equiv+\infty$, on a certainement $f^{**}\leqslant f$. Si f n'a pas de minorante affine, $f^*\equiv+\infty$ par (3.38), puis $f^{**}\equiv-\infty\leqslant f$. Dans les autres cas, f^{**} est l'enveloppe supérieure des minorantes affines de f (point 1 de la proposition 3.46), si bien que $f^{**}\leqslant f$ (quel que soit $x\in\mathbb{E}$, $f^{**}(x)$ est le supremum de valeurs plus petites que $f(x)$).

D'après la proposition 3.40, lorsque $x\in\mathrm{dom}\, f$, on a $f(x)=\bar f(x)$ si, et seulement si, f est s.c.i. en x. Le résultat est alors une conséquence du fait que $\bar f=f^{**}$ lorsque $f\in\mathrm{Conv}(\mathbb{E})$ (point 2 de la proposition 3.46). □

D'après la proposition 3.46, il est naturel d'appeler aussi la biconjuguée f^{**}, l'*enveloppe convexe fermée* de f. La figure 3.7 donne la biconjuguée de la fonction représentée à la figure 3.6.

La proposition suivante montre qu'il ne sert à rien de réitérer le processus de conjugaison au-delà du second ordre : la tri-conjuguée se confond avec la conjuguée.

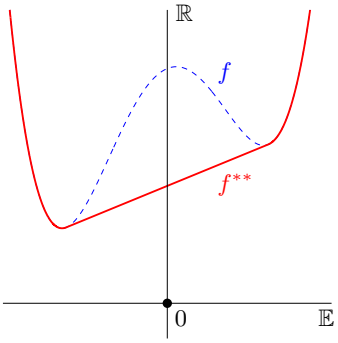


Fig. 3.7. Fonction biconjuguée f^{**} (en trait plein) de la fonction f de la figure 3.6 (en tirets)

Proposition 3.48 (triconjuguée) *Soit $f : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ une fonction propre ayant une minorante affine. Alors $(f^{**})^* = f^*$.*

Démonstration. D'après le corollaire 3.47, $f^{**} \leqslant f$ et donc $(f^{**})^* \geqslant f^*$ (exercice 3.16-(3)). Inversement, pour x^* fixé, $f^{**}(x) \geqslant \langle x^*, x\rangle - f^*(x^*)$ pour tout $x \in \mathbb{E}$, et donc

$$(f^{**})^*(x^*) = \sup_{x\in\mathbb{E}}\Big(\langle x^*, x\rangle - f^{**}(x)\Big) \leqslant f^*(x^*).$$

□

3.5.3 Règles de calcul

Inf-image sous une application linéaire

Rappelons la définition de l'inf-image $f \operatorname{\underline{\vee}} A$ d'une fonction f sous une application linéaire A, introduite à la section 3.2. On se donne deux espaces euclidiens $\mathbb{E}$ et $\mathbb{F}$ (on aura besoin ici d'un produit scalaire sur $\mathbb{E}$ et $\mathbb{F}$, alors que cette structure n'est pas nécessaire dans la définition de $f \operatorname{\underline{\vee}} A$), une fonction $f : \mathbb{E} \to \overline{\mathbb{R}}$ et une application linéaire $A : \mathbb{E} \to \mathbb{F}$. Alors $f \operatorname{\underline{\vee}} A : \mathbb{F} \to \overline{\mathbb{R}}$ est définie en $y \in \mathbb{F}$ par

$$(f \operatorname{\underline{\vee}} A)(y) = \inf_{\substack{x\in\mathbb{E}\\ Ax=y}} f(x). \tag{3.41}$$

Nous donnons à la proposition 3.49 l'expression de la conjuguée de $f \operatorname{\underline{\vee}} A$. La proposition 3.50 énonce des conditions assurant que $f \operatorname{\underline{\vee}} A$ est dans $\mathrm{Conv}(\mathbb{F})$ ou $\mathrm{C\overline{onv}}(\mathbb{F})$.

Proposition 3.49 (conjuguée de $f \operatorname{\underline{\vee}} A$) *Dans les conditions définies ci-dessus, on a*

$$(f \operatorname{\underline{\vee}} A)^* = f^* \circ A^*. \tag{3.42}$$

De plus, si f est propre et a une minorante affine et si l'application linéaire A vérifie

$$\mathcal{R}(A^*) \cap \operatorname{dom} f^* \neq \varnothing,$$

alors $f \operatorname{\underline{\vee}} A$ est propre et a une minorante affine.

Démonstration. Pour $y^* \in \mathbb{F}$, on trouve

$$\begin{aligned}
(f \operatorname{\underline{\vee}} A)^*(y^*) &= \sup_{y\in\mathbb{F}} \left(\langle y^*, y\rangle - \inf_{\substack{x\in\mathbb{E}\\ Ax=y}} f(x) \right) \\
&= \sup_{\substack{(x,y)\in\mathbb{E}\times\mathbb{F}\\ Ax=y}} \Big(\langle y^*, y\rangle - f(x) \Big) \\
&= \sup_{x\in\mathbb{E}} \Big(\langle y^*, Ax\rangle - f(x) \Big) \\
&= \sup_{x\in\mathbb{E}} \Big(\langle A^*y^*, x\rangle - f(x) \Big) \\
&= f^*(A^*y^*).
\end{aligned}$$

Supposons à présent que f est propre et a une minorante affine et que $\mathcal{R}(A^*) \cap \operatorname{dom} f^* \neq \varnothing$. Alors, clairement, $f \operatorname{\underline{\vee}} A \not\equiv +\infty$ (elle prend une valeur finie sur $A(\operatorname{dom} f)$, qui n'est pas vide). D'autre part, par hypothèse, il existe un $x_0^* := A^*y_0^* \in \operatorname{dom} f^*$. Alors l'inégalité $+\infty > f^*(x_0^*) \geqslant \langle x_0^*, x\rangle - f(x)$ valable pour tout $x \in \mathbb{E}$ conduit à

$$+\infty > f(x) \geqslant \langle y_0^*, Ax\rangle - (f^* \circ A^*)(y_0^*), \quad \text{pour tout } x \in \operatorname{dom} f.$$

Évidemment $(f \operatorname{\underline{\vee}} A)(y) = +\infty$ si $y \notin A(\operatorname{dom} f)$. Soit à présent $y \in A(\operatorname{dom} f)$. En prenant l'infimum ci-dessus pour les $x \in \operatorname{dom} f$ vérifiant $Ax = y$, on trouve

$$(f \operatorname{\underline{\vee}} A)(y) \geqslant \langle y_0^*, y\rangle - (f^* \circ A^*)(y_0^*), \quad \text{pour tout } y \in \mathbb{F}.$$

Dès lors $f \operatorname{\underline{\vee}} A$ a une minorante affine.

Proposition 3.50 (régularité de $f \operatorname{\underline{\vee}} A$) *Soient $\mathbb{E}$ et $\mathbb{F}$ deux espaces euclidiens, $f : \mathbb{E} \to \overline{\mathbb{R}}$ une fonction et $A : \mathbb{E} \to \mathbb{F}$ une application linéaire.*

1) *Si $f \in \operatorname{Conv}(\mathbb{E})$ et si A vérifie $\mathcal{R}(A^*) \cap \operatorname{dom} f^* \neq \varnothing$, alors $f \operatorname{\underline{\vee}} A \in \operatorname{Conv}(\mathbb{F})$.*
2) *Si $f \in \overline{\operatorname{conv}}(\mathbb{E})$ et si A vérifie $\mathcal{R}(A^*) \cap (\operatorname{dom} f^*)^\circ \neq \varnothing$, alors $f \operatorname{\underline{\vee}} A \in \overline{\operatorname{conv}}(\mathbb{F})$. De plus, l'infimum dans (3.41) est atteint lorsqu'il est fini.*

Démonstration. 1) Comme $f \in \operatorname{Conv}(\mathbb{E})$, $f \operatorname{\underline{\vee}} A$ est convexe (proposition 3.37). D'autre part, comme $f : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ est propre avec une minorante affine (proposition 3.7) et que $\mathcal{R}(A^*) \cap \operatorname{dom} f^* \neq \varnothing$, $f \operatorname{\underline{\vee}} A$ est également propre avec une minorante affine (proposition 3.49). Donc $f \operatorname{\underline{\vee}} A \in \operatorname{Conv}(\mathbb{F})$.

2) Pour montrer que $f \operatorname{\underline{\vee}} A$ est fermée, on suppose que l'on a des suites $\{y_k\} \subseteq \mathbb{F}$ et $\{\alpha_k\} \subseteq \mathbb{R}$ telles que

$$(f \mathbin{\underline{\vee}} A)(y_k) \leqslant \alpha_k, \qquad y_k \to y \qquad \text{et} \qquad \alpha_k \to \alpha.$$

Il suffit de montrer que $(f \mathbin{\underline{\vee}} A)(y) \leqslant \alpha$. D'après la définition de $f \mathbin{\underline{\vee}} A$, on peut trouver une suite $\{\varepsilon_k\} \downarrow 0$ et une suite $\{x_k\} \subseteq \mathbb{E}$ telles que

$$f(x_k) \leqslant \alpha_k + \varepsilon_k, \qquad Ax_k = y_k, \qquad y_k \to y \qquad \text{et} \qquad \alpha_k \to \alpha.$$

Il s'agit de passer à la limite dans ces relations.

La suite $\{x_k\}$ n'étant pas nécessairement bornée, on cherche à construire une autre suite $\{\bar{x}_k\} \subseteq \mathbb{E}$, bornée elle, qui laisse inchangés f et A. Si l'on se donne $x_1^* \in \mathcal{R}(A^*) \cap \operatorname{dom} f^*$, qui est non vide par hypothèse, il revient au même d'imposer

$$f(\bar{x}_k) - \langle x_1^*, \bar{x}_k\rangle = f(x_k) - \langle x_1^*, x_k\rangle \qquad \text{et} \qquad A\bar{x}_k = Ax_k,$$

puisqu'alors $f(\bar{x}_k) = f(x_k)$ (en effet $x_1^* \in \mathcal{R}(A^*)$ et $A\bar{x}_k = Ax_k$ impliquent que $\langle x_1^*, \bar{x}_k\rangle = \langle x_1^*, x_k\rangle$). Remarquons que $f(\cdot) - \langle x_1^*, \cdot\rangle = f^{**}(\cdot) - \langle x_1^*, \cdot\rangle$ n'est autre que la conjuguée φ^* de la fonction $\varphi \in \overline{\operatorname{Conv}}(\mathbb{E})$ définie par $\varphi(x^*) := f^*(x^* + x_1^*)$. Comme $0 \in \operatorname{dom}\varphi$, $\mathbb{E}_0 := \operatorname{aff}\operatorname{dom}\varphi$ est un sous-espace vectoriel. Soit $\bar{x}_k$ la projection orthogonale de x_k sur $\mathbb{E}_0 - \mathcal{R}(A^*)$. D'une part, $x_k - \bar{x}_k \in \mathcal{R}(A^*)^\perp = \mathcal{N}(A)$, si bien que $A(x_k) = A(\bar{x}_k)$. D'autre part, $x_k - \bar{x}_k \in \mathbb{E}_0^\perp$, si bien que

$$\varphi^*(x_k) = \sup_{x^* \in \mathbb{E}_0} \langle x^*, x_k\rangle - \varphi(x^*) = \sup_{x^* \in \mathbb{E}_0} \langle x^*, \bar{x}_k\rangle - \varphi(x^*) = \varphi^*(\bar{x}_k).$$

On a donc

$$f(\bar{x}_k) \leqslant \alpha_k + \varepsilon_k, \qquad A\bar{x}_k = y_k, \qquad y_k \to y \qquad \text{et} \qquad \alpha_k \to \alpha.$$

Si $\{\bar{x}_k\}$ est bornée, on peut en extraire une sous-suite convergente, disons vers $\bar{x}$. Par la semi-continuité inférieure de f, on a $f(\bar{x}) \leqslant \liminf f(\bar{x}_k)$. En passant à la limite ci-dessus, on a

$$f(\bar{x}) \leqslant \alpha \qquad \text{et} \qquad A\bar{x} = y.$$

Donc $(f \mathbin{\underline{\vee}} A)(y) \leqslant \alpha$ et $(f \mathbin{\underline{\vee}} A)$ est fermée.

Remarquons que si l'on prend ci-dessus $y_k = y$ et $\alpha_k = (f \mathbin{\underline{\vee}} A)(y)$, pour tout k, la suite $\{x_k\}$ devient une suite minimisante du problème de minimisation dans (3.41). Le raisonnement ci-dessus montre alors que ce problème a une solution.

Il reste donc à montrer que $\{\bar{x}_k\}$ est bornée. Observons que $\bar{x}_k \in \mathbb{E}_1 := \mathbb{E}_0 - \mathcal{R}(A^*) = \operatorname{vect}(\operatorname{dom}\varphi - \mathcal{R}(A^*))$. Dès lors, pour montrer que $\{\bar{x}_k\}$ est bornée, il suffit de monter que, quel que soit $z \in \mathbb{E}_1$, $\{\langle z, \bar{x}_k\rangle\}$ est bornée (par l'absurde : on prend pour z, un point d'adhérence de $\{\bar{x}_k/\|\bar{x}_k\|\}$). Soit donc $z \in \mathbb{E}_1$. Par hypothèse, $0 \in (\operatorname{dom} f^*)^\ominus - \mathcal{R}(A^*) = (\operatorname{dom} f^* - \mathcal{R}(A^*))^\ominus = (\operatorname{dom}\varphi - \mathcal{R}(A^*))^\ominus$, puisque $\operatorname{dom} f^* = \operatorname{dom}\varphi - x_1^*$ et que $x_1^* \in \mathcal{R}(A^*)$. Donc $tz \in \operatorname{dom}\varphi - \mathcal{R}(A^*)$ pour $t > 0$ assez petit : il existe $x^* \in \operatorname{dom}\varphi$ et $y^* \in \mathbb{F}$ tels que $tz = x^* - A^*y^*$. En introduisant v^* tel que $x_1^* = A^*v^*$, on trouve

$$\begin{aligned} t\langle z, \bar{x}_k\rangle &= \langle x^*, \bar{x}_k\rangle - \langle A^*y^*, \bar{x}_k\rangle \\ &\leqslant \varphi(x^*) + \varphi^*(\bar{x}_k) - \langle y^*, A\bar{x}_k\rangle \\ &= \varphi(x^*) + \varphi^*(x_k) - \langle y^*, Ax_k\rangle \\ &= \varphi(x^*) + f(x_k) - \langle y^* + v^*, Ax_k\rangle \\ &\leqslant \varphi(x^*) + \alpha_k + \varepsilon_k - \langle y^* + v^*, y_k\rangle, \end{aligned}$$

dont le membre de droite est borné (il converge vers $\varphi(x^*) + \alpha - \langle y^* + v^*, y\rangle \in \mathbb{R}$). □

Pré-composition avec une application linéaire

D'après la proposition 3.32, si $A : \mathbb{E} \to \mathbb{F}$ est une application linéaire et si $g \in \mathrm{C\overline{onv}}(\mathbb{F})$ vérifie $\mathcal{R}(A) \cap \operatorname{dom} g \neq \varnothing$, alors $g \circ A \in \mathrm{C\overline{onv}}(\mathbb{E})$, si bien que $(g \circ A)^{**} = g \circ A$. On s'intéresse ci-dessous au calcul de la biconjuguée $(g \circ A)^{**}$ lorsque la fonction $g \in \mathrm{Conv}(\mathbb{F})$ n'est pas nécessairement fermée, mais vérifie une condition plus forte que la condition minimale $\mathcal{R}(A) \cap \operatorname{dom} g \neq \varnothing$ assurant la propreté de $g \circ A$.

Proposition 3.51 (biconjuguée de $g \circ A$) *Soient $\mathbb{E}$ et $\mathbb{F}$ deux espaces vectoriels, $A : \mathbb{E} \to \mathbb{F}$ une application linéaire et $g \in \mathrm{Conv}(\mathbb{F})$ une fonction vérifiant $\mathcal{R}(A) \cap (\operatorname{dom} g)^{\ominus} \neq \varnothing$. Alors $g \circ A \in \mathrm{Conv}(\mathbb{E})$ et*

$$(g \circ A)^{**} = g^{**} \circ A.$$

Démonstration. On a déjà vu que $g \circ A \in \mathrm{Conv}(\mathbb{E})$. Comme dans la démonstration de la proposition 3.32, on introduit l'application linéaire

$$B : \mathbb{E} \times \mathbb{R} \to \mathbb{F} \times \mathbb{R} : (x, \alpha) \mapsto (Ax, \alpha)$$

qui permet d'écrire $\operatorname{epi}(h \circ A) = B^{-1}(\operatorname{epi} h)$, quelle que soit la fonction $h : \mathbb{F} \to \overline{\mathbb{R}}$. D'après (3.3) et grâce à la condition de propreté renforcée $\mathcal{R}(A) \cap (\operatorname{dom} g)^{\ominus} \neq \varnothing$:

$$\begin{aligned} B^{-1}((\operatorname{epi} g)^{\ominus}) &= B^{-1}\Big(\{(y, \alpha) : y \in (\operatorname{dom} g)^{\ominus},\ g(y) < \alpha\}\Big) \\ &= \{(x, \alpha) : Ax \in (\operatorname{dom} g)^{\ominus},\ g(Ax) < \alpha\} \\ &\neq \varnothing. \end{aligned}$$

Ensuite

$$\begin{aligned} \operatorname{epi}((g \circ A)^{**}) &= \overline{\operatorname{epi}(g \circ A)}, && \text{[proposition 3.46, point 2]} \\ &= \overline{B^{-1}(\operatorname{epi} g)} \\ &= B^{-1}(\overline{\operatorname{epi} g}), && \text{[proposition 2.17 et } B^{-1}((\operatorname{epi} g)^{\ominus}) \neq \varnothing\text{]} \\ &= B^{-1}(\operatorname{epi} g^{**}) \\ &= \operatorname{epi}(g^{**} \circ A). \end{aligned}$$

On a donc bien établi que $(g \circ A)^{**} = g^{**} \circ A$. □

Proposition 3.52 (conjuguée de $g \circ A$) *Soient $\mathbb{E}$ et $\mathbb{F}$ deux espaces euclidiens, $A : \mathbb{E} \to \mathbb{F}$ une application linéaire et $g \in \mathrm{Conv}(\mathbb{F})$.*

1) *Si $g \in \mathrm{C\overline{onv}}(\mathbb{F})$ et si $\mathcal{R}(A) \cap \operatorname{dom} g \neq \varnothing$, alors $g \circ A \in \mathrm{C\overline{onv}}(\mathbb{E})$ et*

$$(g \circ A)^{*} = (g^{*} \mathbin{\overline{\square}} A^{*})^{**}. \tag{3.43}$$

2) *Si g est convexe polyédrique et si $\mathcal{R}(A) \cap \operatorname{dom} g \neq \varnothing$, alors*

$$(g \circ A)^* = g^* \mathbin{\overline{\vee}} A^*. \tag{3.44}$$

De plus, l'infimum dans la définition de $(g^* \mathbin{\overline{\vee}} A^*)(x^*)$ est atteint s'il est fini.

3) *Si $g \in \operatorname{Conv}(\mathbb{F})$ et si $\mathcal{R}(A) \cap (\operatorname{dom} g)^\ominus \neq \varnothing$, alors $g \circ A \in \operatorname{Conv}(\mathbb{E})$ et (3.44) a lieu. De plus, l'infimum dans la définition de $(g^* \mathbin{\overline{\vee}} A^*)(x^*)$ est atteint s'il est fini.*

Démonstration. 1) D'après la proposition 3.32, $g \circ A \in \operatorname{C\overline{onv}}(\mathbb{E})$. On applique alors la proposition 3.49 avec $f := g^*$ et $A := A^*$ (ses hypothèses sont vérifiées, car $g^{**} = g$) : (3.42) donne $(g^* \mathbin{\overline{\vee}} A^*)^* = g \circ A$, qui par conjugaison conduit à (3.43).

2) Si g est convexe polyédrique et si $\mathcal{R}(A) \cap \operatorname{dom} g \neq \varnothing$, alors, d'après le point 1, on a (3.43). Mais g^* est convexe polyédrique (exercice 3.15) et donc aussi $g^* \mathbin{\overline{\vee}} A^*$ (proposition 3.38), qui est donc fermée. Alors (3.43) donne (3.44). La proposition 3.38 nous apprend aussi que l'infimum dans la définition de $(g^* \mathbin{\overline{\vee}} A^*)(x^*)$ est atteint s'il est fini.

3) D'après la proposition 3.32, $g \circ A \in \operatorname{Conv}(\mathbb{E})$. On peut appliquer le point 1 avec $g := g^{**} \in \operatorname{C\overline{onv}}(\mathbb{F})$, car $\mathcal{R}(A) \cap \operatorname{dom} g^{**} \supseteq \mathcal{R}(A) \cap (\operatorname{dom} g^{**})^\ominus = \mathcal{R}(A) \cap (\operatorname{dom} g)^\ominus$ [selon l'exercice 3.18 (2)] $\neq \varnothing$ [par hypothèse]. Ceci donne

$$(g^{**} \circ A)^* = (g^* \mathbin{\overline{\vee}} A^*)^{**}.$$

D'après la proposition 3.51 et grâce à l'hypothèse $\mathcal{R}(A) \cap (\operatorname{dom} g)^\ominus \neq \varnothing$, $g^{**} \circ A = (g \circ A)^{**}$, si bien que $(g^{**} \circ A)^* = (g \circ A)^*$. D'autre part, d'après la proposition 3.50 (2) avec $f := g^*$ et $A := A^*$ et grâce au fait vu ci-dessus que $\mathcal{R}(A) \cap \operatorname{dom} g^{**} \neq \varnothing$, on voit que $g^* \mathbin{\overline{\vee}} A^*$ est fermée (donc $(g^* \mathbin{\overline{\vee}} A^*)^{**} = g^* \mathbin{\overline{\vee}} A^*$) et que l'infimum définissant $g^* \mathbin{\overline{\vee}} A^*$ est atteint lorsqu'il est fini. □

Inf-convolution

Une inf-convolution peut se voir comme une inf-image sous une application linéaire, voir (3.29)–(3.30). En appliquant la proposition 3.49 à cette inf-image, on obtient le résultat suivant.

Proposition 3.53 (conjuguée de l'inf-convolution) *Soient $f_1, \ldots, f_p : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ des fonctions propres telles que*

$$\bigcap_{1 \leqslant i \leqslant p} \operatorname{dom} f_i^* \neq \varnothing.$$

Alors $f_1 \boxplus \cdots \boxplus f_p$ est propre avec une minorante affine et

$$(f_1 \boxplus \cdots \boxplus f_p)^* = f_1^* + \cdots + f_p^*.$$

DÉMONSTRATION. On applique donc la proposition 3.49. On sait que $f_1 \boxplus \cdots \boxplus f_p = f \boxvee A$, avec $f : \mathbb{E}^p \to \mathbb{R} \cup \{+\infty\}$ et $A : \mathbb{E}^p \to \mathbb{E}$ définies par

$$f(x_1, \ldots, x_p) = f_1(x_1) + \cdots + f_p(x_p) \qquad \text{et} \qquad A(x_1, \ldots, x_p) = x_1 + \cdots + x_p.$$

Exiger la propreté de f revient à demander celle des f_i. D'autre part, f a une minorante affine si, et seulement si, les f_i ont une minorante affine. On peut alors calculer f^* et A^* (la somme des $f_i^*(x_i^*)$ ci-dessous peut valoir $+\infty$, mais on est sûr de ne pas additionner $+\infty$ et $-\infty$) :

$$f^*(x_1^*, \ldots, x_p^*) = f_1^*(x_1^*) + \cdots + f_p^*(x_p^*) \qquad \text{et} \qquad A^*(x^*) = (x^*, \ldots, x^*).$$

Enfin la condition $\mathcal{R}(A^*) \cap \operatorname{dom} f^* \neq \varnothing$ s'écrit $\cap_i \operatorname{dom} f_i^* \neq \varnothing$. En conclusion, $f_1 \boxplus \cdots \boxplus f_p$ est propre avec une minorante affine et $(f_1 \boxplus \cdots \boxplus f_p)^* = (f \boxvee A)^* = f^* \circ A^* = f_1^* + \cdots + f_p^*$. □

Somme

Proposition 3.54 (conjuguée d'une somme) *Soient $f_1, \ldots, f_p : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ des fonctions.*

1) *Si les $f_i \in \overline{\mathrm{Conv}}(\mathbb{E})$ et $\cap_i \operatorname{dom} f_i \neq \varnothing$, alors $f_1 + \cdots + f_p \in \overline{\mathrm{Conv}}(\mathbb{E})$, $f_1^* \boxplus \cdots \boxplus f_p^* \in \mathrm{Conv}(\mathbb{E})$ et*

$$(f_1 + \cdots + f_p)^* = (f_1^* \boxplus \cdots \boxplus f_p^*)^{**}. \tag{3.45}$$

2) *Si les $f_i \in \mathrm{Conv}(\mathbb{E})$ et $\cap_i (\operatorname{dom} f_i)^\ominus \neq \varnothing$ (on peut remplacer $(\operatorname{dom} f_i)^\ominus$ par $\operatorname{dom} f_i$ dans cette condition si f_i est polyédrique), alors $f_1 + \cdots + f_p \in \mathrm{Conv}(\mathbb{E})$, $f_1^* \boxplus \cdots \boxplus f_p^* \in \overline{\mathrm{Conv}}(\mathbb{E})$ et*

$$(f_1 + \cdots + f_p)^* = f_1^* \boxplus \cdots \boxplus f_p^*. \tag{3.46}$$

De plus, l'infimum dans la définition de $(f_1^* \boxplus \cdots \boxplus f_p^*)(x^*)$ est atteint lorsqu'il est fini.

DÉMONSTRATION. 1) Supposons que les $f_i \in \overline{\mathrm{Conv}}(\mathbb{E})$ et que $\cap_i \operatorname{dom} f_i \neq \varnothing$. Les fonctions f_i^* vérifient bien les hypothèses demandées sur les f_i dans la proposition 3.53. On en déduit que $f_1^* \boxplus \cdots \boxplus f_p^*$ est propre avec une minorante affine et que $(f_1^* \boxplus \cdots \boxplus f_p^*)^* = f_1 + \cdots + f_p$. Donc $f_1 + \cdots + f_p \in \overline{\mathrm{Conv}}(\mathbb{E})$ (proposition 3.43) et $f_1^* \boxplus \cdots \boxplus f_p^* \in \mathrm{Conv}(\mathbb{E})$ ($f_1^* \boxplus \cdots \boxplus f_p^*$ est aussi convexe par la proposition 3.36). En prenant la conjuguée de la dernière identité, on obtient (3.45).

2a) Si toutes les fonctions f_i sont convexes polyédriques et que l'on a seulement $\cap_i \operatorname{dom} f_i \neq \varnothing$, le point 1 et l'exercice 3.15 donnent le résultat énoncé au point 2. En effet, dans ce cas, les $f_i \in \overline{\mathrm{Conv}}(\mathbb{E})$ comme fonctions convexes polyédriques, les $f_i^* \in \overline{\mathrm{Conv}}(\mathbb{E})$ comme conjuguées de fonctions convexes polyédriques et $f_1^* \boxplus \cdots \boxplus f_p^* \in \overline{\mathrm{Conv}}(\mathbb{E})$ comme inf-convolution de fonctions convexes polyédriques. De plus l'infimum dans la définition de $(f_1^* \boxplus \cdots \boxplus f_p^*)(x^*)$ est atteint lorsqu'il est fini.

2b) Supposons à présent que les $f_i \in \mathrm{Conv}(\mathbb{E})$ et que $\cap_i (\mathrm{dom}\, f_i)^\ominus \neq \varnothing$. On se ramène à la proposition 3.52 (point 2) en écrivant $f_1 + \cdots + f_p = g \circ A$, avec $g \in \mathrm{Conv}(\mathbb{E}^p)$ et $A : \mathbb{E} \to \mathbb{E}^p$ linéaire, définies par

$$g(x_1, \ldots, x_p) = f_1(x_1) + \cdots + f_p(x_p) \qquad \text{et} \qquad A(x) = (x, \ldots, x).$$

Il faut vérifier que $\mathcal{R}(A) \cap (\mathrm{dom}\, g)^\ominus \neq \varnothing$. On a $\mathcal{R}(A) = \{(x, \ldots, x) : x \in \mathbb{E}\}$ et $(\mathrm{dom}\, g)^\ominus = (\mathrm{dom}\, f_1 \times \cdots \times \mathrm{dom}\, f_p)^\ominus = (\mathrm{dom}\, f_1)^\ominus \times \cdots \times (\mathrm{dom}\, f_p)^\ominus$ (proposition 2.17), si bien que $\mathcal{R}(A) \cap (\mathrm{dom}\, g)^\ominus = \{(x, \ldots, x) : x \in (\mathrm{dom}\, f_1)^\ominus \cap \cdots \cap (\mathrm{dom}\, f_p)^\ominus\}$, qui est non vide par hypothèse. Dès lors $f_1 + \cdots + f_p \in \mathrm{Conv}(\mathbb{E})$ et

$$(f_1 + \cdots + f_p)^*(x^*) = (g \circ A)^*(x^*) = (g^* \mathbin{\overline{\vee}} A^*)(x^*) = \inf_{\substack{x_1^*, \ldots, x_p^* \in \mathbb{E} \\ A^*(x_1^*, \ldots, x_p^*) = x^*}} g^*(x_1^*, \ldots, x_p^*).$$

De plus, l'infimum ci-dessus est atteint lorsqu'il est fini. On vérifie aisément que $g^*(x_1^*, \ldots, x_p^*) = f_1^*(x_1^*) + \cdots + f_p^*(x_p^*)$ et que $A^*(x_1^*, \ldots, x_p^*) = x_1^* + \cdots + x_p^*$, ce qui permet de conclure.

2c) Considérons à présent le cas mixte où les $f_i \in \mathrm{Conv}(\mathbb{E})$, où $f_1, \ldots, f_k$ sont polyédriques ($0 < k < p$) et où $(\cap_{1 \leqslant i \leqslant k} \mathrm{dom}\, f_i) \cap (\cap_{k+1 \leqslant i \leqslant p} (\mathrm{dom}\, f_i)^\ominus) \neq \varnothing$. On note

$$g_1 = f_1 + \cdots + f_k \qquad \text{et} \qquad g_2 = f_{k+1} + \cdots + f_p.$$

D'après les points 2a et 2b, on a

$$\begin{aligned} g_1^*(y_1^*) &= \inf\{f_1^*(x_1^*) + \cdots + f_k^*(x_k^*) : x_1^* + \cdots + x_k^* = y_1^*\} \\ g_2^*(y_2^*) &= \inf\{f_{k+1}^*(x_{k+1}^*) + \cdots + f_p^*(x_p^*) : x_{k+1}^* + \cdots + x_p^* = y_2^*\}, \end{aligned}$$

avec des bornes inférieures atteintes lorsqu'elles sont finies. Il suffit donc de montrer que

$$(g_1 + g_2)^*(x^*) = \inf\{g_1^*(y_1^*) + g_2^*(y_2^*) : y_1^* + y_2^* = x^*\} \tag{3.47}$$

et que la borne inférieure est atteinte si elle est finie, car alors

$$\begin{aligned} (f_1 + \cdots + f_p)^*(x^*) &= (g_1 + g_1)^*(x^*) \\ &= \inf_{y_1^* + y_2^* = x^*} \Big(g_1^*(y_1^*) + g_2^*(y_2^*)\Big) \qquad (3.48) \\ &= \inf_{y_1^* + y_2^* = x^*} \Bigg(\inf_{x_1^* + \cdots + x_k^* = y_1^*} \Big(f_1^*(x_1^*) + \cdots + f_k^*(x_k^*)\Big) \\ &\qquad + \inf_{x_{k+1}^* + \cdots + x_p^* = y_2^*} \Big(f_{k+1}^*(x_{k+1}^*) + \cdots + f_p^*(x_p^*)\Big)\Bigg) \\ &= \inf_{x_1^* + \cdots + x_p^* = x^*} \Big(f_1^*(x_1^*) + \cdots + f_p^*(x_p^*)\Big). \end{aligned}$$

De plus, si $(f_1 + \cdots + f_p)^*(x^*)$ est fini, l'infimum en (3.48) est atteint par des y_i^* (c'est ce que l'on démontrera), si bien que les $g_i(y_i^*)$ sont finis et les autres bornes inférieures sont également atteintes. Montrons donc (3.47) et l'affirmation qui suit.

On note $\mathcal{A}_2 := \mathrm{aff}\, \mathrm{dom}\, g_2$. Clairement, $\mathrm{dom}\, g_1 = \cap_{1 \leqslant i \leqslant k} \mathrm{dom}\, f_i$ et $(\mathrm{dom}\, g_2)^\ominus = (\cap_{k+1 \leqslant i \leqslant p} \mathrm{dom}\, f_i)^\ominus = \cap_{k+1 \leqslant i \leqslant p} (\mathrm{dom}\, f_i)^\ominus$, car cette dernière intersection est supposée non vide (proposition 2.17). Alors $(\mathrm{dom}\, g_1) \cap (\mathrm{dom}\, g_2)^\ominus \neq \varnothing$, par hypothèse. Il en découle (voir l'exercice 2.12)

$$(\mathcal{A}_2 \cap \operatorname{dom} g_1)^{\ominus} \cap (\operatorname{dom} g_2)^{\ominus} \neq \varnothing.$$

On voit que l'on peut appliquer le point 2b aux fonctions

$$h_1 = g_1 + \mathcal{I}_{\mathcal{A}_2} \in \operatorname{Conv}(\mathbb{E})$$

et g_2, qui ont un point commun dans l'intérieur relatif de leur domaine. Comme $g_1 + g_2 = h_1 + g_2$, on trouve

$$(g_1 + g_2)^*(x^*) = (h_1 + g_2)^*(x^*) = \inf_{z_1^* + z_2^* = x^*} \Big(h_1^*(z_1^*) + g_2^*(z_2^*) \Big),$$

avec un infimum atteint s'il est fini. On peut aussi calculer h_1^* par le point 2a, car g_1 et $\mathcal{I}_{\mathcal{A}_2}$ sont polyédriques (exercice 3.15) et ont des domaines qui s'intersectent :

$$h_1^*(z_1^*) = \inf_{y_1^* + z^* = z_1^*} \Big(g_1^*(y_1^*) + \mathcal{I}_{\mathcal{A}_2}^*(z^*) \Big),$$

avec un infimum atteint s'il est fini. En injectant ci-dessus, on trouve

$$(g_1 + g_2)^*(x^*) = \inf_{y_1^* + z^* + z_2^* = x^*} \Big(g_1^*(y_1^*) + \mathcal{I}_{\mathcal{A}_2}^*(z^*) + g_2^*(z_2^*) \Big),$$

avec un infimum atteint s'il est fini. On peut maintenant regrouper les deux derniers termes, car $\mathcal{I}_{\mathcal{A}_2} \in \operatorname{Conv}(\mathbb{E})$ et $g_2 \in \operatorname{Conv}(\mathbb{E})$ ont un point commun dans l'intérieur relatif de leur domaine ($(\operatorname{dom} \mathcal{I}_{\mathcal{A}_2})^{\ominus} = \mathcal{A}_2 \supseteq (\operatorname{dom} g_2)^{\ominus}$) :

$$g_2^*(y_2^*) = (\mathcal{I}_{\mathcal{A}_2} + g_2)^*(y_2^*) = \inf_{z^* + z_2^* = y_2^*} \Big(\mathcal{I}_{\mathcal{A}_2}^*(z^*) + g_2^*(z_2^*) \Big),$$

avec un infimum atteint s'il est fini. On conclut

$$\begin{aligned} (g_1 + g_2)^*(x^*) &= \inf_{y_1^* + y_2^* = x^*} \left(g_1^*(y_1^*) + \inf_{z^* + z_2^* = y_2^*} \Big(\mathcal{I}_{\mathcal{A}_2}^*(z^*) + g_2^*(z_2^*) \Big) \right) \\ &= \inf_{y_1^* + y_2^* = x^*} \Big(g_1^*(y_1^*) + g_2^*(y_2^*) \Big), \end{aligned}$$

avec un infimum atteint s'il est fini. □

3.6 Sous-différentiabilité

La notion de dérivée est fondamentale en analyse car elle permet d'approcher localement des fonctions par des modèles linéaires, plus simples à étudier. Ces modèles fournissent des renseignements sur les fonctions qu'ils approchent, si bien que de nombreuses questions d'analyse passent par l'étude des fonctions linéarisées (stabilité, inversibilité locale, optimalité, *etc*). On rencontre beaucoup de fonctions convexes qui ne sont pas différentiables au sens classique (voir l'annexe C pour un rappel sur le calcul différentiel), en particulier lorsque celles-ci résultent de constructions qui n'ont rien pour assurer la différentiabilité des fonctions qu'elles produisent. Il en est ainsi de la fonction duale associée à un problème d'optimisation sous contraintes, pour en

citer un exemple emblématique. Pour ces fonctions, on dispose toutefois de la notion de sous-différentiel qui peut jouer un rôle similaire à celui de la dérivée des fonctions plus régulières, plus lisses. Nous la développons dans cette section.

Le sous-différentiel est un substitut de la notion de gradient pour les fonctions convexes *non différentiables* (section 3.6.1). Ce n'est plus un vecteur de $\mathbb{E}$ comme le gradient, mais un sous-ensemble de $\mathbb{E}$ qui n'est réduit à un point qu'en cas de différentiabilité (proposition 3.68). Nous l'introduirons de trois manières différentes : à partir des dérivées directionnelles, dont on sait qu'elles existent toujours pour les fonctions convexes (proposition 3.15), en utilisant les minorantes affines de f (section 3.2.2) et en faisant usage de la conjuguée f^* (section 3.5).

Une fonction convexe est sous-différentiable si son sous-différentiel est non vide (définition 3.56). La propriété de sous-différentiabilité s'exprime donc sous la forme de résultat d'existence. Appliqué à la fonction valeur d'un problème d'optimisation cela impliquera l'existence de multiplicateurs optimaux (section 4.6.1) et donc de solutions d'un autre problème d'optimisation, le problème dual (chapitre 14). On dispose donc là d'une méthode originale pour montrer qu'un problème d'optimisation a une solution : il suffit de montrer qu'il est le dual d'un problème dont la fonction valeur est sous-différentiable en zéro (cette technique sera par exemple utilisée pour établir le théorème 14.14).

Dans cette section, on suppose que l'on travaille sur un espace euclidien $\mathbb{E}$, dont le produit scalaire est noté $\langle \cdot, \cdot \rangle$.

3.6.1 Définitions

Dans cette section, on se donne une fonction $f : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ convexe avec domaine non vide, c'est-à-dire

$$f \in \mathrm{Conv}(\mathbb{E}).$$

On rappelle que ces fonctions ont aussi une minorante affine et donc $f^* \in \mathrm{C\overline{onv}}(\mathbb{E})$ (proposition 3.43). On rappelle également, qu'en tout point de son domaine, une fonction convexe admet une dérivée directionnelle suivant toute direction, celle-ci pouvant éventuellement prendre les valeurs $-\infty$ ou $+\infty$ (proposition 3.15).

La notion de sous-différentiabilité est fondée sur la proposition suivante, dont la longueur de démonstration est inversement proportionnelle à l'importance, qui est grande.

Proposition 3.55 (définitions du sous-différentiel) *Soient* $f \in \mathrm{Conv}(\mathbb{E})$, $x \in \mathrm{dom}\, f$ *et* $x^* \in \mathbb{E}$. *Alors les propriétés suivantes sont équivalentes :*

(i) $\forall\, d \in \mathbb{E} : f'(x;d) \geqslant \langle x^*, d\rangle$,
(ii) $\forall\, y \in \mathbb{E} : f(y) \geqslant f(x) + \langle x^*, y - x\rangle$,
(iii) $x \in \arg\min_{y\in\mathbb{E}} \big(f(y) - \langle x^*, y\rangle\big) = \arg\max_{y\in\mathbb{E}} \big(\langle x^*, y\rangle - f(y)\big)$,
(iv) $f(x) + f^*(x^*) \leqslant \langle x^*, x\rangle$,
(v) $f(x) + f^*(x^*) = \langle x^*, x\rangle$.

Démonstration. $[(i) \Rightarrow (ii)]$ Par convexité de f et (i) avec $d = y - x$, on obtient (ii) :

$$\forall\, y \in \mathbb{E} : \quad f(y) \geqslant f(x) + f'(x; y - x) \geqslant f(x) + \langle x^*, y - x\rangle.$$

[$(ii) \Rightarrow (iii)$] On récrit (ii) comme suit :

$$\forall\, y \in \mathbb{E} : \quad f(y) - \langle x^*, y\rangle \geqslant f(x) - \langle x^*, x\rangle. \tag{3.49}$$

C'est (iii).

[$(iii) \Rightarrow (iv)$] On récrit (3.49) comme suit :

$$\forall\, y \in \mathbb{E} : \quad \langle x^*, x\rangle - f(x) \geqslant \langle x^*, y\rangle - f(y). \tag{3.50}$$

En prenant le supremum en $y \in \mathbb{E}$ à droite, on obtient (iv).

[$(iv) \Rightarrow (v)$] Par définition de la conjuguée, on a toujours $f(x) + f^*(x^*) \geqslant \langle x^*, x\rangle$.

[$(v) \Rightarrow (i)$] Par (v), on a certainement (3.50), qui avec $y = x + td$ devient

$$f(x + td) - f(x) \geqslant \langle x^*, td\rangle.$$

Après division par $t > 0$ et passage à la limite lorsque $t \downarrow 0$, on obtient (i). □

Définitions 3.56 On dit qu'une fonction $f \in \mathrm{Conv}(\mathbb{E})$ est *sous-différentiable* en un point $x \in \mathrm{dom}\, f$ s'il existe $x^* \in \mathbb{E}$ vérifiant les propriétés équivalentes (i)-(v) de la proposition 3.55. Ces éléments x^* sont appelés les *sous-gradients* de f en x. L'ensemble des sous-gradients de f en x est appelé le *sous-différentiel* de f en x. Cet ensemble est noté $\partial f(x)$. Par convention, $\partial f(x) = \varnothing$ si $x \notin \mathrm{dom}\, f$. □

Chacune des cinq conditions équivalentes (i)-(v) de la proposition 3.55 reflète un aspect particulier des sous-gradients, permettant en particulier de les calculer.

1) La condition (i) nous apprend qu'un sous-gradient est la pente d'une minorante *linéaire* de $\delta_x(\cdot) = f'(x;\cdot)$, forcément exacte en 0 (les deux fonctions s'annulent en zéro). Ce point de vue conduit à la méthode de calcul de $\partial f(x)$ suivante. On calcule d'abord la dérivée directionnelle $f'(x;\cdot)$ de f en x et on en détermine toutes les minorantes linéaires ; ses pentes forment $\partial f(x)$.

2) Selon la condition (ii), dont l'inégalité porte le nom d'*inégalité du sous-gradient*, un sous-gradient est la pente d'une minorante *affine* de f, exacte en x. On utilise parfois cette approche pour montrer qu'une pente particulière $s \in \mathbb{E}$ est un sous-gradient. Elle est par exemple utilisée dans la démonstration de la proposition 14.34.

3) La condition (iii) nous informe aussi que x^* est un sous-gradient de f en x si $f(\cdot) - \langle x^*, \cdot\rangle$ atteint son minimum en x. Ce point de vue conduit à la méthode de calcul de $\partial f(x)$ suivante. On se donne d'abord x^* et on calcule les minimiseurs de $x \mapsto f(x) - \langle x^*, x\rangle$; un tel minimiseur x est tel que $x^* \in \partial f(x)$.

4) Les conditions (iv)-(v) nous disent que x^* est un sous-gradient de f en x si la plus haute minorante affine de f de pente x^*, qui est l'application $y \in \mathbb{E} \mapsto \langle x^*, y\rangle - f^*(x^*)$ (voir la discussion au début de la section 3.5.1), est exacte en x. Ce point de vue conduit à la méthode de calcul de $\partial f(x)$ suivante. On commence par calculer la fonction conjuguée f^* puis on utilise la relation (v). Cette approche est souvent la plus aisée lorsque f^* ne prend que les valeurs 0 ou $+\infty$; des exemples sont donnés aux exercices 3.30, 3.31, 3.34 et 3.36.

Nous donnerons à la section 3.6.3 des règles de calcul du sous-différentiel qui permettent souvent de ne pas devoir recourir aux méthodes de calcul de base décrites ci-dessus.

La figure 3.8 illustre la définition du sous-différentiel.

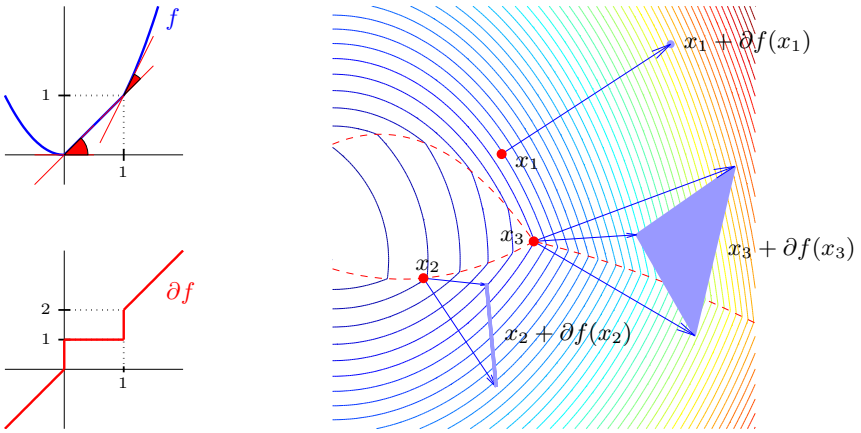


Fig. 3.8. Représentations du sous-différentiel. À gauche : la fonction $x \in \mathbb{R} \mapsto f(x) = \max(x, x^2)$ et la multifonction ∂f. À droite : quelques sous-différentiels dans $\mathbb{R}^2$ de $x \in \mathbb{R}^2 \mapsto \max(q_1(x), q_2(x), q_3(x))$, où les q_i sont quadratiques convexes.

- À gauche, on a représenté la fonction $f : \mathbb{R} \to \mathbb{R}$ définie par $f(x) = \max(x, x^2)$ (en haut) et l'application multivoque $x \in \mathbb{R} \mapsto \partial f(x) \in \mathcal{P}(\mathbb{R})$ (en bas). Comme on le verra à la proposition 3.68, $\partial f(x)$ ne se distingue de $\nabla f(x)$, le gradient de f en x, qu'en des points de non-différentiabilité, c'est-à-dire ici aux points $x = 0$ et $x = 1$. D'après le point (ii) de la proposition 3.55, en un tel point x, $\partial f(x)$ est l'ensemble des pentes x^* telles que $y \mapsto f(x) + \langle x^*, y - x\rangle$ minore f sur $\mathbb{R}$. Ainsi $\partial f(0) = [0, 1]$ et $\partial f(1) = [1, 2]$. On observe le caractère monotone maximal de la multifonction $x \mapsto \partial f(x)$ (proposition 3.67).
- À droite sont représentées les courbes de niveau d'une fonction $f : \mathbb{R}^2 \to \mathbb{R}$, qui est le maximum de trois fonctions quadratiques convexes. Les courbes en tirets sont les lieux de non différentiabilité, que l'on repère par les coins qu'y font les courbes de niveau. On les appelle les *plis* de f. Au point x_1, la fonction est différentiable et le sous-différentiel $\partial f(x_1)$ se confond avec le vecteur gradient $\nabla f(x_1)$ (proposition 3.68), que l'on a représenté ici après avoir translaté l'origine en x_1. En x_2 et x_3 la fonction n'est pas différentiable. Les sous-différentiels en ces points sont alors les enveloppes convexes des vecteurs gradients des fonctions quadratiques « actives » en ces points (proposition 3.78) : $\partial f(x_2)$ est un segment et $\partial f(x_3)$ est un triangle. Les sous-différentiels représentés à la figure 3.8 sont polyédriques, mais ceci est dû à la nature de la fonction considérée (une enveloppe supérieure d'un nombre fini de fonctions convexes différentiables). Il n'en est pas toujours ainsi ; par exemple, le sous-différentiel en zéro de la norme ℓ_2 dans $\mathbb{R}^n$ est la boule unité associée à cette norme (point 3 de l'exercice 3.30).

Voici quelques corollaires de la proposition 3.55.

Corollaire 3.57 (sous-différentiel de la dérivée directionnelle) *Soit $x \in \operatorname{dom} f$ et δ_x l'application dérivée directionnelle définie en (3.10). Alors*

$$\partial f(x) = \partial \delta_x(0).$$

DÉMONSTRATION. Comme $\delta_x(0) = 0$, cette relation se déduit de l'équivalence entre les points (i) et (ii) de la proposition 3.55. □

Les sous-différentiels de f et de sa conjuguée f^* jouissent d'une belle règle de réciprocité, parfois appelée *règle de bascule* [316].

Corollaire 3.58 (règle de bascule)
1) *Si $f \in \operatorname{Conv}(\mathbb{E})$, alors $x^* \in \partial f(x) \implies x \in \partial f^*(x^*)$.*
2) *Si $f \in \operatorname{C\overline{onv}}(\mathbb{E})$, alors $x^* \in \partial f(x) \iff x \in \partial f^*(x^*)$.*

DÉMONSTRATION. 1) Si $x^* \in \partial f(x)$, on a $f(x) + f^*(x^*) = \langle x^*, x\rangle$ (point (v) de la proposition 3.55). Comme $f^{**} \leqslant f$ (corollaire 3.47), on a aussi, $f^{**}(x) + f^*(x^*) \leqslant \langle x^*, x\rangle$, ce qui exprime le fait que $x \in \partial f^*(x^*)$ (point (iv) de la proposition 3.55).
2) On a cette fois

$$\begin{aligned} x^* \in \partial f(x) &\iff f(x) + f^*(x^*) = \langle x^*, x\rangle \qquad [\text{proposition 3.55 } (v)] \\ &\iff f^{**}(x) + f^*(x^*) = \langle x^*, x\rangle \qquad [\text{proposition 3.46}] \\ &\iff x \in \partial f^*(x^*) \qquad [\text{proposition 3.55 } (v) \text{ appliquée à } f^*], \end{aligned}$$

ce qui montre l'équivalence. □

La réciproque n'a pas lieu au point 1, pour la fonction $f \in \operatorname{Conv}(\mathbb{R})$ ci-dessous

$$f(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1 & \text{si } x = 0 \\ +\infty & \text{si } x > 0 \end{cases} \qquad \text{donc} \qquad f^*(x^*) = \begin{cases} +\infty & \text{si } x^* < 0 \\ 0 & \text{si } x^* \geqslant 0, \end{cases}$$

puisque l'on a $0 \notin \partial f(0) = \varnothing$, alors que $0 \in \partial f^*(0) =]-\infty, 0]$.

Corollaire 3.59 (sous-différentiel et f^*) *Si $f \in \operatorname{C\overline{onv}}(\mathbb{E})$, alors*

$$\partial f(x) = \operatorname*{arg\,max}_{x^* \in \mathbb{E}} \big(\langle x^*, x\rangle - f^*(x^*)\big). \tag{3.51}$$

DÉMONSTRATION. Cela se déduit des équivalences suivantes :

$$\begin{aligned} x^* \in \partial f(x) &\iff x \in \partial f^*(x^*) \qquad \text{[règle de bascule]} \\ &\iff x^* \in \arg\max\big(\langle \cdot, x\rangle - f^*(\cdot)\big), \end{aligned}$$

par le point (iii) de la proposition 3.55, appliqué à $f^* \in \overline{\mathrm{Conv}}(\mathbb{E})$ au lieu de f. □

Corollaire 3.60 (sous-différentiels de f et f^{**}) *Soient $f \in \mathrm{Conv}(\mathbb{E})$ et $x \in \mathrm{dom}\, f$. Alors*
1) $\partial f(x) \subseteq \partial f^{**}(x)$,
2) $\partial f(x) \neq \varnothing \implies f(x) = f^{**}(x)$,
3) $f(x) = f^{**}(x) \implies \partial f(x) = \partial f^{**}(x)$.

Démonstration. 1) Si $x^* \in \partial f(x)$, alors $f(x) + f^*(x^*) = \langle x^*, x\rangle$ (point (v) de la proposition 3.55), donc $f^{**}(x) + (f^{**})^*(x^*) \leqslant \langle x^*, x\rangle$ car $f^{**} \leqslant f$ (corollaire 3.47) et $(f^{**})^* = f^*$ (proposition 3.48), si bien que $x^* \in \partial f^{**}(x)$ (point (iv) de la proposition 3.55).

2) On a toujours $f^{**}(x) \leqslant f(x)$ (corollaire 3.47). D'autre part, si $x^* \in \partial f(x)$, on a par définition de f^{**} et par le point (v) de la proposition 3.55 :

$$f^{**}(x) \geqslant \langle x^*, x\rangle - f^*(x^*) = f(x).$$

3) D'après le point 1, il reste à montrer l'inclusion $\partial f^{**}(x) \subseteq \partial f(x)$. Si $x^* \in \partial f^{**}(x)$, on a $f^{**}(x) + (f^{**})^*(x^*) = \langle x^*, x\rangle$. En utilisant l'hypothèse $f(x) = f^{**}(x)$ et la propriété $(f^{**})^*(x^*) = f^*(x^*)$, on obtient $f(x) + f^*(x^*) = \langle x^*, x\rangle$, c'est-à-dire $x^* \in \partial f(x)$. □

Corollaire 3.61 (optimalité I)
1) $\inf f = -f^*(0)$.
2) *Si $f \in \mathrm{Conv}(\mathbb{E})$, alors $\bar{x} \in \arg\min f$ si, et seulement si,* $0 \in \partial f(\bar{x})$.
3) *Si $f \in \overline{\mathrm{Conv}}(\mathbb{E})$, alors* $\arg\min f = \partial f^*(0)$.

Démonstration. 1) $f^*(0) = \sup_x \langle 0, x\rangle - f(x) = -\inf_x f(x)$.

2) C'est une conséquence immédiate du point (ii) de la proposition 3.55 : si x minimise f, on peut prendre $x^* = 0$ dans (ii), donc $0 \in \partial f(x)$; inversement, si $0 \in \partial f(x)$, $x^* = 0$ dans (ii) montre que x minimise f.

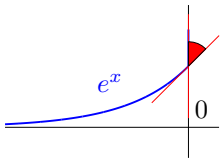
3) On a $\bar{x} \in \arg\min f$ si, et seulement si, $0 \in \partial f(\bar{x})$ (point 2) si, et seulement si, $\bar{x} \in \partial f^*(0)$ (règle de bascule). □

D'après le point 1 du corollaire, trouver un minimiseur de f sur $\mathbb{E}$ revient à trouver les « zéros » de la fonction multivoque $\partial f(\cdot)$. On rapprochera ce fait du résultat selon lequel, lorsque f est convexe différentiable, les minimiseurs de f sont les zéros de son application gradient $\nabla f(\cdot)$.

3.6.2 Quelques propriétés

Une fonction $f \in \mathrm{Conv}(\mathbb{E})$ n'est pas nécessairement sous-différentiable en tout point de son domaine. Comme nous allons le montrer, la difficulté vient essentiellement du fait que la fonction dérivée directionnelle en x peut prendre la valeur $-\infty$. Le fait que $f'(x;d) = -\infty$ empêche que f soit sous-différentiable en x se voit clairement par le point (i) de la proposition 3.55. La réciproque est le sujet de la proposition 3.62 ci-dessous. Une telle situation se présente pour la fonction convexe f définie par (3.9), l'opposée de la racine carrée [441]. Cette fonction n'est pas sous-différentiable en 0, parce que $f'(0;1) = -\infty$. Évidemment, si $f'(x;d) = -\infty$, alors $f'(x;-d) = +\infty$ par (3.8), mais ce n'est pas la valeur $+\infty$ de la dérivée directionnelle qui empêche f d'être sous-différentiable en x. C'est ce que montre l'exemple suivant de fonction convexe f, dont le sous-différentiel en 0 vaut $[1, +\infty[$:

$$f(x) = \begin{cases} e^x & \text{si } x \leqslant 0 \\ +\infty & \text{sinon.} \end{cases}$$



Comme nous l'avons mentionné au début de cette section, la mise en évidence de conditions suffisantes de sous-différentiabilité est importante, car ces conditions deviennent alors un moyen de montrer que certains problèmes d'optimisation ont des solutions. C'est le but de la proposition suivante.

Proposition 3.62 (sous-différentiabilité) *Si $f \in \mathrm{Conv}(\mathbb{E})$ et $x \in \mathrm{dom}\, f$, les propriétés suivantes sont équivalentes :*

(i) $\partial f(x) \neq \varnothing$,
(ii) *il existe* $y \in (\mathrm{dom}\, f)^\circ$ *tel que* $f'(x; y-x) > -\infty$,
(iii) $f'(x;\cdot)$ *est propre* (*ou ne prend pas la valeur* $-\infty$).

Ces propriétés sont vérifiées si $x \in (\mathrm{dom}\, f)^\circ$.

Démonstration. $[(i) \Rightarrow (ii)]$ Soit y un point quelconque de $(\mathrm{dom}\, f)^\circ$, qui est non vide. Alors $f'(x; y-x) > -\infty$ sinon $\partial f(x)$ serait vide (par le point (i) de la proposition 3.55).

$[(ii) \Rightarrow (iii)]$ Comme $f'(x;0) = 0$, il suffit de montrer que $f'(x;\cdot)$ ne prend pas la valeur $-\infty$. On raisonne par l'absurde. Soit $d_1 := y - x$ (qui peut être nul, si $x \in (\mathrm{dom}\, f)^\circ$) et supposons qu'il existe $d_0 \in \mathbb{E}$ tel que $f'(x;d_0) = -\infty$. On pose $d_t := (1-t)d_0 + td_1$. Par convexité de $f'(x;\cdot)$ (point 1 de la proposition 3.16), $f'(x;d_t)$ vaut $-\infty$ pour $t \in [0,1[$ et vaut $+\infty$ pour $t > 1$. Mais, pour $\alpha \in]0,1]$ suffisamment petit, $x + \alpha d_0 \in \mathrm{dom}\, f$ (sinon la dérivée directionnelle suivant d_0 ne vaudrait pas $-\infty$). Alors, par convexité de $\mathrm{dom}\, f$,

$$x + \alpha d_t = (1-t)[x + \alpha d_0] + t[x + \alpha d_1] \in \mathrm{dom}\, f$$

pour $t \in [0,1]$ et même pour $t > 1$ proche de 1 (car $x + \alpha d_1 \in (\mathrm{dom}\, f)^\circ$ par le lemme 2.14). Mais alors, on ne peut avoir $f'(x;d_t) = +\infty$ pour ces $t > 1$ (le quotient différentiel est monotone), comme affirmé précédemment.

[$(iii) \Rightarrow (i)$] Si $f'(x;\cdot)$ est propre, cette fonction convexe a une minorante affine (proposition 3.7) : $\exists x^* \in \mathbb{E}$ et $\alpha \in \mathbb{R}$ tels que pour tout $d \in \mathbb{E}$ on a

$$f'(x;d) \geqslant \langle x^*, d\rangle + \alpha.$$

En remplaçant d par td avec $t \uparrow \infty$, on voit que l'on peut prendre $\alpha = 0$. Par le point (i) de la proposition 3.55, ce $x^* \in \partial f(x)$ qui est donc non vide.

Finalement, si $x \in (\operatorname{dom} f)^{\circ\circ}$, la propriété (ii) est vérifiée en prenant $y = x$. □

On trouvera une autre condition nécessaire et suffisante de sous-différentiabilité à l'exercice 3.26, qui est dans le même esprit que la proposition précédente : il faut et il suffit que f soit « sous-lipschitzienne ».

Corollaire 3.63 (optimalité II) *Si $f \in \Gamma_0\overline{\operatorname{conv}}(\mathbb{E})$ et $0 \in (\operatorname{dom} f^*)^{\circ\circ}$, alors* $\arg\min f$ *est non vide.*

Démonstration. Lorsque $f \in \Gamma_0\overline{\operatorname{conv}}(\mathbb{E})$, $\arg\min f = \partial f^*(0)$ (corollaire 3.61). D'après la proposition, lorsque $0 \in (\operatorname{dom} f^*)^{\circ\circ}$, f^* est sous-différentiable en zéro, donc $\arg\min f \neq \varnothing$. □

Proposition 3.64 (propriétés géométrique et topologique du sous-différentiel) *Soient $f \in \operatorname{Conv}(\mathbb{E})$, $\mathbb{E}_0$ le sous-espace vectoriel parallèle à $\operatorname{aff}(\operatorname{dom} f)$, $\mathrm{P}_{\mathbb{E}_0}$ le projecteur orthogonal sur $\mathbb{E}_0$ et $x \in \operatorname{dom} f$. On note $f|_{x+\mathbb{E}_0}$ la restriction de f à $x + \mathbb{E}_0 = \operatorname{aff}(\operatorname{dom} f)$. Alors*

1) $\partial f(x) = \partial f|_{x+\mathbb{E}_0}(x) + \mathbb{E}_0^\perp$, *en particulier* $\mathrm{P}_{\mathbb{E}_0}\,\partial f(x) = \partial f|_{x+\mathbb{E}_0}(x)$,
2) $\partial f(x)$ *est convexe et fermé (éventuellement vide),*
3) $x \in (\operatorname{dom} f)^{\circ\circ} \iff \mathrm{P}_{\mathbb{E}_0}\,\partial f(x)$ *est non vide et borné,*
4) $x \in (\operatorname{dom} f)^{\circ} \iff \partial f(x)$ *est non vide et borné.*

Démonstration. 1) On note $f_0 : y \in \mathbb{E}_0 \mapsto f(x+y)$ la restriction de f à $x + \mathbb{E}_0$. Soit $x^* \in \partial f(x)$. On peut écrire $x^* = x_0^* + h$, avec $x_0^* = \mathrm{P}_{\mathbb{E}_0}\, x^*$ et $h = x^* - x_0^* \in \mathbb{E}_0^\perp$. Il reste à montrer que $x_0^* \in \partial f_0(0)$. Par définition de x^*, $f'(x;d) \geqslant \langle x^*, d\rangle$, pour tout $d \in \mathbb{E}$. Mais, pour tout $d \in \mathbb{E}_0$: $f'(x;d) = f_0'(0;d)$ et $\langle x^*, d\rangle = \langle x_0^*, d\rangle$, si bien que $f_0'(0;d) \geqslant \langle x_0^*, d\rangle$. On a montré que $x_0^* \in \partial f_0(0)$. Réciproquement, soient $x_0^* \in \partial f_0(0)$ et $h \in \mathbb{E}_0^\perp$. Alors, pour tout $d \in \mathbb{E}_0$: $f_0'(0;d) \geqslant \langle x_0^*, d\rangle$, $f_0'(0;d) = f'(x;d)$ et $\langle h, d\rangle = 0$, si bien que $f'(x;d) \geqslant \langle x_0^* + h, d\rangle$. Comme $f'(x;d) = +\infty$ si $d \notin \mathbb{E}_0$, on a certainement aussi $f'(x;d) \geqslant \langle x_0^* + h, d\rangle$, pour tout $d \in \mathbb{E}$. Donc $x_0^* + h \in \partial f(x)$.

2) Comme on a toujours $f(x) + f^*(x^*) \geqslant \langle x^*, x\rangle$, le sous-différentiel est donné par

$$\partial f(x) = \{x^* : f^*(x^*) - \langle x^*, x\rangle \leqslant -f(x)\}.$$

Mais $x^* \mapsto f^*(x^*) - \langle x^*, x\rangle$ est dans $\Gamma_0\overline{\operatorname{conv}}(\mathbb{E})$ (proposition 3.43). Comme ensemble de sous-niveau de cette fonction, $\partial f(x)$ est donc convexe et fermé.

3) Notons d'abord que $\partial f(x)$ est non vide lorsque $x \in (\operatorname{dom} f)^{\circ\circ}$ (proposition 3.62), donc $\mathrm{P}_{\mathbb{E}_0}\,\partial f(x) \neq \varnothing$. Montrons à présent que cet ensemble est borné. Pour un $\varepsilon > 0$

assez petit, on note $L \geqslant 0$ le module de Lipschitz de f sur $B(x, 2\varepsilon) \cap (\operatorname{dom} f)^{\ominus}$ (proposition 3.14). Soit x^* un élément non nul de $\partial f(x)$ (si le sous-différentiel est réduit à $\{0\}$, il n'y a plus rien à démontrer). En prenant $y = x + \varepsilon x^*/\|x^*\|$ au point (ii) de la proposition 3.55, on obtient $\varepsilon\|x^*\| \leqslant f(x + \varepsilon x^*/\|x^*\|) - f(x) \leqslant L\varepsilon$. Donc $\|x^*\| \leqslant L$.

On démontre l'implication inverse par l'absurde. Supposons que $x \in (\operatorname{dom} f) \setminus (\operatorname{dom} f)^{\ominus}$ et que $x^* \in \partial f(x)$ qui est supposé non vide (si le sous-différentiel est vide, il n'y a plus rien à démontrer). Il faut montrer que $\mathrm{P}_{\mathbb{E}_0} \partial f(x)$ est non borné. Par définition de x^*, on a $f(y) \geqslant f(x) + \langle x^*, y - x\rangle$ pour tout $y \in \mathbb{E}$. Soit $\nu \in \mathbb{E}_0$ une normale non nulle à $\operatorname{dom} f$ en x (elle existe par la proposition 2.29), qui vérifie donc $\langle \nu, y - x\rangle \leqslant 0$ pour tout $y \in \operatorname{dom} f$. Alors, on a $f(y) \geqslant f(x) + \langle x^* + t\nu, y - x\rangle$, pour tout $y \in \mathbb{E}$ et pour tout $t > 0$. Dès lors $x^* + t\nu \in \mathbb{E}_0 \cap \partial f(x)$ pour tout $t > 0$, ce qui démontre que cet ensemble est non borné.

4) Si $x \in (\operatorname{dom} f)^\circ$, $\mathbb{E}_0 = \mathbb{E}$ et le point 3 montre que $\partial f(x)$ est non vide et borné.

Réciproquement, si $\partial f(x)$ est non vide et borné, $\mathbb{E}_0 = \mathbb{E}$ par le point 1. Alors $(\operatorname{dom} f)^{\ominus} = (\operatorname{dom} f)^\circ$ et on voit que $x \in (\operatorname{dom} f)^\circ$ par le point 3. □

Corollaire 3.65 (optimalité III) *Si $f \in \mathrm{C}\overline{\mathrm{onv}}(\mathbb{E})$ et $0 \in (\operatorname{dom} f^*)^\circ$, alors* $\arg\min f$ *est non vide et compact.*

Démonstration. Lorsque $f \in \mathrm{C}\overline{\mathrm{onv}}(\mathbb{E})$, $\arg\min f = \partial f^*(0)$ (corollaire 3.61). D'après les points 2 et 4 de la proposition, lorsque $0 \in (\operatorname{dom} f^*)^\circ$, $\partial f^*(0)$ est non vide et compact. □

Nous avons défini le sous-différentiel à partir de la dérivée directionnelle. La proposition suivante montre que l'on peut retrouver les dérivées directionnelles à partir du sous-différentiel : $f'(x;\cdot)$ est la fonction d'appui de $\partial f(x)$.

Proposition 3.66 (formule du max) *Si $f \in \mathrm{Conv}(\mathbb{E})$ et $x \in (\operatorname{dom} f)^{\ominus}$, alors, pour tout $d \in \mathbb{E}$, on a*

$$f'(x;d) = \sup_{x^* \in \partial f(x)} \langle x^*, d\rangle. \tag{3.52}$$

Le supremum est atteint si $f'(x;d) < +\infty$.

Démonstration. En tant qu'élément de $\mathrm{C}\overline{\mathrm{onv}}(\mathbb{E})$ (proposition 3.16), $f'(x;\cdot)$ est l'enveloppe supérieure de ses minorantes affines (point 1 de la proposition 3.46). Comme $f'(x;\cdot)$ est positivement homogène de degré 1 (proposition 3.16), toute minorante affine est majorée par une minorante linéaire, si bien que $f'(x;\cdot)$ est aussi l'enveloppe supérieure de ses minorantes linéaires : pour tout $d \in \mathbb{E}$,

$$f'(x;d) = \sup_{\substack{x^* \in \mathbb{E} \\ \langle x^*, h\rangle \leqslant f'(x;h),\ \forall\, h \in \mathbb{E}}} \langle x^*, d\rangle = \sup_{x^* \in \partial f(x)} \langle x^*, d\rangle.$$

Montrons que l'on peut remplacer le supremum par un maximum, lorsque $d \in \mathbb{E}_0 := \operatorname{dom} f'(x;\cdot)$, qui est un sous-espace vectoriel lorsque $x \in (\operatorname{dom} f)^\ominus$ (proposition 3.16). On a en effet

$$f'(x;d) = \sup_{x^* \in \partial f(x)} \langle P_{\mathbb{E}_0} x^*, d\rangle = \sup_{x^* \in P_{\mathbb{E}_0}\partial f(x)} \langle x^*, d\rangle,$$

parce que $P_{\mathbb{E}_0}\partial f(x) \subseteq \partial f(x)$ (point 1 de la proposition 3.64). On utilise alors le fait que $P_{\mathbb{E}_0}\partial f(x)$ est compact lorsque $x \in (\operatorname{dom} f)^\ominus$ (points 2 et 3 de la proposition 3.64). □

Le résultat précédent ne tient plus si x est sur la frontière relative du domaine de f. Voici un contre-exemple : f est l'indicatrice de la boule-unité fermée de $\mathbb{R}^2$, pour la norme euclidienne, et $x = (-1, 0)$. Alors $f'(x;0) = 0$ et si $d \neq 0$:

$$f'(x;d) = \begin{cases} 0 & \text{si } d_1 > 0 \\ +\infty & \text{si } d_1 \leqslant 0. \end{cases}$$

Comme la fonction $\delta_x : d \mapsto f'(x;d)$ n'est pas fermée, elle ne peut être la fonction d'appui d'un ensemble (point 1 de la proposition 3.11) et donc certainement pas du sous-différentiel. D'ailleurs, ce dernier s'écrit $\partial f(x) = \{x^* \in \mathbb{R}^2 : x_1^* \leqslant 0,\ x_2^* = 0\}$ et

$$\sigma_{\partial f(x)}(d) = \begin{cases} 0 & \text{si } d_1 \geqslant 0 \\ +\infty & \text{si } d_1 < 0 \end{cases}$$

est l'enveloppe convexe fermée de δ_x. Cette propriété est tout à fait générale pour les fonctions de $\mathrm{C\overline{onv}}(\mathbb{E})$ (voir l'exercice 3.19).

On peut voir $\partial f : \mathbb{E} \multimap \mathbb{E} : x \mapsto \partial f(x)$ comme la multifonction qui à un élément de $\mathbb{E}$ fait correspondre le sous-différentiel $\partial f(x)$, qui est une partie de $\mathbb{E}$. En voici quelques propriétés.

Proposition 3.67 (la multifonction ∂f) *Si $f \in \mathrm{Conv}(\mathbb{E})$, alors*

(i) $(\operatorname{dom} f)^\ominus \subseteq \operatorname{dom} \partial f \subseteq \operatorname{dom} f$,
(ii) $\mathcal{R}(\partial f) \subseteq \operatorname{dom} f^*$,
(iii) $\mathcal{G}(\partial f)$ *est fermé,*
(iv) la multifonction ∂f est monotone.

Si $f \in \mathrm{C\overline{onv}}(\mathbb{E})$, alors

(v) $(\operatorname{dom} f^*)^\ominus \subseteq \mathcal{R}(\partial f)$,
(vi) $(\partial f)^{-1} = \partial f^*$,
(vii) la multifonction ∂f est monotone maximale.

Démonstration. [*(i)*] C'est une conséquence de la proposition 3.62 et de la définition 3.56 du sous-différentiel.

[*(ii)*] Si $x^* \in \partial f(x)$, alors $x \in \operatorname{dom} f$ (par convention) et $f^*(x^*) = \langle x^*, x\rangle - f(x) \in \mathbb{R}$ (par le point *(v)* de la proposition 3.55).

[*(iii)*] Soit $\{x_k, x_k^*\} \subseteq \mathcal{G}(\partial f)$, avec $x_k \to x$ et $x_k^* \to x^*$. Il faut montrer que $(x, x^*) \in \mathcal{G}(\partial f)$. Il suffit en fait de passer à la limite dans

$$\forall\, y \in \mathbb{E} : \quad f(y) \geqslant f(x_k) + \langle x_k^*, y - x_k \rangle.$$

[(iv)] Soient $x_1^* \in \partial f(x_1)$ et $x_2^* \in \partial f(x_2)$. En additionnant les deux inégalités $f(x_2) \geqslant f(x_1) + \langle x_1^*, x_2 - x_1 \rangle$ et $f(x_1) \geqslant f(x_2) + \langle x_2^*, x_1 - x_2 \rangle$, on obtient $\langle x_2^* - x_1^*, x_2 - x_1 \rangle \geqslant 0$, ce qui montre la monotonie de ∂f.

[(v)] Si $x^* \in (\operatorname{dom} f^*)^\ominus$, alors $x^* \in \operatorname{dom} \partial f^*$ (par le point (i)), si bien qu'il existe un $x \in \partial f^*(x^*)$ et par le point 2 du corollaire 3.58 (c'est ici que l'on a besoin d'avoir $f \in \overline{\operatorname{Conv}}(\mathbb{E})$), $x^* \in \partial f(x)$.

[(vi)] C'est le point 2 de la proposition 3.58.

[(vii)] Soit $(y, y^*) \in \mathbb{E}^2$ tel que

$$\langle y^* - x^*, y - x \rangle \geqslant 0, \quad \forall\, (x, x^*) \in \mathcal{G}(\partial f). \tag{3.53}$$

Il suffit de montrer que $(y, y^*) \in \mathcal{G}(\partial f)$. Le problème

$$\min_{x \in \mathbb{E}} \left(f(x) + \frac{1}{2} \|x - (y^* + y)\|^2 \right)$$

a une unique solution $\bar{x}$ (en effet, le critère tend vers l'infini à l'infini car f a une minorante affine ; il est aussi s.c.i. et strictement convexe). Celle-ci vérifie $y^* + y - \bar{x} \in \partial f(\bar{x})$. En prenant $(x, x^*) = (\bar{x}, y^* + y - \bar{x})$ dans (3.53), on trouve que $y = \bar{x}$ et donc que $y^* \in \partial f(\bar{x}) = \partial f(y)$. □

On note que l'on peut très bien avoir $\mathcal{R}(\partial f) \neq \operatorname{dom} f^*$: pour la fonction $x \mapsto f(x) = e^x$, $\mathcal{R}(\partial f) =]0, +\infty[$, alors que $\operatorname{dom} f^* = [0, +\infty[$.

C'est le caractère fermé de f qui rend ∂f monotone *maximale*. La théorie des *inéquations variationnelles* étudie les opérateurs (éventuellement multivoques) qui ne « dérivent pas d'un potentiel » (T n'est pas de la forme ∂f, pour une fonction $f \in \operatorname{Conv}(\mathbb{E})$). Cette théorie englobe donc l'optimisation. Lorsque les opérateurs considérés sont monotones, c'est leur caractère *maximal* qui leur donne une chance d'être surjectif, comme le caractère fermé de f donne une chance au problème $\inf_x (f(x) - \langle u, x \rangle)$, équivalent à $u \in \partial f(x)$, d'avoir une solution. On voit que cela ne suffit pas et qu'une hypothèse de coercivité (équivalente à la croissance de f vers l'infini à l'infini) peut être bien utile.

La proposition suivante établit un lien entre la sous-différentiabilité et la différentiabilité. Par différentiabilité, on entend soit la Gâteaux-différentiabilité, soit la Fréchet-différentiabilité, deux notions identiques pour une fonction convexe (proposition 3.18). Ce résultat peut être utilisé pour montrer qu'un fonction convexe est différentiable ; il suffit de montrer que son sous-différentiel est un singleton.

Proposition 3.68 (lien avec la différentiabilité) *Soient $f \in \operatorname{Conv}(\mathbb{E})$ et $x \in (\operatorname{dom} f)^\circ$. Si f est différentiable en x, alors f est sous-différentiable en x et $\partial f(x) = \{\nabla f(x)\}$. Inversement, si $\partial f(x)$ est le singleton $\{x^*\}$, alors f est différentiable en x et $\nabla f(x) = x^*$.*

Démonstration. Si f est différentiable en x, le point (ii) de la proposition 3.55 est vérifiée avec $x^* = \nabla f(x)$; donc $\nabla f(x) \in \partial f(x)$ et f est sous-différentiable. D'autre

part, si $x_0^* \in \partial f(x)$, le point (i) de la proposition 3.55 montre que $f'(x) \cdot d \geqslant \langle x_0^*, d\rangle$, $\forall\, d \in \mathbb{E}$. La linéarité de $f'(x)$ implique que $f'(x) \cdot d = \langle x_0^*, d\rangle$, $\forall\, d \in \mathbb{E}$ et donc que $x_0^* = \nabla f(x)$. On en déduit que $\partial f(x) = \{\nabla f(x)\}$.

Si $\partial f(x)$ est le singleton $\{x^*\}$, la formule du max (3.52) donne pour tout $d \in \mathbb{E}$: $f'(x; d) = \langle x^*, d\rangle$. Donc f est (Gâteaux-)différentiable en x et $\nabla f(x) = x^*$. □

Soit f une fonction convexe différentiable. Alors f' est lipschitzienne si, et seulement si, f^* est fortement convexe. C'est ce que montre la proposition suivante [317], qui nous sera utile dans l'analyse des algorithmes quasi-newtoniens (chapitre 11).

Proposition 3.69 (fonction convexe avec gradient lipschitzien) *Soit $f : \mathbb{E} \to \mathbb{R}$ une fonction convexe et $\mathcal{C}^{1,1}$. On note ∇f le gradient de f pour le produit scalaire $\langle \cdot, \cdot \rangle$ et $L > 0$ la constante de Lipschitz de ∇f pour la norme $\|\cdot\|$ associée à ce produit scalaire. Alors, pour tout $x, y \in \mathbb{E}$, on a*

$$\langle \nabla f(y) - \nabla f(x), y - x\rangle \geqslant \frac{1}{L}\|\nabla f(y) - \nabla f(x)\|^2.$$

Démonstration. Soient $x, y \in \mathbb{E}$. On note $x^* = \nabla f(x)$ et $y^* = \nabla f(y)$. Alors $x \in \partial f^*(x^*)$ et $y \in \partial f^*(y^*)$. Il suffit de montrer la forte convexité de f^* (proposition 3.24), c'est-à-dire

$$f^*(y^*) \geqslant f^*(x^*) + \langle y^* - x^*, x\rangle + \frac{1}{2L}\|y^* - x^*\|^2.$$

Par définition de f^*, on a

$$f^*(y^*) = \sup_{y \in \mathbb{E}} \Big(\langle y^*, y\rangle - f(y) \Big),$$

si bien que pour obtenir l'inégalité ci-dessus, il est utile de trouver une majoration de $f(y)$. En utilisant le développement de Taylor avec reste intégral, la constante de Lipschitz de ∇f et le point (v) de la proposition 3.55, on obtient

$$\begin{aligned} f(y) &= f(x) + \langle \nabla f(x), y - x\rangle + \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x\rangle \, \mathrm{d}t \\ &\leqslant f(x) + \langle x^*, y - x\rangle + \frac{L}{2}\|y - x\|^2 \\ &= -f^*(x^*) + \langle x^*, y\rangle + \frac{L}{2}\|y - x\|^2. \end{aligned}$$

Alors

$$f^*(y^*) \geqslant f^*(x^*) + \sup_{y \in \mathbb{E}} \Big(\langle y^* - x^*, y\rangle - \frac{L}{2}\|y - x\|^2 \Big).$$

La borne supérieure du membre de droite est atteinte pour $y = x + \frac{1}{L}(y^* - x^*)$. Après substitution, on obtient l'inégalité recherchée. □

3.6.3 Règles de calcul

Quelques règles du calcul sous-différentiel sont données dans les propositions suivantes (on en trouvera d'autres dans [317 ; 1993 ; section VI.4]). Les fonctions considérées sont parfois supposées ne prendre que des valeurs finies, mais on trouvera au corollaire 14.15 une extension de certaines règles de calcul au cas de fonctions pouvant prendre la valeur $+\infty$.

Combinaison conique

Proposition 3.70 (multiplication par un scalaire positif) *Soit $\alpha \geqslant 0$, $f \in \mathrm{Conv}(\mathbb{E})$ et $x \in \mathbb{E}$. Alors*

$$\partial(\alpha f)(x) = \alpha\, \partial f(x). \tag{3.54}$$

Démonstration. C'est une conséquence immédiate de la définition du sous-différentiel. □

Proposition 3.71 (somme) *Soient $f_1, \ldots, f_p \in \mathrm{Conv}(\mathbb{E})$ et $x \in \mathbb{E}$. Alors*

$$\partial(f_1 + \cdots + f_p)(x) \supseteq \partial f_1(x) + \cdots + \partial f_p(x). \tag{3.55}$$

L'égalité a lieu ci-dessus si

$$\bigcap_{1 \leqslant i \leqslant p} (\mathrm{dom}\, f_i)^{\circ\circ} \neq \varnothing. \tag{3.56}$$

Dans la condition (3.56), on peut remplacer $(\mathrm{dom}\, f_i)^{\circ\circ}$ par $\mathrm{dom}\, f_i$ si f_i est polyédrique.

Démonstration. On note $f = f_1 + \cdots + f_p$.

L'inclusion (3.55) se déduit directement de la définition du sous-différentiel: si $x_i^* \in \partial f_i(x)$ pour $i = 1, \ldots, p$, alors $x \in \cap_i (\mathrm{dom}\, f_i) = \mathrm{dom}\, f$ et, pour tout $h \in \mathbb{E}$, $f'(x;h) = f_1'(x;h) + \cdots + f_p'(x;h) \geqslant \langle x_1^* + \cdots + x_p^*, h\rangle$, si bien que $x_1^* + \cdots + x_p^* \in \partial f(x)$.

Supposons maintenant que (3.56) a lieu et que $x^* \in \partial f(x)$; donc $x \in \mathrm{dom}\, f$. Il s'agit de décomposer x^* en des sous-gradients des f_i. On passe par la conjuguée de f donnée par la proposition 3.54, point 2 : puisque $f^*(x^*)$ est fini, on peut écrire

$$f^*(x^*) = (f_1^* \boxplus \cdots \boxplus f_p^*)(x^*) = f_1^*(x_1^*) + \cdots + f_p^*(x_p^*), \quad \text{avec } x^* = x_1^* + \cdots + x_p^*.$$

Alors, d'après la définition du sous-différentiel (point (v) de la proposition 3.55):

$$\langle x_1^* + \cdots + x_2^*, x\rangle = f_1(x) + \cdots + f_p(x) + f_1^*(x_1^*) + \cdots + f_p^*(x_p^*).$$

Mais on a toujours $f_2(x)+\cdots+f_p(x)+f_2^*(x_2^*)+\cdots+f_p^*(x_p^*)\geqslant\langle x_2^*+\cdots+x_p^*,x\rangle$, si bien que l'identité ci-dessus donne $\langle x_1^*,x\rangle\geqslant f_1(x)+f_1^*(x_1^*)$. Comme l'inégalité inverse a toujours lieu, on obtient $\langle x_1^*,x\rangle=f_1(x)+f_1^*(x_1^*)$, c'est-à-dire que $x_1^*\in\partial f_1(x_1)$. De même pour les autres indices : $x_i^*\in\partial f_i(x_i)$, pour tout i. Donc $x^*\in\partial f_1(x)+\cdots+\partial f_p(x)$. □

On ne peut pas se passer de la condition (3.56) pour avoir égalité en (3.55). Par exemple, si f_1 est la fonction $-\sqrt{x}$ définie en (3.9) et $f_2=\mathcal{I}_{]-\infty,0]}$, la somme $f=f_1+f_2$ a son domaine réduit à $\{0\}$ et $\partial f(0)=\mathbb{R}$, alors que $\partial f_1(0)+\partial f_2(0)=\varnothing$, parce que $\partial f_1(0)=\varnothing$.

Composition

Le cadre est le suivant. On dispose d'une fonction affine $a:\mathbb{E}\to\mathbb{F}$ entre deux espaces euclidiens $\mathbb{E}$ et $\mathbb{F}$. Celle-ci est supposée être définie en $x\in\mathbb{E}$ par

$$a(x)=Ax+b,$$

où $A:\mathbb{E}\to\mathbb{F}$ est linéaire et $b\in\mathbb{F}$. On note A^* l'application linéaire adjointe de A pour les produits scalaires que l'on s'est donné sur $\mathbb{E}$ et $\mathbb{F}$. L'application affine a est composée avec une application $g:\mathbb{F}\to\overline{\mathbb{R}}$.

Proposition 3.72 (pré-composition par une fonction affine) *Dans le cadre défini ci-dessus, si $g\in\mathrm{Conv}(\mathbb{F})$, alors pour tout $x\in\mathbb{E}$:*

$$\partial(g\circ a)(x)\supseteq A^*\Big(\partial g(a(x))\Big),\tag{3.57}$$

avec égalité si l'une des conditions suivantes est vérifiée :
- $\mathcal{R}(a)\cap(\operatorname{dom}g)^\ominus\neq\varnothing$,
- $\mathcal{R}(a)\cap\operatorname{dom}g\neq\varnothing$ *et g est aussi polyédrique.*

Démonstration. Démontrons (3.57). Soit $y^*\in\partial g(a(x))$, supposé non vide. Alors, pour tout $x'\in\mathbb{E}$, on a $(g\circ a)(x')=g(a(x'))\geqslant g(a(x))+\langle y^*,A(x'-x)\rangle=(g\circ a)(x)+\langle A^*y^*,x'-x\rangle$, ce qui montre que $A^*y^*\in\partial(g\circ a)(x)$.

Démontrons l'égalité en (3.57) lorsque $\mathcal{R}(A)\cap(\operatorname{dom}g)^\ominus\neq\varnothing$ [resp. lorsque g est convexe polyédrique et $\mathcal{R}(A)\cap(\operatorname{dom}g)\neq\varnothing$]. Dans ce cas, $g\circ a\in\mathrm{Conv}(\mathbb{E})$ (proposition 3.32). Soit $x^*\in\partial(g\circ a)(x)$ avec $x\in\operatorname{dom}(g\circ a)$, ce qui s'écrit (proposition 3.55)

$$(g\circ a)^*(x^*)+(g\circ a)(x)=\langle x^*,x\rangle.$$

En introduisant $g_0\in\mathrm{Conv}(\mathbb{F})$ par $g_0(y)=g(y+b)$, on peut écrire $g\circ a=g_0\circ A$. La relation ci-dessus implique que $(g_0\circ A)^*(x^*)$ est fini. D'autre part, $\mathcal{R}(A)\cap(\operatorname{dom}g_0)^\ominus=(\mathcal{R}(a)-b)\cap(\operatorname{dom}g-b)^\ominus=(\mathcal{R}(a)\cap(\operatorname{dom}g)^\ominus)-b$, qui est donc non vide [on trouve que $\mathcal{R}(A)\cap\operatorname{dom}g_0$ est non vide dans le cas polyédrique]. D'après la proposition 3.52, il existe alors un $y^*\in\mathbb{F}$ tel que $A^*y^*=x^*$ et $(g_0\circ A)^*(x^*)=g_0^*(y^*)$. On en déduit que $(g\circ a)^*(x^*)=g_0^*(y^*)=g^*(y^*)-\langle y^*,b\rangle$. La relation ci-dessus devient

$$g^*(y^*)+g(a(x))=\langle A^*y^*,x\rangle+\langle y^*,b\rangle=\langle y^*,a(x)\rangle.$$

Elle exprime que $y^* \in \partial g(a(x))$ (proposition 3.55) ; donc $x^* \in A^*(\partial g(a(x)))$. □

La démonstration du corollaire suivant est proposée à l'exercice 3.27. Le résultat doit être différencié du point 1 du théorème 3.64.

Corollaire 3.73 (restriction à un sous-espace affine) *Soient $\mathbb{E}_0$ un sous-espace vectoriel de $\mathbb{E}$, $\mathrm{P}_{\mathbb{E}_0}$ le projecteur orthogonal de $\mathbb{E}$ sur $\mathbb{E}_0$, $x_0 \in \mathbb{E}$ et $f \in \mathrm{Conv}(\mathbb{E})$. On note $f|_{x_0+\mathbb{E}_0} : x \in \mathbb{E}_0 \mapsto f(x_0+x)$ la restriction de f à $x_0+\mathbb{E}_0$. Alors en $x \in \mathbb{E}_0$, on a*

$$\partial(f|_{x_0+\mathbb{E}_0})(x) \supseteq \mathrm{P}_{\mathbb{E}_0}\Big(\partial f(x_0+x)\Big), \tag{3.58}$$

avec égalité si l'une des conditions suivantes est vérifiée :
- $(x_0+\mathbb{E}_0) \cap (\mathrm{dom}\, f)^\circ \neq \varnothing$,
- $(x_0+\mathbb{E}_0) \cap (\mathrm{dom}\, f) \neq \varnothing$ *et f est aussi polyédrique.*

Replaçons-nous dans le cadre défini à la section 3.4.1. Soient $\mathbb{E}$ un espace vectoriel, $F : \mathbb{E} \to (\mathbb{R} \cup \{+\infty\})^m$ une fonction convexe (composante par composante) et $g : \mathbb{R}^m \to \overline{\mathbb{R}}$ une fonction convexe *croissante* (pour l'ordre habituel de $\mathbb{R}^m$). On sait qu'alors la fonction composée $(g \circ F)$ définie en $x \in \mathbb{E}$ par

$$(g \circ F)(x) = \begin{cases} g(F(x)) & \text{si } F(x) \in \mathbb{R}^m \\ +\infty & \text{sinon.} \end{cases}$$

est convexe (proposition 3.33). On s'intéresse ici à sa sous-différentiabilité. Pour rendre les notations plus compactes, on écrira

$$\partial F(x) := \partial F_1(x) \times \cdots \times \partial F_m(x) \subseteq \mathbb{E}^m.$$

Rappelons que si F et g sont à valeurs finies et différentiables, on a

$$(g \circ F)'(x) = g'(x) \circ F'(x).$$

Le résultat suivant généralise cette identité.

Proposition 3.74 (post-composition par une fonction convexe croissante) *Dans le cadre défini ci-dessus, le sous-différentiel de $g \circ F$ en $x \in \mathbb{E}$ est donné par*

$$\partial(g \circ F)(x) = \left\{ \sum_{i=1}^m \gamma_i x_i^* : \gamma \in \partial g(F(x)),\ x_i^* \in \partial F_i(x),\ \textit{for } i \in [1:n] \right\}.$$

Démonstration. □

Ce résultat a de nombreuses conséquences pratiques. On en déduit par exemple que le carré (ou la puissance $p \geqslant 2$) d'une fonction convexe *positive* est convexe.

Fonction marginale

Rappelons le cadre introduit à la section 3.4.3. On dispose de deux espaces euclidiens $\mathbb{E}$ et $\mathbb{F}$ et d'une fonction $\varphi : \mathbb{E} \times \mathbb{F} \to \overline{\mathbb{R}}$. La fonction marginale $f : \mathbb{E} \to \overline{\mathbb{R}}$ associée à φ est définie en $x \in \mathbb{E}$ par

$$f(x) = \inf_{y \in \mathbb{F}} \varphi(x, y). \tag{3.59}$$

Le sous-différentiel de f dépend de celui de φ qui est supposé calculé pour le produit scalaire produit (A.10) sur $\mathbb{E} \times \mathbb{F}$.

Proposition 3.75 (fonction marginale) *Dans le cadre défini ci-dessus, supposons que $\varphi \in \mathrm{Conv}(\mathbb{E} \times \mathbb{F})$ et que $f \in \mathrm{Conv}(\mathbb{E})$. Si $x \in \mathrm{dom}\, f$ et $f(x) = \varphi(x, y_x)$ pour un certain $y_x \in \mathbb{F}$ (i.e., l'infimum en (3.59) est atteint en y_x), alors*

$$\partial f(x) = \{x^* : (x^*, 0) \in \partial\varphi(x, y_x)\}.$$

Démonstration. On a les équivalences suivantes :

$$\begin{aligned}
& x^* \in \partial f(x) \\
& \iff f(x') \geqslant f(x) + \langle x^*, x' - x\rangle, \quad \forall\, x' \in \mathbb{E} \qquad \text{[proposition 3.55 } (ii)\text{]} \\
& \iff \varphi(x', y') \geqslant \varphi(x, y_x) + \langle (x^*, 0), (x', y') - (x, y_x)\rangle, \quad \forall\, (x', y') \in \mathbb{E} \times \mathbb{F} \qquad (3.60) \\
& \iff (x^*, 0) \in \partial\varphi(x, y_x) \qquad \text{[proposition 3.55 } (ii)\text{]}.
\end{aligned}$$

L'implication directe $\boxed{\Rightarrow}$ dans (3.60) vient de l'inégalité $f(x') \leqslant \varphi(x', y')$ qui est vérifiée pour tout $y' \in \mathbb{F}$ et de l'égalité $f(x) = \varphi(x, y_x)$. L'implication réciproque $\boxed{\Leftarrow}$ dans (3.60) vient de ce que le membre de droite est indépendant de y', si bien que l'on peut donc prendre la borne inférieure en y' dans le membre de gauche sans modifier l'inégalité. □

Remarques 3.76 1) Il faut bien noter que, si la borne inférieure $\inf_y \varphi(x, y)$ est atteinte en plusieurs y_x, $\{x^* : (x^*, 0) \in \partial\varphi(x, y_x)\}$ ne dépend pas du minimiseur y_x choisi. C'est une conséquence de la démonstration.

On a un autre éclairage sur cette indépendance par rapport à y_x en observant que φ est constante sur l'ensemble $M(x) := \{(x, y_x) : y_x \text{ minimise } \varphi(x, \cdot)\}$, si bien que $\partial\varphi$ est aussi constant sur l'*intérieur relatif* de $M(x)$ (exercice 3.38). Cependant $\partial\varphi(x, y_x)$ peut varier lorsque (x, y_x) passe de l'intérieur relatif de $M(x)$ à sa frontière relative. C'est le cas de la fonction définie par $\varphi(x, y) = \max(0, |y|-1)$, dont la fonction marginale est nulle :

$$M(x) = \{x\} \times [-1, 1] \qquad \text{et} \qquad \partial\varphi(x, y_x) = \begin{cases} \{0\} \times [-1, 0] & \text{si } y = -1 \\ \{(0, 0)\} & \text{si } -1 < y < 1 \\ \{0\} \times [0, 1] & \text{si } y = 1. \end{cases}$$

Malgré cela, quel que soit le y_x choisi dans $[-1, 1]$, le sous différentiel de f en x est toujours réduit à $\{0\}$.

2) D'autre part, si φ est différentiable en (x, y_x), où y_x est un minimiseur quelconque de $\varphi(x,\cdot)$, alors f est également différentiable en x (car son sous-différentiel est un singleton) et on a
$$\nabla f(x) = \nabla_x \varphi(x, y_x).$$
C'est comme s'il y avait un minimiseur unique $y_x := \eta(x)$, fonction différentiable de x, que l'on écrivait $f(x) = \varphi(x, \eta(x))$ et que l'on calculait $\nabla f(x)$ par une dérivation en chaîne :
$$\nabla f(x) = \nabla_x \varphi(x, \eta(x)) + \eta'(x)^* \nabla_y \varphi(x, \eta(x)).$$
On retrouverait le résultat ci-dessus en observant que $\nabla_y \varphi(x, \eta(x)) = 0$ car $\eta(x)$ minimise $\varphi(x,\cdot)$.

3) Le fait que $\varphi(x,\cdot)$ ait un minimum unique n'implique nullement la différentiabilité de la fonction marginale en x. Par exemple, f est la fonction marginale de φ définie par $\varphi(x, y_x) = f(x) + y_x^2$. Cette dernière a un minimum $y_x = 0$ unique en y quel que soit x, alors que f peut ne pas être différentiable.

Enveloppe supérieure

Considérons à présent le cas de l'enveloppe supérieure
$$f := \sup_{i \in I} f_i$$
de fonctions convexes $f_i : \mathbb{E} \to \overline{\mathbb{R}}$, pour i dans un ensemble quelconque I, dont les éléments sont appelés indices ci-dessous. Rappelons de la section 3.4.2, qu'elle est définie en $x \in \mathbb{E}$ par
$$f(x) = \sup_{i \in I} \Big(f_i(x) \Big).$$
On aura besoin de désigner l'ensemble des indices des fonctions f_i valant $f(x)$ en $x \in \mathbb{E}$:
$$I^0(x) := \{ i \in I : f(x) = f_i(x) \}.$$

Proposition 3.77 (enveloppe supérieure I) *Dans le cadre défini ci-dessus, si les f_i sont convexes et $f \in \mathrm{Conv}(\mathbb{E})$, alors pour tout $x \in \mathrm{dom}\, f$, on a*
$$\partial f(x) \supseteq \overline{\mathrm{co}} \left(\bigcup_{i \in I^0(x)} \partial f_i(x) \right). \tag{3.61}$$

Démonstration. Soient $x \in \mathrm{dom}\, f$, $i \in I^0(x)$ et $x_i^* \in \partial f_i(x)$. Comme un sous-différentiel est un ensemble convexe fermé (proposition 3.64), il suffit de montrer que $x_i^* \in \partial f(x)$. Quel que soit $y \in \mathbb{E}$, on a

$$\begin{aligned} f(y) &\geqslant f_i(y) \quad [\text{par définition de } f] \\ &\geqslant f_i(x) + \langle x_i^*, y - x\rangle \quad [\text{car } x_i^* \in \partial f_i(x)] \\ &= f(x) + \langle x_i^*, y - x\rangle \quad [\text{car } i \in I^0(x)]. \end{aligned}$$

Ceci montre que $x_i^* \in \partial f(x)$ (point (ii) de la proposition 3.55). □

La proposition 3.77 montre en particulier, que $\partial f(x) \neq \varnothing$, si l'un des $\partial f_i(x) \neq \varnothing$, avec $i \in I^0(x)$. Il se pourrait cependant que $I^0(x) = \varnothing$ ou que $\partial f_i(x) = \varnothing$ pour tout $i \in I^0(x)$, auxquels cas cette proposition ne donne pas d'information. Ce dernier cas se présente, par exemple, lorsque les fonctions f_i sont obtenues en divisant par $i \in \mathbb{N}^*$ (ou en mulitpliant par $i \in]0,1]$) la fonction définie en (3.9) (l'opposée de la racine carrée) : $I^0(0) = \mathbb{N}^*$, mais tous les $\partial f_i(0) = \varnothing$, alors que f est l'indicatrice de $\mathbb{R}_+$ et a donc un sous-différentiel $\partial f(0) =]-\infty, 0[$ non vide.

L'identification de conditions pour avoir l'égalité en (3.61) est un problème délicat. Un des résultats les plus simples est le suivant. Pour des résultats plus fins, on pourra consulter la section VI.4.4 dans [317 ; 1993].

Proposition 3.78 (enveloppe supérieure II) *Dans le cadre défini ci-dessus, supposons que*

- *les f_i sont convexes,*
- *I est compact,*
- *$\forall\, x \in \operatorname{dom} f$, $i \mapsto f_i(x)$ est semi-continue supérieurement,*

alors pour tout $x \in (\operatorname{dom} f)^\circ$:

$$\partial f(x) = \operatorname{co}\left(\bigcup_{i \in I^0(x)} \partial f_i(x)\right). \tag{3.62}$$

Démonstration. □

3.7 Proximalité

3.7.1 Opérateur proximal

Soient $\mathbb{E}$ un espace euclidien (produit scalaire $\langle\cdot,\cdot\rangle$ et norme associée $\|\cdot\|$) et $f \in \overline{\operatorname{Conv}}(\mathbb{E})$. Pour x donné dans $\mathbb{E}$, on considère le problème

$$\inf_{y \in \mathbb{E}} \left(f(y) + \frac{1}{2}\|y - x\|^2\right). \tag{3.63}$$

Il s'agit donc de l'inf-convolution de f et $\frac{1}{2}\|\cdot\|^2$, mais il n'est pas très important de le savoir. Il sera utile d'introduire une notation pour le critère de ce problème, qui dépend de $x \in \mathbb{E}$:

$$\varphi_x(y) := f(y) + \frac{1}{2}\|y - x\|^2.$$

D'après la proposition 3.7, f a une minorante affine, si bien que $\varphi_x(y) \to \infty$ lorsque $\|y\| \to \infty$; de plus φ_x est s.c.i. et strictement convexe. Dès lors le problème ci-dessus a une solution unique. On la note

$$x_p \text{ solution unique de (3.80)}$$

et on l'appelle le *point proximal* de x (celui-ci dépend de f et du produit scalaire utilisé). Il est clair que le point proximal $x_p \in \operatorname{dom} f$. On appelle *opérateur proximal*, l'application

$$P_f : \mathbb{E} \to \mathbb{E} : x \mapsto x_p.$$

Le point proximal est caractérisé par la condition d'optimalité de (3.80) : $0 \in \partial\varphi_x(x_p) = \partial f(x_p) + x_p - x$. Dès lors

$$\tilde{x} = P_f(x) \quad \Longleftrightarrow \quad \exists\, \tilde{g} \in \partial f(\tilde{x}) : \tilde{x} = x - \tilde{g}. \tag{3.64}$$

On note que l'unicité de x_p implique celle de

$$g_p := x - x_p. \tag{3.65}$$

On verra à la proposition 3.79 que ce sous-gradient n'est autre que $P_{f^*}(x)$. L'algorithme qui, en x considéré comme l'itéré courant, choisit x_p comme itéré suivant est appelé l'algorithme proximal. Nous l'étudierons à la section 7.2.

On trouvera à la figure 3.9 une illustration du concept de point proximal dans

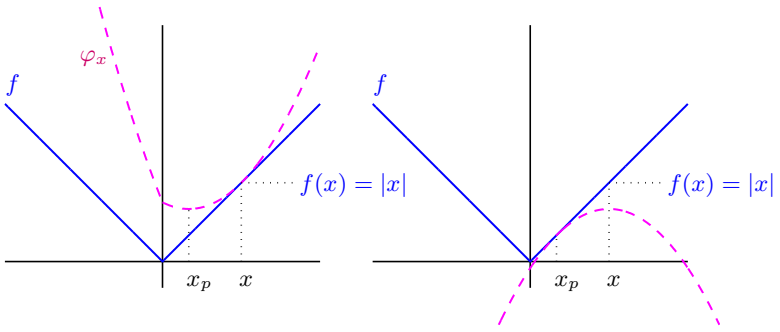


Fig. 3.9. Deux points de vue sur le point proximal associé à la fonction $f = |\cdot|$

le cas où $\mathbb{E} = \mathbb{R}$ et f est la valeur absolue. Le graphique de gauche représente une première manière de trouver le point proximal x_p, vu comme solution de (3.80). On y voit la fonction f et la fonction φ_x (en tirets) obtenue en ajoutant à f une fonction quadratique. Le point x_p minimise φ_x. La seconde manière de trouver le point proximal de x utilise l'équation d'optimalité (3.64) : la fonction quadratique concave $y \mapsto f(x_p) + \frac{1}{2}\|x_p - x\|^2 - \frac{1}{2}\|y - x\|^2$, représentée en tirets, minore f et a en $y = x_p$ la même valeur que f et un gradient valant $x - x_p$, qui est donc égal au sous-gradient

g_p de f en x_p. Les deux graphes sont donc tangents et x_p est l'abscisse de l'unique point de contact entre ces deux graphes [54].

On vérifie facilement que si f est l'indicatrice $\mathcal{I}_X$ d'une partie convexe fermée non vide X de $\mathbb{E}$, P_f est la projection sur X. On peut donc voir l'opérateur proximal comme une généralisation de ce concept. Comme le montre la proposition suivante, il a d'ailleurs des propriétés analogues à celle d'une projection (comparez avec la proposition 2.26) : P_f est *monotone* et *contractante*. La proposition affirme aussi la contractilité de l'opérateur

$$R_f = P_f - P_{f^*}.$$

Proposition 3.79 (opérateur proximal) *Pour une fonction $f \in \overline{\mathrm{Conv}}(\mathbb{E})$, les propriétés suivantes ont lieu :*
(i) $I - P_f = P_{f^*}$,
(ii) pour tout x et $y \in \mathbb{E}$,

$$\langle P_f(y) - P_f(x), y - x\rangle \geqslant \|P_f(y) - P_f(x)\|^2, \tag{3.66}$$

$$\|P_f(y) - P_f(x)\|^2 + \|P_{f^*}(y) - P_{f^*}(x)\|^2 \leqslant \|y - x\|^2, \tag{3.67}$$

$$\|R_f(y) - R_f(x)\| \leqslant \|y - x\|. \tag{3.68}$$

Démonstration. [(i)] On a $(I - P_f)(x) = x - x_p = g_p$, avec $g_p \in \partial f(x_p)$. D'après le corollaire 3.58 et $f \in \overline{\mathrm{Conv}}(\mathbb{E})$, on a $x_p \in \partial f^*(g_p)$. Alors, l'égalité $g_p = x - x_p$ et la caractérisation (3.64) montrent que $g_p = P_{f^*}(x)$.

[(ii)-(3.66)] On a $x_p = x - g_p^x$, avec $g_p^x \in \partial f(x_p)$, et $y_p = y - g_p^y$, avec $g_p^y \in \partial f(y_p)$. En retranchant

$$y - x = y_p - x_p + (g_p^y - g_p^x). \tag{3.69}$$

En multipliant scalairement cette identité par $(y_p - x_p)$ et en utilisant la monotonie du sous-différentiel (proposition 3.67), on trouve

$$\langle y_p - x_p, y - x\rangle = \|y_p - x_p\|^2 + \underbrace{\langle g_p^y - g_p^x, y_p - x_p\rangle}_{\geqslant 0} \geqslant \|y_p - x_p\|^2.$$

[(ii)-(3.67)] L'inégalité s'obtient en élevant (3.69) au carré et en tenant compte de la monotonie du sous-différentiel.

[(ii)-(3.68)] En utilisant à nouveau la monotonie du sous-différentiel, on a

$$\begin{aligned}
\|R_f(y) - R_f(x)\|^2 &= \|(y_p - g_p^y) - (x_p - g_p^x)\|^2 \\
&= \|y_p - x_p\|^2 - 2\underbrace{\langle g_p^y - g_p^x, y_p - x_p\rangle}_{\geqslant 0} + \|g_p^y - g_p^x\|^2 \\
&\leqslant \|y_p - x_p\|^2 + \|g_p^y - g_p^x\|^2 \\
&\leqslant \|y - x\|^2 \quad [(3.67)].
\end{aligned}$$

□

On peut appliquer le résultat précédent en prenant $f^* \in \mathrm{C}\overline{\mathrm{onv}}(\mathbb{E})$ au lieu de f. Le résultat (ii) s'écrit alors avec $g_p^x = P_{f^*}(x) \in \partial f(x_p)$ et $g_p^y = P_{f^*}(y) \in \partial f(y_p)$:

$$\langle g_p^y - g_p^x, y - x\rangle \geqslant \|g_p^y - g_p^x\|^2,$$

qui n'est pas sans rappeler la proposition 3.69. Il y a deux différences toutefois : il n'y a pas de facteur $1/L$ et g_p^x n'est pas un sous-gradient en x mais en x_p.

3.7.2 Régularisée de Moreau-Yosida

Soient toujours $\mathbb{E}$ un espace euclidien (produit scalaire $\langle\cdot,\cdot\rangle$ et norme associée $\|\cdot\|$) et $f \in \mathrm{C}\overline{\mathrm{onv}}(\mathbb{E})$. La *régularisée de Moreau-Yosida* de f est la fonction $\tilde{f} : \mathbb{E} \to \mathbb{R}$ définie par

$$\tilde{f}(x) = \inf_{y\in\mathbb{E}} \left(f(y) + \frac{1}{2}\|y - x\|^2 \right) = f(x_p) + \frac{1}{2}\|x_p - x\|^2, \tag{3.70}$$

où x_p est le point proximal de x. On observe que $\tilde{f}$ est bien à valeurs réelles car $x_p \in \operatorname{dom} f$ et $f(x_p) > -\infty$ (f ne prend pas la valeur $-\infty$).

Cette notion est illustrée en dimension 1 à la figure 3.10 lorsque f est la valeur

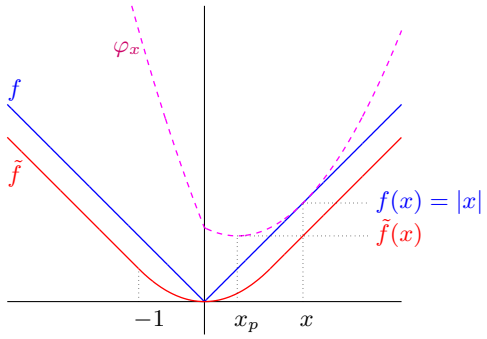


Fig. 3.10. Régularisée de Moreau-Yosida $\tilde{f}$ de $f : x \mapsto |x|$

absolue et $\|\cdot\| = |\cdot|$. La ligne en tirets représente la fonction φ_x dont on calcule le minimum pour obtenir $\tilde{f}(x)$.

La proposition suivante donne quelques propriétés intéressantes de la régularisée de Moreau-Yosida. En particulier, elle montre que $\tilde{f}$ est toujours $\mathcal{C}^{1,1}$ et régularise donc f si celle-ci n'est pas différentiable. Cette régularisation se fait sans changer les minima de f, ni sa valeur minimale.

Proposition 3.80 (régularisée de Moreau-Yosida) *Soit $f \in \mathrm{C}\overline{\mathrm{onv}}(\mathbb{E})$. Alors*

(*i*) $f \circ P_f \leqslant \tilde{f} \leqslant f$.
(*ii*) $\inf \tilde{f} = \inf f$.
(*iii*) $\arg\min \tilde{f} = \arg\min f = \{x \in \mathbb{E} : \tilde{f}(x) = f(x)\} = \{x \in \mathbb{E} : x = x_p\}$.
(*iv*) $\tilde{f} \in \mathrm{C}\overline{\mathrm{onv}}(\mathbb{E})$ *et* $\tilde{f}$ *est* $\mathcal{C}^{1,1}$ *avec* $\nabla \tilde{f}(x) = x - x_p$.

Démonstration. [(*i*)] La première inégalité se déduit du fait que $\tilde{f}(x) = f(x_p) + \frac{1}{2}\|x_p - x\|^2 \geqslant f(x_p)$. Pour la seconde, on prend $y = x$ comme argument de l'infimum définissant $\tilde{f}$.

[(*ii*)] Du point (*i*), on a $\inf \tilde{f} \leqslant \inf f$. Par le même point (*i*), on a $\tilde{f}(x) \geqslant f(x_p) \geqslant \inf f$; donc $\inf \tilde{f} \geqslant \inf f$.

[(*iii*)] Montrons d'abord les équivalences :

$$f(x) = \tilde{f}(x) \quad \Longleftrightarrow \quad x = x_p \quad \Longleftrightarrow \quad x \in \arg\min f. \tag{3.71}$$

Si $f(x) = \tilde{f}(x)$, l'infimum dans la définition de $\tilde{f}$ est atteint pour $y = x$, donc $x_p = x$; réciproquement, si $x_p = x$, l'expression (3.70) de $\tilde{f}$ montre que $f(x) = \tilde{f}(x)$. Considérons à présent la seconde équivalence. Si $x_p = x$, $g_p = x - x_p = 0$ et donc $0 \in \partial f(x)$ (équivalence (3.64)), ce signifie que $x \in \arg\min f$; réciproquement, $0 \in \partial f(x)$ et $x = x - 0$ impliquent que $x_p = x$ (par l'équivalence (3.64)).

Il reste à montrer que

$$\arg\min f = \arg\min \tilde{f}.$$

Si $x \in \arg\min f$, on a par (3.71) puis le point (*ii*) que $\tilde{f}(x) = f(x) = \inf f = \inf \tilde{f}$, donc $x \in \arg\min \tilde{f}$. Inversement, si $x \in \arg\min \tilde{f}$, on a par les points (*i*) et (*ii*) : $f(x_p) \leqslant \tilde{f}(x) = \inf \tilde{f} = \inf f \leqslant f(x_p)$, si bien que l'on a en fait égalité partout. Alors $\tilde{f}(x) = f(x_p)$ implique que $x_p = x$ (voir la formule de $\tilde{f}$) et donc $x \in \arg\min f$ par 3.71.

[(*iv*)] La régularisée $\tilde{f}$ est la fonction marginale de $(x, y) \mapsto f(y) + \frac{1}{2}\|x - y\|^2$, qui est dans $\mathrm{Conv}(\mathbb{E} \times \mathbb{E})$. Cette dernière est bornée inférieurement en y à x fixé ; donc la régularisée est dans $\mathrm{Conv}(\mathbb{E})$ (proposition 3.35) et même dans $\mathrm{C}\overline{\mathrm{onv}}(\mathbb{E})$ car elle est partout finie (proposition 3.14). D'après la proposition 3.75 :

$$s \in \partial \tilde{f}(x) \quad \Longleftrightarrow \quad \begin{pmatrix} s \\ 0 \end{pmatrix} = \begin{pmatrix} x - x_p \\ g_p + x_p - x \end{pmatrix},$$

pour un $g_p \in \partial f(x_p)$. Dès lors, $\partial \tilde{f}(x)$ est le singleton $\{x - x_p\}$, ce qui implique la différentiabilité de $\tilde{f}$, avec $\nabla \tilde{f}(x) = x - x_p$ (proposition 3.68). Comme $x - x_p = P_{f^*}(x)$, la caractère lipschitzien de $\nabla \tilde{f}$ se déduit de celui de P_{f^*} (proposition 3.79, formule (3.67)). □

3.8 Point-selle et convexité-concavité

3.8.1 Point-selle

La notion de point-selle (n dit aussi point-col) est très générale, comme l'est celle d'un minimum d'une fonction. Elle est attachée à des fonctions définies sur un produit d'ensembles n'ayant, a priori, pas de structure particulière.

Définitions 3.81 Soient X et Y deux ensembles et $\varphi : X \times Y \to \overline{\mathbb{R}}$. On dit que $(\bar{x}, \bar{y}) \in X \times Y$ est un *point-selle* de φ sur $X \times Y$ si

$$\forall (x, y) \in X \times Y : \qquad \varphi(\bar{x}, y) \leqslant \varphi(\bar{x}, \bar{y}) \leqslant \varphi(x, \bar{y}). \tag{3.72}$$

Dans les conditions ci-dessus, $\varphi(\bar{x}, \bar{y})$ est appelée la *valeur-selle* de φ. □

Autrement dit, $(\bar{x}, \bar{y})$ est un point-selle de φ si, et seulement si, $x \mapsto \varphi(x, \bar{y})$ atteint un minimum en $\bar{x}$ et $y \mapsto \varphi(\bar{x}, y)$ atteint un maximum en $\bar{y}$. Rien n'est requis en dehors de la « croix » $(\{\bar{x}\} \times Y) \cup (X \times \{\bar{y}\})$, si bien l'image de la selle ou du col à laquelle fait réfeerence l'appellation peut être trompeuse comme lorsque $\varphi : \mathbb{R}^2 \to \mathbb{R}$ est définie par $\varphi(x, y) = x^2 y^2$ (tous les points de l'axe des ordonnées sont des points-selles). Enfin, on pourra souvent se ramener à la définition précédente par un changement variable. Par exemple, le point $(0, 0) \in \mathbb{R}^2$ n'est pas un point-selle de la fonction $(x, y) \in \mathbb{R}^2 \mapsto xy + (x^3 + y^3)$, au sens de la définition ci-dessus, mais le devient *localement* après le changement de variables $(x, y) \curvearrowright (\tilde{x}, \tilde{y})$ avec $\tilde{x} := x+y$ et $\tilde{y} := x-y$ (voir le tracé de droite à la figure 3.11).

La définition est illustrée à la figure 3.11 dans le cas où $X = Y = \mathbb{R}$. On y a

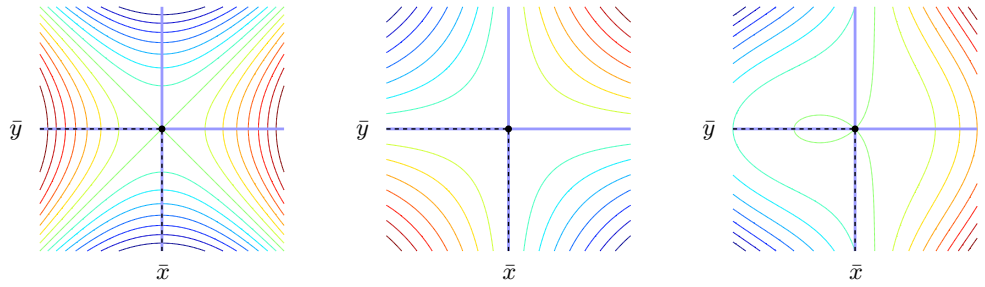


Fig. 3.11. Courbes de niveau de fonctions φ avec point-selle $(\bar{x}, \bar{y}) = (0, 0) : \varphi(x, y) = x^2 - y^2$ à gauche, $\varphi(x, y) = xy$ au milieu et $\varphi(x, y) = (x + y)(x - y) + (x + y)^3 + (x - y)^3$ à droite.

tracé les courbes de niveaux de trois fonctions φ. Le tracé de gauche correspond à la fonction $\varphi(x, y) = x^2 - y^2$ dont le graphe se présente comme une selle, strictement convexe suivant l'axe des abscisses et strictement concave suivant l'axe des ordonnées : il y a un unique point-selle $(\bar{x}, \bar{y}) = (0, 0)$. Le tracé du millieu est un peu moins lié à l'idée que l'on peut se faire de point-selle et correspond à la fonction $\varphi(x, y) = xy$: on a toujours le même unique point-selle $(\bar{x}, \bar{y}) = (0, 0)$. Le tracé de droite correspond à une fonction non quadratique et le point-selle $(\bar{x}, \bar{y}) = (0, 0)$ n'est que local en x.

La notion de point-selle joue un rôle important en dualité min-max (section 14.1) ; en particulier par le résultat énoncé au téorème 14.3.

Notes

Bien que long, ces deux premiers chapitres ne donnent qu'un aperçu, utile à l'optimisation, de l'immense théorie qu'est l'Analyse Convexe. On en apprendra davantage en travaillant les monographies de Rockafellar [517 ; 1970], d'Hiriart-Urruty et Lemaréchal [317 ; 1993] et de Borwein et Lewis [78 ; 2000]. Elles traitent toutes les trois de la théorie en dimension finie et sont orientées vers l'optimisation. Elles ont été nos trois sources d'information principales. L'ouvrage de Rockafellar est très complet, mais difficile. On trouve chez Hiriart-Urruty et Lemaréchal à peu près tout ce dont on a besoin d'analyse convexe en optimisation et on y apprend à manier chaque notion avec de multiples points de vue. Le livre de Borwein et Lewis se concentre sur l'essentiel et laisse au lecteur la possibilité de mettre à l'épreuve ses connaissances et sa maîtrise du sujet sur de nombreux exercices instructifs et stimulants. Mentionnons aussi la monographie d'Auslender et Teboulle [29] consacrée à l'extension des notions de cônes et fonctions asymptotiques à l'optimisation non convexe et à l'étude des inéquations variationnelles.

Les livres d'Ekeland et Temam [191 ; 1974] et de Barbu et Precupanu [36 ; 1975] sont des classiques de la dimension infinie. Pour des traitements plus récents, on pourra consulter Aubin et Ekeland [25 ; 1984], Phelps [480 ; 1993] qui synthétise les questions touchant aux liens entre la différentiabilité des fonctions convexes et la structure des espaces de Banach sur lesquels elles sont définies (*espaces d'Asplund* en particulier) Bonnans et Shapiro [74 ; 2000] et Borwein et Vanderwerff [79 ; 2010].

La notion de fonction conjuguée fut introduite par Mandelbrojt [403 ; 1939] pour une fonction d'une seule variable réelle ; il montre la relation $f^{**} = f$ pour une fonction convexe ne prenant que des valeurs finies. L'opération de conjugaison est précisée et améliorée par Fenchel [203 ; 1949], qui l'étend aux fonctions convexes dépendant d'un nombre fini de variables et qui introduit la notation f^*. La conjugaison généralise une transformation de fonction introduite bien plus tôt par Legendre [385 ; 1787]. L'extension aux espaces vectoriels topologiques est due à Brondsted [93 ; 1964], Moreau [442 ; 1967] et Rockafellar.

L'expression « règle de bascule » (corollaire 3.58) a été empruntée à Hiriart-Urruty [316] et celle « formule du max » (proposition 3.66) à Borwein et Lewis [78]. Le caractère monotone maximal du sous-différentiel a été mis en évidence par Minty [423 ; 1964] et Moreau [441 ; 1965].

Le lecteur trouvera des compléments sur la régularisée de Moreau-Yosida dans le très bel article de Moreau [441], sur lequel nous nous sommes en partie fondés. Cette notion s'étend aux opérateurs monotones maximaux (voir par exemple [90, 91]). On peut aussi prendre des pénalisations autres que la pénalisation quadratique pour construire cette régularisée [576, 186 ; 1992-93].

Exercices

3.1. *Épigraphe d'une fonction.* Soit $f : \mathbb{E} \to \overline{\mathbb{R}}$. Alors

1) $\operatorname{aff}(\operatorname{epi} f) = (\operatorname{aff}(\operatorname{dom} f)) \times \mathbb{R}$,
2) f est convexe si, et seulement si, son épigraphe stricte est convexe,
3) si f est convexe, alors $(\operatorname{epi} f)^\circ = \{(x, \alpha) : x \in (\operatorname{dom} f)^\circ,\ f(x) < \alpha\}$.

3.2. *Autre définition d'une fonction convexe.* Soit $f : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ une fonction. Alors f est convexe si, et seulement si,

$$\forall\, x, y \in \operatorname{dom} f, \quad \forall\, t \notin [0,1] : \qquad f((1-t)x + ty) \geqslant (1-t)f(x) + tf(y). \tag{3.73}$$

3.3. *Fonction convexe propre.* Soient $\mathbb{E}$ un espace vectoriel et $f : \mathbb{E} \to \overline{\mathbb{R}}$ une fonction convexe. S'il existe $x \in (\operatorname{dom} f)^{\circ}$ tel que $f(x) > -\infty$, alors f est propre.

3.4. *Inégalités de convexité.* Soient $\mathbb{E}$ un espace vectoriel (de dimension finie).

1) *Inégalité de Jensen* [335, 336]. Une fonction $f : \mathbb{E} \to \overline{\mathbb{R}}$ est convexe si, et seulement si, pour tout m-uplet $(t_1, \dots, t_m) \in \Delta_m$ et pour tout $(x_1, \dots, x_m) \in (\operatorname{dom} f)^m$, on a
$$f\left(\textstyle\sum_{i=1}^m t_i x_i\right) \leqslant \sum_{i=1}^m t_i f(x_i). \tag{3.74}$$
2) *Inégalité de Jensen, version intégrale.* Une fonction $f : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ est convexe si, et seulement si, pour tout espace de probabilité $(\Omega, \mathcal{A}, \mu)$ et toute fonction $(\mathcal{A}, \mu)$-intégrable $F : \Omega \to \mathbb{E}$ telle que $\int_\Omega F \,\mathrm{d}\mu \in \operatorname{dom} f$, on a
$$f\left(\textstyle\int_\Omega F \,\mathrm{d}\mu\right) \leqslant \int_\Omega (f \circ F) \,\mathrm{d}\mu. \tag{3.75}$$
3) On suppose que f est convexe et on se donne x_0, $x_1 \in \operatorname{dom} f$ et $t > 1$ tels que $(1-t)x_0 + tx_1 \in \operatorname{dom} f$. Alors $f((1-t)x_0 + tx_1) \geqslant (1-t)f(x_0) + tf(x_1)$.
4) *Inégalité géométrico-arithmétique.* La moyenne géométrique est inférieure à la moyenne arithmétique : si $\{a_i\}_{i=1}^n$ sont n nombres positifs, on a
$$\left(\textstyle\prod_{i=1}^n a_i\right)^{1/n} \leqslant \frac{1}{n} \sum_{i=1}^n a_i. \tag{3.76}$$
Remarque. Pour retenir le sens de cette inégalité, il suffit de constater que l'inégalité inverse n'a pas lieu si un des a_i est nul et un autre ne l'est pas.
5) Soient p et $p' \in]0,1[$ reliés par $\frac{1}{p} + \frac{1}{p'} = 1$ (*nombres conjugués*). Pour tout a, $b \in \mathbb{R}_+$, on a
$$ab \leqslant \tfrac{1}{p} a^p + \tfrac{1}{p'} b^{p'}, \tag{3.77}$$
avec égalité si, et seulement si, $a^p = b^{p'}$.

3.5. *Fonction asymptotique d'une pré-composition par une fonction affine.* Démontrez la proposition 3.30.

3.6. *Inf-convolution.* Démontrez la proposition 3.36 sur l'inf-convolution, ainsi que les propriétés suivantes :

1) $f \leqslant g \Longrightarrow f \boxplus h \leqslant g \boxplus h$,
2) $f \boxplus \mathcal{I}_{\{0\}} = f$,
3) $\mathcal{I}_A \boxplus \mathcal{I}_B = \mathcal{I}_{A+B}$.

3.7. *Courbes de niveau d'une fonction convexe.* Chaque dessin A, B et C de la figure 3.12 représente 3 courbes de niveau (iso-valeurs) d'une fonction f définie sur $\mathbb{R}^2$ à valeurs dans $\mathbb{R}$. De manière plus précise, la courbe étiquetée « i » ($i = 1, 2, 3$) représente l'ensemble $\{x \in \mathbb{R}^2 : f(x) = i\}$. Ces fonctions peuvent-elles être convexes ?

3.8. *Fonction quadratique convexe.* On considère la fonction réelle définie sur $\mathbb{R}^n$ par $f(x) = g^{\mathsf{T}} x + \frac{1}{2} x^{\mathsf{T}} H x$, où $g \in \mathbb{R}^n$ et H est une matrice d'ordre n symétrique.

1) Montrez que f est convexe si, et seulement si, $H \succcurlyeq 0$.
2) Montrez que f est strictement convexe si, et seulement si, $H \succ 0$.
3) Calculez le cône asymptotique commun aux ensembles de sous-niveau non vides de f lorsque $H \succcurlyeq 0$ (voir la proposition 3.29).

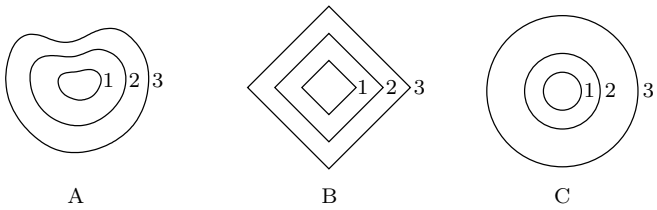


Fig. 3.12. Courbes de niveau d'une fonction convexe ?

3.9. *Exemples d'inf-convolution.* Soient A et B deux opérateurs auto-adjoints sur un espace euclidien $\mathbb{E}$. On note $x \mapsto f(x) = \langle Ax, x\rangle$ et $x \mapsto g(x) = \langle Bx, x\rangle$ les formes quadratiques associées. Calculez $f \boxplus g$.

3.10. *Dérivée directionnelle de la fonction max.* Pour $i \in [1\,{:}\,m]$, on suppose données des fonctions $f_i : \mathbb{E} \to \mathbb{R}$ qui en $x \in \mathbb{E}$ sont continues et ont des dérivées directionnelles. On note $f_{\max} : R \to \mathbb{R}$ la fonction max, définie par $f_{\max}(x) = \max_{1\leqslant i\leqslant m} f_i(x)$, et $I(x) := \{i \in [\![1, m]\!] : f_i(x) = f_{\max}(x)\}$. Alors

$$f'_{\max}(x; d) = \max_{i\in I(x)} f'_i(x; d).$$

3.11. *Caractérisation de la forte convexité d'une fonction.* Démontrez la proposition 3.24.

3.12. *Problème d'optimisation sans solution.* Soient $F : \mathbb{R}^n \to \mathbb{R}^m$, $X \subseteq \mathbb{R}^n$ et $\|\cdot\|$ une norme sur $\mathbb{R}^m$. On considère le problème $\min\{\|F(x)\| : x \in X\}$. On suppose que $0 \notin F(X)$ et que $F(X)$ est un ouvert. Montrez que :

1) ce problème n'a pas de solution,
2) si F est continue et $\{x_k\}$ est une suite minimisante, ou bien $\|x_k\| \to \infty$, ou bien $\liminf \mathrm{d}_{X^c}(x_k) = 0$ ($\mathrm{d}_{X^c}(x_k)$ est la distance de x_k au complémentaire de X, supposée infinie si $X = \mathbb{R}^n$).

Remarque. Ce résultat s'applique par exemple au problème $\min\{e^x : x \in \mathbb{R}\}$.

3.13. *Minimisation concave.* Supposons qu'un problème de minimisation d'une *fonction concave* f sur un ensemble *convexe* C ait une solution x_*. Alors f prend sa valeur minimale $f(x_*)$ sur la face de C engendrée par x_*. En particulier, si C est fermé et a un point extrême, le problème a un minimum en un point extrême de C.

3.14. *Optimisation multicritère, Pareto optimalité* [332]. Soient X un ensemble et $f : X \to \mathbb{R}^m$ une fonction. On considère les composantes de f comme m critères à minimiser (on parle d'*optimisation multicritère*). On dit que $x_* \in X$ est *Pareto optimal* s'il n'existe pas de $x \in X$ vérifiant $f_i(x) < f_i(x_*)$ pour tout i, ce qui s'écrit aussi

$$\forall\, x \in X, \quad \max_{1\leqslant i\leqslant m} \Big(f_i(x) - f_i(x_*)\Big) \geqslant 0.$$

On dit que $x_* \in X$ est *moyennement optimal* s'il existe un $r \in \mathbb{R}^m_+$, *non nul*, tel que x_* minimise $x \mapsto r^{\mathsf{T}} f(x)$ sur X.

1) Montrez que x_* est Pareto optimal si, et seulement si, $(f(x_*) - \mathbb{R}^m_{++}) \cap f(X) = \varnothing$.
2) On a représenté à la figure 3.13, dans le cas où $m = 2$, un exemple d'ensemble $f(X)$. Déterminez dans celui-ci l'image par f des points Pareto optimaux et l'image par f des points moyennement optimaux.
3) Montrez qu'un point moyennement optimal est Pareto optimal.
4) Montrez que, lorsque $f(X) + \mathbb{R}^m_+$ est convexe (a fortiori lorsque $f(X)$ est convexe), un point Pareto optimal est moyennement optimal.

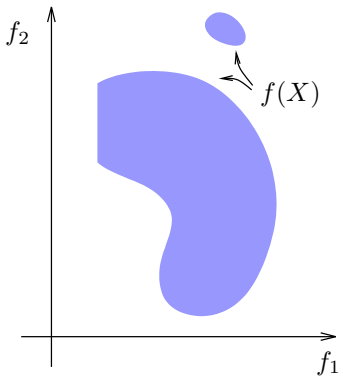


Fig. 3.13. Un ensemble $f(X)$ dans $\mathbb{R}^2$

5) Montrez que $f(X) + \mathbb{R}^m_+$ est convexe si X est convexe et si les fonctions f_i sont convexes ; alors que $f(X)$ ne l'est pas nécessairement.

3.15. *Propriétés particulières aux fonctions convexes polyédriques.* Soient f, f_1 et f_2 des fonctions convexes polyédriques. Alors

1) $f_1 + f_2$ est polyédrique,
2) $f_1 \boxplus f_2$ est polyédrique, $\operatorname{epi} f_1 \boxplus f_2 = \operatorname{epi} f_1 + \operatorname{epi} f_2$ et l'infimum dans la définition de $(f_1 \boxplus f_2)(x)$ est atteint pour tout $x \in \operatorname{dom}(f_1 \boxplus f_2) = \operatorname{dom} f_1 + \operatorname{dom} f_2$,
3) f^* est polyédrique.

3.16. *Calcul de fonctions conjuguées.* On note $\langle \cdot, \cdot \rangle$ le produit scalaire utilisé pour définir la conjuguée.

1) *Fonction quadratique strictement convexe.* Soient Q une matrice auto-adjointe définie positive (pour le produit scalaire $\langle \cdot, \cdot \rangle$), $q \in \mathbb{R}^n$ et $f : \mathbb{R}^n \to \mathbb{R}$ la fonction quadratique définie par

$$f(x) = \frac{1}{2}\langle Qx, x\rangle + \langle q, x\rangle.$$

Alors, pour tout $x^* \in \mathbb{R}^n$,

$$f^*(x^*) = \frac{1}{2}\langle Q^{-1}(x^* - q), (x^* - q)\rangle.$$

Si Q est seulement semi-définie positive, on a $f^*(Qx+q) = \frac{1}{2}\langle Qx, x\rangle$ pour tout $x \in \mathbb{R}^n$ et $f^*(x^*) = +\infty$ si $x^* \notin \mathcal{R}(Q) + q$.

2) Si $g(x) = f(x) + \alpha$, alors $g^*(x^*) = f^*(x^*) - \alpha$.

3) *Monotonie.* Si f, $g : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ sont propres avec une minorante affine et $f \leqslant g$, alors $f^* \geqslant g^*$.

4) *Enveloppe inférieure.* Soient I un ensemble d'indices quelconque et, pour $i \in I$, des fonctions $f_i : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ propres ayant une minorante affine commune, c.-à-d., $\exists\, x_0^* \in \mathbb{E}$ tel que $\sup_{i\in I} f_i^*(x_0^*) < +\infty$. Alors leur *enveloppe inférieure* $f_{\inf} : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ définie en x par

$$f_{\inf}(x) = \inf_{i\in I} f_i(x)$$

est propre avec une minorante affine et on a

$$f^*_{\inf} = \sup_{i\in I} f_i^*.$$

5) *Enveloppe supérieure.* Soient I un ensemble d'indices quelconque et, pour $i \in I$, des fonctions $f_i \in \mathrm{C}\overline{\mathrm{onv}}(\mathbb{E})$ telles que leur *enveloppe supérieure* $f_{\sup} : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ définie en x par

$$f_{\sup}(x) = \sup_{i\in I} f_i(x)$$

ait un domaine non vide. Alors $f_{\sup} \in \mathrm{C}\overline{\mathrm{onv}}(\mathbb{E})$ et on a

$$f^*_{\sup} = \left(\inf_{i\in I} f_i^*\right)^{**}.$$

3.17. *Minimum de f et f^{**}.* Soient $\mathbb{E}$ un espace euclidien et $f : \mathbb{E} \to \overline{\mathbb{R}}$ une fonction non nécessairement convexe. Montrez que

1) $\inf f = \inf f^{**}$,
2) $\overline{\mathrm{co}}(\arg\min f) \subseteq \arg\min f^{**}$ (l'égalité n'a pas nécessairement lieu),
3) si f est propre et fermée, si $\arg\min f \neq \varnothing$ et si $\arg\min f^{**}$ est borné, alors $\overline{\mathrm{co}}(\arg\min f) = \arg\min f^{**}$.

3.18. *Enveloppe convexe fermée de fonctions convexes.* Soient f, f_1 et $f_2 \in \mathrm{Conv}(\mathbb{E})$.

1) Si $x_0 \in (\mathrm{dom}\, f)^\circ$ et $x \in \mathbb{E}$, alors $f^{**}(x) = \lim_{t\uparrow 1} f((1-t)x_0 + tx)$.
2) $\mathrm{aff}\,\mathrm{dom}\, f = \mathrm{aff}\,\mathrm{dom}\, f^{**}$, $(\mathrm{dom}\, f)^\circ = (\mathrm{dom}\, f^{**})^\circ$ et $\mathrm{adh}\,\mathrm{dom}\, f = \mathrm{adh}\,\mathrm{dom}\, f^{**}$.
3) Si f_1 et $f_2 \in \mathrm{C}\overline{\mathrm{onv}}(\mathbb{E})$ et $(\mathrm{dom}\, f_1) \cap (\mathrm{dom}\, f_2) \neq \varnothing$, alors $f_1 + f_2 \in \mathrm{C}\overline{\mathrm{onv}}(\mathbb{E})$.
Si f_1 et $f_2 \in \mathrm{Conv}(\mathbb{E})$ et $(\mathrm{dom}\, f_1)^\circ \cap (\mathrm{dom}\, f_2)^\circ \neq \varnothing$, alors $(f_1 + f_2)^{**} = f_1^{**} + f_2^{**}$.

3.19. *Fonction d'appui du sous-différentiel.* Soient $f \in \mathrm{C}\overline{\mathrm{onv}}(\mathbb{E})$ et $x \in \mathbb{E}$ tel que $\partial f(x)$ soit non vide. Alors la fonction d'appui de $\partial f(x)$ est l'enveloppe convexe fermée de $\delta_x(\cdot) = f'(x;\cdot) : \sigma_{\partial f(x)} = \delta_x^{**} \leqslant \delta_x$.

3.20. *Allure du sous-différentiel.* Soit $f \in \mathrm{Conv}(\mathbb{E})$.

1) Soient $\bar{x} \in (\mathrm{dom}\, f)^\circ$ et $\varepsilon > 0$. Montrez que $\bar{B}(0,\varepsilon) \subseteq \partial f(\bar{x})$ si, et seulement si, $f'(\bar{x};h) \geqslant \varepsilon\|h\|$ pour tout $h \in \mathbb{E}$.
2) Supposons que f soit sous-différentiable en $x \in \mathbb{E}$ et que $t \in \mathbb{R} \mapsto f(x+th)$ soit différentiable en 0. Alors $\partial f(x) \subseteq f'(x;h)h/\|h\|^2 + h^\perp$.

3.21. *Représentation du sous-différentiel.* Soit $f : \mathbb{R}^2 \to \mathbb{R}$ une fonction convexe, dont on note $\partial f(x)$ le sous-différentiel en x pour le produit scalaire euclidien. La figure 3.14 donne quatre dessins : A, B, C et D. Dans chacun d'eux, on a noté x un point de $\mathbb{R}^2$, $\mathcal{C}$ une *partie* d'une courbe de niveau de f (c.-à-d., f est constante sur une courbe contenant $\mathcal{C}$) et S un ensemble. Pour chacun de ces dessins :

(a) dites si S peut être l'ensemble $x + \partial f(x)$,
(b) dans l'affirmative en (a) et si $S = x + \partial f(x)$, dites si x est un minimiseur de f,
(c) dans l'affirmative en (a) et (b), et si $S = x + \partial f(x)$, dites si x est l'*unique* minimiseur de f.

3.22. *Sous-différentiel de fonction fortement convexe* [523 ; 1976, proposition 6]. Soient $\mathbb{E}$ un espace euclidien dont le produit scalaire est noté $\langle\cdot,\cdot\rangle$, $f \in \mathrm{Conv}(\mathbb{E})$ et $\alpha > 0$. Alors les propriétés suivantes sont équivalentes :

(*i*) la fonction f est fortement convexe de module α,
(*ii*) la multifonction ∂f est fortement monotone de module α,
(*iii*) $\forall\, x \in \mathrm{dom}\, f$, $\forall\, x^* \in \partial f(x)$, $\forall\, y \in \mathbb{E}$, on a $f(y) \geqslant f(x) + \langle x^*, y - x\rangle + (\alpha/2)\|y - x\|^2$.

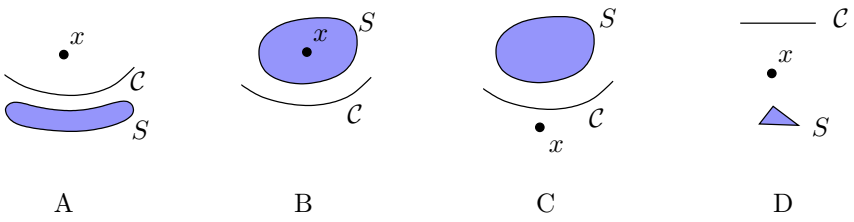


Fig. 3.14. Représentation du sous-différentiel d'une fonction convexe ?

3.23. *Sous-différentiel d'une fonction définie sur un espace produit.* Soient $\mathbb{E}_1$ et $\mathbb{E}_2$ deux espaces euclidiens, $\mathbb{E}_1 \times \mathbb{E}_2$ muni du produit scalaire produit (A.10) et $f \in \mathrm{Conv}(\mathbb{E}_1 \times \mathbb{E}_2)$. On introduit les fonctions partielles $f^1_{x_2}(\cdot) = f(\cdot, x_2)$ et $f^2_{x_1}(\cdot) = f(x_1, \cdot)$. Alors

$$\partial f(x_1, x_2) \subseteq \partial f^1_{x_2}(x_1) \times \partial f^2_{x_1}(x_2).$$

Donnez un exemple de fonction pour laquelle l'égalité n'a pas lieu.

3.24. *Minimum saillant* [484 ; § 5.2.3]. Soient $\mathbb{E}$ un espace euclidien dont la norme est notée $\|\cdot\|$, $f \in \mathrm{Conv}(\mathbb{E})$, $\bar{B}$ la boule unité fermée de $\mathbb{E}$, $\bar{x} \in \mathrm{dom}\, f$ et $\alpha > 0$ (un réel). Montrez que les propriétés suivantes sont équivalentes :

(*i*) $\forall\, x \in \mathbb{E} :\ f(x) \geqslant f(\bar{x}) + \alpha\|x - \bar{x}\|$,
(*ii*) $\forall\, d \in \mathbb{E} :\ f'(\bar{x}; d) \geqslant \alpha\|d\|$,
(*iii*) $\alpha\bar{B} \subseteq \partial f(\bar{x})$.

On dit qu'un point $\bar{x}$ vérifiant ces propriétés est un *minimum saillant* de $f \in \mathrm{Conv}(\mathbb{E})$.

3.25. *Ensemble saillant de minimiseurs.* Soient $\mathbb{E}$ un espace euclidien de produit scalaire noté $\langle\cdot,\cdot\rangle$ (norme associée $\|\cdot\|$), $\bar{B} := \{x \in \mathbb{E} : \|x\| \leqslant 1\}$ la boule unité fermée de $\mathbb{E}$, $f \in \overline{\mathrm{conv}}(\mathbb{E})$ et $f_{\min} := \inf\{f(x) : x \in \mathbb{E}\}$. On suppose que $f_{\min}$ est fini et que f a un ensemble de minimiseurs $S := \{x \in \mathbb{E} : f(x) = f_{\min}\}$ non vide.

1) Montrez que, quel que soit $x \in \mathbb{E}$, le problème $\inf\{\|x - y\| : y \in S\}$ a une solution et une seule. On la note $\mathrm{P}_S(x)$.

On note $\mathrm{dist}(x, S) := \|x - \mathrm{P}_S(x)\|$ la distance de x à S, pour la norme $\|\cdot\|$. On dit que l'ensemble des minimiseurs S est *saillant* s'il existe $\alpha > 0$ tel que

$$\forall x \in \mathbb{E} :\quad f(x) \geqslant f_{\min} + \alpha\,\mathrm{dist}(x, S). \tag{3.78}$$

Dans une première partie, on se propose de montrer que cette propriété est *équivalente* à l'inclusion suivante (pour le même $\alpha > 0$)

$$S + \alpha\bar{B} \subseteq \bigcup_{x \in S} \Big(x + \partial f(x)\Big), \tag{3.79}$$

où le sous-différentiel $\partial f(x)$ de f en $x \in \mathbb{E}$ est calculé pour le produit scalaire $\langle\cdot,\cdot\rangle$.
Supposons dans un premier temps que (3.78) ait lieu et l'on se donne pour objectif de démontrer (3.79). Soient $\bar{x}_0 \in S$, $u \in \bar{B}$ et $\bar{x} := \mathrm{P}_S(\bar{x}_0 + \alpha u)$. On introduit $g := \bar{x}_0 + \alpha u - \bar{x}$.

2) Montrez que $\|g\| \leqslant \alpha$.
3) Montrez que $\forall x \in \mathbb{E}$, $f(x) \geqslant f(\bar{x}) + \langle g, x - \bar{x}\rangle$.

4) Montrez que (3.79) est vérifiée.

Supposons à présent que (3.79) ait lieu et l'on se donne pour objectif de démontrer (3.78). Soient $x \in \mathbb{E}\backslash S$ et $\bar{x} := \mathrm{P}_S(x)$.

5) Montrez qu'il existe un $\bar{x}_1 \in S$ et $g_1 \in \partial f(\bar{x}_1)$ tels que

$$\bar{x} + \alpha \frac{x - \bar{x}}{\|x - \bar{x}\|} = \bar{x}_1 + g_1.$$

6) Montrez que $f(x) \geqslant f_{\min} + \langle g_1, x - \bar{x}_1 \rangle$ et en déduire (3.78).

Dans une seconde partie, on considère l'algorithme de minimisation de f, qui calcule en $x \in \mathbb{E}$ le nouvel itéré x_+ comme solution de

$$\inf_{y \in \mathbb{E}} \left(f(y) + \frac{1}{2}\|y - x\|^2 \right). \tag{3.80}$$

7) Sachant que les fonctions de $\mathrm{C}\overline{\mathrm{onv}}(\mathbb{E})$ ont une minorante affine, montrez que le problème (3.80) a une solution et une seule.
8) Supposons que f ait un ensemble saillant de minimiseurs. Montrez que si x est suffisamment proche de S, alors $x_+ \in S$.

3.26. *Sous-différentiabilité et sous-lipschitzianité* [236 ; lemme 2]. Soit $f \in \mathrm{Conv}(\mathbb{E})$ et $x \in \operatorname{dom} f$. Montrez que $\partial f(x) \neq \varnothing$ si, et seulement si, il existe une constante $L \geqslant 0$ telle que

$$\forall\, y \in \mathbb{E} : \quad f(y) \geqslant f(x) - L\|y - x\|.$$

3.27. Démontrez le corollaire 3.73.

3.28. *Fonction indicatrice*. Soient P une partie non vide d'un espace vectoriel $\mathbb{E}$ et $\mathcal{I}_P$ son indicatrice.

1) Convexité : démontrez la proposition 3.6.

Supposons à présent que $\mathbb{E}$ soit un espace euclidien, que C soit un convexe non vide de $\mathbb{E}$, que $x \in C$ et que K est un cône non vide de $\mathbb{E}$.

2) Dérivée directionnelle : $\mathcal{I}'_X(x; d) = \mathcal{I}_{\mathrm{T}^{\mathrm{a}}_C(x)}(d)$.
3) Conjuguée : $\mathcal{I}^*_P = \sigma_P$ et $\mathcal{I}^*_K = \mathcal{I}_{K^-}$.
4) Sous-différentiel : $\partial \mathcal{I}_C(x) = \mathrm{N}_C(x)$.

3.29. *Fonction d'appui*. Soit C un convexe fermé non vide d'un espace euclidien $(\mathbb{E}, \langle \cdot, \cdot \rangle)$. On rappelle la définition (3.7) de la fonction d'appui $\sigma_C : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$, qui est une fonction convexe propre et fermée.

1) Conjuguée : $\sigma^*_C = \mathcal{I}_C$.
2) Sous-différentiel : $\partial \sigma_C(x^*) = \arg\max\{\langle x^*, x \rangle : x \in C\}$.

3.30. *Norme*. Soient $\mathbb{E}$ un espace euclidien dont le produit scalaire est noté $\langle \cdot, \cdot \rangle$ et $f(\cdot) := \|\cdot\|$ une norme (pas nécessairement celle associée au produit scalaire). On note $\|\cdot\|_{\mathrm{D}}$ la norme duale de $\|\cdot\|$ par rapport à ce produit scalaire (voir (A.8)) et $\bar{B}_{\mathrm{D}}$ la boule-unité fermée pour la norme duale.

1) La norme $\|\cdot\|$ est une fonction convexe propre fermée, avec une minorante affine.
2) $f^* = \mathcal{I}_{\mathrm{D}}$.
3) Soit $\|\cdot\|_{\mathrm{DD}}$ la *norme biduale* de $\|\cdot\|$, c'est-à-dire la norme duale de $\|\cdot\|_{\mathrm{D}}$ par rapport au produit scalaire donné :

$$\|u\|_{\mathrm{DD}} := \sup_{\|v\|_{\mathrm{D}} \leqslant 1} \langle v, u \rangle.$$

Montrez que $\|\cdot\|_{\text{DD}} = \|\cdot\|$ et que

$$\forall\, x \in \mathbb{E}, \quad \exists\, x^* \in \mathbb{E} : \quad \|x^*\|_{\text{D}} = 1 \quad \text{et} \quad \langle x, x^*\rangle = \|x\|. \tag{3.81}$$

4) Le sous-différentiel de la norme s'écrit

$$\partial f(x) = \operatorname*{arg\,max}_{\|y\|_{\text{D}} \leqslant 1} \langle y, x\rangle = \{x^* \in \mathbb{E} : \|x^*\|_{\text{D}} \leqslant 1 \text{ et } \langle x^*, x\rangle = \|x\|\}. \tag{3.82}$$

En particulier :
- $\partial f(0) = \bar{B}_D$,
- $\|x^*\|_{\text{D}} = 1$ si $x^* \in \partial f(x)$ et $x \neq 0$,
- si $\mathbb{E} = \mathbb{R}^n$, $\langle\cdot,\cdot\rangle$ est le produit scalaire euclidien, $x \neq 0$ et $p \in]1, \infty[$, on a

$$\partial(\|\cdot\|_1)(x) = \{x^* \in \mathbb{R}^n : x_i^* = -1 \text{ si } x_i < 0, \quad x_i^* \in [-1, +1] \text{ si } x_i = 0, \quad x_i^* = +1 \text{ si } x_i > 0\}, \tag{3.83a}$$

$$\nabla(\|\cdot\|_p)(x) = \{\operatorname{sgn}(x_i)(|x_i|/\|x\|_p)^{p-1}\}_{i\in[1\,:\,n]}, \tag{3.83b}$$

$$\partial(\|\cdot\|_\infty)(x) = \operatorname{co}\{\operatorname{sgn}(x_i)e^i : i \in I\}, \tag{3.83c}$$

où e^i est le i-ième vecteur de base de $\mathbb{R}^n$ et $I := \{i \in [1\,:\,n] : |x_i| = \|x\|_\infty\}$.

5) On déduit de la formule du max (3.52) que la dérivée directionnelle de la norme en $x \in \mathbb{E}$ dans la direction $h \in \mathbb{E}$ s'écrit

$$(\|\cdot\|)'(x; h) = \max_{\substack{x^* \in \bar{B}_D \\ \langle x^*, x\rangle = \|x\|}} \langle x^*, h\rangle. \tag{3.84}$$

Démontrez :
- (3.84) en utilisant directement la définition (C.1) de la dérivée directionnelle,
- $(\|\cdot\|)'(0; h) = \|h\|$,
- si $\mathbb{E} = \mathbb{R}^n$, $\langle\cdot,\cdot\rangle$ est le produit scalaire euclidien, $x \neq 0$ et $p \in]1, \infty[$, on a

$$(\|\cdot\|_1)'(x; h) = \tag{3.85a}$$

$$(\|\cdot\|_p)'(x; h) = \tag{3.85b}$$

$$(\|\cdot\|_\infty)'(x; h) = \max_{i:\ |x_i| = \|x\|_\infty} (\operatorname{sgn} x_i) h_i, \tag{3.85c}$$

3.31. *Puissance de norme.* Soient $\mathbb{E}$ un espace euclidien dont le produit scalaire est noté $\langle\cdot,\cdot\rangle$ et $\|\cdot\|$ une norme (pas nécessairement celle associée au produit scalaire). On note $\|\cdot\|_{\text{D}}$ la norme duale de $\|\cdot\|$ par rapport à ce produit scalaire (voir (A.8)) et on considère la fonction $f : \mathbb{E} \to \mathbb{R}$ définie en $x \in \mathbb{E}$ par

$$f(x) = \frac{1}{p}\|x\|^p,$$

où $p > 1$ est un réel. On note $p' = p/(p-1)$ le *nombre conjugué* de p.

1) La fonction f est convexe propre et fermée.

2) La conjuguée $f^* : \mathbb{E} \to \mathbb{R}$ de f prend en $s \in \mathbb{E}$ la valeur

$$f^*(s) = \frac{1}{p'}\|s\|_{\text{D}}^{p'}.$$

Le sous-différentiel de f en $x \in \mathbb{E}$ s'écrit (pour le sous-différentiel de la norme, voir l'exercice 3.30) :

$$\partial f(x) = \{s \in \mathbb{E} : \langle s, x\rangle = \|x\|^p = \|s\|_{\text{D}}^{p'}\} = \|x\|^{p-1}\, \partial(\|\cdot\|)(x).$$

Remarque : la dernière expression du sous-différentiel $\partial f(x)$ rappelle la dérivation en chaîne de la composition de $x \in \mathbb{E} \mapsto \|x\|$ et de $t \in \mathbb{R} \mapsto t^p/p$.

3.32. *Compteur de composante non nulle.* On note $\|\cdot\|_0$ le *compteur de composante non nulle* d'un vecteur (donc $\|x\|_0$ est le nombre de composantes non nulles du vecteur $x \in \mathbb{R}^n$, voir aussi la section 19.4). On note également $\|x\|_p$ la norme ℓ_p de $x \in \mathbb{R}^n$ ($p = 1$ ou ∞) et $\bar{B}_1 := \{x \in \mathbb{R}^n : \|x\|_1 \leqslant 1\}$ la boule unité fermée de $\mathbb{R}^n$ pour la norme ℓ_1. On considère la fonction $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ définie en $x \in \mathbb{R}^n$ par

$$f(x) = \|x\|_0 + \mathcal{I}_{\bar{B}_1}(x),$$

où $\mathcal{I}_{\bar{B}_1}$ est l'indicatrice de $\bar{B}_1$ ($\mathcal{I}_{\bar{B}_1}(x) = 0$ si $x \in \bar{B}_1$ et $\mathcal{I}_{\bar{B}_1}(x) = +\infty$ si $x \notin \bar{B}_1$). Montrez que sa conjuguée f^* et la biconjuguée f^{**} prennent les valeurs suivantes :

$$f^*(x^*) = (\|x^*\|_\infty - 1)^+. \tag{3.86}$$
$$f^{**}(x) = \|x\|_1 + \mathcal{I}_{\bar{B}_1}(x). \tag{3.87}$$

Remarque : on a montré que, sur la boule unité de la norme ℓ_1, $\|\cdot\|_1$ est la plus grande fonction convexe qui minore la fonction non convexe $\|\cdot\|_0$ (alors que sur $\mathbb{R}^n$ cette plus grande fonction convexe est nulle).

3.33. *Fonction maximale.* On suppose que $\mathbb{R}^n$ est muni du produit scalaire euclidien et on considère la *fonction maximale* $f : \mathbb{R}^n \to \mathbb{R}$ définie en $x \in \mathbb{R}^n$ par $f(x) = \max_{1 \leqslant i \leqslant n} x_i$.

1) $f \in \mathrm{C}\overline{\mathrm{onv}}(\mathbb{R}^n)$.
2) La conjuguée f^* de f prend en $x^* \in \mathbb{R}^n$ la valeur
$$f^*(x^*) = \begin{cases} 0 & \text{si } x^* \geqslant 0 \text{ et } \sum_{i=1}^n x_i^* = 1, \\ +\infty & \text{sinon.} \end{cases}$$
3) Le sous-différentiel $\partial f(x)$ de f en $x \in \mathbb{R}^n$ s'écrit
$$\partial f(x) = \mathrm{co}\{e^i : i \in I\},$$
où e^i est le i-ième vecteur de base de $\mathbb{R}^n$ et $I = \{j \in [1:n] : x_j = \max_i x_i\}$.
4) La fonction f est différentiable en x si, et seulement si, le maximum des x_i est atteint pour un seul indice i. Dans ce cas, $\nabla f(x) = e^i$.

3.34. *Valeur propre maximale.* On note $\mathcal{S}^n$ [resp. $\mathcal{S}^n_+$] l'ensemble des matrices d'ordre n symétriques [resp. symétriques semi-définies positives], que l'on munit du produit scalaire $\langle A, B\rangle = \operatorname{tr} AB$. On note $\lambda_{\max} : \mathcal{S}^n \to \mathbb{R}$ l'application valeur propre maximale.

1) La fonction $\lambda_{\max}$ est convexe.
2) La conjuguée de $\lambda_{\max}$ est donnée par
$$\lambda^*_{\max}(A^*) = \begin{cases} 0 & \text{si } A^* \in \mathcal{S}^n_+ \text{ et } \operatorname{tr} A^* = 1 \\ +\infty & \text{sinon.} \end{cases}$$
3) Le sous-différentiel de $\lambda_{\max}$ est le compact donné par
$$\partial\lambda_{\max}(A) = \mathrm{co}\{vv^\mathsf{T} : \|v\|_2 = 1,\ Av = \lambda_{\max}(A)v\}.$$
4) Déduire de la formule du sous-différentiel que
   a) $\lambda_{\max}(\cdot)$ est différentiable en A si $\lambda_{\max}(A)$ est simple, et que dans ce cas $\nabla\lambda_{\max}(A) = vv^\mathsf{T}$, où $\pm v$ sont les uniques vecteurs propres unitaires correspondants à la valeur propre maximale ; retrouvez ce dernier résultat en utilisant le théorème des fonctions implicites (théorème C.14) ;

b) $\partial\lambda_{\max}(0) = \{S \in \mathcal{S}^n_+ : \operatorname{tr} S = 1\}$;
c) la dérivée directionnelle de $\lambda_{\max}$ en $A \in \mathcal{S}^n$ dans la direction $D \in \mathcal{S}^n$ s'écrit

$$\lambda'_{\max}(A; D) = \lambda_{\max}(V^{\mathsf{T}} D V),$$

où V est une matrice dont les colonnes forment une base orthonormale de l'espace propre associé à $\lambda_{\max}(A)$.

3.35. *Valeur propre maximale d'une matrice hermitienne.* On note $\mathcal{H}^n$ [resp. $\mathcal{H}^n_+$] l'ensemble des matrices d'ordre n hermitiennes [resp. hermitiennes semi-définies positives], que l'on munit du produit scalaire $\langle A, B\rangle = \operatorname{tr} AB$. L'écriture $A \succcurlyeq 0$ signifie que $A \in \mathcal{H}^n_+$. On note $\lambda_{\max} : \mathcal{H}^n \to \mathbb{R}$ l'application valeur propre maximale, qui est bien définie car les valeurs propres d'une matrice hermitenne sont réelles. Les résultats ci-dessous s'appliquent bien sûr aux matrices réelles symétriques, qui sont des matrices hermitiennes avec une partie imaginaire nulle.

1) Démontrez (B.18) et en déduire que la fonction $\lambda_{\max}$ est convexe.
2) La conjuguée de $\lambda_{\max}$ est donnée par

$$\lambda^*_{\max}(A^*) = \begin{cases} 0 & \text{si } A^* \succcurlyeq 0 \text{ et } \operatorname{tr} A^* = 1 \\ +\infty & \text{sinon.} \end{cases} \tag{3.88}$$

3) Le sous-différentiel de $\lambda_{\max}$ est le compact donné par

$$\partial\lambda_{\max}(A) = \operatorname{co}\{vv^{\mathsf{H}} : \|v\| = 1,\ Av = \lambda_{\max}(A)v\}. \tag{3.89}$$

3.36. *Distance à un convexe.* Soient $\mathbb{E}$ un espace vectoriel normé (norme notée $\|\cdot\|$) et C un ensemble convexe fermé non vide de $\mathbb{E}$. On considère la fonction $\mathrm{d}_C : \mathbb{E} \to \mathbb{R}$, la *distance à* C, définie par $\mathrm{d}_C(x) = \inf_{y\in C} \|x - y\|$.

1) d_C est convexe.

On suppose dorénavant que $\mathbb{E}$ est un espace euclidien (produit scalaire noté $\langle\cdot,\cdot\rangle$, mais la norme $\|\cdot\|$ n'est pas nécessairement celle associée à ce produit scalaire). On note $\bar{B}_D$ la boule-unité fermée pour la norme duale, $\mathrm{N}_C(x)$ le cône normal à C en x et $\bar{x}$ une projection d'un point x sur C (c'est-à-dire une solution du problème $\inf\{\|x-y\| : y \in C\}$, qui n'est pas nécessairement unique car la norme n'est pas nécessairement associée à un produit scalaire).

2) $\mathrm{d}^{\infty}_C = \|\,\mathrm{P}_{(C^{\infty})^-}(\cdot)\|$
3) $\mathrm{d}^*_C = \mathcal{I}_{\bar{B}_D} + \sigma_C$.
4) $\partial\,\mathrm{d}_C(x) = \big\{x^* \in \bar{B}_D \cap \mathrm{N}_C(\bar{x}) : \langle x^*, x - \bar{x}\rangle = \|x - \bar{x}\|\big\}$.

On suppose à présent que la norme $\|\cdot\|$ est celle associée au produit scalaire $\langle\cdot,\cdot\rangle$.

5) Si $x \notin C$, alors d_C est différentiable en x et $\nabla\,\mathrm{d}_C(x) = (x-\bar{x})/\|x-\bar{x}\|$;
si $x \in C^\circ$ (l'intérieur de C), alors d_C est différentiable en x et $\nabla\,\mathrm{d}_C(x) = 0$;
si $x \in \partial C = C \setminus C^\circ$ (la frontière de C), alors $\partial\,\mathrm{d}_C(x) = \bar{B} \cap \mathrm{N}_C(x)$.
6) Montrez par un contre-exemple que si $P \subseteq \mathbb{E}$ n'est pas convexe, d_P n'est pas nécessairement différentiable sur le complémentaire de P.

3.37. *Carré de la distance euclidienne à un convexe.* Soit $\mathbb{E}$ un espace euclidien dont la norme associée au produit scalaire est notée $\|\cdot\|$, C un convexe fermé non vide de $\mathbb{E}$ et P_C le projecteur orthognal sur C. On note $\mathrm{e}_C : \mathbb{E} \to \mathbb{R}$ la *moitié de la distance à C au carré*, qui est donc définie en $x \in \mathbb{E}$ par

$$\mathrm{e}_C(x) = \inf_{z\in C} \tfrac{1}{2}\|z - x\|^2 = \tfrac{1}{2}\|\,\mathrm{P}_C(x) - x\|^2.$$

Montrez que

1) la fonction e_C est convexe,
2) pour tout $(x,h)\in\mathbb{E}^2$, $0\leqslant\langle \mathrm{P}_C(x)-x, \mathrm{P}_C(x+h)-\mathrm{P}_C(x)\rangle\leqslant\|h\|^2$,
3) la fonction e_C est différentiable en tout point $x\in\mathbb{E}$ et $\nabla \mathrm{e}_C(x)=x-\mathrm{P}_C(x)$,
4) le projecteur P_C est le gradient d'une fonction dérivable.

3.38. *Constance du sous-différentiel.* Soit $f:\mathbb{E}\to\mathbb{R}$ une fonction convexe et supposons que f soit constante sur un ensemble A. Alors ∂f est constant sur $A^{\diamond}$ et ce sous-différentiel commun est inclus dans celui de tout point de A.

3.39. *Ensembles de sous-niveau bornés.* Soient $\mathbb{E}$ un espace euclidien de dimension finie et $f:\mathbb{E}\to\mathbb{R}\cup\{+\infty\}$ une application dont on considère les propriétés suivantes :

(i) f vérifie la *condition de croissance positive à l'infini* suivante

$$\liminf_{\|x\|\to\infty}\frac{f(x)}{\|x\|}:=\lim_{r\to\infty}\inf_{x\notin B_r}\frac{f(x)}{\|x\|}>0,$$

(ii) les ensembles de sous-niveau de f sont bornés,
(iii) f^* est continue en 0,
(iv) $\cup_{x\in\mathbb{E}}\partial f(x)$ est un voisinage de 0.

Montrez que

1) (i) $\Longrightarrow$ (ii), mais que la réciproque n'est pas nécessairement vraie,
2) si $f\in\mathrm{Conv}(\mathbb{E})$, alors ($i$) $\Longleftrightarrow$ (ii) $\Longleftrightarrow$ (iii),
3) si $f\in\mathrm{C}\overline{\mathrm{onv}}(\mathbb{E})$, alors ($i$) $\Longleftrightarrow$ (ii) $\Longleftrightarrow$ (iii) $\Longleftrightarrow$ (iv).

3.40. *Proximité d'une solution et petitesse d'un sous-gradient.* Soient $\mathbb{E}$ un espace euclidien de dimension finie et $f\in\mathrm{C}\overline{\mathrm{onv}}(\mathbb{E})$. On suppose que l'ensemble $\mathcal{S}$ des minimiseurs de f est non vide et borné. Montrez que, quel que soit $r>0$, il existe un voisinage V de $0\in\mathbb{E}$ tel que $\mathrm{d}_{\mathcal{S}}(x)\leqslant r$ lorsque $x^*\in\partial f(x)\cap V$. Montrez par un contre-exemple que ce résultat ne tient plus si $\mathcal{S}$ n'est pas borné.

Remarque. Il n'y a pas de réciproque : x peut être proche de $\mathcal{S}$ sans qu'il y ait un petit sous-gradient en x. La valeur absolue $f=|\cdot|$ sur $\mathbb{R}$ en est un contre-exemple : $|f'(x)|=1$ quel que soit $x\neq 0$.

3.41. *Croissance quadratique locale* (adapté de [523, 524]). Soient $\mathbb{E}$ et $\mathbb{F}$ deux espaces normés. On dit qu'une multifonction $T:\mathbb{E}\rightrightarrows\mathbb{F}$ est *localement radialement lipschitzienne de module* $L\geqslant 0$ en un point $x_0\in\mathbb{E}$ s'il existe un voisinage V de x_0 tel que, pour tout $x\in V$ et pour tout $y\in T(x)$, on a $\mathrm{d}_{T(x_0)}(y)\leqslant L\|x-x_0\|$. On considère à présent, un espace euclidien $\mathbb{E}$ de dimension finie et une fonction $f\in\mathrm{C}\overline{\mathrm{onv}}(\mathbb{E})$ ayant un ensemble de minimiseurs $\mathcal{S}$ non vide et borné. Montrez que les propriétés suivantes sont équivalentes ($L>0$).

(i) ∂f^{-1} est localement radialement lipschitzienne de module L en 0,
(ii) il existe un voisinage V de $0\in\mathbb{E}$ tel que $\mathrm{d}_{\mathcal{S}}(x)\leqslant L\|x^*\|$ lorsque $x^*\in\partial f(x)\cap V$,
(iii) il existe une constante $\alpha>0$ et un rayon $r>0$ tels que pour tout $x\in\mathcal{S}+\bar{B}_r$, on a $f(x)\geqslant\inf f+\alpha(\mathrm{d}_{\mathcal{S}}(x))^2$.

Si (i) ou (ii) a lieu et s'il y a un minimum unique, on peut prendre $\alpha=1/(2L)$ dans (iii). Si (iii) a lieu, on peut prendre $L=1/\alpha$ dans (i) ou (ii).

3.42. *Point proximal limite.* Soient $\mathbb{E}$ un espace euclidien de dimension finie, $f\in\mathrm{C}\overline{\mathrm{onv}}(\mathbb{E})$ et $r>0$. On note $x_p(r)$ le point proximal de $x\in\mathbb{E}$, défini comme l'unique solution de

$$\inf_{y\in\mathbb{E}}\left(f(y)+\frac{1}{2r}\|y-x\|^2\right).$$

Montrez que lorsque $r \uparrow \infty$:

(*i*) $\|x_p(r) - x\|$ croît,
(*ii*) $f(x_p(r))$ décroît,
(*iii*) si $\arg\min f \neq \varnothing$, $x_p(r)$ converge vers la projection de x sur $\arg\min f$.

3.43. *Proximalité pondérée.* Cet exercice propose d'étendre la notion de point proximal au cas d'une pénalisation quadratique générale. Soient $\mathbb{E}$ un espace euclidien (produit scalaire $\langle\cdot,\cdot\rangle$ et norme associée $\|\cdot\|$) et $f \in \mathrm{C}\overline{\mathrm{onv}}(\mathbb{E})$. On se donne un opérateur auto-adjoint défini positif M auquel on associe un produit scalaire $\langle\cdot,\cdot\rangle_M = \langle M\cdot,\cdot\rangle$ et une norme $\|\cdot\|_M = \langle\cdot,\cdot\rangle_M^{1/2}$. Pour x donné dans $\mathbb{E}$, le point proximal, toujours noté x_p, est ici défini comme l'unique solution de

$$\inf_{y\in\mathbb{E}} \left(f(y) + \frac{1}{2}\|y-x\|_M^2 \right). \tag{3.90}$$

On note $P_{f,M} : \mathbb{E} \to \mathbb{E} : x \mapsto x_p$ l'application proximale pondérée par M. Le sous-différentiel ∂f est calculé par rapport au produit scalaire non pondéré $\langle\cdot,\cdot\rangle$. Montrez que

(*i*) $\tilde{x} = P_{f,M}(x) \iff \exists\, \tilde{g} \in \partial f(\tilde{x}) : \tilde{x} = x - M^{-1}\tilde{g}$,
(*ii*) $I - P_{f,M} = M^{-1} \circ P_{f^*,M^{-1}} \circ M$,
(*iii*) pour tout x et $y \in \mathbb{E}$, $\langle y_p - x_p, y - x\rangle_M \geqslant \|y_p - x_p\|_M^2$,
(*iv*) pour tout x et $y \in \mathbb{E}$, $\|y_p - x_p\|_M^2 + \|g_p^y - g_p^x\|_{M^{-1}}^2 \leqslant \|y-x\|_M^2$, où $g_p^x = M^{-1}(x - x_p)$ et $g_p^y = M^{-1}(y - y_p)$.

3.44. *Régularisée de Moreau-Yosida d'une fonction quadratique convexe.* On suppose que $f : x \in \mathbb{E} \mapsto f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle$ avec A symétrique définie positive et on considère sa régularisée de Moreau-Yosida

$$x \in \mathbb{E} \mapsto \tilde{f}(x) = \inf_{y\in\mathbb{E}} \left(f(y) + \frac{1}{2}\|y-x\|_M^2 \right),$$

où $\|\cdot\|_M = \langle M\cdot,\cdot\rangle^{1/2}$ et M est une matrice symétrique définie positive. Montrez que $\tilde{f}$ est quadratique et que sa hessienne s'écrit $\nabla^2\tilde{f} = (A^{-1} + M^{-1})^{-1}$.

4 Conditions d'optimalité

[Cette méthode] *ne trompe jamais et peut s'étendre à des questions très belles.*

P. de Fermat (1601-1665).

On ne trouvera point de Figures dans cet Ouvrage. Les méthodes que j'y expose ne demandent ni constructions, ni raisonnements géométriques ou méchaniques, mais seulement des opérations algébriques, assujetties à une marche régulière et uniforme. Ceux qui aiment l'Analyse verront avec plaisir la Méchanique en devenir une nouvelle branche, et me sauront gré d'en avoir étendu ainsi le domaine.

J.-L. de la Grange. *Méchanique Analytique.* [372 ; 1788, pages XI-XII]

Considérons le problème d'optimisation sous forme générale

$$(P_X) \quad \begin{cases} \min f(x) \\ x \in X, \end{cases}$$

dans lequel on minimise une fonction $f : X \to \mathbb{R}$ définie sur une partie X d'un espace euclidien $\mathbb{E}$. On note $\langle \cdot, \cdot \rangle$ le produit scalaire de cet espace vectoriel et

$$n := \dim \mathbb{E}$$

sa dimension (finie).

Rappelons quelques notions relatives au problème (P_X), qui ont été introduites aux sections 1.1 et 1.2 et qui seront continuellement utilisées dans ce chapitre. L'ensemble X sur lequel on minimise f est appelé l'*ensemble admissible* du problème. Un *minimum* (*global*) de ce problème est un point $x_* \in X$ tel que

$$\forall x \in X : \qquad f(x_*) \leqslant f(x). \tag{4.1}$$

On dit que $x_* \in X$ est un *minimum local* de (P_X) s'il existe un voisinage V de x_* tel que

$$\forall x \in X \cap V : \qquad f(x_*) \leqslant f(x). \tag{4.2}$$

On parle de minimum global (resp. local) *strict* si l'on a inégalité stricte dans l'expression (4.1) (resp. (4.2)), lorsque $x \neq x_*$.

Les *conditions d'optimalité* sont des équations, des inéquations ou des propriétés que vérifient les solutions de (P_X) (*conditions nécessaires*, CN) ou qui assurent à un point d'être solution de (P_X) (*conditions suffisantes*, CS). Elles traduisent ainsi l'expression (4.2) de l'optimalité locale en une forme analytique plus directement utilisable. Ces conditions sont utiles pour de nombreuses raisons, par exemple :

- pour vérifier l'optimalité éventuelle d'un point $x \in X$, voir si c'est un minimum, un maximum ou un point stationnaire (voir plus loin pour la définition de ce terme),
- pour calculer analytiquement des solutions de (P_X),
- pour mettre en œuvre des méthodes numériques permettant de trouver des solutions (P_X),
- pour définir des tests d'arrêt des itérations dans les algorithmes de résolution de (P_X).

On parlera de *conditions du premier ordre* (notées CN1 ou CS1, selon qu'il s'agit de conditions nécessaires ou suffisantes) lorsque celles-ci ne font intervenir que les dérivées premières de f et des fonctions définissant l'ensemble admissible. Quant aux *conditions du second ordre* (notées CN2 ou CS2), elles ne font intervenir que les dérivées premières et secondes de ces fonctions.

Nous commencerons par une condition nécessaire d'optimalité du premier ordre très générale puisqu'elle concerne le problème (P_X) ; elle est de nature géométrique (section 4.1). C'est cette condition qui sera ensuite traduite en une forme analytique pour les problèmes sans contrainte (section 4.2), les problèmes avec contraintes d'égalité (section 4.3), les problèmes avec contraintes d'égalité et d'inégalité (section 4.4) et enfin les problèmes avec contraintes générales (section 4.5). Dans ces deux dernières sections, c'est le lemme de Farkas qui permet de faire cette traduction. Les résultats obtenus dans ce dernier cas permettent de retrouver les résultats précédents. Dès lors, le lecteur pressé et courageux pourra passer directement de la section 4.1 à la section 4.5. Rien ne sera perdu quant au contenu, mais on gagnera en compréhension en se penchant sur les situations plus simples décrites aux sections 4.2, 4.3 et 4.4.

Connaissances supposées. Notion de cône dual et lemme de Farkas. Les conditions du deuxième ordre pour le problème avec contraintes d'égalité et d'inégalité (section 4.4.4) requièrent la connaissance de la dualité linéaire, qui ne sera vue qu'au chapitre 17 ! Cette dernière section peut être passée en première lecture.

4.1 Une condition nécessaire d'optimalité géométrique

Ce sont les conditions nécessaires d'optimalité qui sont les plus difficiles à établir. Ceci est dû au fait que l'on procède par linéarisation de l'ensemble admissible, ce qui conduit au *cône tangent* (section 4.1.1) – jusqu'ici tout va bien – mais que ce cône tangent est *approché* par le *cône linéarisant*, obtenu en linéarisant les fonctions qui décrivent l'ensemble admissible. Cette approximation bien utile n'est pas toujours adéquate et il faut ce que l'on appellera des *conditions de qualification* pour que cette démarche puisse aboutir.

Les conditions suffisantes d'optimalité s'obtiennent, quant à elles, généralement sans difficulté ; soit à partir des conditions nécessaires du premier ordre, lorsque le

problème est convexe (il faudra spécifier ce que l'on entend par là), soit en renforçant légèrement les conditions nécessaires du second ordre dans le cas général.

4.1.1 Cônes tangent et normal

Dans cette section, on suppose que $\mathbb{E}$ désigne un espace vectoriel normé de dimension finie. Sa norme est notée $\|\cdot\|$.

Le cône tangent est un concept essentiel qui aide à l'écriture des conditions d'optimalité. Il sert à linéariser l'ensemble admissible au point optimal x_*, comme on le fait avec le critère en optimisation sans contrainte pour exprimer l'optimalité de x_* par l'équation $f'(x_*) = 0$. Comme nous le verrons (proposition 4.3), la notion de cône tangent définie ci-après étend aux ensembles quelconques celle introduite pour les ensembles convexes à la section 2.5.6. Nous utiliserons essentiellement la notion de tangence suivante. On note

$$\mathbb{R}_{++} := \{t \in \mathbb{R} : t > 0\}.$$

Définitions 4.1 (cône tangent) Soient $X \subseteq \mathbb{E}$ et $x \in \mathbb{E}$. On dit que le vecteur $d \in \mathbb{E}$ est *tangent* à X en x s'il existe une suite $\{d_k\} \subseteq \mathbb{E}$ et une suite $\{t_k\} \subseteq \mathbb{R}_{++}$ telles que

$$d_k \to d, \quad t_k \downarrow 0, \quad x + t_k d_k \in X. \tag{4.3}$$

On note $\mathrm{T}_x X$ ou $\mathrm{T}_X(x)$ l'ensemble des vecteurs tangents à X en x et on l'appelle le *cône tangent*. □

Pour le distinguer d'autres notions de cône tangent, celui de la définition précédente est parfois appelé le *cône contingent* [74 ; § 2.2.4].

La définition précédente a une interprétation géométrique simple : pour que la direction d soit tangente à X en x, il faut qu'elle soit limite de directions d_k, telles que $x + \mathbb{R}_{++} d_k$ rencontre X en un point qui se rapproche de x lorsque $k \to \infty$. Le passage à la limite est essentiel, sous peine d'appauvrir radicalement $\mathrm{T}_x X$ et de le rendre ainsi inutilisable. Il ne suffirait pas en effet de prendre comme cône tangent, ce que l'on nomme le *cône des directions admissibles* en x, qui est défini et noté

$$\mathrm{T}_x^a X := \{d \in \mathbb{E} : x + td \in X \text{ pour tout } t > 0 \text{ petit}\}. \tag{4.4}$$

Si ce cône nous a servi de concept de départ pour définir le cône tangent à un ensemble convexe (définition 2.52), il est de peu d'utilité ici. Par exemple, si X est la sphère unité de $\mathbb{R}^n$ et x un point de celle-ci, $\mathrm{T}_x^a X = \varnothing$ alors que $\mathrm{T}_x X$ est un sous-espace vectoriel de dimension $n - 1$. En l'absence de convexité, le cône tangent diffère donc en général de $\mathrm{T}_x^a X$ ou de son adhérence.

La définition 4.1 d'une direction tangente est facile à interpréter, mais elle n'est pas la plus souvent utilisée. Il s'avérera plus commode de faire jouer le rôle principal à la suite $\{x_k\}$ définie par $x_k = x + t_k d_k$, plutôt qu'à la suite $\{d_k\}$. On voit aisément que

$$d \in \mathrm{T}_x X \quad \Longleftrightarrow \quad \exists\, \{x_k\} \subseteq X, \quad \exists\, \{t_k\} \downarrow 0 : \quad \frac{x_k - x}{t_k} \to d. \tag{4.5}$$

Forcément, $x_k \to x$. Voici un premier exemple d'utilisation de cette définition équivalente.

Proposition 4.2 (cône tangent fermé) *Le cône tangent est un fermé.*

DÉMONSTRATION. Soit $\{d_k\} \subseteq \mathrm{T}_x X$, avec $d_k \to d$. Il faut montrer que $d \in \mathrm{T}_x X$. Pour tout $k \geqslant 1$, il existe une suite $\{x_{k,j}\}_j \subseteq X$ et une suite $\{t_{k,j}\}_j \downarrow 0$ telles que $(x_{k,j} - x)/t_{k,j} \to d_k$ lorsque $j \to \infty$. On construit à présent des suites par le procédé diagonal. Pour tout $i \geqslant 1$, on détermine d'abord k_i tel que $\|d_{k_i} - d\| \leqslant \frac{1}{i}$ et ensuite j_i tel que $\|(x_{k_i,j_i} - x)/t_{k_i,j_i} - d_{k_i}\| \leqslant \frac{1}{i}$ et $t_{k_i,j_i} \leqslant \frac{1}{i}$. Clairement $\{x_{k_i,j_i}\}_i \subseteq X$, $t_{k_i,j_i} \downarrow 0$ et $(x_{k_i,j_i} - x)/t_{k_i,j_i} \to d$ lorsque $i \to \infty$, si bien que $d \in \mathrm{T}_x X$. □

La proposition suivante montre que la notion de cône tangent introduite ici généralise aux ensembles quelconques celle introduite à la section 2.5.6 pour les ensembles convexes.

Proposition 4.3 (cône tangent à un convexe) *On a*

$$\mathrm{T}_x X \subseteq \overline{\mathbb{R}_+(X - x)}, \tag{4.6}$$

avec égalité si X est un ensemble convexe et $x \in X$, auquel cas $\mathrm{T}_x X$ est convexe.

DÉMONSTRATION. L'inclusion est claire, car une direction tangente est limite d'éléments de la forme $(x_k - x)/t_k \in \mathbb{R}_{++}(X - x)$.

Supposons à présent que X soit convexe et que $x \in X$. Comme le cône tangent est fermé (proposition 4.2), l'égalité sera prouvée en (4.6) si l'on montre que $\mathbb{R}_+(X - x) \subseteq \mathrm{T}_x X$. Soient $y \in X$ et $\alpha \geqslant 0$; il s'agit de montrer que $\alpha(y - x) \in \mathrm{T}_x X$. On prend une suite $\{t_k\} \downarrow 0$ formée de scalaires inférieurs à $1/\alpha$ (à $+\infty$ si $\alpha = 0$). Alors $x_k := x + t_k\alpha(y - x) = (1 - t_k\alpha)x + t_k\alpha y \in X$ parce que x et $y \in X$, ensemble supposé convexe. Il reste à observer que $(x_k - x)/t_k$ est le vecteur constant $\alpha(y - x)$, si bien que celui-ci est dans $\mathrm{T}_x X$.

La convexité de $\mathrm{T}_x X$ lorsque X est convexe et $x \in X$ se déduit de celle de $\overline{\mathbb{R}_+(X - x)}$. □

La proposition précédente montre que la définition 4.1 du cône tangent coïncide avec la définition 2.53 lorsque X est convexe et $x \in X$. Cependant, si X n'est pas convexe, $\mathrm{T}_x X$ n'est pas nécessairement convexe. Voici un exemple.

Exemple 4.4 ($\mathrm{T}_x X$ non convexe) L'ensemble non convexe $X := \{(x_1, x_2) \in \mathbb{R}^2 : |x_1| \leqslant |x_2|\}$ a son cône tangent en $(0, 0)$ qui s'écrit $\mathrm{T}_{(0,0)} X = \{d \in \mathbb{R}^2 : |d_1| \leqslant |d_2|\}$, Celui-ci n'est pas convexe. □

On vérifie que

$$\begin{aligned} x \notin \overline{X} &\implies \mathrm{T}_x X = \varnothing, \\ x \in \operatorname{int} X &\implies \mathrm{T}_x X = \mathbb{E}, \\ x \in \operatorname{intr} X &\implies \mathrm{T}_x X = (\operatorname{aff} X) - (\operatorname{aff} X). \end{aligned}$$

Les cas non triviaux se présentent donc lorsque x est sur la frontière de X. On a représenté à la figure 4.1 quelques exemples de cônes tangents.

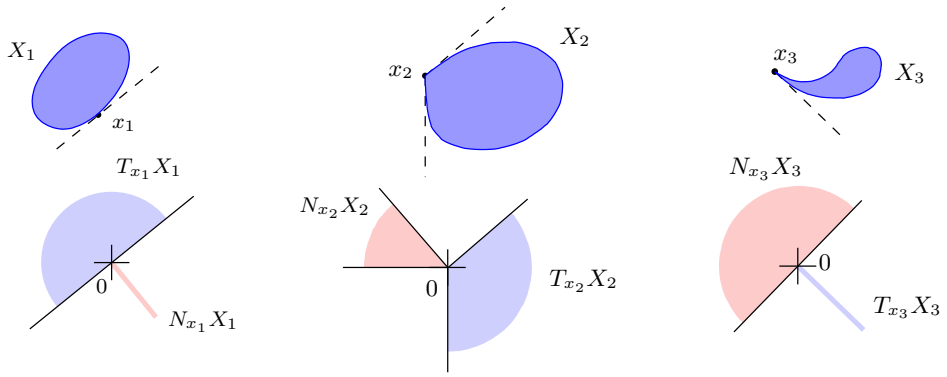


Fig. 4.1. Cônes tangent (en bleu clair) et normal (en rose) à quelques ensembles (en haut)

On verra que si X est donné par des contraintes fonctionnelles d'égalité et d'inégalité et si des *conditions de qualification de contraintes* sont vérifiées, le cône tangent se calcule aisément en utilisant les gradients des contraintes fonctionnelles actives en x (voir la section 4.4.2).

On note que $\mathrm{T}_x X$ ne dépend de X que dans un voisinage de x : si l'on modifie X en dehors d'un voisinage de x, on ne change pas $\mathrm{T}_x X$ (il suffit de ne retenir dans la propriété (4.5) que les x_k qui sont dans ce voisinage). D'autre part, $\mathrm{T}_x X$ ne dépend pas de la manière utilisée pour définir X. En particulier, si X est défini par des contraintes fonctionnelles (voir plus loin), le cône tangent ne dépend pas du choix des fonctions utilisées pour le définir.

Définitions 4.5 (cône normal) Soient $\mathbb{E}$ un espace euclidien (produit scalaire noté $\langle \cdot, \cdot \rangle$), $X \subseteq \mathbb{E}$ et $x \in \mathbb{E}$. On dit que $p \in \mathbb{E}$ est *normal* à X en x si

$$\forall d \in \mathrm{T}_x X \ : \quad \langle p, d \rangle \leqslant 0. \tag{4.7}$$

On note $\mathrm{N}_x X$ l'ensemble des vecteurs normaux à X en x et on l'appelle le *cône normal*. □

Cette définition du cône normal est simple mais un peu restrictive selon Rockafellar [526 ; 1993 ; p. 193], qui l'appelle *cône normal régulier* ; elle suffira pour notre propos.

Clairement, le cône normal est le dual négatif du cône tangent :

$$\mathrm{N}_x X = (\mathrm{T}_x X)^- = -(\mathrm{T}_x X)^+ \tag{4.8}$$

et, contrairement à $\mathrm{T}_x X$, $\mathrm{N}_x X$ est donc toujours convexe (point 1 de la proposition 2.43). D'après le point 5 de la proposition 2.43, un cône est contenu dans son bidual et donc

$$\mathrm{T}_x X \subseteq (\mathrm{N}_x X)^-. \tag{4.9}$$

On montre que (voir l'exercice 2.34)

$$X \text{ convexe} \quad \Longrightarrow \quad \mathrm{T}_x X = (\mathrm{N}_x X)^-.$$

Si X n'est pas convexe, $\mathrm{T}_x X$ peut ne pas être convexe (exemple 4.4) ; donc l'égalité ci-dessus peut ne pas avoir lieu car $(\mathrm{N}_x X)^-$ est toujours convexe (point 1 de la proposition 2.43). Remarquons aussi que, lorsque X est convexe, on retrouve le cône normal défini à la section 2.5.2 (voir l'exercice 4.2).

4.1.2 Condition nécessaire de Peano-Kantorovitch

Commençons donc par considérer le problème général

$$(P_X) \quad \left\{ \begin{array}{l} \min f(x) \\ x \in X, \end{array} \right. \tag{4.10}$$

où X est une partie de l'espace vectoriel $\mathbb{E}$. La condition nécessaire d'optimalité du premier ordre suivante met en évidence l'utilité des notions de cônes tangent et *normal*.

Théorème 4.6 (CN1 de Peano-Kantorovitch) *Si x_* est un minimum local de (P_X) et si f est dérivable en x_*, on a*

$$\forall d \in \mathrm{T}_{x_*} X \ : \quad f'(x_*) \cdot d \geqslant 0. \tag{4.11}$$

Ceci s'écrit aussi

$$\boxed{\nabla f(x_*) \in (\mathrm{T}_{x_*} X)^+} \quad \textit{ou} \quad \boxed{\nabla f(x_*) + \mathrm{N}_{x_*} X \ni 0,} \tag{4.12}$$

où $\nabla f(x_*)$ est le gradient de f en x_* pour le produit scalaire de $\mathbb{E}$, $(\cdot)^+$ désigne le dual pour ce même produit scalaire et $\mathrm{N}_{x_*} X$ est le cône normal à X en x_*.

Démonstration. Soit $d \in \mathrm{T}_{x_*} X$ non nul (la relation (4.11) est triviale si $d = 0$). Alors il existe des suites $\{x_k\} \subseteq X$ et $\{t_k\} \downarrow 0$ telles que $d_k := \frac{x_k - x_*}{t_k} \to d$. Pour k grand, $x_k = x_* + t_k d_k \in X$ est voisin de x_*, si bien que par l'optimalité locale,

$$\text{pour } k \text{ grand} : \quad f(x_* + t_k d_k) \geqslant f(x_*).$$

Si f est dérivable en x_*, $f(x_* + t_k d_k) = f(x_*) + f'(x_*) \cdot (t_k d_k) + o(\|t_k d_k\|)$, donc

$$0 \leqslant f'(x_*) \cdot d_k + \frac{o(\|t_k d_k\|)}{t_k} = f'(x_*) \cdot d_k + \frac{o(\|t_k d_k\|)}{\|t_k d_k\|} \|d_k\|.$$

En passant à la limite quand $k \to \infty$, on obtient (4.11). Le reste s'en déduit. □

La démarche qui sera suivie dans ce chapitre pour obtenir des conditions nécessaires d'optimalité dans le cas où X est donné par des contraintes fonctionnelles sera de trouver une expression de (4.12) plus accessible au calcul. Dans chaque cas, nous aurons à calculer le cône normal $\mathrm{N}_{x_*} X$ et donc le cône tangent $\mathrm{T}_{x_*} X$.

4.1.3 Problème avec convexité

Lorsque l'ensemble admissible est convexe, on peut écrire une condition d'optimalité du premier ordre sans utiliser la notion de cône tangent ; c'est ce que nous faisons dans cette section. Cette condition ressemble très fort à la CN1 générale (4.11), quoique plus simple à écrire et à démontrer. Elle ne fait pas intervenir le cône tangent, comme nous l'annoncions, mais le *cône des directions admissibles* défini par (4.4). Bien que ce dernier soit plus petit que $\mathrm{T}_{x_*} X$, il n'y a pas de perte d'information lorsque f est dérivable en x_*, parce que son adhérence est $\mathrm{T}_{x_*} X$ (voir la définition 2.53) et que $d \mapsto f'(x_*) \cdot d$ est continue.

Proposition 4.7 (CN1 et CS1 en présence de convexité) *Supposons que X soit convexe et que f ait des dérivées directionnelles en un point $x_* \in X$. Si x_* est un minimum local de (P_X), on a*

$$\forall x \in X \ : \quad f'(x_*; x - x_*) \geqslant 0. \tag{4.13}$$

Inversement, si f est convexe sur le convexe X et si (4.13) a lieu, alors x_* est un minimum global de (P_X).

Démonstration. Pour $t > 0$ petit et $x \in X$, $x_* + t(x - x_*) \in X$ (convexité de X) et si x_* est un minimum local de (P_X), on a

$$\frac{f(x_* + t(x - x_*)) - f(x_*)}{t} \geqslant 0.$$

En passant à la limite lorsque $t \downarrow 0$, on obtient (4.13).

Si f est convexe sur X, on a pour tout $x \in X$ (proposition 3.19) :

$$f(x) \geqslant f(x_*) + f'(x_*; x - x_*).$$

Si (4.13) est vérifiée, on a alors $f(x) \geqslant f(x_*)$, ce qui montre que x_* est un minimum global de f sur X. □

La discussion et le résultat de cette section conduisent naturellement à la définition suivante de problème (P_X) convexe.

Définition 4.8 ((P_X) convexe) On dit qu'un problème d'optimisation de la forme (P_X), défini en (4.10), est *convexe* si son ensemble admissible X est une partie convexe de l'espace vectoriel $\mathbb{E}$ et si son critère f est convexe sur X. □

4.2 Problème sans contrainte

Le problème considéré dans cette section ne présente pas de contrainte. On l'écrit

$$\boxed{\min_{x \in \mathbb{E}} f(x),} \tag{4.14}$$

où $f : \mathbb{E} \to \mathbb{R}$ est supposée définie sur une espace euclidien $\mathbb{E}$.

4.2.1 Condition de Fermat

Pour le problème sans contrainte (4.14), l'ensemble admissible $X = \mathbb{E}$ est convexe. On déduit alors immédiatement de la proposition 4.7 une condition nécessaire d'optimalité du premier ordre et, pour les problèmes convexes, une condition suffisante d'optimalité du premier ordre.

Théorème 4.9 (CN1 de Fermat, CS1) *Supposons que $f : \mathbb{E} \to \mathbb{R}$ soit dérivable en $x_* \in \mathbb{E}$. Si f a un minimum local en x_*, alors*

$$f'(x_*) = 0 \quad ou \quad \nabla f(x_*) = 0. \tag{4.15}$$

Inversement, si f est convexe et si (4.15) a lieu, alors x_* est un minimum global de f sur $\mathbb{E}$.

4.2.2 Conditions d'optimalité du second ordre

Les deux résultats suivants donnent des conditions nécessaires puis des conditions suffisantes du second ordre. Dans ces propositions, la hessienne $\nabla^2 f(x_*)$ et sa (semi-)définie positivité sont associées au produit scalaire que l'on s'est donné sur $\mathbb{E}$. Une solution x_* vérifiant les conditions suffisantes d'optimalité du second ordre de la proposition 4.11 est appelé un *minimum fort* du problème (4.14).

Proposition 4.10 (CN2) *Supposons que x_* soit un minimum local de f sur $\mathbb{E}$ et que f soit dérivable dans un voisinage de x_* et deux fois dérivable en x_*. Alors*

$$\nabla f(x_*) = 0 \quad et \quad \nabla^2 f(x_*) \text{ est semi-définie positive.}$$

Démonstration. Soient $d \in \mathbb{E}$, $\{t_k\} \subseteq \mathbb{R}$ une suite telle que $t_k \downarrow 0$ et $x_k := x_* + t_k d$. Avec la régularité supposée de f on a (voir le théorème C.18)

$$f(x_k) = f(x_*) + f'(x_*) \cdot (x_k - x_*) + \frac{1}{2} f''(x_*) \cdot (x_k - x_*)^2 + o(\|x_k - x_*\|^2). \tag{4.16}$$

Comme $f'(x_*) = 0$ et que $f(x_k) \geqslant f(x_*)$ pour k suffisamment grand, on voit que

$$0 \leqslant \frac{1}{2} f''(x_*) \cdot (t_k d)^2 + o(\|t_k d\|^2), \quad \text{pour } k \text{ grand.}$$

En divisant par $t_k^2 > 0$ et en faisant tendre $k \to \infty$, on en déduit que $0 \leqslant \frac{1}{2} f''(x_*) \cdot d^2 = \langle \nabla^2 f(x_*) d, d \rangle$. Donc $\nabla^2 f(x_*)$ est semi-définie positive. □

Proposition 4.11 (CS2) *Si f est dérivable dans le voisinage d'un point x_* et deux fois dérivable en x_* et si*

$$\nabla f(x_*) = 0 \quad \textit{et} \quad \nabla^2 f(x_*) \textit{ est définie positive,}$$

alors x_* est un minimum local strict de f.

Démonstration. On raisonne par l'absurde. Si le résultat n'est pas vrai, il existe une suite $\{x_k\} \subseteq \mathbb{E}$ telle que $x_k \to x_*$, $x_k \neq x_*$ et $f(x_k) \leqslant f(x_*)$. En extrayant une sous-suite au besoin, on peut supposer qu'avec $t_k := \|x_k - x_*\| \neq 0$,

$$\frac{x_k - x_*}{t_k} \to d.$$

Clairement, $d \neq 0$.

De (4.16) et $\nabla f(x_*) = 0$, on déduit que

$$0 \geqslant \frac{1}{2} f''(x_*) \cdot (x_k - x_*)^2 + o(\|x_k - x_*\|^2).$$

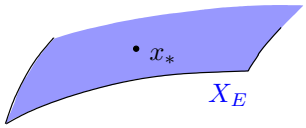
En divisant par t_k^2 et en passant à la limite, on obtient $0 \geqslant f''(x_*) \cdot d^2 = \langle \nabla^2 f(x_*) d, d \rangle$. Comme $d \neq 0$, ceci est en contradiction avec la définie positivité supposée de $\nabla^2 f(x_*)$. □

On peut donner une démonstration plus courte du résultat précédent. L'avantage de celle que nous avons utilisée est d'encore fonctionner pour les problèmes avec contraintes. Ce fut donc un bon entraînement.

On peut donner une condition suffisante d'optimalité *diffuse* (voir l'exercice 4.4 pour un énoncé précis) : si $\nabla f(x_*) = 0$ et $\nabla^2 f(x)$ est semi-définie positive pour x voisin de x_*, alors x_* est un minimum local de (4.14). On peut donc relaxer l'hypothèse de définie positivité de $\nabla^2 f(x_*)$ pourvu que l'on ait la *semi*-définie positivité de la hessienne *dans un voisinage* de x_*. Sous ces hypothèses, x_* n'est plus nécessairement un minimum local *strict*, comme le montre le cas d'un fonction constante.

4.3 Problème avec contraintes d'égalité

On considère dans cette section un problème d'optimisation dans lequel l'ensemble admissible n'est pas l'espace $\mathbb{E}$ tout entier mais une partie X_E de celui-ci, définie par un nombre fini de contraintes d'égalité :



$$X_E := \{x \in \mathbb{E} : c(x) = 0\}.$$

Ces contraintes sont donc spécifiées au moyen d'une fonction

$$c : \mathbb{E} \to \mathbb{F},$$

où $\mathbb{F}$ est également un espace euclidien (de dimension finie) dont le produit scalaire est aussi noté $\langle \cdot, \cdot \rangle$. Les dimensions des espaces sont notées

$$n := \dim \mathbb{E} \qquad \text{et} \qquad m := \dim \mathbb{F}.$$

Il sera souvent approprié de supposer qu'en la solution x_* recherchée, la *jacobienne* $c'(x_*)$ de la contrainte vérifie

$$c'(x_*) \text{ est surjective.}$$

Ceci requiert certainement d'avoir $m \leqslant n$, c'est-à-dire d'avoir moins de contraintes que de variables à optimiser. Lorsque cette hypothèse est vérifiée, X_E est, dans un voisinage de x_*, une *variété* (concept de base de la géométrie différentielle que l'on peut voir comme une surface ayant des propriétés de représentation particulières) de dimension $n - m$ et l'image qui en est donnée dans la figure ci-dessus en est une idéalisation acceptable lorsque $n = 3$ et $m = 1$. Le problème d'optimisation est quant à lui noté

$$\boxed{(P_E) \quad \begin{cases} \min f(x) \\ c(x) = 0, \end{cases}}$$

où $f : \mathbb{E} \to \mathbb{R}$ en est le critère.

L'adaptation au problème (P_E) de la définition 4.8 de problème (P_X) convexe conduit à la définition suivante.

Définition 4.12 ((P_E) convexe) On dit que le problème (P_E) est *convexe* si son ensemble admissible X_E est convexe et si son critère f est convexe sur X_E. □

Il est plus courant cependant de requérir l'affinité de X_E pour que le problème (P_E) soit convexe, ce que n'assure pas la définition précédente. Comme le montre la proposition suivante, c'est ce qui se produit lorsque c_E est affine.

Proposition 4.13 (convexité de X_E) *Si c est affine, alors X_E est un sous-espace affine (donc un convexe).*

Démonstration. Soient x_0 et $x_1 \in X_E$ et $t \in \mathbb{R}$. Il s'agit de montrer que $(1-t)x_0 + tx_1 \in X_E$, c'est-à-dire que $c((1-t)x_0 + tx_1) = 0$. On a

$$\begin{aligned} c((1-t)x_0 + tx_1) &= (1-t)c(x_0) + tc(x_1) \qquad [\text{affinité de } c] \\ &= 0 \qquad [c(x_0) = 0 \text{ et } c(x_1) = 0]. \end{aligned}$$

□

Mais X_E peut être un sous-espace affine sans que c soit affine. C'est le cas, par exemple, de $X_E = \{0\} \subseteq \mathbb{R}$ défini par la contrainte $c : x \in \mathbb{R} \mapsto x + x^3 \in \mathbb{R}$.

4.3.1 Conditions de Lagrange

Pour expliciter la condition d'optimalité (4.12) au cas du problème (P_E), il faut calculer le cône tangent à son ensemble admissible X_E.

Proposition 4.14 (cône tangent à X_E) *Supposons que $c : \mathbb{E} \to \mathbb{F}$ soit dérivable en $x \in X_E$. Alors*

$$\mathrm{T}_x\, X_E \subseteq \mathcal{N}(c'(x)). \tag{4.17}$$

Si, de plus, $c'(x)$ est surjective et si c est C^1 dans un voisinage de x, alors l'égalité a lieu en (4.17).

Démonstration. Soit $d \in \mathrm{T}_x\, X_E$. Alors il existe des suites $\{x_k\} \subseteq X_E$ et $\{t_k\} \subseteq \mathbb{R}_{++}$ telles que l'on ait (4.5). Comme c est dérivable en $x \in X_E$, on a

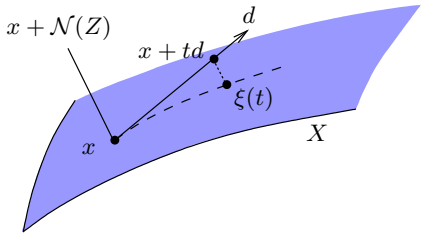
$$0 = c(x_k) = c'(x) \cdot (x_k - x) + o(\|x_k - x\|).$$

En divisant par t_k et en passant à la limite lorsque $t_k \downarrow 0$, on trouve que $d \in \mathcal{N}(c'(x))$.

Supposons à présent que $c'(x)$ soit surjective et que c soit C^1 dans un voisinage de x. Soit $d \in \mathcal{N}(c'(x))$. Pour montrer que $d \in \mathrm{T}_x\, X_E$, il faut construire des suites $\{x_k\} \subseteq X_E$ et $\{t_k\} \subseteq \mathbb{R}$ vérifiant (4.5). La suite $\{x_k\}$ s'obtiendra en échantillonnant un chemin $\xi : t \mapsto \xi(t)$ de X_E passant par x et admettant d comme tangente en x. Soit $A := c'(x)$ la jacobienne de c en x. Comme A est surjective, il existe un espace vectoriel $\mathbb{G}$ et une application linéaire $Z : \mathbb{E} \to \mathbb{G}$, tels que

$$\begin{pmatrix} A \\ Z \end{pmatrix} : \mathbb{E} \to \mathbb{F} \times \mathbb{G} \text{ soit bijective.} \tag{4.18}$$

On peut, par exemple, prendre pour Z le projecteur orthogonal de $\mathbb{E}$ sur $\mathbb{G} = \mathcal{N}(A)$. Observons que $\mathcal{N}(A)$ et $\mathcal{N}(Z)$ sont alors deux sous-espaces vectoriels *supplémentaires*. On détermine $\xi(t)$ dans X_E, qui vérifie donc $c(\xi(t)) = 0$, en l'écrivant comme la somme de $x + td \in x + \mathcal{N}(A)$ et d'un déplacement dans $\mathcal{N}(Z)$, donc $Z(\xi(t) - x - td) = 0$; voir la figure ci-dessous. Ceci justifie le fait de vouloir annuler la fonction $F : \mathbb{E} \times \mathbb{R} \to \mathbb{F} \times \mathbb{G}$ définie par



$$F(\xi, t) = \begin{pmatrix} c(\xi) \\ Z(\xi - x - td) \end{pmatrix}.$$

Clairement $F(x, 0) = 0$. Par ailleurs, F est de classe C^1 dans un voisinage de $(x, 0)$ et $F'_\xi(x, 0)$ est l'application linéaire de (4.18), qui est inversible. On peut alors appliquer le théorème des fonctions implicites (théorème C.14) : il existe une fonction $t \in \mathbb{R} \mapsto \xi(t) \in \mathbb{E}$ définie dans un voisinage de 0, de classe C^1, telle que

$$\begin{cases} F(\xi(t),t)=0, & \text{pour tout } t \text{ voisin de } 0\\ \xi(0)=x. \end{cases}$$

En dérivant $F(\xi(t),t)=0$ en $t=0$ (théorèmes C.5 et C.6), on trouve

$$\begin{pmatrix} A\\ Z \end{pmatrix} \xi'(0) = \begin{pmatrix} 0\\ Zd \end{pmatrix} = \begin{pmatrix} A\\ Z \end{pmatrix} d,$$

car $Ad=0$. L'inversibilité de la matrice de (4.18) permet d'en déduire que $\xi'(0)=d$, une propriété que l'on recherchait. On peut alors construire les suites désirées :

$$t_k \downarrow 0, \quad x_k := \xi(t_k) \in X_E \to x, \quad \frac{x_k - x}{t_k} = \frac{\xi(t_k)-\xi(0)}{t_k} \to \xi'(0) = d.$$

Donc $d \in \mathrm{T}_x X_E$. □

Avoir l'égalité en (4.17) est souhaitable, comme le montrera l'établissement des conditions d'optimalité dans le théorème 4.17 ci-dessous. Ceci motive l'introduction de la notion suivante.

Définition 4.15 On dit que la contrainte c est *qualifiée* en $x \in X_E$ pour représenter X_E si c est dérivable en x et si

$$\mathrm{T}_x X_E = \mathcal{N}(c'(x)). \tag{4.19}$$

□

Remarques 4.16 1) D'après la proposition 4.14, la surjectivité de $c'(x)$ est une condition *suffisante* de qualification de la contrainte de (P_E). D'autres conditions suffisantes seront données à la section 4.4.2.

2) Insistons sur la signification de (4.19). Dans son membre de gauche on trouve un ensemble qui ne dépend que de X_E, pas de la contrainte c qui a été utilisée pour le décrire. C'est l'inverse dans le membre de droite, qui fait directement intervenir c. L'ensemble admissible $X_E = \{x \in \mathbb{E} : c(x)=0\}$ peut être représenté par différentes fonctions c. On peut donc voir la condition de qualification (4.19) comme un critère permettant de sélectionner les fonctions c qui représentent correctement X_E.

3) Par exemple, au lieu d'utiliser c, on pourrait utiliser

$$\tilde{c} : \mathbb{E} \to \mathbb{R}, \qquad \text{définie par } \tilde{c}(x) = \frac{1}{2}\|c(x)\|^2,$$

puisque $\tilde{c}(x)=0$ si, et seulement si, $c(x)=0$. Ceci paraît attrayant puisque l'on a remplacé toutes les contraintes d'égalité, en nombre potentiellement grand, par une seule contrainte. Cependant, la contrainte $\tilde{c}$ a encore moins de chance d'être qualifiée que c puisque $\nabla\tilde{c}(x) = c'(x)^* c(x) = 0$ en un point $x \in X_E$ et donc $\mathcal{N}(\tilde{c}'(x)) = \mathbb{E}$, qui est le plus souvent trop grand.

4) Il y a en fait un lien subtil entre la condition suffisante de qualification de la contrainte c, qu'est la surjectivité de sa jacobienne $c'(x)$, et la *stabilité* de X_E par rapport à de petites perturbations de c. Ainsi, avec la fonction $\tilde{c}$ définie au

point 3, si $X_E \neq \varnothing$, $X_{E,\varepsilon} := \{x \in \mathbb{E} : \tilde{c}(x) = \varepsilon\}$ a des chances d'être non vide si $\varepsilon \geqslant 0$, alors que $X_{E,\varepsilon} = \varnothing$ si $\varepsilon < 0$, si bien que l'on peut dire que $\tilde{c}$ ne fournit pas une représentation stable de X_E. Plus généralement, si c est de classe C^1, l'hypothèse de surjectivité de la jacobienne $c'(x)$ assure que, pour des $p \in \mathbb{R}^{m_E}$ voisins de zéro, l'ensemble perturbé $\{x : \mathbb{R}^n : c(x)+p = 0\}$ n'est pas vide (c'est une conséquence du théorème des fonctions implicites, le théorème C.14). On dit alors que l'ensemble X_E est *stable* par rapport à des perturbations de la contrainte c.

5) La non-surjectivité de $c'(x)$ peut avoir diverses origines. Cela peut provenir de contraintes redondandes, superflues, qui peuvent parfois être éliminées, mais qui n'empêchent pas pour autant la qualification de la contrainte. Par exemple, si une contrainte $c : \mathbb{E} \to \mathbb{F}$ est qualifiée pour représenter X_E et si on la double en $\tilde{c} := (c, c) : \mathbb{E} \to \mathbb{F} \times \mathbb{F}$, la contrainte $\tilde{c}(x) = 0$ est encore qualifiée pour représenter X_E puisque $\mathcal{N}(\tilde{c}'(x)) = \mathcal{N}(c'(x)) = \mathrm{T}_x X_E$, alors que $\tilde{c}'(x)$ n'est pas surjective.

   Plus malencontreux est le cas où le terme linéaire de c est nul, comme au point 3 ci-dessus ou, plus simplement, comme lorsque l'ensemble admissible $X_E = \{0_{\mathbb{R}}\}$ est défini par $c(x) = 0$ avec $c(x) = x^2$. Cette contrainte c n'est tout simplement pas qualifiée en $x = 0$: $\mathrm{T}_0 X_E = \{0\} \neq \mathcal{N}(c'(0)) = \mathbb{R}$. □

Nous avons à présent établi tout ce qu'il faut pour traduire l'*expression géométrique de l'optimalité*, celle donnée par (4.12), en des conditions nécessaires d'optimalité du premier ordre plus aisément utilisables, ce que l'on peut voir comme une *expression analytique de l'optimalité*.

Théorème 4.17 (CN1 de Lagrange) *Soit x_* une solution locale de (P_E). Supposons que f et c soient dérivables en x_* et que la contrainte c soit qualifiée en x_* au sens de la définition 4.15. Alors, il existe un vecteur $\lambda_* \in \mathbb{F}$ tel que*

$$\nabla f(x_*) + c'(x_*)^* \lambda_* = 0, \tag{4.20}$$

où $\nabla f(x_*)$ est le gradient de f en x_* et $c'(x_*)^* : \mathbb{F} \to \mathbb{E}$ est l'opérateur adjoint de la jacobienne $c'(x_*)$ pour les produits scalaires donnés sur $\mathbb{E}$ et $\mathbb{F}$. Le vecteur λ_* vérifiant (4.20) est unique si $c'(x_*)$ est surjective.

Démonstration. On a successivement

$$\begin{aligned}
\nabla f(x_*) &\in (\mathrm{T}_{x_*} X_E)^+ && [(4.12)] \\
&= \mathcal{N}(c'(x_*))^+ && [\text{qualification des contraintes en } x_*] \\
&= \mathcal{N}(c'(x_*))^\perp && [\mathcal{N}(c'(x_*)) \text{ est un sous-espace vectoriel}] \\
&= \mathcal{R}(c'(x_*)^*) && [(A.6) \text{ ou la proposition } 2.45].
\end{aligned}$$

Il existe donc un vecteur $\lambda_* \in \mathbb{F}$ tel que l'on ait (4.20). Lorsque $c'(x_*)$ est surjective, $c'(x_*)^*$ est injective et il ne peut y avoir deux λ_* différents vérifiant (4.20). □

Le vecteur λ_* est appelé le *multiplicateur de Lagrange* et la formule (4.20) porte le nom de *règle du multiplicateur de Lagrange*. On notera bien que le multiplicateur est

un élément de $\mathbb{F}$, l'espace d'arrivée de c (en dimension infinie, λ appartiendrait au dual de $\mathbb{F}$, qui est ici identifié à $\mathbb{F}$ lui-même grâce au produit scalaire dont il est muni). On dit aussi que λ_* est la *solution duale* du problème (P_E). Alors x_* est appelée *solution primale* et (x_*, λ_*) *solution primale-duale*. On appelle *point stationnaire* du problème (P_E) un point x_* vérifiant ses conditions nécessaires d'optimalité du premier ordre :

$$\begin{cases} \nabla f(x_*) + c'(x_*)^* \lambda_* = 0 \\ c(x_*) = 0, \end{cases} \tag{4.21}$$

pour un certain multiplicateur $\lambda_* \in \mathbb{F}$. La valeur $f(x_*)$ du critère en un point stationnaire x_* est appelée une *valeur critique* du problème.

En pratique, il est souvent commode de retrouver le système d'optimalité (4.21) en introduisant le *lagrangien* du problème qui est la fonction

$$\ell : \mathbb{E} \times \mathbb{F} \to \mathbb{R}$$

définie par

$$\ell(x, \lambda) := f(x) + \langle \lambda, c(x) \rangle.$$

Le nom de multiplicateur attribué à λ vient d'ailleurs du fait qu'il multiplie les contraintes dans le lagrangien (par l'intermédiaire du produit scalaire de $\mathbb{F}$). Cette fonction jouera un rôle essentiel dans tout cet ouvrage, et par conséquent il en sera de même du multiplicateur de Lagrange. Les conditions nécessaires d'optimalité du premier ordre s'écrivent alors

$$\begin{cases} \nabla_x \ell(x_*, \lambda_*) = 0 \\ c(x_*) = 0 \end{cases} \qquad \text{ou} \qquad \nabla_{(x,\lambda)} \ell(x_*, \lambda_*) = 0. \tag{4.22}$$

Il est utile de spécifier ces conditions d'optimalité lorsque $\mathbb{E} = \mathbb{R}^n$, $\mathbb{F} = \mathbb{R}^m$ et que l'on munit ces espaces du produit scalaire euclidien $\langle u, v \rangle = u^\mathsf{T} v = \sum_i u_i v_i$. Il y a alors n variables $x_1, \ldots, x_n$ à optimiser et les m contraintes du problème (P_E) sont données explicitement au moyen de m fonctions $c_i : \mathbb{R}^n \to \mathbb{R}$:

$$c_1(x_1, \ldots, x_n) = 0, \quad \ldots, \quad c_m(x_1, \ldots, x_n) = 0.$$

Le lagrangien du problème s'écrit

$$\ell(x, \lambda) = f(x) + \lambda^\mathsf{T} c(x) = f(x) + \sum_{i=1}^{m} \lambda_i c_i(x).$$

Observons que le multiplicateur de Lagrange $\lambda \in \mathbb{R}^m$ a autant de composantes qu'il y a de contraintes ; chacune des composantes λ_i étant associée à une contrainte c_i ; on dit d'ailleurs que le multiplicateur λ_i est associé à la contrainte c_i. Si l'on note $A_* := c'(x_*)$ la *jacobienne* de c en x_*, qui est la matrice $m \times n$ définie par

$$(A_*)_{ij} := \frac{\partial c_i(x)}{\partial x_j}, \quad 1 \leqslant i \leqslant m, \quad 1 \leqslant j \leqslant n,$$

et $\nabla \varphi(x)$ le gradient de φ ($\varphi = f$ ou c_i) pour le produit scalaire euclidien, c'est-à-dire le vecteur de ses dérivées partielles, les conditions d'optimalité (4.22) s'écrivent

$$\begin{cases} \nabla f(x_*) + A_*^{\mathsf{T}} \lambda_* = 0 \\ c(x_*) = 0 \end{cases} \quad \text{ou} \quad \begin{cases} \nabla f(x_*) + \sum_{i=1}^m (\lambda_*)_i \nabla c_i(x_*) = 0 \\ c(x_*) = 0. \end{cases} \tag{4.23}$$

Les multiplicateurs de Lagrange sont modifiés par un changement du produit scalaire de $\mathbb{F}$, mais pas par un changement du produit scalaire de $\mathbb{E}$. En effet, supposons que l'on prenne sur $\mathbb{E}$ et $\mathbb{F}$ les produits scalaires suivants

$$(x, x') \in \mathbb{E} \times \mathbb{E} \mapsto \langle\!\langle x, x' \rangle\!\rangle_{\mathbb{E}} = \langle S_{\mathbb{E}} x, x' \rangle_{\mathbb{E}},$$
$$(y, y') \in \mathbb{F} \times \mathbb{F} \mapsto \langle\!\langle y, y' \rangle\!\rangle_{\mathbb{F}} = \langle S_{\mathbb{F}} y, y' \rangle_{\mathbb{F}},$$

où $S_{\mathbb{E}} : \mathbb{E} \to \mathbb{E}$ [resp. $S_{\mathbb{F}} : \mathbb{F} \to \mathbb{F}$] est un opérateur auto-adjoint et défini positif pour le produit scalaire $\langle \cdot, \cdot \rangle_{\mathbb{E}}$ [resp. $\langle \cdot, \cdot \rangle_{\mathbb{F}}$]. Le gradient de f en x_* dans (4.20) devient $S_{\mathbb{E}}^{-1} \nabla f(x_*)$ par (C.7) et l'adjoint de la jacobienne de c en x_* devient $S_{\mathbb{E}}^{-1} c'(x_*)^* S_{\mathbb{F}}$ par (A.11), si bien que les multiplicateurs λ_* sont modifiés par ces changements de produits scalaires comme suit

$$S_{\mathbb{F}}^{-1} \lambda_*.$$

Les multiplicateurs sont donc modifiés par un changement de produit scalaire sur $\mathbb{F}$ comme le sont les gradients d'une fonction définie sur $\mathbb{F}$ (voir (C.7)), ce qui anticipe l'interprétation marginaliste des multiplicateurs optimaux qui sera donnée à la section 4.6.1.

Remarques 4.18 1) L'hypothèse (4.19) de qualification des contraintes se retrouvera plus loin sous des versions adaptées à des ensembles admissibles différemment définis. On ne peut pas s'en passer, comme le montre l'exemple illustré à la figure 4.2, dans lequel l'ensemble admissible est réduit à l'unique point

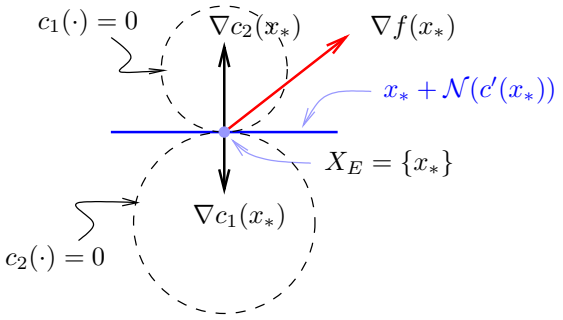


Fig. 4.2. Contraintes d'égalité non qualifiées

d'intersection de deux cercles tangents. Le problème de minimiser une fonction sur un tel ensemble admissible a évidemment une solution (c'est l'unique point admissible), mais la contrainte n'est pas qualifiée, car $\mathrm{T}_{x_*} X_E = \{0\}$, alors que $\mathcal{N}(c'(x_*)) = \mathcal{N}(\nabla c_1(x_*)^{\mathsf{T}}) \cap \mathcal{N}(\nabla c_2(x_*)^{\mathsf{T}})$ est un sous-espace vectoriel de dimension un.

Par ailleurs, si le critère a en x_* un gradient $\nabla f(x_*)$ comme celui représenté dans la figure, ce gradient n'est pas combinaison linéaire des gradients des contraintes

(ceux-ci n'engendrent pas $\mathbb{R}^2$ tout entier car ils sont colinéaires). Dans ce cas, on ne peut pas trouver un multiplicateur $\lambda_* \in \mathbb{R}^2$ qui vérifie la première équation du système de droite dans (4.23). Dès lors, il n'y a pas de système d'optimalité.

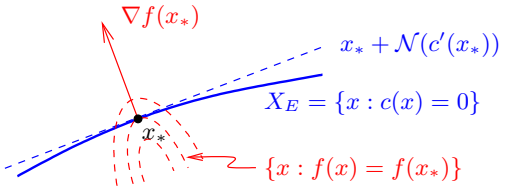
Les problèmes d'optimisation sans système d'optimalité sont difficiles à résoudre et la plupart des algorithmes se fourvoient en s'y attaquant. Nous avons dit que ce manque de qualification n'est pas intrinsèquement lié à l'ensemble admissible, mais à la façon de le décrire par la contrainte c. Quand on se trouve en face d'un problème à la contrainte non qualifiée, il faut donc essayer d'en trouver une autre équivalente et qui soit qualifiée.

Dans l'exemple trivial de la figure 4.2, il suffirait évidemment de prendre la contrainte $\tilde{c}(x) = x - x_*$, qui est qualifiée puisque $\mathcal{N}(\tilde{c}'(x_*)) = \mathcal{N}(I) = \{0\}$. La contrainte $\tilde{c}(x) = 0$, équivalente à $x = x_*$, manifesterait alors clairement la trivialité de ce problème.

2) On peut retrouver la condition nécessaire d'optimalité du premier ordre des problèmes sans contrainte en utilisant le théorème 4.17 pour le problème équivalent

$$\begin{cases} \min f(x) + x_{n+1} \\ x_{n+1} = 0. \end{cases}$$

3) Comme on l'a vu dans la démonstration du théorème 4.17, en présence de qualification, l'expression géométrique (4.12) de l'optimalité s'exprime par



$$\nabla f(x_*) \in \mathcal{N}(c'(x_*))^\perp.$$

Autrement dit, le gradient de f en x_* pour le produit scalaire de $\mathbb{E}$ est orthogonal (au sens du même produit scalaire) à l'espace tangent à l'ensemble admissible $X_E = \{x \in \mathbb{E} : c(x) = 0\}$ en x_*. On peut aussi dire que l'espace tangent en x_* à la variété $\{x \in \mathbb{E} : f(x) = f(x_*)\}$ (qui en est bien une si $\nabla f(x_*) \neq 0$) contient l'espace tangent en x_* à l'ensemble admissible $\{x \in \mathbb{E} : c(x) = 0\}$. Cette expression géométrique se traduit directement dans l'expression analytique (4.20) de l'optimalité, grâce à la relation d'algèbre linéaire

$$\mathcal{N}(A)^\perp = \mathcal{R}(A^*),$$

où A est un opérateur linéaire (voir (A.6), que l'on peut d'ailleurs déduire du lemme de Farkas – proposition 2.45). □

L'ensemble des multiplicateurs optimaux λ_* associés à un point stationnaire x_* est l'ensemble noté et défini par

$$\Lambda_* := \{\lambda_* \in \mathbb{F} : \nabla f(x_*) + c'(x_*)^* \lambda_* = 0\}.$$

Il s'agit donc d'un sous-espace affine (non vide par définition de la stationnarité). Celui-ci est réduit à un singleton (unicité du multiplicateur optimal associé à x_*) si, et seulement si, $c'(x_*)$ est surjective.

Proposition 4.19 (unicité du multiplicateur optimal) *Soit x_* un point stationnaire du probleme (P_E), dont on note Λ_* l'ensemble des multiplicateurs optimaux associés. Alors Λ_* est un singleton si, et seulement si, $c'(x_*)$ est surjective.*

Comme dans le cas sans contrainte, on obtient facilement une condition suffisante d'optimalité si le problème (P_E) est convexe.

Proposition 4.20 (CS1 pour problème convexe) *Supposons que le problème (P_E) soit convexe au sens de la définition 4.12, que $x_* \in \mathbb{E}$ vérifie la contrainte de (P_E), que f soit dérivable en x_* et qu'il existe un multiplicateur $\lambda_* \in \mathbb{F}$ tel que (x_*, λ_*) vérifie (4.21). Alors, x_* est un minimum global de (P_E).*

Démonstration. Comme f est convexe et l'ensemble admissible de (P_E) est convexe, il suffit, d'après la proposition 4.7, de montrer que

$$\forall\, x \in X_E : \quad \langle \nabla f(x_*), x - x_* \rangle \geqslant 0.$$

En utilisant, la première condition d'optimalité de (4.21), cela revient à montrer que

$$\forall\, x \in X_E : \quad \langle \lambda_*, c'(x_*)(x - x_*) \rangle \leqslant 0.$$

On conclut en observant que, pour $x \in X_E$:

$$c'(x_*)(x - x_*) = \lim_{t \downarrow 0} \frac{1}{t} \left[c(x_* + t(x - x_*)) - c(x_*) \right] = 0,$$

car $c(x_*) = 0$ (parce que $x_* \in X_E$) et $c(x_* + t(x - x_*)) = 0$ (parce que $x_* + t(x - x_*) = (1-t)x_* + tx \in X_E$ pour $t \in]0, 1]$, par la convexité supposée de X_E, x et $x_* \in X_E$). □

4.3.2 Conditions d'optimalité du second ordre

Commençons par donner des conditions *nécessaires* d'optimalité du second ordre, lorsque l'on suppose l'existence de multiplicateurs. Si x_* est un minimum local de (P_E), on note

$$\Lambda_* := \{\lambda_* \in \mathbb{F} : (x_*, \lambda_*) \text{ vérifie } (4.20)\}$$

l'ensemble des multiplicateurs optimaux associés a x_* et, pour $\lambda_* \in \Lambda_*$, on note

$$L_* := \nabla^2_{xx} \ell(x_*, \lambda_*),$$

la hessienne du lagrangien en (x_*, λ_*) par rapport à x (le couple optimal (x_*, λ_*) utilisé dans la définition de L_* est alors spécifié dans le contexte).

Théorème 4.21 (CN2) *Soit $x_* \in \mathbb{E}$ un minimum local de (P_E). Supposons que f et c soient dérivables dans un voisinage de x_* et deux fois dérivables en x_*. On suppose également que $\Lambda_* \neq \varnothing$. Alors*

$$\forall d \in \mathrm{T}_{x_*} X_E, \quad \forall \lambda_* \in \Lambda_* : \qquad \langle L_* d, d\rangle \geqslant 0. \tag{4.24}$$

Démonstration. Soient $d \in \mathrm{T}_{x_*} X_E$ et $\lambda_* \in \Lambda_*$. Comme d est une direction tangente à X en x_*, par définition, il existe des suites $\{x_k\} \subseteq \mathbb{E}$ et $\{t_k\} \subseteq \mathbb{R}_{++}$ telles que

$$x_k \in X_E, \qquad t_k \downarrow 0 \qquad \text{et} \qquad \frac{x_k - x}{t_k} \to d.$$

D'après la régularité de f et c et (4.20), on a

$$\ell(x_k, \lambda_*) = \ell(x_*, \lambda_*) + \frac{1}{2}\ell''_{xx}(x_*, \lambda_*) \cdot (x_k - x_*)^2 + o(\|x_k - x_*\|^2).$$

Mais $\ell(x_*, \lambda_*) = f(x_*)$ et $\ell(x_k, \lambda_*) = f(x_k) \geqslant f(x_*)$, pour k assez grand, si bien que pour k grand :

$$0 \leqslant \frac{1}{2}\ell''_{xx}(x_*, \lambda_*) \cdot (x_k - x_*)^2 + o(\|x_k - x_*\|^2).$$

En divisant par t_k^2 et en passant à la limite lorsque $k \to \infty$, on trouve l'inégalité recherchée : $\ell''_{xx}(x_*, \lambda_*) \cdot d^2 \geqslant 0$ ou $\langle L_* d, d\rangle \geqslant 0$. □

Contre-exemple 4.22 (CN2 sans qualification) L'inégalité dans (4.24) ne tient plus nécessairement pour des $d \in \mathcal{N}(c'(x_*)) \setminus \mathrm{T}_{x_*} X_E$ (donc sans qualification de la contrainte en x_*), même si la condition de Lagrange (4.20) est vérifiée pour un certain $\lambda_* \in \mathbb{F}$. Ce cas se présente si $\mathbb{E} = \mathbb{F} = \mathbb{R}$, $f(x) = -x^2$ et $c(x) = x^2$ (l'unique point admissible est $x_* = 0$, qui est donc solution). La condition de Lagrange est vérifiée avec le multiplicateur $\lambda_* = 0$ (en particulier), mais, pour ce choix, $L_* = -2$ n'est pas semi-définie positive sur $\mathcal{N}(c'(x_*)) = \mathbb{R}$, mais bien sur $\mathrm{T}_{x_*} X_E = \{0\}$. □

La condition (4.24) n'est pas très utile si l'on ne dispose pas d'une forme explicite de $\mathrm{T}_{x_*} X_E$. On peut en obtenir une en ajoutant l'hypothèse de qualification de la contrainte (4.19). Dans ce cas, l'existence du multiplicateur est assurée et cela conduit au résultat suivant, qui est celui qui est le plus souvent utilisé. Évidemment cette hypothèse supplémentaire est satisfaite si $c'(x_*)$ est surjective (proposition 4.14).

Corollaire 4.23 *Soit $x_* \in \mathbb{E}$ un minimum local de (P_E). Supposons que f et c soient dérivables dans un voisinage de x_* et deux fois dérivables en x_*. On suppose également que la contrainte c est qualifiée en x_* au sens de la définition 4.15, ce qui implique que $\Lambda_* \neq \varnothing$. Alors*

$$\forall d \in \mathcal{N}(c'(x_*)), \quad \forall\, \lambda_* \in \Lambda_* \ : \qquad \langle L_* d, d\rangle \geqslant 0. \tag{4.25}$$

Les conditions nécessaires du second ordre du corrolaire précédent présentent deux éléments nouveaux par rapport à celles de la proposition 4.10, relatives aux problèmes sans contrainte :

- d'une part, ce n'est pas la hessienne de f qui y intervient, mais celle du lagrangien,
- d'autre part, cette hessienne n'est pas « semi-définie positive dans tout l'espace », mais seulement « semi-définie positive dans un sous-espace vectoriel », à savoir le noyau de $c'(x_*)$.

Comme l'a montré la démonstration du théorème 4.21, ceci est dû, d'une part, au fait que c'est $\nabla_x \ell(x_*, \lambda_*)$ qui s'annule et pas $\nabla f(x_*)$ (si bien que c'est sur $\nabla^2_{xx} \ell(x_*, \lambda_*)$ que l'on peut avoir de l'information et pas sur $\nabla^2 f(x_*)$) et, d'autre part, au fait que l'on n'a de l'information sur la minimalité de f que pour des points proches de x_* dans l'ensemble admissible (si bien que la semi-définie positivité n'a lieu que dans le cône tangent à la contrainte, pas dans l'espace entier).

Le résultat suivant donne des conditions *suffisantes* d'optimalité du second ordre pour le problème (P_E), qui sont à peine plus fortes que les CN2 : le lien entre les CN2 et les CS2 rappelle celui que l'on avait en optimisation sans contrainte (comparez avec les propositions 4.10 et 4.11). On peut donc dire que les CN2 obtenues au théorème 4.21 ne sont pas trop faibles ; il n'y en a pas trop peu.

Une solution x_* vérifiant les CS2 du théorème 4.24 est parfois appelée un *minimum fort* du problème (P_E).

Théorème 4.24 (CS2) *Supposons que f et c soient dérivables dans un voisinage de $x_* \in \mathbb{E}$ et deux fois dérivables en x_*. Supposons que $c(x_*) = 0$, que $\Lambda_* \neq \varnothing$ et que*

$$\forall d \in \mathrm{T}_{x_*} X_E \setminus \{0\}, \quad \exists\, \lambda_* \in \Lambda_* \ : \qquad \langle L_* d, d\rangle > 0. \tag{4.26}$$

Alors x_* est un minimum local strict de (P_E).

Démonstration. On raisonne par l'absurde. Si le résultat est faux, il existe une suite $\{x_k\} \subseteq X_E$ telle que $x_k \to x_*$, $x_k \neq x_*$ et $f(x_k) \leqslant f(x_*)$. En extrayant une sous-suite au besoin, on peut supposer qu'avec $t_k := \|x_k - x_*\| \neq 0$, on a

$$\frac{x_k - x_*}{t_k} \to d.$$

Donc $d \in \mathrm{T}_{x_*} X_E \setminus \{0\}$. Soit alors $\lambda_* \in \Lambda_*$ associé à ce d par les hypothèses de la proposition.

Comme dans la démonstration du théorème 4.21, on peut écrire

$$f(x_k) = f(x_*) + \frac{1}{2} \ell''_{xx}(x_*, \lambda_*) \cdot (x_k - x_*)^2 + o(\|x_k - x_*\|^2).$$

L'inégalité $f(x_k) \leqslant f(x_*)$ implique alors que

$$0 \geqslant \frac{1}{2}\ell''_{xx}(x_*, \lambda_*) \cdot (x_k - x_*)^2 + o(\|x_k - x_*\|^2).$$

En passant à la limite dans l'inégalité ci-dessus, après avoir divisé par t_k^2, on trouve

$$\langle L_* d, d\rangle \leqslant 0,$$

ce qui contredit (4.26). □

Remarques 4.25 1) Le jeu des quantificateurs dans (4.26) peut paraître plus permissif que dans (4.24) ou (4.25). Il n'est rien, car si, dans les conditions du théorème 4.24, (4.26) a lieu, alors on a aussi

$$\forall d \in \mathrm{T}_{x_*} X_E \setminus \{0\}, \quad \forall\, \lambda_* \in \Lambda_* : \qquad \langle L_* d, d\rangle > 0. \tag{4.27}$$

En réalité, $\langle L(x_*, \lambda_*)d, d\rangle$ ne dépend pas du multiplicateur λ_* choisi dans Λ_*. En effet, Λ_* est un sous espacce affine et $\lambda_* \mapsto \langle L(x_*, \lambda_*)d, d\rangle$ est linéaire. Par conséquent, si $\langle L(x_*, \lambda_*^1)d, d\rangle \neq \langle L(x_*, \lambda_*^2)d, d\rangle$ pour deux multiplicateurs optimaux λ_*^1 et $\lambda_*^2 \in \Lambda_*$, on pourrait trouver $\lambda_* \in \Lambda_*$ (combinaison affine de λ_*^1 et λ_*^2) tel que $\langle L(x_*, \lambda_*)d, d\rangle < 0$. Ceci est une contradiction avec le fait que (4.26) implique que x_* est une minimum local de (P_E) et donc, par le théorème 4.21, que $\langle L(x_*, \lambda_*)d, d\rangle \geqslant 0$, quel que soit $\lambda_* \in \Lambda_*$.

2) Bien sûr, la conclusion du théorème 4.24 reste vraie si l'on remplace (4.26) par la condition plus forte

$$\forall d \in \mathcal{N}(c'(x_*)) \setminus \{0\} : \quad \exists\, \lambda_* \in \Lambda_* : \qquad \langle L_* d, d\rangle > 0. \tag{4.28}$$

C'est principalement cette condition plus forte qui est utilisée. D'après la première remarque, il est équivalent de mettre $\exists\, \lambda_* \in \Lambda_*$ ou $\forall\, \lambda_* \in \Lambda_*$ dans (4.28), car $\langle L_* d, d\rangle$ ne dépend pas du multiplicateur λ_* choisi dans Λ_*.

3) On peut donner une forme matricielle compacte de (4.28) en introduisant une base de $\mathbb{E}$, que l'on peut alors identifier à $\mathbb{R}^n$, telle que le produit scalaire induit sur $\mathbb{R}^n$ soit le produit scalaire euclidien et se donnant une matrice Z_*^- dont les colonnes forment une base de $\mathcal{N}(c'(x_*))$ (cette matrice sera de type $n \times (n-m)$, si $c'(x_*)$ est surjective). Alors (4.28) revient à dire que la matrice

$$Z_*^{-\mathsf{T}} L_* Z_*^- \succ 0.$$

La matrice $Z_*^{-\mathsf{T}} L_* Z_*^-$ est appelée la *hessienne réduite du lagrangien*. □

4.3.3 Calcul pratique des solutions de (P_E)

Le théorème 4.17 nous apprend que, sous certaines conditions, une solution locale de (P_E) est un point stationnaire de ce problème. Pour calculer les solutions de (P_E), on pourra donc, dans un premier temps, calculer les solutions du système d'optimalité

(4.21). On notera que ce système est formé de $(n+m)$ équations aux $(n+m)$ inconnues (x_*, λ_*) et qu'il y a donc un sens à en calculer les solutions.

Grâce aux conditions d'optimalité, on a transformé le problème d'optimisation (P_E) en un problème de résolution d'un système d'équations non linéaires, ce qui nous est plus familier. Dans certain cas (par exemple lorsque le problème est de petite taille), on pourra chercher à le résoudre analytiquement, mais le plus souvent, il faudra utiliser des algorithmes spécifiques qui tiennent compte de la structure de (4.21). C'est ce que fait par exemple la méthode newtonienne du chapitre 15.

Cependant, toutes les solutions de (4.21) ne sont pas solutions de (P_E). Par définition, ce ne sont que des points stationnaires. Pour déterminer si un point stationnaire est solution de (P_E), on utilisera les conditions d'optimalité du second ordre, de la manière suivante :

- si la condition nécessaire (4.25) n'est pas vérifiée au point stationnaire, alors celui-ci n'est pas une solution locale de (P_E) (ou une autre hypothèse du corollaire 4.23 n'est pas vérifiée) ;
- si la condition suffisante (4.26) ou (4.28) est vérifiée au point stationnaire (ainsi que les autres hypothèses du théorème 4.24), alors celui-ci est un minimum local strict de (P_E).

Ces deux cas recouvrent un grand nombre de situations, mais pas toutes, car les conditions (4.25) et (4.26) ne sont pas identiques. Le cas est indéterminé lorsqu'en un point stationnaire on a (4.25), mais pas (4.26). Alors les résultats vus ci-dessus ne sont pas suffisants et il faudra recourir à des conditions d'optimalité d'ordre supérieur pour pouvoir dire si le point stationnaire est solution de (P_E).

L'approche décrite ci-dessus est systématique (voir l'épigraphe de Lagrange en début de ce chapitre) et donc très souvent utilisée en pratique. Cependant, il ne faudra pas négliger d'utiliser l'intuition ou la familiarité avec un problème particulier pour en déterminer les solutions de manière plus rapide, plus astucieuse.

4.4 Problème avec contraintes d'égalité et d'inégalité

Dans cette section, on considère le problème

$$(P_{EI}) \quad \begin{cases} \min f(x) & \\ c_i(x) = 0, & i \in E \\ c_i(x) \leqslant 0, & i \in I, \end{cases}$$

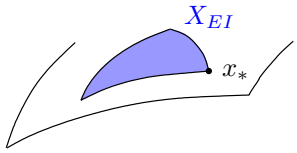
où f et les c_i sont des fonctions définies sur un espace euclidien $\mathbb{E}$ (produit scalaire noté $\langle \cdot, \cdot \rangle$) à valeurs dans $\mathbb{R}$ et où E et I forment une partition de $\{1, \ldots, m\}$ ($E \cup I = [1:m]$ et $E \cap I = \varnothing$). On note

$$m_E := |E| \qquad \text{et} \qquad m_I := |I|$$

le cardinal des ensembles E et I (donc $m = m_E + m_I$). Les contraintes sont donc données par une fonction $c : \mathbb{E} \to \mathbb{R}^m$. On supposera que $\mathbb{R}^m$ est muni du produit scalaire euclidien. On ne prend pas un espace euclidien arbitraire pour l'espace d'arrivée de c

car il faudrait donner un sens à l'inégalité de (P_{EI}), ce qui compliquerait singulièrement l'analyse. Si $v \in \mathbb{R}^m$, on note v_E (resp. v_I) le vecteur de $\mathbb{R}^{m_E}$ (resp. $\mathbb{R}^{m_I}$) formé des composantes v_i de v avec $i \in E$ (resp. $i \in I$). De même on notera c_E et c_I les fonctions définissant les contraintes d'égalité et d'inégalité, respectivement.

Dans cette section, on note



$$X_{EI} := \{x \in \mathbb{E} : c_E(x) = 0, c_I(x) \leqslant 0\} \qquad (4.29)$$

l'ensemble admissible de (P_{EI}). Si en un point admissible x_*, $c'(x_*)$ est surjective, cet ensemble se présente autour de x_* comme une partie de la variété $\{x \in \mathbb{E} : c_E(x) = 0\}$ formée des points qui vérifient aussi l'inégalité $c_I(x) \leqslant 0$. Une image idéalisée de cet ensemble pourrait être celle donnée dans la figure ci-dessus, qui correspond à $n = 3$, $m_E = 1$ et $m_I = 3$. Notons que rien n'impose que la solution soit sur la frontière de X_{EI} comme dans la figure, mais c'est dans ce cas que le problème est le plus complexe à analyser.

L'adaptation au problème (P_{EI}) de la définition 4.8 de problème (P_X) convexe conduit à la définition suivante.

Définition 4.26 ((P_{EI}) convexe) On dit que le problème (P_{EI}) est *convexe* si son ensemble admissible X_{EI} est convexe et si son critère f est convexe sur X_{EI}. □

Proposition 4.27 (convexité de X_{EI}) *Si la contrainte d'égalité c_E est affine, si les contraintes d'inégalité c_i $(i \in I)$ sont convexes, alors X_{EI} est convexe.*

Démonstration. L'ensemble admissible s'écrit comme une intersection de convexes (il est donc convexe). En effet,

$$X_{EI} = \{x \in \mathbb{E} : c_E(x) = 0\} \cap \{x \in \mathbb{E} : c_I(x) \leqslant 0\}.$$

Le premier est un sous-espace affine par l'affinité de c_E (proposition 4.13), donc un convexe. Le second est l'intersection des ensembles $\{x \in \mathbb{E} : c_i(x) \leqslant 0\}$, pour $i \in I$, qui sont convexes comme ensembles de sous-niveau de fonctions convexes. □

Mais X_{EI} peut être convexe sans que c_E soit affine et les $\{c_i\}_{i\in I}$ soient convexes. Par exemple $\{x \in \mathbb{R} : x + x^3 \leqslant 0\}$ est convexe, alors que $x \mapsto x + x^3$ n'est pas convexe.

Définition 4.28 On dit que c_i est *active* en x si $c_i(x) = 0$. On note

$$I^0(x) := \{i \in I : c_i(x) = 0\}$$

l'ensemble des indices des contraintes d'inégalité actives en x. On adopte les notations simplifiées $I^0_x := I^0(x)$ et $I^0_* := I^0(x_*)$. □

Les problèmes d'optimisation avec contraintes d'inégalité sont considérablement plus difficiles à analyser et à résoudre numériquement qu'un problème avec contraintes d'égalité. Un calcul analytique, sur papier, est rarement possible et d'ailleurs souvent difficile lui aussi. Lorsqu'il n'y a que des contraintes d'égalité, la compréhension du problème repose sur l'analyse mathématique classique, en particulier, nous l'avons vu, sur le théorème des fonctions implicites (théorème C.14), alors que la présence d'inégalité requiert l'utilisation d'outils spécifiques, essentiellement, nous le verrons, ceux de l'analyse convexe. Par ailleurs, numériquement, la difficulté principale provient du fait que, d'une manière ou d'une autre, le calcul de la solution détermine forcément les contraintes qui y sont actives. Si celles-ci étaient connues, on pourrait se ramener au cas des problèmes avec seulement des contraintes d'égalité lisses. Or il y a 2^{m_I} manières de rendre les m_I contraintes d'inégalité actives. C'est à ce nombre exponentiel que l'on fait allusion lorsque l'on parle de la *combinatoire* des problèmes avec contraintes d'inégalité. Celle-ci est redoutable et en rapport direct avec la *conjecture* $\mathsf{P} = \mathsf{NP}$, puisqu'un problème d'optimisation quadratique non convexe (c.-à-d., un problème avec un critère quadratique non convexe et des contraintes affines) est NP-ardu (section 5.2.3). On est donc en présence d'un problème pour lequel il est vraisemblable que le *principe de conservation des ennuis* s'applique ; on veut dire par là que la difficulté du problème ne peut être éliminée en lui trouvant une autre formulation équivalente. Ainsi, on pourrait penser simplifier le problème en reformulant les contraintes d'inégalité par l'une des contraintes d'égalité apparemment plus simples et équivalentes suivantes

$$\max(0, c_I(x)) = 0 \qquad \text{ou} \qquad \|\max(0, c_I(x))\|^2 = 0.$$

Cependant le première contrainte est non lisse et la seconde, bien qu'une fois différentiable, n'est en général pas qualifiée dans un sens discuté ci-dessus (après la définition 4.15).

4.4.1 Conditions de Karush, Kuhn et Tucker

Pour particulariser la condition nécessaire d'optimalité du théorème 4.6 et sa formule $\nabla f(x_*) \in (\mathrm{T}_x\, X_{EI})^+$ au cas où l'ensemble admissible X_{EI} est donné par des contraintes fonctionnelles d'égalité et d'inégalité, on doit préciser comment s'écrit le *cône tangent*

$$\mathrm{T}_x\, X_{EI}$$

à partir des fonctions c_E et c_I définissant les contraintes. La proposition 4.29 montre que $\mathrm{T}_x\, X_{EI}$ est inclus dans le cône obtenu par linéarisation des contraintes actives en x, appelé *cône linéarisant* et noté

$$\mathrm{T}'_x\, X_{EI} := \{d \in \mathbb{E} : c'_E(x) \cdot d = 0,\ c'_{I^0_x}(x) \cdot d \leqslant 0\}.$$

Proposition 4.29 (cône tangent à X_{EI}) *Supposons que $c_{E\cup I_x^0}$ soit dérivable en $x \in X_{EI}$. Alors*

$$\mathrm{T}_x\, X_{EI} \subseteq \mathrm{T}'_x\, X_{EI}. \tag{4.30}$$

Démonstration. Soit $d \in \mathrm{T}_x\, X_{EI}$. Alors il existe des suites $\{x_k\} \subseteq X_{EI}$ et $\{t_k\} \downarrow 0$ telles que

$$\frac{x_k - x}{t_k} \to d.$$

Pour $i \in E \cup I_x^0$, on a $c_i(x) = 0$ et donc

$$c_i(x_k) = c'_i(x) \cdot (x_k - x) + o(\|x_k - x\|).$$

Pour $i \in E$, $c_i(x_k) = 0$ et donc

$$c'_i(x) \cdot \frac{(x_k - x)}{t_k} = \frac{o(\|x_k - x\|)}{t_k}.$$

À la limite, on trouve $c'_E(x) \cdot d = 0$. De la même manière, pour $i \in I_x^0$, en utilisant le fait que $c_i(x_k) \leqslant 0$, on trouve : $c'_i(x) \cdot d \leqslant 0$. □

En général, on n'a pas égalité en (4.30). Ceci avait déjà été observé lorsqu'il n'y avait que des contraintes d'égalité (voir les commentaires sur la figure 4.2). La figure 4.3 donne un exemple avec des contraintes d'inégalité seulement : on y a

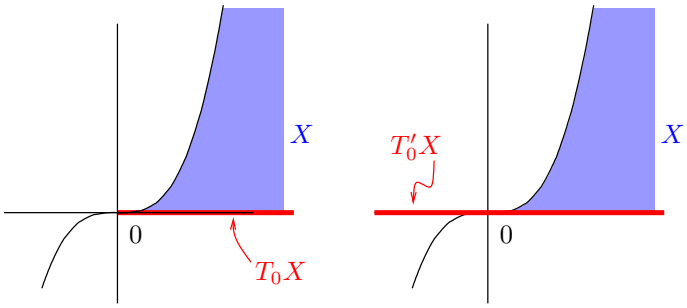


Fig. 4.3. Cône tangent $\mathrm{T}_x\, X$ (gauche) et cône linéarisant $\mathrm{T}'_x\, X$ (droite)

représenté l'ensemble $X = \{x \in \mathbb{R}^2 : 0 \leqslant x_2 \leqslant x_1^3\}$ et les deux cônes $\mathrm{T}_x\, X$ et $\mathrm{T}'_x\, X$ en $x = (0,0)$. Le vecteur $d = (-1,0)$ est dans $\mathrm{T}'_{(0,0)}\, X$ mais pas dans $\mathrm{T}_{(0,0)}\, X$. Le cas se présente également lorsque $X = \{x \in \mathbb{R}^2 : x_1^2 \leqslant x_2^2\}$ (exemple 4.4), puisque $\mathrm{T}_{(0,0)}\, X = \{d \in \mathbb{R}^2 : |d_1| \leqslant |d_2|\}$ alors que $\mathrm{T}'_{(0,0)}\, X = \mathbb{R}^2$. On notera d'ailleurs que $\mathrm{T}'_x\, X$ est convexe (c'est un polyèdre convexe), tandis que $\mathrm{T}_x\, X$ ne l'est pas nécessairement.

Par ailleurs, l'égalité $\mathrm{T}_x\, X_{EI} = \mathrm{T}'_x\, X_{EI}$ est souhaitable car, comme on le verra, le calcul de $(\mathrm{T}'_x\, X_{EI})^+$ est aisé alors qu'en toute généralité celui de $(\mathrm{T}_x\, X_{EI})^+$ ne l'est

pas. Comme c'est ce dernier qui intervient dans la condition d'optimalité (4.12) que l'on veut exprimer en termes plus faciles à utiliser, on a tout intérêt à ce que l'égalité $\mathrm{T}_{x_*} X_{EI} = \mathrm{T}'_{x_*} X_{EI}$ ait lieu. Ceci conduit à la notion suivante.

Définition 4.30 (qualification de contrainte) On dit que la contrainte c est *qualifiée* en $x \in X_{EI}$ pour représenter X_{EI} si $c_{E \cup I^0_x}$ est dérivable en x et si

$$\text{(QC)} \quad \mathrm{T}_x X_{EI} = \mathrm{T}'_x X_{EI}, \tag{4.31}$$

c'est-à-dire si le cône tangent $\mathrm{T}_x X_{EI}$ à l'ensemble admissible X_{EI} en x est égal au cône $\mathrm{T}'_x X_{EI}$, défini en (4.30), obtenu par linéarisation des contraintes actives en x. □

Remarques 4.31 On peut adapter à cette définition les remarques 4.16 faites à propos de la qualification des contraintes d'égalité et en ajouter d'autres.

1) Comme le membre de gauche de 4.31 ne dépend que de X_{EI}, pas de c, et que le membre de droite dépend de X_{EI} par l'intermédiaire de c, il faut voir la condition (4.31) comme un moyen de sélectionner les bonnes fonctions c représentant l'ensemble admissible X_{EI}.

2) Par exemple, on ne change pas l'ensemble admissible de (P_{EI}) en remplaçant ses contraintes par une contrainte d'égalité

$$\tilde{c} : \mathbb{E} \to \mathbb{R}, \qquad \text{définie par } \tilde{c}(x) = \frac{1}{2}\|c_E(x)\|_2^2 + \frac{1}{2}\|c_I(x)^+\|_2^2,$$

puisque $x \in X_{EI}$ si, et seulement si, $\tilde{c}(x) = 0$ (on a noté v^+ le vecteur dont la i-ième composante est $\max(0, v_i)$). Si cela paraît attrayant de prime abord, puisque l'on a remplacé des contraintes d'égalité $c_E(x) = 0$ et d'inégalité $c_I(x) \leqslant 0$, en nombre potentiellement grand, par une unique contrainte d'égalité $\tilde{c}(x) = 0$, cette dernière contrainte a l'inconvénient de n'être presque jamais qualifiée. On a en effet $\nabla\tilde{c}(x) = c'_E(x)^* c_E(x) + c'_I(x)^* c_I(x)^+$, qui est nul en tout point admissible. Dès lors, pour cette contrainte $\tilde{c}$, le membre de droite de (4.31) vaut $\mathbb{E}$, qui est presque toujours trop grand.

3) Comme autre exemple de transformation à ne pas faire, mentionnons celle qui consiste à remplacer les inégalités $c(x) \leqslant 0$ par des égalités $\tilde{c}(x, s) = 0$, en introduisant des *écarts en carrés*:

$$\tilde{c} : \mathbb{E} \times \mathbb{R}^m \to \mathbb{R}^m, \qquad \text{définie par } \tilde{c}(x, s) = c(x) + s \cdot s,$$

où l'on a utilisé le *produit de Hadamard* ($(s \cdot s)_i = s_i^2$). L'ensemble admissible $X := \{x \in \mathbb{E} : c(x) \leqslant 0\}$ est alors la projection canonique sur $\mathbb{E}$ de l'ensemble admissible $\tilde{X} := \{(x, s) \in \mathbb{E} \times \mathbb{R}^m : \tilde{c}(x, s) = 0\}$. Cette pratique, parfois encore rencontrée, parrait attrayante parce que l'on s'est débarrassé des contraintes d'inégalité à la combinatoire importante. Cependant, la nouvelle contrainte $\tilde{c}$ ne vérifie pas la condition suffisante de qualification, qu'est sa surjectivité (voir la proposition 4.14) en un point $(x, s) \in \tilde{X}$ pour lequel les gradients $\nabla c_i(x)$ des contraintes actives ne sont pas linéairement indépendants (une hypothèse très forte comme on le verra à la section 4.4.2), ce qui conduit alors à une représentation par $\tilde{c}$ instable [513].

4) La plupart des algorithmes se fourvoient lorsqu'ils doivent trouver la solution d'un problème dont les contraintes ne sont pas qualifiées en la solution. Il est donc préférable, dans ce cas, de changer la description de l'ensemble admissible en utilisant d'autres fonctions, avant de chercher à résoudre le problème.

5) Il y a en fait un lien subtil entre la qualification de la contrainte c et la *stabilité* de X_{EI} par rapport à de petites perturbations de c. On montrera en effet (proposition 4.78) que les conditions de Mangasarian-Fromovitz (QC-MF) énoncées plus loin, qui qualifient la contrainte c représentant X_{EI}, assurent que, pour des $p \in \mathbb{R}^m$ et des $t \in \mathbb{R}$ voisins de zéro, l'ensemble perturbé $X_{EI}^{tp} := \{x : \mathbb{R}^n : c_E(x) + tp_E = 0, c_I(x) + tp_I \leqslant 0\}$ n'est pas vide. On dit alors que l'ensemble X_{EI} est *stable* par rapport à des perturbations de la contrainte c. □

Nous verrons à la section 4.4.2 des conditions suffisantes de qualification des contraintes, plus faciles à vérifier que l'égalité $\mathrm{T}_x\, X_{EI} = \mathrm{T}'_x\, X_{EI}$.

Venons-en à présent à l'un des résultats les plus importants de ce chapitre, celui énonçant les conditions d'optimalité du première ordre du problème d'optimisation différentiable avec contraintes. Nous introduisons avec lui la notation

$$0 \leqslant u \perp v \leqslant 0, \tag{4.32}$$

pour signifier que les composantes d'un vecteur $u \in \mathbb{R}^p$ doivent être positives, que celles d'un vecteur $v \in \mathbb{R}^p$ doivent être négatives et que ces deux vecteurs doivent être orthogonaux pour le produit scalaire euclidien : $u^{\mathsf{T}} v = 0$. Cette notation est communément utilisée pour exprimer les problèmes de complémentarité [141, 142].

Théorème 4.32 (CN1 de Karush-Kuhn-Tucker) *Soit x_* un minimum local de (P_{EI}). Supposons que f et $c_{E \cup I_*^0}$ soient dérivables en x_* et que les contraintes soient qualifiées en x_* au sens de la définition 4.30. Alors, il existe $\lambda_* \in \mathbb{R}^m$ tel que l'on ait*

$$(KKT) \quad \left\{ \begin{array}{ll} (a) & \nabla f(x_*) + c'(x_*)^* \lambda_* = 0 \\ (b) & c_E(x_*) = 0 \\ (c) & 0 \leqslant (\lambda_*)_I \perp c_I(x_*) \leqslant 0, \end{array} \right. \tag{4.33}$$

où $\nabla f(x_*)$ est le gradient de f en x_* et $c'(x_*)^* : \mathbb{R}^m \to \mathbb{E}$ est l'opérateur adjoint de la jacobienne $c'(x_*)$ pour le produit scalaire donné sur $\mathbb{E}$.

Démonstration. On a successivement

$$\begin{array}{rcll} \nabla f(x_*) & \in & (\mathrm{T}_{x_*}\, X_{EI})^+ & [\text{par } (4.12)] \\ & = & (\mathrm{T}'_{x_*}\, X_{EI})^+ & [\text{qualification des contraintes en } x_*] \\ & = & \{-c'_E(x_*)^* y - c'_{I_*^0}(x_*)^* z : y \in \mathbb{R}^{m_E},\ z \in \mathbb{R}_+^{|I_*^0|}\} & [\text{corollaire } 2.46]. \end{array}$$

Il existe donc des vecteurs $y \in \mathbb{R}^{m_E}$ et $z \in \mathbb{R}_+^{|I_*^0|}$ tels que

$$\nabla f(x_*) = -c'_E(x_*)^* y - c'_{I_*^0}(x_*)^* z.$$

On obtient le résultat en introduisant $\lambda_* \in \mathbb{R}^m$ défini par

$$(\lambda_*)_i := \begin{cases} y_i & \text{si } i \in E \\ z_i & \text{si } i \in I_*^0 \\ 0 & \text{si } i \in I \setminus I_*^0. \end{cases}$$ □

Les conditions d'optimalité (4.33) sont appelées *conditions de Karush-Kuhn-Tucker* et le vecteur λ_* est encore appelé *multiplicateur de Lagrange* associé aux contraintes c et à la solution x_*. Un couple (x_*, λ_*) vérifiant (4.33) est appelé *solution primale-duale* de (P_{EI}) : x_* est la solution primale et λ_* est la solution duale (on comprendra l'origine de cette appellation au chapitre 14 sur la dualité). Un point x_* pour lequel il existe un multiplicateur $\lambda_* \in \mathbb{R}^m$ tel que (4.33) soit vérifiée est qualifié de *stationnaire*. La valeur $f(x_*)$ du critère en un point stationnaire x_* est appelée une *valeur critique* du problème.

Comme pour les problèmes avec contraintes d'égalité, le *lagrangien associé au problème* (P_{EI}) est la fonction

$$\ell : \mathbb{E} \times \mathbb{R}^m \to \mathbb{R}$$

définie en $(x, \lambda) \in \mathbb{E} \times \mathbb{R}^m$ par

$$\ell(x, \lambda) := f(x) + \lambda^{\mathsf{T}} c(x) = f(x) + \sum_{i=1}^{m} \lambda_i c_i(x). \tag{4.34}$$

Le multiplicateur λ a autant de composantes qu'il y a de contraintes, chacune de ses composantes multipliant la contrainte à laquelle elle est associée dans le lagrangien. La première équation d'optimalité du système (4.33) est alors le gradient de ℓ en (x_*, λ_*) par rapport à x :

$$\nabla_x \ell(x_*, \lambda_*) = 0 \qquad \text{ou} \qquad \nabla f(x_*) + \sum_{i=1}^{m} (\lambda_*)_i \nabla c_i(x_*) = 0,$$

les gradients de f et des c_i étant ceux associés au produit scalaire de $\mathbb{E}$.

Si $\mathbb{E} = \mathbb{R}^n$ et si l'on choisit le produit scalaire euclidien sur cet espace, le gradient $\nabla f(x_*)$ et la matrice jacobienne $c'(x_*)$ ont leurs composantes données par les dérivées partielles

$$(\nabla f(x_*))_i = \frac{\partial f}{\partial x_i}(x_*) \qquad \text{et} \qquad (c'(x_*))_{ij} = \frac{\partial c_i}{\partial x_j}(x_*)$$

et l'adjointe de $c'(x_*)$ devient sa transposée. La première équation de (4.33) peut donc alors s'écrire

$$\nabla f(x_*) + c'(x_*)^{\mathsf{T}} \lambda_* = 0.$$

Compte tenu du signe de $(\lambda_*)_I$ et de $c_I(x_*)$, chaque terme du produit scalaire dans $(\lambda_*)_I^{\mathsf{T}} c_I(x_*) = 0$, équation dérivant de (4.33)-(c) est négatif, si bien que cette condition revient à exprimer que chacun de ces termes est nul :

$$\forall i \in I \ : \quad (\lambda_*)_i \, c_i(x_*) = 0.$$

Ces relations portent le nom de *conditions de complémentarité*: soit $(\lambda_*)_i = 0$, soit $c_i(x_*) = 0$. Étant donné le signe de $(\lambda_*)_I$, on peut les écrire de manière équivalente comme suit :

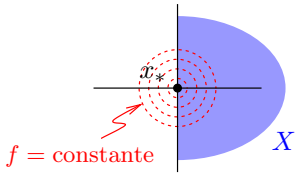
$$\forall i \in I \ : \quad \Big(c_i(x_*) < 0 \implies (\lambda_*)_i = 0\Big), \tag{4.35}$$

ce qui peut s'exprimer en disant qu'un multiplicateur associé à une contrainte d'inégalité inactive ($c_i(x_*) < 0$) est nul. On peut aussi dire que les relations de complémentarité dans (4.33)-(c) sont l'expression analytique du fait que les contraintes inactives en x_* ne sont pas utiles pour exprimer l'optimalité *locale* d'une solution. On dit qu'il y a *complémentarité stricte* si

$$\forall i \in I \ : \quad \Big(c_i(x_*) < 0 \iff (\lambda_*)_i = 0\Big). \tag{4.36}$$

Comme les multiplicateurs ne sont pas nécessairement déterminés de manière unique, on peut avoir complémentarité stricte pour un jeu de multiplicateur et ne pas avoir cette propriété pour un autre jeu. Une solution x_* sans complémentarité stricte se manifeste par le fait qu'une contrainte d'inégalité active en x_* ne joue pas de rôle dans le lagrangien et donc dans le système d'optimalité (KKT). Par conséquent, on peut enlever cette contrainte sans changer l'optimalité de x_* (au premier ordre). Géométriquement, cela peut se voir à partir des courbes de niveau de f dans le voisinage de x_*, comme dans le problème trivial en $x \in \mathbb{R}^2$ suivant

$$\begin{cases} \min \|x\|_2^2 \\ x_1 \geqslant 0. \end{cases}$$



La solution est $x_* = 0$ et le multiplicateur optimal est nul, bien que la contrainte soit active en la solution. Par ailleurs, on peut manifestement enlever la contrainte, sans changer la solution du problème.

La condition (4.33)-(c) peut se résumer par diverses équations (non différentiables). La plus simple utilise la fonction « min », composante par composante :

$$\min(\lambda_I, -c_I(x)) = 0.$$

L'équivalence vient du fait que pour des réels a et b, l'équation $\min(a, b) = 0$ revient à dire que $a \geqslant 0$, $b \geqslant 0$ et $ab = 0$.

Remarques

1. Le signe des multiplicateurs correspondant aux contraintes d'inégalité peut changer si le problème n'est pas écrit sous la forme canonique qui est celle de (P_{EI}). Ainsi le signe change si au lieu de minimiser on maximise, ou si à la place des contraintes de négativité ($c_i(x) \leqslant 0$) on trouve des contraintes de positivité ($c_i(x) \geqslant 0$), ou bien sûr si l'on remplace le signe $+$ par le signe $-$ dans le lagrangien, comme le font certains auteurs.

2. La présence de conditions de *qualification des contraintes* dans les hypothèses du théorème 4.32 gène parfois au premier abord. Il arrive que l'on pose la question suivante. Comment peut-on savoir si les contraintes sont qualifiées en x_*, alors que l'on ne connaît pas x_* et que, pour le calculer, on va utiliser (4.33), qui suppose précisément que les contraintes sont qualifiées en x_* ? Il semble que le raisonnement tourne en rond. Voilà ce que l'on peut dire à ce sujet.
   - On peut très bien connaître un point stationnaire x_* du problème (P_{EI}) sans le calculer par le système (4.33). Dans ce cas, la question toute théorique de savoir s'il existe des multiplicateurs tels que l'on ait (4.33) a tout son sens. On s'intéresse alors à l'existence de multiplicateurs pour d'autres raisons que le calcul d'une solution ; par exemple pour une analyse au second ordre (section 4.4.4) ou une analyse de sensibilité (section 4.6).
   - D'autre part, si l'on ne connaît pas x_*, on peut très bien essayer de chercher une solution primale-duale (x_*, λ_*) de (4.33). Si l'on en trouve une, tant mieux, on aura forcément l'existence de multiplicateurs ! Si l'on n'en trouve pas, on pourra chercher à savoir pourquoi. L'absence de qualification des contraintes peut en être la cause, mais cela peut aussi venir du fait que l'autre hypothèse du théorème 4.32 n'est pas vérifiée : les fonctions f et c ne sont pas régulières en x_*. De ce point de vue, la qualification des contraintes en x_* n'est pas plus étrange que la régularité de f et c en x_*.
   - Finalement, il arrive parfois que l'on ait qualification des contraintes en tout point de l'ensemble admissible X_{EI}, comme on peut avoir régularité de f et c sur X_{EI}. L'existence de multiplicateurs est alors assurée.
3. On peut retrouver dans les conditions de KKT, l'expression géométrique de l'optimalité dont on est parti pour les obtenir, à savoir $\nabla f(x_*) \in (\mathrm{T}_{x_*} X_{EI})^+$. Supposons, pour simplifier l'illustration, que $E = \varnothing$. En tenant compte des conditions de complémentarité dans (4.33)-(c), la relation (4.33)-(a) s'écrit

$$\nabla f(x_*) = \sum_{i \in I_*^0} \underbrace{(\lambda_*)_i}_{\geqslant 0} (-\nabla c_i(x_*)).$$

Regardons ce que cette expression exprime dans la figure 4.4, où l'on a représenté

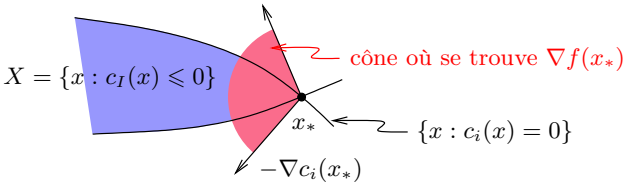


Fig. 4.4. Conditions d'optimalité de KKT

une partie de l'ensemble admissible X_{EI} et supposé qu'une solution locale x_* se trouve sur le bord de X_{EI}. Observons d'abord que l'opposé du gradient de c_i en x_*, $-\nabla c_i(x_*)$, est bien orienté comme dans la figure (tous les vecteurs ont été

translatés de x_*), car $c_i(x)$ est nul sur la frontière de X_{EI} et négatif dans X_{EI}. La relation ci-dessus exprime que le gradient $\nabla f(x_*)$ est dans le *cône* engendré par ces $-\nabla c_i(x_*)$, avec $i \in I_*^0$ (on ne considère donc que les contraintes actives). On se convaincra sans peine que ce cône est bien le dual du cône tangent $\mathrm{T}_{x_*} X_{EI}$, si bien que l'on retrouve le résultat attendu.

4. La figure 4.4 montre à l'évidence que l'optimisation différentiable repose sur l'analyse convexe, pas seulement sur l'algèbre linéaire (un cône n'est pas un objet de cette dernière).

La proposition suivante montre que les conditions de Karush-Kuhn-Tucker (4.33) sont des conditions suffisantes d'optimalité lorsque le problème (P_{EI}) est convexe. On peut en conclure, qu'il n'y a dans le système d'optimalité (4.33), compliqué, ni trop, ni trop peu de relations.

Proposition 4.33 (CS1 pour problème convexe) *Considérons le problème (P_{EI}), que l'on suppose convexe au sens de la définition 4.26. Soit x_* un point vérifiant les contraintes de (P_{EI}). Si f et c sont dérivables en x_* et s'il existe $\lambda_* \in \mathbb{R}^m$ tel que les conditions de Karush-Kuhn-Tucker (4.33) soient vérifiées, alors x_* est un minimum global de (P_{EI}).*

Démonstration. Comme f est convexe, comme l'ensemble admissible de (P_{EI}) est convexe et comme $x_* \in X_{EI}$, il suffit, d'après la proposition 4.7, de montrer que

$$\forall\, x \in X_{EI} : \quad \langle \nabla f(x_*), x - x_* \rangle \geqslant 0.$$

En utilisant (4.33)-(a), cela revient à montrer que

$$\forall\, x \in X_{EI} : \quad (\lambda_*)^{\mathsf{T}} \left(c'(x_*)(x - x_*)\right) \leqslant 0.$$

On a vu à la démonstration de la proposition 4.20 que l'admissibilité de x et x_* impliquait que $c'_E(x_*)(x - x_*) = 0$. Comme $(\lambda_*)_I \geqslant 0$ et $(\lambda_*)_{I \setminus I_*^0} = 0$ par (4.33)-(c), il suffit de montrer que $c'_i(x_*)(x - x_*) \leqslant 0$ pour $i \in I_*^0$. Pour ces $i \in I_*^0$, on a en effet

$$c'_i(x_*)(x - x_*) = \lim_{t \downarrow 0} \frac{1}{t} \left[c_i(x_* + t(x - x_*)) - c_i(x_*)\right] \leqslant 0,$$

car $c_i(x_*) = 0$ (parce que $i \in I_*^0$) et $c_i(x_* + t(x - x_*)) = c_i((1 - t)x_* + tx) \leqslant 0$ pour $t \in \,]0, 1]$ (parce que $(1 - t)x_* + tx \in X_{EI}$ par la convexité supposée de X_{EI}, x et $x_* \in X_{EI}$). □

4.4.2 Qualification des contraintes

S'il n'y a qu'une contrainte fonctionnelle d'égalité c, la proposition 4.14 donne une condition suffisante pour que cette contrainte c définissant X_E soit qualifiée en x : il suffit que c soit C^1 dans un voisinage de x et que $c'(x)$ soit surjective. Le but de cette section est d'établir des conditions similaires lorsque l'ensemble admissible est défini au moyen de contraintes d'égalité et d'inégalité. Les conditions suffisantes de qualification les plus utilisées sont celles désignées par (QC-A), (QC-S), (QC-IL) ou (QC-MF) ci-dessous.

Affinité locale (QC-A)

Définition 4.34 (QC-A) La contrainte c définissant X_{EI} vérifie (QC-A) en $x \in X_{EI}$ si

(i) $c_{I\setminus I_x^0}$ est continue en x,

(ii) $c_{E\cup I_x^0}$ est affine dans un voisinage de x. □

On s'en doute, cette condition est utilisée lorsque l'ensemble admissible X_{EI} est défini par des contraintes affines. Il en est ainsi en optimisation linéaire (chapitre 17) ou, plus généralement, en optimisation quadratique (chapitre **??**). Dans ces disciplines, on ne parle jamais de qualification de contraintes, car celles-ci le sont nécessairement, comme le montre le résutat suivant.

Proposition 4.35 ((QC-A) est une CS de qualification) *Si la contrainte c définissant X_{EI} vérifie les hypothèses (QC-A) de la définition 4.34 en $x \in X_{EI}$, alors c est qualifiée en x au sens de la définition 4.30.*

Démonstration. D'après la proposition 4.29, il suffit de montrer que $\mathrm{T}'_x\, X_{EI} \subseteq \mathrm{T}_x\, X_{EI}$. Soit $d \in \mathrm{T}'_x\, X_{EI}$. On introduit une suite $\{x_k\}$ par $x_k := x + t_k d$, avec $t_k \downarrow 0$. Il suffit de montrer que $x_k \in X_{EI}$ pour k grand. Comme les contraintes actives sont affines dans un voisinage de x, on a pour k grand :

$$c_i(x_k) = c_i(x) + t_k c_i'(x) \cdot d, \quad \text{pour } i \in E \cup I_x^0.$$

D'après les propriétés de d, on en déduit que pour k grand : $c_E(x_k) = 0$ et $c_{I_x^0}(x_k) \leqslant 0$. D'autre part, comme $c_{I\setminus I_x^0}$ est continue en x, $c_{I\setminus I_x^0}(x_k) < 0$ pour k grand. Ceci montre que $x_k \in X_{EI}$ pour k grand. □

Notons encore que, s'il n'y a que des contraintes d'égalité, la condition (QC-A) est nouvelle ; elle n'avait pas en effet été mentionnée à la section 4.3 comme une condition suffisante de qualification de la contrainte définissant X_E.

Slater (QC-S)

Les conditions de qualification de Slater (QC-S) énoncées ci-dessous sont typiquement utilisées en optimisation convexe. L'hypothèse la plus singulière (iii) repose sur les contraintes d'inégalité actives en x car, par leur continuité, les autres ne jouent pas de rôle localement.

Définition 4.36 (QC-S) La contrainte c définissant X_{EI} vérifie les conditions de Slater (QC-S) en $x \in X_{EI}$ si

(i) c_E est *affine*,

(ii) $c_{I\setminus I_x^0}$ est continue en x et $c_{I_x^0}$ est dérivable en x,

(iii) les composantes de $c_{I_x^0}$ sont convexes et on peut trouver un point $\hat{x} \in X_{EI}$ tel que $c_{I_x^0}(\hat{x}) < 0$. □

Parfois, toutes les composantes de c_I sont convexes et (iii) se simplifie en

(iii') les composantes de c_I sont *convexes* et on peut trouver un point $\hat{x}'$ tel que $c_E(\hat{x}') = 0$ et $c_I(\hat{x}') < 0$.

C'est souvent sous cette forme que l'hypothèse est présentée. Il semble que, même lorsque toutes les composantes de c_I sont convexes, l'hypothèse (iii) soit plus générale car on n'y demande pas que $c_{I\setminus I_x^0}(\hat{x}) < 0$. Ce n'est pas le cas, car si (iii) a lieu, le point $\hat{x}' = (x+\hat{x})/2 \in X_{EI}$ vérifie (iii') grâce à la convexité des composantes de c_I puisque $c_{I\setminus I_x^0}(x) < 0$ et $c_{I_x^0}(\hat{x}) < 0$, si bien que $c_I(\hat{x}') \leqslant (c_I(x)+c_I(\hat{x}))/2 < 0$. Dès lors l'intérêt de (iii) par rapport à (iii') est de ne requérir que la convexité des contraintes actives en x.

Proposition 4.37 ((QC-S) est une CS de qualification) *Si la contrainte c définissant X_{EI} vérifie les hypothèses (QC-S) de la définition 4.36 en $x \in X_{EI}$, alors c est qualifiée en x au sens de la définition 4.30.*

Démonstration. D'après la proposition 4.29, il suffit de montrer que $\mathrm{T}'_x\, X_{EI} \subseteq \mathrm{T}_x\, X_{EI}$. Soit $d \in \mathrm{T}'_x\, X_{EI}$. On définit

$$d_k := (1-\tfrac{1}{k})\, d + \tfrac{1}{k}\,(\hat{x}-x),$$

où $\hat{x}$ est donné par (QC-S).

- Pour $i \in I_x^0$ et $k \geqslant 1$, il vient de la linéarité de l'opérateur dérivée $c_i'(x)$ et de la convexité de c_i :

$$c_i'(x)\cdot d_k = (1-\tfrac{1}{k})\underbrace{c_i'(x)\cdot d}_{\leqslant 0} + \tfrac{1}{k}\underbrace{c_i'(x)\cdot(\hat{x}-x)}_{\leqslant c_i(\hat{x})-c_i(x)<0} < 0,$$

  car $c_i'(x)\cdot d \leqslant 0$ ($d \in \mathrm{T}'_x\, X_{EI}$), $c_i(x) = 0$ ($x \in X_{EI}$) et $c_i(\hat{x}) < 0$. Dès lors, il existe un réel $t_k \in\,]0, \frac{1}{k}]$ tel que, pour tout $i \in I_x^0$: $c_i(x+t_k d_k) \leqslant c_i(x) = 0$. On définit $x_k := x + t_k d_k$ qui forme une suite convergeant vers x.
- Pour $i \in I \setminus I_x^0$, la continuité de c_i implique que $c_i(x_k) \leqslant 0$ pour tout k assez grand.
- Pour $i \in E$, l'affinité de c_i implique que $c_i'(x)\cdot(\hat{x}-x) = c_i(\hat{x}) - c_i(x) = 0$, si bien que $c_i'(x)\cdot d_k = c_i'(x)\cdot d = 0$ ($d \in \mathrm{T}'_x\, X_{EI}$) et $c_i(x_k) = c_i(x) + t_k\, c_i'(x)\cdot d_k = 0$.

Dès lors, $x_k \in X_{EI}$ pour tout k assez grand. Par ailleurs, lorsque $k \to \infty$, $t_k \downarrow 0$ et $(x_k - x)/t_k = d_k \to d$. Donc $d \in \mathrm{T}_x\, X_{EI}$. □

Indépendance linéaire (QC-IL)

Définition 4.38 (QC-IL) La contrainte c définissant X_{EI} vérifie les conditions (QC-IL) en $x \in X_{EI}$ si

(i) $c_{E\cup I_x^0}$ est C^1 dans un voisinage de x et $c_{I\setminus I_x^0}$ est continue en x,

(ii) les gradients des contraintes actives en x, à savoir les $\nabla c_i(x)$ pour $i \in E \cup I_x^0$, sont linéairement indépendants, ce qui s'écrit

$$\sum_{i\in E\cup I_x^0} \alpha_i \nabla c_i(x) = 0 \qquad \Longrightarrow \qquad \alpha_i = 0 \text{ pour tout } i \in E \cup I_x^0.$$

□

Commençons par donner des formulations équivalentes évidentes de la condition (ii) de la définition précédente.

Proposition 4.39 (autres formulations de (QC-IL)) *Supposons que $c_{E\cup I_x^0}$ soit dérivable en x. Alors les propriétés suivantes sont équivalentes :*
(i) *les gradients $\nabla c_i(x)$, pour $i \in E \cup I_x^0$, sont linéairement indépendants,*
(ii) *$c'_{E\cup I_x^0}(x)$ est surjective,*
(iii) *pour tout $g \in \mathbb{E}$, l'ensemble*

$$\{\lambda \in \mathbb{R}^m : g + c'_{E\cup I_x^0}(x)^* \lambda_{E\cup I_x^0},\ \lambda_{I\setminus I_x^0} = 0\}$$

est soit vide soit un singleton.

Démonstration. La condition (i) exprime l'injectivité de $c'_{E\cup I_x^0}(x)^*$, car

$$\begin{aligned}
\sum_{i\in E\cup I_x^0} \alpha_i \nabla c_i(x) = 0 \quad &\iff \quad \sum_{i\in E\cup I_x^0} \alpha_i \langle \nabla c_i(x), d\rangle = 0, \quad \forall\, d \in \mathbb{E} \\
&\iff \quad \alpha_{E\cup I_x^0}^{\mathsf{T}} \left(c'_{E\cup I_x^0}(x) \cdot d \right) = 0, \quad \forall\, d \in \mathbb{E} \\
&\iff \quad \langle c'_{E\cup I_x^0}(x)^* \alpha_{E\cup I_x^0}, d\rangle = 0, \quad \forall\, d \in \mathbb{E} \\
&\iff \quad c'_{E\cup I_x^0}(x)^* \alpha_{E\cup I_x^0} = 0.
\end{aligned}$$

Cette injectivité est bien sûr équivalente à (ii). Par ailleurs, (iii) exprime, de manière compliquée, que $c'_{E\cup I_x^0}(x)^*$ est injective. □

La condition (iii) n'est mentionnée que pour faciliter la comparaison avec une condition similaire exprimant (QC-MF) ci-dessous. La condition (ii) montre que (QC-IL) étend aux inégalités, de manière un peu excessive comme nous allons le voir, la condition suffisante de qualification de contraintes d'égalité de la proposition 4.14. Le fait qu'il s'agisse d'une condition suffisante de qualification se démontre d'ailleurs directement en utilisant la proposition 4.14, comme on va le voir ci-dessous.

Proposition 4.40 ((QC-IL) est une CS de qualification) *Si la contrainte c définissant X_{EI} vérifie les hypothèses (QC-IL) de la définition 4.38 en $x \in X_{EI}$, alors c est qualifiée en x au sens de la définition 4.30.*

Démonstration. D'après la proposition 4.29, il suffit de montrer que $\mathrm{T}'_x X_{EI} \subseteq \mathrm{T}_x X_{EI}$. Soit $d \in \mathrm{T}'_x X_{EI}$. On note $J \subseteq I_x^0$ tel que

$$c'_J(x) \cdot d = 0 \qquad \text{et} \qquad c'_{I_x^0\setminus J}(x) \cdot d < 0.$$

Comme $c'_J(x)$ est surjective, par le même argument que celui utilisé dans la démonstration de la proposition 4.14, on peut trouver des suites $\{x_k\} \subseteq \mathbb{E}$ et $\{t_k\} \subseteq \mathbb{R}_{++}$ telles que $c_{E\cup J}(x_k) = 0$, $t_k \downarrow 0$ et $(x_k - x)/t_k \to d$. Forcément, $x_k \to x$. Alors,

- par $c_{I_x^0\setminus J}(x) = 0$ et $c'_{I_x^0\setminus J}(x) \cdot d < 0$, on voit que $c_{I_x^0\setminus J}(x_k) \leqslant 0$ pour k assez grand,

- par la continuité de $c_{I\setminus I_x^0}$ et $c_{I\setminus I_x^0}(x) < 0$, $c_{I\setminus I_x^0}(x_k) \leqslant 0$ pour k assez grand.

Dès lors $\{x_k\} \subseteq X_{EI}$ pour k assez grand et donc $d \in \mathrm{T}_x\, X_{EI}$. □

Mangasarian-Fromovitz (QC-MF)

Définition 4.41 (QC-MF) La contrainte c définissant X_{EI} vérifie les conditions de Mangasarian-Fromovitz (QC-MF) en $x \in X_{EI}$ si

(*i*) $c_{E\cup I_x^0}$ est dérivable en x et $c_{I\setminus I_x^0}$ est continue en x,

(*ii*) les gradients des contraintes actives en x, à savoir les $\nabla c_i(x)$ pour $i \in E \cup I_x^0$, vérifient

$$\sum_{i\in E\cup I_x^0} \alpha_i \nabla c_i(x) = 0 \quad \text{et} \quad \alpha_{I_x^0} \geqslant 0 \qquad \Longrightarrow \qquad \alpha_{E\cup I_x^0} = 0.$$

□

Les conditions de qualifications (QC-IL) et (QC-MF) méritent d'être comparées.

- La différence entre (QC-IL) et (QC-MF) provient essentiellement des conditions (*ii*) dans les définitions 4.38 et 4.41. Après examen, on voit clairement que (QC-MF) est plus faible, plus souvent vérifiée, que (QC-IL), puisque qu'avec (QC-MF) la combinaison linéaire ne sera pas nécessairement trivialement nulle si ses coefficients α_i, avec $i \in I_x^0$, ne sont pas positifs. On peut résumer cette observation par l'implication

$$\boxed{\text{(QC-IL)} \quad \Longrightarrow \quad \text{(QC-MF)}.} \tag{4.37}$$

- Les deux conditions sont illustrées à la figure 4.5, dans laquelle l'ensemble admissible est défini par deux contraintes : $c_1(x) \leqslant 0$ et $c_2(x) \leqslant 0$. On a repris à gauche le cas déjà discuté de la figure 4.2 : aucune des deux conditions (QC-IL) et (QC-MF) n'y est vérifiée et l'existence de multiplicateurs n'est pas assurée,

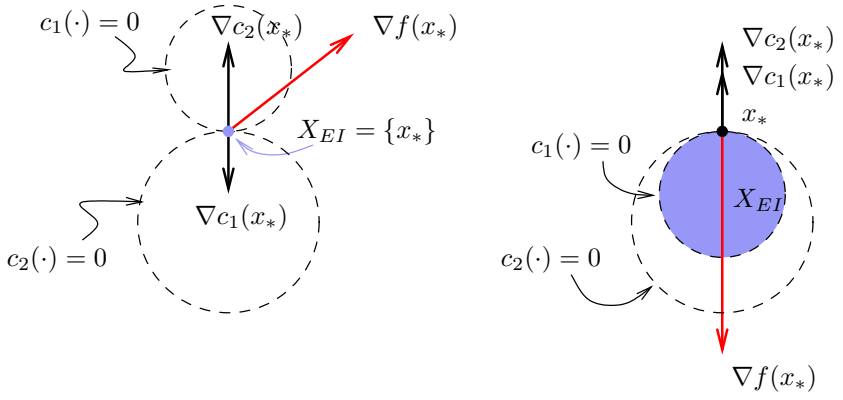


Fig. 4.5. Comparaison des conditions de qualification (QC-IL) et (QC-MF)

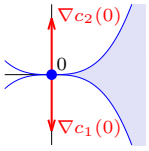
bien que le problème d'optimisation ait une solution (l'ensemble admissible est réduit à un seul point). À droite, (QC-MF) est vérifiée, mais pas (QC-IL). D'après la proposition 4.43 ci-dessous, les contraintes sont alors qualifiées et, d'après le théorème 4.32, il existe des multiplicateurs de Lagrange tel que l'on ait (4.33). Dans le cas présent, ceci implique que $\nabla f(x_*)$ doit être colinéaire et opposé à $\nabla c_1(x_*)$ et $\nabla c_2(x_*)$.

- On peut aussi dire que (QC-IL) traduit en particulier le fait que les sous-espaces vectoriels $\mathcal{R}(c'_E(x)^*)$ et $\mathcal{R}(c'_{I_0(x)}(x)^*)$ ne peuvent se rencontrer qu'en 0, alors que (QC-MF) implique que le sous-espace vectoriel $\mathcal{R}(c'_E(x)^*)$ ne rencontre le *cône* $\mathrm{cone}\{\nabla c_i(x) : i \in I_0(x)\}$ qu'en 0. Dans ce sens, (QC-IL) est une condition qui ne fait intervenir que des objets de l'algèbre linéaire, alors que (QC-MF) requiert des concepts d'analyse convexe.

Ces remarques montrent clairement l'intérêt de la condition de qualification (QC-MF). Malgré sa complexité supplémentaire, il faudra faire l'effort de la comprendre, de l'assimiler, de l'utiliser.

Une bonne manière de vérifier que (QC-MF) *n'*est *pas* vérifiée est de montrer que l'ensemble admissible *n'*est *pas* stable pour de petites perturbations. On veut dire par là que l'ensemble $X^p_{EI} := \{x \in \mathbb{R} : c_E(x) + p_E = 0,\ c_I(x) + p_I \leqslant 0\}$ peut être vide pour certaines perturbations $p \in \mathbb{R}^m$ aussi petites que l'on veut (dans le cas contraire, l'on dit que la description de X_{EI} par c est *stable*) ; dans ce cas donc (QC-MF) n'est pas vérifié. On montrera en effet, dans un cadre plus général (corollaire 4.60), que si (QC-MF) a lieu, alors sa description par c est stable. Ainsi, toujours à la figure 4.5, on voit qu'il n'y a pas de stabilité dans le tracé de gauche (l'ensemble admissible est vide pour des perturbations p, vérifiant $p_1 = p_2 > 0$), alors qu'il est stable à droite. On comprend que numériquement, cette propriété de stabilité de l'ensemble admissible est importante et donc qu'il importe que (QC-MF) soit vérifié lorsqu'on cherche à résoudre le problème numériquement. Notons enfin que l'implication ne va que dans un sens : on peut très bien avoir une description de l'ensemble admissible qui est stable alors que (QC-MF) n'est pas vérifié. C'est le cas de l'ensemble

$$\{x \in \mathbb{R}^2 : -x_1^3 \leqslant x_2 \leqslant x_1^3\}, \qquad (4.38)$$



qui reste non vide, quel que soit la perturbation des contraintes, alors que (QC-MF) n'a pas lieu en $x = 0$ (car les gradients des contraintes $c_1(x) := -x_1^3 - x_2$ et $c_2(x) := -x_1^3 + x_2$ y sont colinéaires et opposés).

Avant de montrer que (QC-MF) implique la qualification de c en x pour représenter X_{EI}, donnons des conditions équivalentes à (QC-MF) qui nous seront utiles. Elles sont à rapprocher de celles énoncées pour (QC-IL) à la proposition 4.39.

- Comparons d'abord les conditions (ii) des propositions 4.39 et 4.42. Si la condition (QC-IL) peut s'exprimer par la surjectivité de la jacobienne $c'_{E\cup I^0_x}(x)$, la condition de Mangasarian-Fromovitz, plus faible que (QC-IL), peut se voir comme une propriété de *surjectivité faible* de $c'_{E\cup I^0_x}(x)$. Le qualificatif *faible* vient du fait que l'on n'a pas nécessairement mieux que l'inégalité pour les indices dans I^0_x.

- On peut aussi comparer les conditions (iii) de la proposition 4.39 et (iv) de la proposition 4.42. L'ensemble considéré au point (iii) de la proposition 4.39 est un sous-espace affine (éventuellement vide) contenant au plus un point, tandis que celui considéré au point (iv) de la proposition 4.42 est un polyèdre convexe (éventuellement vide), pouvant contenir plus d'un point, mais borné. Ce dernier résultat permet de montrer aisément que l'ensemble des multiplicateurs optimaux d'un problème d'optimisation est borné si ses contraintes sont qualifiées au sens de Mangasarian-Fromovitz (proposition 4.45).

Proposition 4.42 (autres formulations de (QC-MF)) *Soit $x \in X_{EI}$ et supposons que la fonction $c_{E\cup I_x^0} : \mathbb{R}^n \to \mathbb{R}^{|E\cup I_x^0|}$ soit dérivable en x. Alors les propriétés suivantes sont équivalentes :*

(i) (QC-MF) a lieu en x,

(ii) $\forall v \in \mathbb{R}^m$, $\exists\, d \in \mathbb{E}$ tel que l'on ait $c'_E(x)\cdot d = v_E$ et $c'_{I_x^0}(x)\cdot d \leqslant v_{I_x^0}$,

(iii) $c'_E(x)$ est surjective et il existe $d \in \mathbb{E}$ tel que $c'_E(x)\cdot d = 0$ et $c'_{I_x^0}(x)\cdot d < 0$,

(iv) pour tout $g \in \mathbb{E}$, le polyèdre convexe

$$\{\lambda \in \mathbb{R}^m : g + c'_E(x)^*\lambda_E + c'_{I_x^0}(x)^*\lambda_{I_x^0} = 0,\ \lambda_{I_x^0} \geqslant 0,\ \lambda_{I\setminus I_x^0} = 0\}$$

est soit vide soit borné.

On peut inverser les inégalités dans (ii) et (iii), en jouant avec l'opposé de d et de v.

Démonstration. $[(i) \Rightarrow (ii)]$ D'après (QC-MF) :

$$\{(\lambda,\mu) : c'_E(x)^*\lambda + c'_{I_x^0}(x)^*\mu = 0,\ \mu \geqslant 0\} = \{(0,0)\}.$$

En prenant le dual de ces ensembles, on obtient grâce à la proposition 2.45 (lemme de Farkas généralisé)

$$\left\{ \begin{pmatrix} c'_E(x) \\ c'_{I_x^0}(x) \end{pmatrix} d + \begin{pmatrix} 0 \\ I \end{pmatrix} h : d \in \mathbb{R}^n,\ h \in \mathbb{R}_+^{|I_x^0|} \right\} = \mathbb{R}^{m_E} \times \mathbb{R}^{|I_x^0|}.$$

On en déduit que, quel que soit $v \in \mathbb{R}^{m_E} \times \mathbb{R}^{|I_x^0|}$, on peut trouver $d \in \mathbb{R}^n$ et $h \in \mathbb{R}_+^{|I_x^0|}$ tels que $c'_E(x)\cdot d = v_E$ et $c'_{I_x^0}(x)\cdot d + h = v_{I_x^0}$, ce qui n'est autre que (ii).

$[(ii) \Rightarrow (iii)]$ La surjectivité de $c'_E(x)$ découle immédiatement de (ii). D'autre part, en prenant $v \in \mathbb{R}^{m_E} \times \mathbb{R}^{|I_x^0|}$ dans (ii) tel que $v_E = 0$ et $v_{I_x^0} = -(1,\dots,1)$, on obtient la deuxième condition de (iii).

$[(iii) \Rightarrow (i)]$ Soit $(\lambda,\mu) \in \mathbb{R}^{m_E} \times \mathbb{R}^{|I_x^0|}$ vérifiant

$$c'_E(x)^*\lambda + c'_{I_x^0}(x)^*\mu = 0 \qquad \text{et} \qquad \mu \geqslant 0.$$

En multipliant scalairement cette identité par le vecteur d donné au point (iii), on trouve

$$\mu^{\mathsf{T}} c'_{I_x^0}(x) d = 0.$$

Comme $\mu \geqslant 0$ et $c'_{I_x^0}(x)d < 0$, ceci implique que $\mu = 0$. Alors la surjectivité de $c'_E(x)$ et $c'_E(x)^*\lambda = 0$ impliquent que $\lambda = 0$.

[$(i) \Leftrightarrow (iv)$] Désignons par Λ le polyèdre convexe de (iv), que l'on suppose non vide, sinon l'implication $(i) \Rightarrow (iv)$ est triviale et l'implication réciproque n'apporte rien. Alors, selon le corollaire 2.8, Λ est borné si, et seulement si, son cône asymptotique Λ^∞ est réduit à $\{0\}$. D'après l'exercice 2.20, on a

$$\Lambda^\infty = \{\mu \in \mathbb{R}^m : c'_E(x)^*\mu_E + c'_{I_x^0}(x)^*\mu_{I_x^0} = 0,\ \mu_{I_x^0} \geqslant 0,\ \mu_{I\setminus I_x^0} = 0\},$$

si bien que ces propriétés sont équivalentes à (QC-MF). □

La proposition 4.42 explicite un aspect algébrique des conditions de Mangasarian-Fromovitz (QC-MF). Celles-ci renferment également un aspect topologique que nous allons préciser. On sait que si A est une matrice surjective $m \times n$ et $v \in \mathbb{R}^m$, on peut trouver $d \in \mathbb{R}^n$ tel que $Ad = v$ (par définition même de la surjectivité, bien sûr). La proposition précédente généralise cela pour des inégalités : si A_E et A_J sont des matrices $m_E \times n$ et $m_J \times n$, vérifiant la condition de régularité ($\alpha_E \in \mathbb{R}^{m_E}$ et $\alpha_J \in \mathbb{R}^{m_J}$)

$$A_E^\mathsf{T}\alpha_E + A_J^\mathsf{T}\alpha_J = 0, \quad \alpha_J \geqslant 0 \qquad \Longrightarrow \qquad (\alpha_E, \alpha_J) = 0, \tag{4.39}$$

alors, quel que soit $v \in \mathbb{R}^{m_E+m_J}$, on peut trouver $d \in \mathbb{R}^n$ tel que $A_E d = v_E$ et $A_J d \leqslant v_J$. Ces deux résultats ne disent rien sur la grandeur de d et on peut se demander si lorsque v tend vers 0, il en de même de d. Ce n'est certainement pas le cas pour une solution d arbitraire car, dans le premier cas, celle-ci n'est définie qu'à un élément du noyau de A près et, si celui-ci n'est pas réduit à $\{0\}$, on pourra vérifier $Ad = v$ avec des d arbitrairement grands. On peut montrer cependant que l'on peut *choisir* d tel qu'il en soit ainsi : *si A_E et A_J vérifient la condition (4.39), il existe une constante C telle que pour tout $v \in \mathbb{R}^{m_E+m_J}$, on peut trouver $d \in \mathbb{R}^n$ tel que $A_E d = v_E$, $A_J d \leqslant v_J$ et $\|d\| \leqslant C\|v\|$*. Le résultat est évident s'il n'y a que des égalités à vérifier : il suffit de choisir $d = A^\mathsf{T}(AA^\mathsf{T})^{-1}v$ (AA^T est inversible si A est surjective !). Le cas où il y a également des inégalités à vérifier est proposé à l'exercice 4.17.

Montrons à présent que si l'une des conditions (QC-IL) ou (QC-MF) est vérifiée alors les contraintes sont qualifiées en x.

Proposition 4.43 (conditions suffisantes de qualification des contraintes) *Soit x vérifiant les contraintes de (P_{EI}) et supposons que les hypothèses ?? soient vérifiées. Alors les contraintes de (P_{EI}) sont qualifiées en x au sens de la définition 4.30, si l'une des conditions (QC-IL) ou (QC-MF) est satisfaite. Plus précisément, on a le diagramme suivant :*

$$\begin{array}{ccc} \textit{(QC-MF)} & \Longleftarrow & \textit{(QC-IL)} \\ \Downarrow & & \\ \textit{(QC)} & & \end{array}$$

De plus, si c_E est affine et les composantes de $c_{I_x^0}$ sont convexes, alors

$$\textit{(QC-S)} \quad \Longleftarrow \quad \textit{(QC-MF)}.$$

DÉMONSTRATION. 2) Montrons à présent que (QC-S) implique le point (iii) de la proposition 4.42 et donc que (QC-MF) a lieu. Soient $\hat{x}$ le point admissible donné dans (QC-S) et $d := \hat{x} - x$. Du fait que c_E est affine et que $c_E(x) = c_E(\hat{x}) = 0$, on a $c'_E(x) \cdot d = 0$. Ensuite, pour $i \in I^0_x$, grâce à la convexité de c_i :

$$c'_i(x) \cdot d \leqslant c_i(\hat{x}) - c_i(x) = c_i(\hat{x}) < 0.$$

Montrons la réciproque lorsque c_E est affine (dernière partie de la proposition). Avec la direction d donnée par le point (iii) de la proposition 4.42, on voit que le point $\hat{x} = x + td$ vérifie $c_E(\hat{x}) = 0$, $c_{I^0_x}(\hat{x}) < 0$ et $c_{I \setminus I^0_x}(\hat{x}) \leqslant 0$, si t est pris assez petit.

3) On a déjà fait remarquer que (QC-IL) implique (QC-MF) et donc le fait démontré ci-dessous que (QC-MF) implique la qualification des contraintes en x suffit à montrer qu'il en est de même avec (QC-IL). On peut toutefois en donner une preuve directe rapide. Soit $d \in \mathrm{T}'_x\, X_{EI}$. Puisque $c'_{E \cup I^0_x}(x)$ est surjective, on peut comme dans la démonstration de la proposition 4.14 construire un chemin $t \mapsto \xi(t)$, défini pour t voisin de $0 \in \mathbb{R}$, tel que

$$c_{E \cup I^0_x}(\xi(t)) = 0 \text{ pour } t \text{ voisin de } 0, \quad \xi(0) = x \quad \text{et} \quad \xi'(0) = d.$$

Alors la suite $\{x_k\}$ définie par $x_k = \xi(t_k)$, avec $t_k \downarrow 0$, est la suite recherchée. Elle vérifie en effet $c_{E \cup I^0_x}(x_k) = 0$ et $c_{I \setminus I^0_x}(x_k) \leqslant 0$ (parce que $c_{I \setminus I^0_x}(x) < 0$ et $c_{I \setminus I^0_x}$ est continue en x). Donc $x_k \in X_{EI}$ et comme $(x_k - x)/t_k = (\xi(t_k) - x)/t_k \to d$, on a bien $d \in \mathrm{T}_x\, X_{EI}$.

4) Il reste à démontrer que (QC-MF) implique la qualification des contraintes en x. Soit $d \in \mathrm{T}'_x\, X_{EI}$, vérifiant donc $c'_E(x) \cdot d = 0$ et $c'_{I^0_x}(x) \cdot d \leqslant 0$. Il faut montrer que $d \in \mathrm{T}_x\, X_{EI}$. Soit

$$D(x) := \{d : c'_E(x) \cdot d = 0,\ c'_{I^0_x}(x) \cdot d < 0\} \subseteq \mathrm{T}'_x\, X_{EI}.$$

Montrons que $\overline{D(x)} = \mathrm{T}'_x\, X_{EI}$. Comme $\mathrm{T}'_x\, X_{EI}$ est fermé, on a $\overline{D(x)} \subseteq \mathrm{T}'_x\, X_{EI}$. Inversement, soit $d_0 \in D(x)$, qui est non vide lorsque (QC-MF) a lieu (proposition 4.42). Si $d \in \mathrm{T}'_x\, X_{EI}$, la suite $\{d_k\}$ définie par $d_k = d + \frac{1}{k} d_0$ est dans $D(x)$ et converge vers d.

Montrons que $D(x) \subseteq \mathrm{T}_x\, X_{EI}$. Soit $d_0 \in D(x)$. Comme $c'_E(x) \cdot d_0 = 0$ et $c'_E(x)$ est surjective, d_0 est tangent à $\{\tilde{x} \in \mathbb{E} : c_E(\tilde{x}) = 0\}$ en x (proposition 4.14). Donc il existe des suites $\{x_k\} \subseteq \mathbb{E}$ et $\{t_k\} \downarrow 0$ telles que

$$c_E(x_k) = 0 \qquad \text{et} \qquad \frac{x_k - x}{t_k} \to d_0.$$

Pour conclure, il suffit de montrer que $c_I(x_k) \leqslant 0$ pour k assez grand. Si $i \in I^0_x$, on a

$$c_i(x_k) = c'_i(x) \cdot (x_k - x) + o(\|x_k - x\|).$$

En divisant par t_k et en notant que $c'_i(x) \cdot d_0 < 0$, on voit que $c_i(x_k) < 0$ pour k assez grand. Si $i \in I \setminus I^0_x$, $c_i(x) < 0$ et comme $x_k \to x$, $c_i(x_k) < 0$ pour k assez grand.

On peut à présent conclure puisque d'après ce qui précède et la proposition 4.29,

$$D(x) \subseteq \mathrm{T}_x\, X_{EI} \subseteq \mathrm{T}'_x\, X_{EI}.$$

En prenant l'adhérence et en se rappelant que $\overline{D(x)} = \mathrm{T}'_x\, X_{EI}$ et que $\mathrm{T}_x\, X_{EI}$ est fermé, on obtient $\mathrm{T}_x\, X_{EI} = \mathrm{T}'_x\, X_{EI}$. □

La démonstration du point 4 (reprise de [315 ; 1996]) peut aussi se faire par une approche semblable à celle utilisée au point 3, mais elle demande plus de régularité sur f et c. Elle consiste à construire un chemin $t \mapsto \xi(t)$ dans X_{EI} pour $t \geqslant 0$ petit qui accepte d comme tangente en $t = 0$ (cette construction sera faite dans la démonstration du lemme 4.47 qui conduit aux conditions nécessaires du deuxième ordre).

Notons pour conclure, que les conditions de Mangasarian-Fromovitz ne sont pas nécessaires pour avoir la qualification des contraintes. Un contre-exemple est donné à l'exercice 4.15.

4.4.3 Ensemble des multiplicateurs optimaux

Soit $\lambda_* \in \mathbb{R}^m$ un multiplicateur optimal (ou solution duale) associé à une solution $x_* \in \mathbb{E}$ du problème (P_{EI}). Il sera utile d'introduire les ensembles d'indices suivants :

$$I_*^{0+} := \{i \in I_*^0 : (\lambda_*)_i > 0\} \qquad \text{et} \qquad I_*^{00} := \{i \in I_*^0 : (\lambda_*)_i = 0\}. \tag{4.40}$$

Ce sont donc des ensembles d'indices qui dépendent à la fois de x_* et λ_*. Les contraintes d'indices $i \in I_*^{0+}$ sont dites *fortement actives* et celles d'indices $i \in I_*^{00}$ sont dites *faiblement actives*. Ces dernières, bien qu'actives ($c_i(x_*) = 0$), peuvent être ôtées du problème sans modifier la stationnarité de x_* ($(\lambda_*)_i = 0$).

Il peut y avoir plus d'une solution duale λ_* associée à une solution primale x_*. L'ensemble des multiplicateurs associés à x_* est noté

$$\Lambda_* := \Lambda(x_*) := \{\lambda_* \in \mathbb{R}^m : (x_*, \lambda_*) \text{ est solution primale-duale de } (P_{EI})\}.$$

La solution x_* étant fixée, les conditions d'optimalité (4.33) montrent que Λ_* est un polyèdre convexe de $\mathbb{R}^m$:

$$\Lambda_* = \{\lambda \in \mathbb{R}^m : \nabla f(x_*) + c'(x_*)^*\lambda = 0,\ \lambda_{I_*^0} \geqslant 0,\ \lambda_{I \setminus I_*^0} = 0\}.$$

En particulier, Λ_* est fermé. Il est non vide si les contraintes sont qualifiées en x_* (théorème 4.32). En raisonnant comme dans la démonstration de la proposition 2.21, on voit que $\lambda_* \in \Lambda_*$ est un sommet de Λ_* si $c'_{E \cup I_*^{0+}}(x_*)$ est surjective. En particulier, Λ_* n'a pas de sommet si $c'_E(x_*)$ n'est pas surjective.

Cet ensemble Λ_* est clairement réduit à un seul multiplicateur si les conditions de qualification (QC-IL) (indépendance linéaire des gradients des contraintes actives) sont vérifiées en x_*. En effet, $(\mathrm{KKT})_{(a)}$ s'écrit aussi

$$\nabla f(x_*) + c'_{E \cup I_*^0}(x_*)^*(\lambda_*)_{E \cup I_*^0} = 0,$$

puisque $(\lambda_*)_{I \setminus I_*^0} = 0$. Par (QC-IL), $c'_{E \cup I_*^0}(x_*)^*$ est injective et il n'existe donc qu'un seul $(\lambda_*)_{E \cup I_*^0}$ vérifiant $(\mathrm{KKT})_{(a)}$.

L'unicité du multiplicateur optimal a lieu, en fait, sous une condition plus faible que (QC-IL), mais plus forte que (QC-MF), que donne le résultat suivant [370]. Cette

condition est aussi nécessaire et suffisante. La condition (4.41) est parfois appelée la *condition de qualification de Mangasarian-Fromovitz stricte* (QC-MFS).

Proposition 4.44 (unicité du multiplicateur) *Soit x_* un point stationnaire de (P_{EI}) (donc f et $c_{E\cup I^0_*}$ sont différentiables en x_* et $\Lambda_* \neq \varnothing$). Alors $\Lambda_* = \{\lambda_*\}$ si, et seulement si,*

$$\left.\begin{array}{l} \displaystyle\sum_{i\in E\cup I^0_*} \alpha_i \nabla c_i(x_*) = 0 \\ \alpha_i \geqslant 0 \text{ pour tout } i \in I^{00}_* \end{array}\right\} \implies \alpha_{E\cup I^0_*} = 0. \tag{4.41}$$

Démonstration. [⇒] On montre la contraposée. Supposons qu'il existe des coefficients α_i non tous nuls tels que $\sum_{i\in E\cup I^0_*} \alpha_i \nabla c_i(x_*) = 0$ et $\alpha_i \geqslant 0$ pour tout $i \in I^{00}_*$ (c'est par cet ensemble d'indices que λ_* intervient). On pose $\alpha_i = 0$ pour $i \in I \setminus I^0_*$. Alors $\lambda_* + \varepsilon\alpha \in \Lambda_*$ pour $\varepsilon > 0$ petit et Λ_* ne serait pas réduit à un seul multiplicateur.

[⇐] Soit $\lambda'_* \in \Lambda_*$ et $\alpha := \lambda'_* - \lambda_*$. Il suffit de montrer que $\alpha = 0$. Clairement : $\sum_{i\in E\cup I^0_*} \alpha_i \nabla c_i(x_*) = 0$ et, pour $i \in I^{00}_*$, on a $\alpha_i = (\lambda'_*)_i \geqslant 0$. D'après (4.41), $\alpha = 0$. □

Mais en toute généralité, Λ_* peut ne pas être réduit à un point. En ce qui concerne son caractère borné, on a le résultat suivant [231].

Proposition 4.45 (bornitude des multiplicateurs) *Soient (x_*, λ_*) un couple vérifiant les conditions de KKT (4.33) et Λ_* l'ensemble convexe fermé non vide des multiplicateurs optimaux associés à x_*. Alors Λ_* est borné si, et seulement si, les conditions de qualification de Mangasarian-Fromovitz (QC-MF) ont lieu en x_*.*

Démonstration. C'est une conséquence de l'équivalence $(i) \Leftrightarrow (iv)$ de la proposition 4.42. □

4.4.4 Conditions d'optimalité du second ordre

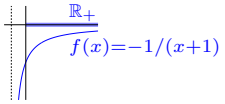
Les conditions d'optimalité du second ordre en présence de contraintes d'inégalité ne s'obtiennent ni ne s'écrivent aussi aisément que lorsqu'il n'y a que des contraintes d'égalité. D'abord, ce n'est pas le cône tangent qui se manifeste dans ces conditions comme dans le théorème 4.21, mais un cône plus petit que l'on appelle le *cône critique*. Par ailleurs, le multiplicateur optimal qui intervient dans la hessienne du lagrangien doit être déterminé en fonction de la direction choisie dans le cône critique. Nous commencerons l'étude des conditions du deuxième ordre par deux sous-sections donnant un premier éclairage sur ces changements importants. Ensuite viendront les conditions nécessaires puis les conditions suffisantes du second ordre.

L'approche suivie ici pour établir les conditions *nécessaires* d'optimalité du second ordre supposera d'emblée la qualification de Mangasarian-Fromovitz. La condition nécessaire d'optimalité du théorème 4.48 reste pourtant valable et peut même être renforcée si les contraintes sont localement affines (elle l'est clairement si (QC-IL) a lieu, car (QC-MF) en découle). Cela peut se voir par la même approche que celle utilisée dans la démonstration du théorème 4.21. L'exercice 4.21 montre ainsi que si la condition de qualification (QC-A) ou (QC-IL) est vérifiée, on a des conditions du second ordre fortes (voir plus loin pour la définition de cette expression).

Le cône critique

Il est tentant d'essayer de généraliser le théorème 4.21 au problème (P_{EI}), en cherchant à montrer qu'en une solution primale-duale (x_*, λ_*), on doit avoir $\langle L_* d, d\rangle$ positif pour toute direction tangente $d \in \mathrm{T}_{x_*} X_{EI}$; on a noté $L_* := \nabla^2_{xx}\ell(x_*, \lambda_*)$ la hessienne du lagrangien au point stationnaire considéré. Ce résultat n'est pas correct, car le cône tangent $\mathrm{T}_{x_*} X_{EI}$ n'est pas celui qui convient, comme le montre le problème suivant

$$\min\{-1/(x+1) : x \in \mathbb{R}_+\}. \tag{4.42}$$



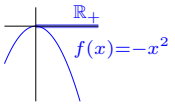
Ce problème a pour unique solution primale-duale $(x_*, \lambda_*) = (0, 1)$ et le cône tangent en la solution s'écrit $\mathrm{T}_{x_*} X_{EI} = \mathbb{R}_+$ si bien que l'on peut prendre $d = 1$ comme direction tangente, mais $\langle L_* d, d\rangle = -2$ n'est pas positif. Nous verrons que $\langle L_* d, d\rangle$ est positif, mais pour des directions d dans un cône plus petit que le cône tangent.

À la recherche d'un cône plus petit, on peut observer que, comme toute solution x_* du problème (P_{EI}) minimise aussi f *localement* sur

$$X^{=}_{EI} := \{x \in \mathbb{E} : c_{E\cup I^0_*}(x) = 0,\ c_{I\setminus I^0_*}(x) < 0\},$$

inclus dans X_{EI}, le théorème 4.21 nous apprend que $\langle L_* d, d\rangle$ est positif pour toute direction $d \in \mathrm{T}_{x_*} X^{=}_{EI}$ et toute solution duale $\lambda_* \in \Lambda_*$ (celles-ci sont aussi solutions duales du problème minimisant f sur $X^{=}_{EI}$). Nous verrons cependant que le cône $\mathrm{T}_{x_*} X^{=}_{EI}$ est trop petit, dans le sens où il ne permet pas d'établir des conditions suffisantes d'optimalité du second ordre. Considérons en effet le problème

$$\min\{-x^2 : x \in \mathbb{R}_+\}. \tag{4.43}$$



Si $x_* = 0$, $X^{=}_{EI} = \{0\}$, si bien que $\mathrm{T}_{x_*} X^{=}_{EI} = \{0\}$ et la hessienne du lagrangien $L_* = -2$ est bien définie positive sur $\mathrm{T}_{x_*} X^{=}_{EI} \setminus \{0\} = \varnothing$, mais x_* n'est pas un minimum local du problème.

Le bon cône s'avérera être le cône linéarisant $\mathrm{T}'_{x_*} X_{EI,f}$ de l'ensemble

$$X_{EI,f} := \{x \in X_{EI} : f(x) \leqslant f(x_*)\}, \tag{4.44}$$

sur lequel f est également minimisée en une solution x_* de (P_{EI}). Ce cône est plus petit que le cône tangent à l'ensemble admissible en x_*, mais suffisamment grand pour permettre d'avoir des conditions suffisantes d'optimalité du second ordre (théorème 4.49). On l'appelle le cône critique du problème.

Définition 4.46 On appelle *cône critique* du problème (P_{EI}) en un point admissible $x \in X_{EI}$, l'ensemble noté et défini par

$$C(x) := \{d \in \mathbb{E} : c'_E(x) \cdot d = 0,\ c'_{I^0_*}(x) \cdot d \leqslant 0,\ f'(x) \cdot d \leqslant 0\}. \tag{4.45a}$$

Une direction $d \in C(x)$ est appelée *direction critique* en x. On utilise la notation simplifiée $C_* := C(x_*)$. □

Dans l'exemple (4.42), $C_* = \{0\}$ est plus petit que le cône tangent $\mathrm{T}_{x_*} X_{EI} = \mathbb{R}_+$. Dans l'exemple (4.43), $C_* = \mathbb{R}_+$ est plus grand que le cône tangent $\mathrm{T}_{x_*} X^{=}_{EI} = \{0\}$. Il est remarquable que l'optimalité au second ordre puisse être synthétisée au moyen de l'unique cône critique, alors que les deux problèmes précédents recouvrent des situations très différentes.

En un point stationnaire x_*, de multiplicateur λ_*, le cône critique en x_* s'écrit aussi

$$\begin{aligned} C_* &= \{d \in \mathbb{E} : c'_E(x_*) \cdot d = 0,\ c'_{I^0_*}(x_*) \cdot d \leqslant 0,\ f'(x_*) \cdot d = 0\}, && \text{(4.45b)} \\ &= \{d \in \mathbb{E} : c'_{E \cup I^{0+}_*}(x_*) \cdot d = 0,\ c'_{I^{00}_*}(x_*) \cdot d \leqslant 0\}, && \text{(4.45c)} \end{aligned}$$

où les ensembles d'indices I^{0+}_* et I^{00}_* ont été introduits en (4.40) et dépendent du multiplicateur λ_*. Ces expressions s'obtiennent en utilisant les conditions d'optimalité (4.33). Observons enfin que si les conditions de complémentarité stricte (4.36) sont satisfaites, le cône critique s'écrit simplement

$$C_* = \{d \in \mathbb{E} : c'_{E \cup I^0_*}(x_*) \cdot d = 0\}, \tag{4.45d}$$

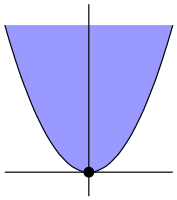
qui est donc le cône linéarisant $\mathrm{T}'_{x_*} X^{=}_{EI}$ (un sous-espace vectoriel). En l'absence de complémentarité stricte, le sous-espace vectoriel (4.45d) est contenu dans C_*, lui-même contenu dans $\mathrm{T}'_{x_*} X_{EI}$.

Trois exemples instructifs

Une autre difficulté dans l'établissement des conditions d'optimalité du second ordre du problème (P_{EI}) provient du fait que l'on doit prendre le multiplicateur optimal λ_* intervenant dans la hessienne du lagrangien $L_* := \nabla^2_{xx} \ell(x_*, \lambda_*)$ en fonction de la direction critique choisie. Autrement dit, la démonstration du théorème 4.21, fondée sur le développement du lagrangien $\ell(\cdot, \lambda_*)$ pour un multiplicateur λ_* fixé indépendamment de la direction critique, ne fonctionne pas pour une solution du problème (P_{EI}) ne vérifiant que la qualification de Mangasarian-Fromovitz (QC-MF) (voir l'exercice 4.21 pour les qualifications (QC-A) et (QC-IL), qui permettent d'utiliser cette technique). Trois exemples vont nous permettre de mieux comprendre pourquoi il en est ainsi et d'apprendre à sélectionner correctement les quantificateurs qui s'appliquent à $d \in C_*$ et $\lambda_* \in \Lambda_*$.

Considérons d'abord le problème à deux variables réelles

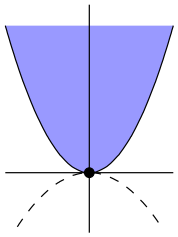
$$\begin{cases} \min\ x_2 \\ x_2 \geqslant x_1^2, \end{cases} \tag{4.46}$$



dont l'ensemble admissible est représenté à droite ci-dessus et dont la solution est $x_* = (0,0)$. Il y a un unique multiplicateur optimal associé à la contrainte, valant $\lambda_* = 1$. La contrainte étant active, (x_*, λ_*) est aussi la solution primale-duale du problème avec contrainte d'égalité $x_2 = x_1^2$, si bien que la hessienne du lagrangien $L_* := \nabla^2_{xx}\ell(x_*, \lambda_*) = \mathrm{diag}(2,0)$ doit être semi-définie positive dans l'espace tangent à la contrainte (corollaire 4.23) : $\langle L_* d, d\rangle \geqslant 0$ pour toute direction d dans $\{d \in \mathbb{R}^2 : d_2 = 0\}$. C'est le cas le plus simple qui peut se présenter. On dira plus tard que l'on a des *conditions d'optimalité du deuxième ordre fortes* : quel que soit le multiplicateur optimal (il n'y en a qu'un seul ici), L_* est semi-défini positif dans le cône critique. Ces conditions sont vérifiées s'il y a un unique multiplicateur (voir la proposition 4.44 pour une condition nécessaire et suffisante pour que cette propriété ait lieu), comme ici ou comme lorsque la qualification (QC-A) ou (QC-IL) a lieu (voir l'exercice 4.21).

Considérons à présent une variante du problème (4.46) où l'on ajoute une contrainte superflue

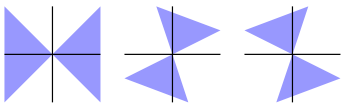
$$\begin{cases} \min\ x_2 \\ x_2 \geqslant x_1^2 \\ x_2 \geqslant -\frac{1}{2}x_1^2. \end{cases} \qquad (4.47)$$



La seconde contrainte ne modifie pas la solution primale qui est toujours $x_* = (0,0)$, mais il y a maintenant plusieurs multiplicateurs optimaux formant l'ensemble $\Lambda_* = \{\lambda \in \mathbb{R}^2_+ : \lambda_1 + \lambda_2 = 1\}$. En prenant comme multiplicateur $\lambda_* = (1,0)$, un sommet de Λ_*, on ignore la seconde contrainte (comme il se doit) et on a le résultat précédent sur la semi-définie positivité de $L_* = \mathrm{diag}(2,0)$ dans $\{d \in \mathbb{R}^2 : d_2 = 0\}$. Par contre, avec $\lambda_* = (0,1)$, l'autre sommet de Λ_*, la hessienne du lagrangien $L_* = \mathrm{diag}(-1,0)$ est définie *négative* dans $\{d \in \mathbb{R}^2 : d_2 = 0\}$. C'est normal ; le lagrangien ne voit que la seconde contrainte, ignorant la première, et $(0,0)$ n'est qu'un point stationnaire du problème $\min\{x_2 : x_2 \geqslant -\frac{1}{2}x_1^2\}$, pas un minimum local. On dira plus tard que l'on a des *conditions d'optimalité du deuxième ordre semi-fortes* : il existe un multiplicateur optimal tel que L_* est semi-défini positif dans le cône critique.

Considérons enfin le problème à trois variables

$$\begin{cases} \min\ x_3 \\ x_3 \geqslant (x_1 + x_2)(x_1 - x_2) \\ x_3 \geqslant (x_2 + 3x_1)(2x_2 - x_1) \\ x_3 \geqslant (2x_2 + x_1)(x_2 - 3x_1). \end{cases} \qquad (4.48)$$



Les trois figures à droite représentent, pour chacune des trois contraintes, les coordonnées (x_1, x_2) qui donnent une valeur positive à leur membre de droite. On voit que, quel que soit (x_1, x_2) non nul, un des membres de droite des contraintes est strictement positif. Dès lors, l'unique solution de ce problème est $x_* = 0$. D'autre part, l'ensemble des multiplicateurs optimaux est le simplexe unitaire $\Lambda_* = \{\lambda \in \mathbb{R}^3_+ : \lambda_1 + \lambda_2 + \lambda_3 = 1\}$. Enfin, la hessienne du lagrangien s'écrit

$$L(x,\lambda) = \begin{pmatrix} 2\lambda_1 - 6(\lambda_2 + \lambda_3) & 5(\lambda_2 - \lambda_3) & 0 \\ 5(\lambda_2 - \lambda_3) & -2\lambda_1 + 4(\lambda_2 + \lambda_3) & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Quelle que soit la valeur de $\lambda_* \in \Lambda_*$, L_* n'est pas semi-définie positive sur le cône critique, qui est ici le sous-espace $\{d \in \mathbb{R}^3 : d_3 = 0\}$ (en effet, l'élément $(1,1)$ vaut $8\lambda_1 - 6$ et l'élément (2,2) vaut $4 - 6\lambda_1$, si bien qu'il faudrait que λ_1 soit supérieur à $3/4$ et inférieur à $2/3$). On dira plus tard que l'on a des *conditions d'optimalité du deuxième ordre faibles*: pour toute direction critique d, il existe un multiplicateur optimal λ_* (dépendant de d) tel que $\langle L_* d, d\rangle \geqslant 0$.

Conditions nécessaires du second ordre

Le théorème 4.48 ci-dessous énonce des conditions nécessaires d'optimalité du deuxième ordre en supposant que les conditions de qualification de Mangasarian-Fromovitz (QC-MF) ont lieu, ce qui n'est pas très restrictif (proposition 4.43).

Ce résultat se fonde sur l'examen de l'allure du critère f le long de chemins $t \mapsto \xi(t)$ émanant en $t = 0$ de la solution considérée x_* et restant dans l'ensemble admissible X_{EI} pour de petites valeurs *positives* du paramètre t. Plus on considère de tels chemins, plus on aura d'information. La démarche était la même dans le cas où il n'y avait que des contraintes d'égalité, mais il s'est avéré que tous les chemins ayant la même tangente à l'origine donnaient la même information au deuxième ordre (on utilisait en fait que des suites associées à de telles directions tangentes). Dans le cas où l'on a des contraintes d'inégalité, pour une direction tangente $d \in \mathrm{T}_{x_*} X_{EI}$, il va falloir distinguer les chemins ayant $d = \xi'(0)$ comme tangente à l'origine, mais ayant des courbures $h = \xi''(0)$ différentes. Tangence et courbure doivent vérifier des relations de compatibilité de manière à ce que le chemin rentre dans l'ensemble admissible. On cherche donc des chemins $t \mapsto \xi(t)$ tels que

$$\xi(0) = x_*, \quad \xi'(0) = d, \quad \xi''(0) = h, \quad \xi(t) \in X_{EI}, \text{ pour } t \geqslant 0 \text{ petit}. \tag{4.49}$$

Voyons quelles sont les conditions que doivent vérifier d et h pour que l'on puisse en effet construire un chemin vérifiant (4.49). Des conditions nécessaires sont aisées à obtenir, le lemme 4.47 montrera qu'elles sont presque suffisantes.

L'application ξ doit vérifier $c_E(\xi(t)) = 0$, pour tout $t > 0$ petit. En dérivant par rapport à t en $t = 0$, on voit qu'il faut que $c'_E(x_*)\cdot d = 0$. Si l'on suppose leur continuité, les contraintes d'inégalité inactives en x_* le resteront pour de petites valeurs de $t > 0$. En ce qui concerne les contraintes d'inégalité actives en x_*, on doit avoir pour $t > 0$ petit :

$$0 \geqslant c_{I^0_*}(\xi(t)) = c_{I^0_*}(x_*) + t\, c'_{I^0_*}(x_*) \cdot d + o(t).$$

Comme $c_{I^0_*}(x_*) = 0$, un passage à la limite lorsque $t \downarrow 0$ après division par $t > 0$ donne $c'_{I^0_*}(x_*) \cdot d \leqslant 0$. En résumé, l'analyse au premier ordre montre qu'il faut que d appartienne au cône tangent $\mathrm{T}_{x_*} X_{EI}$ (sous l'hypothèse de qualification (QC-MF), voir la proposition 4.29). C'est la seule condition qui portera sur d.

En dérivant deux fois $c_E(\xi(t)) = 0$ par rapport à t en $t = 0$ (ce qui demandera d'avoir un chemin deux fois dérivable en $t = 0$), on voit que l'on doit avoir

$$c''_E(x_*) \cdot d^2 + c'_E(x_*) \cdot h = 0. \tag{4.50}$$

C'est la première condition sur h. Pour les indices $i \in I^0_*$ des contraintes d'inégalité actives, en utilisant le fait que $c_i(x_*) = 0$ et $c'_i(x_*) \cdot d \leqslant 0$, on obtient

$$\begin{aligned} c_i(\xi(t)) &= c_i(x_*) + tc_i'(x_*)\cdot d + \frac{t^2}{2}\left(c_i''(x_*)\cdot d^2 + c_i'(x_*)\cdot h\right) + o(t^2) \\ &\leqslant \frac{t^2}{2}\left(c_i''(x_*)\cdot d^2 + c_i'(x_*)\cdot h\right) + o(t^2). \end{aligned}$$

Pour que cette quantité soit négative pour $t > 0$ petit, on va demander la stricte négativité des termes entre parenthèses en se donnant $\epsilon > 0$ et en imposant

$$c_{I_*^0}''(x_*)\cdot d^2 + c_{I_*^0}'(x_*)\cdot h + \epsilon \leqslant 0. \tag{4.51}$$

C'est la deuxième condition sur h. Notons que, lorsque les conditions de qualification de Mangasarian-Fromovitz (QC-MF) ont lieu, on peut toujours trouver un h vérifiant (4.50) et (4.51) (proposition 4.42).

Lemme 4.47 (existence d'un chemin admissible sous (QC-MF)) *Soit $x_* \in X_{EI}$. On suppose que c_E est C^2 dans un voisinage de x_*, que $c_{I_*^0}$ est deux fois dérivable en x_* et que $c_{I\setminus I_*^0}$ est continue en x_*. On se donne $d \in \mathrm{T}_{x_*} X_{EI}$. On suppose également que les conditions de Mangasarian-Fromovitz (QC-MF) ont lieu en x_* et donc que pour $\epsilon > 0$ donné, on peut trouver $h \in \mathbb{E}$ vérifiant (4.50) et (4.51). Alors il existe un chemin $t \mapsto \xi(t)$ de classe C^2, défini pour $|t|$ suffisamment petit, tel que l'on ait (4.49).*

Démonstration. La technique de démonstration est semblable à celle utilisée dans la démonstration de la proposition 4.14. Soit $A_E = c_E'(x_*)$ la jacobienne de c_E en x_*. Celle-ci étant surjective, on peut trouver une matrice Z, $(n - m_E) \times n$, telle que

$$\begin{pmatrix} A_E \\ Z \end{pmatrix} \text{ soit inversible.}$$

Pour $d \in \mathrm{T}_{x_*} X_{EI}$ et $h \in \mathbb{E}$ donnés dans l'énoncé, on considère la fonction $F : \mathbb{E}\times\mathbb{R} \to \mathbb{R}^n$ définie par

$$F(\xi, t) = \begin{pmatrix} c_E(\xi) \\ Z(\xi - x_* - td - \frac{t^2}{2}h) \end{pmatrix}.$$

Elle est C^2 dans un voisinage de $(x_*, 0)$, $F(x_*, 0) = 0$ et $F_\xi'(x_*, 0)$ est inversible. Par le théorème des fonctions implicites (théorème C.14), il existe alors une fonction $t \mapsto \xi(t)$, définie pour $|t|$ petit, de classe C^2, telle que $F(\xi(t), t) = 0$ pour tout $|t|$ petit et $\xi(0) = x_*$. En dérivant une première fois $F(\xi(t), t) = 0$ en $t = 0$, on trouve

$$\begin{pmatrix} A_E \\ Z \end{pmatrix} \xi'(0) = \begin{pmatrix} 0 \\ Zd \end{pmatrix} = \begin{pmatrix} A_E \\ Z \end{pmatrix} d,$$

car $A_E d = 0$. Donc $\xi'(0) = d$. En dérivant deux fois $F(\xi(t), t) = 0$ en $t = 0$, on trouve

$$\begin{pmatrix} A_E \\ Z \end{pmatrix} \xi''(0) = \begin{pmatrix} -c_E''(x_*)\cdot d^2 \\ Zh \end{pmatrix} = \begin{pmatrix} A_E \\ Z \end{pmatrix} h,$$

par (4.50). Donc $\xi''(0) = h$. Enfin, $\xi(t) \in X_{EI}$ pour $t \geqslant 0$ petit, grâce au choix de h vérifiant (4.50) et (4.51). □

Théorème 4.48 (CN2) *Soit x_* une solution locale de (P_{EI}). Supposons que f et c_E soient C^2 dans un voisinage de x_*, que $c_{I^0_*}$ soit deux fois dérivable en x_* et que $c_{I\setminus I^0_*}$ soit continue en x_*. Supposons également que les conditions de qualification de Mangasarian-Fromovitz (QC-MF) aient lieu en x_*. Alors*

$$\forall\, d \in C_* : \quad \max_{\lambda_* \in \Lambda_*} \langle L_* d, d\rangle \geqslant 0. \tag{4.52}$$

Démonstration. Soit $d \in C_* \subseteq \mathrm{T}_{x_*} X_{EI}$. Pour chaque choix de $h \in \mathbb{E}$ vérifiant (4.50) et (4.51), on a un chemin $t \mapsto \xi(t)$ de classe C^2 satisfaisant (4.49). Voyons comment choisir h. Un développement de $t \mapsto f(\xi(t))$ au deuxième ordre donne

$$f(\xi(t)) = f(x_*) + t f'(x_*)\cdot d + \frac{t^2}{2}\left(f''(x_*)\cdot d^2 + f'(x_*)\cdot h\right) + o(t^2).$$

Par optimalité $f(\xi(t)) \geqslant f(x_*)$, pour $t \geqslant 0$ petit (car alors $\xi(t) \in X_{EI}$), et $f'(x_*)\cdot d \leqslant 0$ (car $d \in C_*$). Dès lors

$$0 \leqslant \frac{t^2}{2}\left(f''(x_*)\cdot d^2 + f'(x_*)\cdot h\right) + o(t^2). \tag{4.53}$$

Pour obtenir le plus d'informations possible, on a avantage à trouver un h minimisant les termes entre parenthèses.

Ceci nous conduit à considérer le problème linéaire en $h \in \mathbb{E}$ suivant

$$\begin{cases} \min \left(f'(x_*)\cdot h + f''(x_*)\cdot d^2\right) \\ c_i'(x_*)\cdot h + c_i''(x_*)\cdot d^2 = 0, \quad \text{pour } i \in E \\ c_i'(x_*)\cdot h + c_i''(x_*)\cdot d^2 + \epsilon \leqslant 0, \quad \text{pour } i \in I^0_*. \end{cases} \tag{4.54}$$

Il est réalisable (son ensemble admissible est non vide) grâce à la qualification de Mangasarian-Fromovitz (proposition 4.42). Son dual est le problème linéaire en $\lambda \in \mathbb{R}^m$ suivant

$$\begin{cases} \max \left(\langle L(x_*,\lambda)d, d\rangle + \epsilon\|\lambda_I\|_1\right) \\ \nabla_x \ell(x_*,\lambda) = 0 \\ \lambda_{I^0_*} \geqslant 0 \\ \lambda_{I\setminus I^0_*} = 0, \end{cases}$$

dont l'ensemble admissible est Λ_*, également non vide par les conditions de Mangasarian-Fromovitz. D'après le théorème 17.11 de dualité forte, ces deux problèmes ont alors une solution et il n'y a pas de saut de dualité : si h est une solution de (4.54),

$$f'(x_*)\cdot h + f''(x_*)\cdot d^2 = \max_{\lambda_* \in \Lambda_*} \left(\langle L_* d, d\rangle + \epsilon\|(\lambda_*)_I\|_1\right). \tag{4.55}$$

On peut à présent conclure. Introduisons la suite $\{x_k\}$ définie par $x_k = \xi(t_k)$, avec $t_k \downarrow 0$. Par construction, $x_k \in X_{EI}$ pour k assez grand. D'autre part, par (4.53) et (4.55), on a

$$0 \leqslant \frac{t_k^2}{2}\Big(\max_{\lambda_* \in \Lambda_*} \langle L_* d, d\rangle + \epsilon \|(\lambda_*)_I\|_1\Big) + o(t_k^2).$$

En passant à la limite lorsque $t_k \downarrow 0$, après avoir diviser par t_k^2, on obtient

$$0 \leqslant \max_{\lambda_* \in \Lambda_*} \Big(\langle L_* d, d\rangle + \epsilon \|(\lambda_*)_I\|_1\Big).$$

Comme $\epsilon > 0$ est arbitraire et que Λ_* est borné (proposition 4.45), on obtient le résultat. □

Sous (QC-MF), Λ_* est compact, si bien que la condition (4.52) s'écrit aussi

$$\forall\, d \in C_*,\ \exists\, \lambda_* \in \Lambda_* : \quad \langle L_* d, d\rangle \geqslant 0. \tag{4.56}$$

On dit alors que l'on a des *conditions nécessaires d'optimalité du second ordre faibles*. Elles sont vérifiées sous les conditions du théorème 4.48. C'est donc le cas dans l'exemple 4.48. Si l'on a les conditions plus fortes suivantes

$$\exists\, \lambda_* \in \Lambda_*,\ \forall\, d \in C_* : \quad \langle L_* d, d\rangle \geqslant 0, \tag{4.57}$$

on dit que l'on a des *conditions nécessaires d'optimalité du second ordre semi-fortes*. Elles sont vérifiées dans l'exemple 4.47. Si l'on a les conditions encore plus fortes suivantes

$$\forall\, \lambda_* \in \Lambda_*,\ \forall\, d \in C_* : \quad \langle L_* d, d\rangle \geqslant 0, \tag{4.58}$$

on dit que l'on a des *conditions nécessaires d'optimalité du second ordre fortes*. Elles sont vérifiées dans l'exemple 4.46. Ces dernières conditions sont vérifiées s'il y a un unique multiplicateur, par exemple lorsque (QC-IL) a lieu ou lorsque les conditions de la proposition 4.44 s'appliquent. Il est montré à l'exercice 4.21, que ces conditions fortes ont aussi lieu sous l'hypothèse de qualification (QC-A).

La vérification numérique des conditions nécessaires d'optimalité du second ordre n'est pas aisée. Déjà, lorsque les conditions semi-fortes (4.57) ont lieu pour un multiplicateur optimal λ_*, il s'agit de vérifier que la forme quadratique $d \mapsto \langle L_* d, d\rangle$ associée à la hessienne du lagrangien est semi-définie positive sur le cône C_*, qui est polyédrique, c'est-à-dire que L_* est C_*-*copositive* [325, 49, 318]. En toute généralité, une telle vérification est un problème NP-ardu [447, 170]. Maintenant, s'il y a aussi complémentarité stricte, le cône critique devient le sous-espace vectoriel (4.45d) et la vérification de la semi-définie positivité de $d \mapsto \langle L_* d, d\rangle$ sur ce sous-espace est alors une opération simple d'algèbre linéaire.

Conditions suffisantes du second ordre

La proposition suivante donne des conditions suffisantes d'optimalité du second ordre pour le problème (P_{EI}). On notera qu'elles ne font pas intervenir d'hypothèse de qualification de contrainte. Le fait que le cône critique intervienne aussi dans ces conditions suffisantes d'optimalité est une garantie de sa pertinence. Une solution x_* vérifiant ces conditions est appelé un *minimum faible* du problème (P_{EI}). L'inégalité (4.61) est connue sous le nom de *propriété de croissance quadratique*. Elle montre que f croît au moins quadratiquement lorsque'on se déplace de x_* vers l'« intérieur » de X_{EI}.

Théorème 4.49 (CS2) *Supposons que f et $c_{E\cup I^0_*}$ soient dérivables dans un voisinage d'un point $x_* \in \mathbb{E}$ et deux fois dérivables en x_*. Supposons également que l'ensemble Λ_* des multiplicateurs λ_* tels que (x_*, λ_*) vérifie les conditions d'optimalité de KKT (4.33) ne soit pas vide. Supposons enfin que*

$$\forall d \in C_* \setminus \{0\},\ \exists\, \lambda_* \in \Lambda_* : \quad \langle L_* d, d\rangle > 0, \tag{4.59}$$

ou de manière équivalente ($\|\cdot\|$ est une norme arbitraire)

$$\exists\, \bar\gamma > 0,\ \forall d \in C_*,\ \exists\, \lambda_* \in \Lambda_* : \quad \langle L_* d, d\rangle \geqslant \bar\gamma \|d\|^2. \tag{4.60}$$

Alors, pour tout $\gamma \in [0, \bar\gamma[$, il existe un voisinage V de x_* tel que pour tout $x \in X_{EI} \cap V$, différent de x_* :

$$f(x) > f(x_*) + \frac{\gamma}{2}\|x - x_*\|^2. \tag{4.61}$$

En particulier, x_* est un minimum local strict de (P_{EI}).

Démonstration. Montrons d'abord que (4.59) et (4.60) sont équivalentes. Il est clair que (4.60) implique (4.59). Inversement, supposons que l'on ait (4.59), mais pas (4.60). Alors il existerait une suite $\{d_k\} \subseteq C_*$ telle que $\|d_k\| = 1$ et $\sup_{\lambda\in\Lambda_*} \langle L(x_*, \lambda)d_k, d_k\rangle \downarrow 0$. On pourrait extraire une sous-suite $d_k \to d \in C_* \setminus \{0\}$, car C_* est fermé. Si l'on note λ_* le multiplicateur associé à d par (4.59), on voit que cela conduirait à une contradiction :

$$0 \leftarrow \sup_{\lambda\in\Lambda_*} \langle L(x_*, \lambda)d_k, d_k\rangle \geqslant \langle L_* d_k, d_k\rangle \to \langle L_* d, d\rangle > 0.$$

Pour la suite, on raisonne à nouveau par l'absurde. Si le résultat n'est pas vrai, on peut trouver un $\gamma \in [0, \bar\gamma[$ et une suite $\{x_k\} \subseteq X_{EI}$ telle que $x_k \to x_*$, $x_k \neq x_*$ et $f(x_k) \leqslant f(x_*) + \frac{\gamma}{2}\|x_k - x_*\|^2$. En extrayant une sous-suite au besoin, on peut supposer qu'avec $t_k := \|x_k - x_*\|$, on a

$$\frac{x_k - x_*}{t_k} \to d.$$

Donc $d \in \mathrm{T}_{x_*} X_{EI} \setminus \{0\}$. D'autre part, de $f(x_*) + \frac{\gamma}{2}\|x_k - x_*\|^2 \geqslant f(x_k) = f(x_*) + f'(x_*)\cdot(x_k - x_*) + o(\|x_k - x_*\|)$, on déduit $f'(x_*)\cdot d \leqslant 0$ et donc $d \in C_* \setminus \{0\}$.

Il reste à exhiber une contradiction, ce que l'on obtient en développant le lagrangien $\ell(\cdot, \lambda_*)$, où λ_* est le multiplicateur associé à d par (4.60) :

$$\ell(x_k, \lambda_*) = \ell(x_*, \lambda_*) + \frac{1}{2}\ell''_{xx}(x_*, \lambda_*)\cdot(x_k - x_*)^2 + o(\|x_k - x_*\|^2).$$

On a $\ell(x_*, \lambda_*) = f(x_*)$ et $\ell(x_k, \lambda_*) \leqslant f(x_k) \leqslant f(x_*) + \frac{\gamma}{2}\|x_k - x_*\|^2$. Donc

$$\frac{\gamma}{2}\|x_k - x_*\|^2 \geqslant \frac{1}{2}\langle L_*(x_k - x_*), x_k - x_*\rangle + o(\|x_k - x_*\|^2).$$

En divisant par t_k^2 et en passant à la limite, on trouve

$$\langle L_* d, d\rangle \leqslant \gamma \|d\|^2,$$

ce qui contredit (4.60), puisque $\gamma < \bar{\gamma}$ et $d \in C_* \setminus \{0\}$. □

Le résultat de la proposition 4.49 reste vrai si dans (4.59), λ_* peut être pris indépendamment de d :

$$\exists \lambda_* \in \Lambda_*,\ \forall d \in C_* \setminus \{0\} : \quad \langle L_* d, d\rangle > 0. \tag{4.62}$$

On dit alors que des *conditions suffisantes d'optimalité du second ordre semi-fortes* ont lieu. Le résultat reste *a fortiori* vrai si λ_* peut être pris arbitraire :

$$\forall \lambda_* \in \Lambda_*,\ \forall d \in C_* \setminus \{0\} : \quad \langle L_* d, d\rangle > 0. \tag{4.63}$$

On parle alors de *conditions suffisantes d'optimalité du second ordre fortes*. On peut montrer que si dans la proposition 4.49 on remplace (4.59) par (4.63), si les conditions de qualification (QC-MF) ont lieu et si les fonctions f et c sont un peu plus régulières, il n'y a pas d'autre point stationnaire que x_* dans un voisinage de x_* : c'est un point stationnaire isolé (voir l'exercice 4.20).

4.4.5 Calcul pratique des solutions de (P_{EI})

La proposition 4.32 nous dit que, sous certaines conditions, en particulier de qualification des contraintes, une solution locale de (P_{EI}) est un point stationnaire de ce problème, ce qui veut dire qu'elle vérifie l'ensemble des relations de (4.33). Pour calculer les solutions de (P_{EI}), on pourra donc, dans un premier temps, calculer les solutions du système d'optimalité (4.33). Toutefois, ce calcul est beaucoup plus difficile que lorsqu'il n'y a que des contraintes d'égalité. La difficulté vient de la présence d'inégalités et, en particulier, des conditions de complémentarité. Comme nous allons le voir, nous y retrouverons la *combinatoire des problèmes d'optimisation avec contraintes d'inégalité*, confirmant ainsi le *principe de conservation des ennuis* dont nous avons parlé dans l'introduction de la section 4.4.

En général, il faut utiliser des algorithmes spécifiques pour résoudre ce système d'optimalité (c'est ce à quoi est consacré une grande partie de cet ouvrage !). Dans certains cas, cependant, en particulier pour des problèmes de petite taille ayant peu de contraintes d'inégalité ou des problèmes structurés, on pourra chercher les solutions analytiquement en considérant l'une après l'autre toutes les manières de satisfaire les contraintes d'inégalité. Ainsi, dans chaque cas considéré, on suppose qu'un certain nombre de contraintes d'inégalité sont actives et que les autres ne le sont pas. Soit $J \subseteq I$, l'ensemble des indices des contraintes d'inégalité supposées actives en la solution : $c_{E\cup J}(x_*) = 0$ et $c_{J^c}(x_*) < 0$ (on a noté $J^c = I \setminus J$). Du fait de la complémentarité, $(\lambda_*)_{J^c} = 0$ et on est donc conduit à chercher les solutions du système de $n + |E \cup J|$ équations à $n + |E \cup J|$ inconnues suivant

$$\begin{cases} \nabla f(x_*) + c'_{E\cup J}(x_*)^*(\lambda_*)_{E\cup J} = 0 \\ c_{E\cup J}(x_*) = 0. \end{cases}$$

Si une solution $(x_*, (\lambda_*)_{E\cup J})$ de ce système vérifie $c_{J^c}(x_*) < 0$ et $(\lambda_*)_J \geqslant 0$, on satisfait les hypothèses du cas considéré et $(x_*, ((\lambda_*)_{E\cup J}, 0_{J^c}))$ est une solution de

(4.33). Sinon cette solution doit être écartée. En examinant ainsi tous les ensembles $J \subseteq I$ possibles, on peut trouver tous les points stationnaires du problème.

La méthode présentée ci-dessus est fastidieuse et n'est utilisée que dans de rares cas. On notera en effet qu'avec m_I contraintes d'inégalité, il y a 2^{m_I} cas à examiner, et donc 2^{m_I} systèmes non linéaires à résoudre. C'est une tâche considérable dès que m_I dépasse quelques unités ! C'est ici que l'on retrouve la *combinatoire des problèmes d'optimisation*. Le but des algorithmes d'optimisation pour problèmes avec contraintes d'inégalité est précisément de trouver des solutions du système d'optimalité (4.33), en gérant de manière efficace cette combinatoire, c'est-à-dire en évitant d'explorer toutes les possibilités. L'algorithme du simplexe (voir le chapitre 17) en a été un des premiers exemples.

On notera aussi que, si le calcul de points stationnaires est une tâche difficile, potentiellement NP-ardue, la *vérification* de la stationnarité d'un point $x_* \in \mathbb{E}$ est beaucoup plus simple. Il suffit en effet de voir si l'on peut trouver un multiplicateur $\lambda_* \in \mathbb{R}^m$ tel que

$$\nabla f(x_*) + c'_{E \cup I^0_*}(x_*)^*(\lambda_*)_{E \cup I^0_*} = 0 \qquad \text{et} \qquad (\lambda_*)_{I^0_*} \geqslant 0, \tag{4.64}$$

où l'on a noté comme d'habitude $I^0_* := \{i \in I : c_i(x_*) = 0\}$. Ce système est linéaire en λ_* (avec des inégalités toutefois) et l'on peut vérifier s'il a une solution par l'optimisation linéaire (en adaptant le problème (17.20) à la situation présente si l'on a besoin de démarrer un algorithme de résolution en un point admissible). Si le système linéaire (4.64) en λ_* n'a pas de solution, soit x_* n'est pas stationnaire, soit les contraintes ne sont pas qualifiées en x_*.

Comme pour les problèmes d'optimisation sans contrainte ou avec contraintes d'égalité, tous les points stationnaires (les solutions de (4.33)) ne sont pas solutions de (P_{EI}). Pour déterminer si un point stationnaire est solution de (P_{EI}), on pourra utiliser les conditions d'optimalité du second ordre, de la manière suivante :

- si (4.52) n'est pas vérifiée au point stationnaire, alors celui-ci n'est pas une solution locale de (P_{EI}) (ou une autre hypothèse du théorème 4.48 n'est pas vérifiée) ;
- si (4.59) est vérifiée au point stationnaire (ainsi que les autres hypothèses du théorème 4.49), alors celui-ci est un minimum local strict de (P_{EI}).

Ces deux cas recouvrent un grand nombre de situations, mais pas toutes, car les conditions (4.52) et (4.59) ne sont pas identiques. Le cas est indéterminé lorsqu'un point stationnaire vérifie (4.52), mais pas (4.59). Alors les résultats donnés ci-dessus ne sont pas suffisants et il faudra recourir à des conditions d'optimalité d'ordre supérieur pour pouvoir dire si le point stationnaire est solution de (P_{EI}).

4.5 Problème avec contraintes générales

Dans cette section, on cherche à établir des conditions d'optimalité de problèmes d'optimisation plus généraux que (P_{EI}), lequel a des contraintes fonctionnelles d'égalité et d'inégalité, à savoir des problèmes qui s'écrivent de la manière suivante :

$$(P_G) \quad \begin{cases} \min f(x) \\ c(x) \in G, \end{cases}$$

où $f : \mathbb{E} \to \mathbb{R}$ et $c : \mathbb{E} \to \mathbb{F}$ sont des fonctions *différentiables* et G est un *convexe fermé* non vide de $\mathbb{F}$ (pas nécessairement un *cône*). Dans cette description et dans toute cette section, $\mathbb{E}$ et $\mathbb{F}$ sont deux espaces euclidiens dont les produits scalaires sont tous deux notés $\langle \cdot, \cdot \rangle$. L'indice G de (P_G) faire référence au mot « général ». On désigne par

$$X_G := \{x \in \mathbb{E} : c(x) \in G\} = c^{-1}(G) \tag{4.65}$$

l'ensemble admissible du problème (P_G).

Le problème (P_G) est bien une généralisation du problème (P_{EI}), puisqu'on retrouve ce dernier en prenant

$$\mathbb{F} = \mathbb{R}^m \quad \text{et} \quad G = \{0_{\mathbb{R}^{m_E}}\} \times \mathbb{R}^{m_I}_{-}. \tag{4.66}$$

Le cas où l'on a en plus de la contrainte « $c(x) \in G$ », une contrainte « $x \in Q$ » d'appartenance à un convexe fermé non vide Q se déduit de (P_G) et est considéré à l'exercice 4.23.

Un des intérêts de cette formulation générale est d'avoir tout son sens en dimension infinie, ce qui n'est pas le cas de (P_{EI}). Il est en effet malaisé de spécifier ce qu'est un ensemble infini de contraintes d'inégalité à valeurs dans un espace de dimension infinie. On peut donc voir la généralisation proposée dans cette section comme un premier pas dans l'étude des problèmes d'optimisation de dimension infinie. Les résultats obtenus sont cependant aussi utiles pour résoudre certains problèmes de dimension finie à la structure différente de celle de (P_{EI}), comme les problèmes d'optimisation semi-définie positive du chapitre 20. Un autre avantage de cette abstraction est de mieux faire ressortir la structure des objets manipulés, des conditions d'optimalité obtenues, ainsi que celle des raisonnements employés.

La contrainte de (P_G) peut aussi s'exprimer comme l'inclusion $0 \in T(x)$ où $T : \mathbb{E} \multimap \mathbb{F}$ est la multifonction définie en $x \in \mathbb{E}$ par $T(x) := c(x) - G$. Ce formalisme sera souvent utilisé et permettra d'avoir un point de vue différent sur les problèmes considérés, permettant d'obtenir de nouveaux résultats.

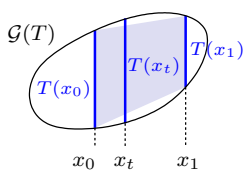
Définition 4.50 (multifonction convexe) On dit qu'une multifonction $T : \mathbb{E} \multimap \mathbb{F}$ est *convexe* si son graphe $\mathcal{G}(T) := \{(x, y) \in \mathbb{E} \times \mathbb{F} : y \in T(x)\}$ est convexe dans $\mathbb{E} \times \mathbb{F}$. Il revient au même de dire que pour tout $(x_0, x_1) \in \mathbb{E}^2$ et tout $t \in [0, 1]$, on a

$$T((1-t)x_0 + tx_1) \supseteq (1-t)T(x_0) + t\,T(x_1). \tag{4.67}$$

Cette propriété s'exprime plus généralement au moyen des parties de $\mathbb{E}$: T est convexe si, et seulement si, pour toutes parties P_1 et $P_2 \subseteq \mathbb{E}$ et tout $t \in [0, 1]$, on a

$$T((1-t)P_0 + tP_1) \supseteq (1-t)T(P_0) + t\,T(P_1). \tag{4.68}$$

La figure ci-contre montre le graphe d'une multifonction convexe de $\mathbb{R} \multimap \mathbb{R}$. La convexité de son graphe implique qu'il contient la zone ombragée représentant la réunion, lorsque t parcourt $[0, 1]$, des ensembles $\{x_t\} \times \{(1-t)\,T(x_0) + t\,T(x_1)\} = (1-t)[\{x_0\} \times T(x_0)] + t[\{x_1\} \times T(x_1)]$ (on y a noté $x_t := (1-t)x_0 + tx_1$). C'est ce qu'exprime la condition (4.67). □



Une multifonction convexe *univoque* (c'est-à-dire telle que, quel que soit $x \in \mathbb{E}$, $T(x)$ est un singleton) est *affine*, parce que l'on a alors nécessairement l'égalité dans (4.67). Donc si $f : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ est une *fonction* convexe, $x \mapsto \{f(x)\}$ n'est en général pas une *multifonction* convexe, mais la multifonction $x \mapsto [f(x), +\infty[\subseteq \mathbb{R}$ est convexe (son graphe est l'épigraphe de f, qui est convexe).

Par ailleurs, en notant que, pour $C \subseteq \mathbb{E}$, $T(C) := \cup_{x \in C} T(x)$ est la projection canonique de $\mathcal{G}(T) \cap (C \times \mathbb{F})$ sur $\mathbb{F}$, on voit que

$$\left.\begin{array}{r} T : \mathbb{E} \multimap \mathbb{F} \text{ est convexe} \\ C \text{ est convexe dans } \mathbb{E} \end{array}\right\} \quad \Longrightarrow \quad T(C) \text{ est convexe dans } \mathbb{F}. \tag{4.69}$$

L'adaptation au problème (P_G) de la définition 4.8 de problème (P_X) convexe conduit à la définition suivante.

Définition 4.51 ((P_G) convexe) On dit que le problème (P_G) est *convexe* si son ensemble admissible X_G est convexe et si son critère f est convexe sur X_G. □

Proposition 4.52 (convexité de X_G) *Si la multifonction $x \mapsto c(x) - G$ est convexe, alors X_G est convexe.*

Démonstration. Soit $T : \mathbb{E} \multimap \mathbb{F} : x \mapsto c(x) - G$. Alors

$$x \in X_G \iff c(x) \in G \iff 0 \in c(x) - G = T(x) \iff (x, 0) \in \mathcal{G}(T),$$

si bien que X_G est la projection canonique de $\mathcal{G}(T) \cap (\mathbb{E} \times \{0\})$ sur $\mathbb{E}$. Comme $\mathcal{G}(T)$ est convexe, par hypothèse, il en est de même de X_G. □

Dans le cas du problème (P_{EI}), $G = \{0_{\mathbb{R}^{m_E}}\} \times \mathbb{R}^{m_I}_-$ et la convexité de $x \mapsto c(x) - G$ revient à dire que c_E est affine et c_I a toutes ses composantes convexes. La proposition 4.52 est donc une généralisation de la proposition 4.27.

4.5.1 Conditions d'optimalité du premier ordre

Des conditions d'optimalité du premier ordre du problème (P_G) peuvent s'obtenir en utilisant la même approche que pour les problèmes précédents, par l'intermédiaire de la condition du premier ordre générale de Peano-Kantorovitch du théorème 4.6. Cette condition requiert le calcul du cône tangent $\mathrm{T}_x X_G$ à l'ensemble admissible X_G. Comme précédemment, on cherche à exprimer ce cône tangent en « linéarisant » les objets c et G qui définissent l'ensemble admissible. Cette linéarisation se fait en x pour c, qui est définie sur $\mathbb{E}$, et en $c(x)$ pour G, qui est défini sur $\mathbb{F}$. Cette opération conduit au cône $\mathrm{T}'_x X_G$, plus grand que $\mathrm{T}_x X_G$, que l'on appelle le *cône linéarisant* de X_G.

Proposition 4.53 (cônes tangent et linéarisant) *Si c est dérivable en $x \in X_G$, alors*

$$\mathrm{T}_x\, X_G \subseteq \mathrm{T}'_x\, X_G := \{d \in \mathbb{E} : c'(x)d \in \mathrm{T}_{c(x)}G\}. \tag{4.70}$$

Démonstration. Soit $d \in \mathrm{T}_x\, X_G$. Il existe alors une suite $\{x_k\} \subseteq X_G$ convergeant vers x et une suite $\{t_k\} \downarrow 0$ telles que $d_k := (x_k - x)/t_k \to d$. D'autre part, c étant dérivable en x : $c(x_k) = c(x) + c'(x)(x_k - x) + o(\|x_k - x\|)$. On utilise alors le fait que $c(x) \in G$, $c(x_k) \in G$ et $(c(x_k) - c(x))/t_k \to c'(x)d$ pour en déduire que $c'(x)d \in \mathrm{T}_{c(x)}\, G$. □

On n'a pas nécessairement l'égalité en (4.70), car $\mathrm{T}'_x\, X_G$ est convexe (c'est l'image réciproque par l'application linéaire $c'(x)$ du cône tangent $\mathrm{T}_{c(x)}G$, qui est convexe par la convexité de G) alors que $\mathrm{T}_x\, X_G$ ne l'est pas nécessairement (on n'a pas imposé à la fonction c définissant X_G d'être affine et donc l'ensemble admissible X_G n'est pas nécessairement convexe). C'est gênant, car c'est le cône tangent $\mathrm{T}_x\, X_G$ qui intervient dans la condition nécessaire d'optimalité générale de Peano-Kantorovitch (4.12) alors que le cône linéarisant $\mathrm{T}'_x\, X_G$ a l'avantage d'avoir une expression analytique que l'on aimerait pouvoir exploiter. Comme pour le problème (P_{EI}), la notion de qualification des contraintes définissant X_G est liée au fait d'avoir l'égalité en (4.70), *mais pas seulement*. Rappelons-nous la technique de démonstration de la proposition 4.32 conduisant aux conditions d'optimalité du premier ordre de Karush, Kuhn et Tucker, technique qui sera aussi utilisée pour démontrer la proposition 4.56 ci-dessous. Le but est de montrer que le gradient $\nabla f(x_*)$ appartient à un cône que l'on peut expliciter. Deux ingrédients interviennent dans cette approche :

- l'égalité entre le cône tangent et le cône linéarisant, qui permet ainsi d'avoir une expression exploitable du premier,
- la polyédricité du cône linéarisant, qui permet d'éliminer la prise de l'adhérence après application du lemme de Farkas.

Ici $\mathrm{T}'_x\, X_G$ n'est pas polyédrique, parce que l'on ne veut pas imposer cette propriété restrictive de polyédricité à G. De manière à sélectionner les problèmes non convexes pour lesquels on peut utiliser l'approche proposée pour établir les conditions d'optimalité du premier ordre, on introduit une hypothèse, dite *de qualification*, qui assure précisément l'égalité entre le cône tangent et le cône linéarisant (c'est (4.71b) ci-dessous, rien de neuf par rapport à la définition 4.30), mais aussi le caractère fermé de l'image par $c'(x)^*$ du dual du cône linéarisant (c'est la condition (4.71b) ci-dessous).

Définition 4.54 (qualification de c pour représenter X_G) On dit que la contrainte c est *qualifiée* pour représenter X_G en $x \in X_G$ si c est dérivable en x et si les deux conditions suivantes sont vérifiées :

$$\mathrm{T}_x\, X_G = \mathrm{T}'_x\, X_G, \tag{4.71a}$$

$$c'(x)^*[(\mathrm{T}_{c(x)}\, G)^+] \text{ est fermé}, \tag{4.71b}$$

où $c'(x)^* : \mathbb{F} \to \mathbb{E}$ désigne l'opérateur linéaire adjoint de $c'(x) : \mathbb{E} \to \mathbb{F}$. □

La vérification des conditions de qualification (4.71) est une tâche complexe. De manière à la simplifier, on dispose d'une condition suffisante de qualification que l'on présentera et étudiera à la section 4.5.3.

La qualification des contraintes permet, comme pour les problèmes (P_E) et (P_{EI}), d'obtenir des conditions d'optimalité du premier ordre, celles du théorème 4.56 ci-dessous. Lorsque $G = K$ est un cône, on utilisera la notation de complémentarité

$$K^- \ni u \perp v \in K$$

pour exprimer de manière compacte que des vecteurs u et v de $\mathbb{F}$ vérifient les trois propriétés $u \in K^-$, $v \in K$ et $\langle u, v\rangle = 0$. Nous utiliserons l'équivalence suivante.

Lemme 4.55 *Si K est un cône convexe d'un espace euclidien $\mathbb{E}$, on a*

$$u \in \mathrm{N}_v K \quad \Longleftrightarrow \quad K^- \ni u \perp v \in K. \tag{4.72}$$

Démonstration. En effet, on peut supposer que $v \in K$, puisque cette propriété est présente des deux côtés de l'équivalence (elle l'est à gauche car $\mathrm{N}_v K \neq \varnothing$). Alors

$$\begin{aligned} u \in \mathrm{N}_v K &\iff \langle u, v' - v\rangle \leqslant 0 \text{ pour tout } v' \in K \\ &\iff \langle u, v\rangle = 0 \text{ et } \langle u, v'\rangle \leqslant 0 \text{ pour tout } v' \in K \\ &\iff \langle u, v\rangle = 0 \text{ et } u \in K^-, \end{aligned}$$

où, pour obtenir la deuxième équivalence, on a choisi $v' = v/2$ et $v' = 2v$ (c'est ici qu'intervient la conicité de K). □

Théorème 4.56 (CN1) *Soit x_* un minimum local de (P_G). Supposons que f et c soient dérivables en x_* et que c soit qualifiée pour représenter X_G en x_*, au sens de la définition 4.54. Alors, il existe $\lambda_* \in \mathbb{F}$ tel que l'on ait*

$$\nabla f(x_*) + c'(x_*)^* \lambda_* = 0 \tag{4.73a}$$

$$\lambda_* \in \mathrm{N}_{c(x_*)} G. \tag{4.73b}$$

Si $G = K$ est un cône convexe, (4.73b) se récrit

$$K^- \ni \lambda_* \perp c(x_*) \in K. \tag{4.73c}$$

Démonstration. Structurellement, la démonstration est semblable à celle du théorème 4.32, dans un cadre un peu plus abstrait toutefois. On a successivement

$$\begin{aligned}
\nabla f(x_*) &\in (\mathrm{T}_{x_*} X_G)^+ && \text{[Peano-Kantorovitch (4.12)]}\\
&= (\mathrm{T}'_{x_*} X_G)^+ && \text{[qualification des contraintes (4.71a) en } x_*\text{]}\\
&= \{d \in \mathbb{E} : c'(x_*)d \in \mathrm{T}_{c(x_*)} G\}^+ && \text{[formule (4.70) de } \mathrm{T}'_{x_*} X_G\text{]}\\
&= \mathrm{adh}\big[c'(x_*)^*((\mathrm{T}_{c(x_*)} G)^+)\big]\\
&\qquad \text{[proposition 2.45 et } \mathrm{T}_{c(x_*)} G \text{ est un cône convexe fermé]}\\
&= c'(x_*)^*((\mathrm{T}_{c(x_*)} G)^+)\\
&\qquad \text{[qualification des contraintes (4.71b) en } x_*\text{]}.
\end{aligned}$$

On a donc montré l'existence d'un $\lambda_* \in -(\mathrm{T}_{c(x_*)} G)^+ = \mathrm{N}_{c(x_*)}G$ (c'est (4.73b)) tel que l'on ait (4.73a). La relation de complémentarité (4.73c) résulte de (4.73b) et de l'équivalence (4.72), valable lorsque $G = K$ est un cône convexe. □

Lorsque $G = K$ est un cône convexe, les conditions d'optimalité du premier ordre (4.73a) et (4.73c) du problème d'optimisation (P_G) ressemblent structurellement beaucoup aux conditions d'optimalité de Karush, Kuhn et Tucker (4.33) du problème (P_{EI}), avec quelques différences toutefois. Les commentaires ci-dessous sont destinés à rendre ces conditions plus naturelles pour le lecteur déjà familiarisé avec les conditions d'optimalité vues précédemment.

1. D'abord, on reconnaît dans (4.73a) le gradient par rapport à x du lagrangien $\ell : \mathbb{E} \times \mathbb{F} \to \mathbb{R}$ défini en $(x, \lambda) \in \mathbb{E} \times \mathbb{F}$ par

$$\ell(x, \lambda) = f(x) + \langle \lambda, c(x) \rangle.$$

   Ce lagrangien ne fait intervenir que la partie fonctionnelle de la contrainte $c(x) \in K$. L'ensemble K est pris en compte par la seconde condition, à savoir (4.73c).

2. Dans le cas du problème (P_{EI}), K est le cône polyédrique $\{0_{\mathbb{R}^{m_E}}\} \times \mathbb{R}^{m_I}_-$, si bien que les conditions d'optimalité du premier ordre s'expriment par (4.73a) et (4.73c). Comme $K^- = \mathbb{R}^{m_E} \times \mathbb{R}^{m_I}_+$ la complémentarité (4.73c) n'intervient que sur les contraintes d'inégalité et requiert que $0 \leqslant (\lambda_*)_I \perp c_I(x_*) \leqslant 0$. On retrouve bien les conditions de KKT (4.33).

Le pendant de la condition suffisante du premier ordre (CS1) de la proposition 4.33 est donné ci-dessous. Comme précédemment, cette CS1 ne requiert pas de condition de qualification de contrainte.

Proposition 4.57 (CS1 pour problème convexe) *Si le problème (P_G) est convexe au sens de la définition 4.51, si f et c sont différentiables en $x_* \in X_G$ et s'il existe un multiplicateur $\lambda_* \in \mathbb{F}$ tel que (x_*, λ_*) vérifie les conditions du premier ordre (4.73a)-(4.73b), alors x_* est un minimum global de (P_G).*

Démonstration. Comme f est convexe, comme l'ensemble admissible X_G est convexe et comme $x_* \in X_G$, il suffit, d'après la proposition 4.7, de montrer que

$$\forall\, x \in X_G : \quad \langle \nabla f(x_*), x - x_* \rangle \geqslant 0.$$

En utilisant (4.73a), cela revient à montrer que

$$\forall\, x \in X_G : \quad \langle \lambda_*, c'(x_*)(x - x_*)\rangle \leqslant 0.$$

Par (4.73b), $\lambda_* \in \mathrm{N}_{c(x_*)}G = (\mathrm{T}_{c(x_*)}\, G)^-$, si bien qu'il suffit de montrer que

$$c'(x_*)(x - x_*) \in \mathrm{T}_{c(x_*)}\, G.$$

Par définition de la dérivée

$$c'(x_*)(x - x_*) = \lim_{t\downarrow 0} \frac{1}{t}\big[c(x_* + t(x - x_*)) - c(x_*)\big].$$

Pour $t \in [0,1]$, $c(x_* + t(x - x_*)) = c((1-t)x_* + tx) \in G$ (parce que x_* et x sont dans le convexe X_G), si bien que le membre de droite est dans l'adhérence de $\mathbb{R}_+(G - c(x_*))$ qui est $\mathrm{T}_{c(x_*)}\, G$. □

4.5.2 Condition de Robinson

La condition suffisante de qualification de contrainte au sens de la définition 4.54, que nous allons analyser et qui étend la condition de Mangasarian-Fromovitz (QC-MF) du problème (P_{EI}) au problème (P_G), est la *condition de Robinson* suivante [513], écrite ici en un point $x_0 \in \mathbb{E}$:

$$(\mathrm{QC\text{-}R})_{x_0} \qquad 0 \in \mathrm{int}\{c(x_0) + c'(x_0)\mathbb{E} - G\}. \tag{4.74}$$

Cette condition peut paraître hermétique au premier abord, mais cela vaut la peine de faire l'effort de s'y familiariser, car elle est utile à plus d'un titre. Sa clarification passera par le résultat important suivant, qui fait le lien entre cette condition et, comme nous le verrons, la notion de régularité métrique de la multifonction $x \in \mathbb{E} \mapsto c(x) - G \subseteq \mathbb{F}$, associée à la contrainte de (P_G).

Théorème 4.58 (qualification de Robinson et régularité métrique) *Si c est continûment différentiable dans un voisinage de $x_0 \in X_G$, alors les propriétés suivantes sont équivalentes :*

(i) la condition de qualification de Robinson $(QC\text{-}R)_{x_0}$ a lieu en x_0,

(ii) il existe une constante positive μ, telle que $\forall\,(x,y) \in \mathbb{E}\times\mathbb{F}$ proche de $(x_0, 0)$, on a

$$\mathrm{dist}(x, c^{-1}(y + G)) \leqslant \mu\ \mathrm{dist}(c(x), y + G). \tag{4.75}$$

Avant de donner une démonstration de ce théorème, qui sera longue, nous commençons par en déduire deux corollaires immédiats, qui permettront d'apporter un premier éclairage sur sa signification.

Le premier corollaire donne une borne d'erreur pour X_G. Une *borne d'erreur* est une estimations de la distance à un ensemble par une quantité plus aisément calculable. En général, X_G est un ensemble plus « complexe » que G, ne fût-ce que parce que G est convexe alors que ce n'est pas nécessairement le cas de $X_G := c^{-1}(G)$

qui peut être rendu très compliqué par l'intermédiaire de la fonction c. La borne d'erreur de Robinson (4.76) ci-dessous permet d'avoir une estimation de la distance $\operatorname{dist}(x, X_G)$, qui peut donc être difficile à évaluer, au moyen de la distance $\operatorname{dist}(c(x), G)$ de $c(x)$ (supposé aisément calculable) au convexe G.

Corollaire 4.59 (borne d'erreur de Robinson pour X_G) *Si c est continûment différentiable dans un voisinage de $x_0 \in X_G$ et si la condition de qualification de Robinson (QC-R)$_{x_0}$ a lieu en x_0, alors il existe une constante positive μ, telle que pour tout x voisin de x_0, on a*

$$\operatorname{dist}(x, X_G) \leqslant \mu \operatorname{dist}(c(x), G). \tag{4.76}$$

Démonstration. Il suffit de prendre $y = 0$ dans (4.75) et d'utiliser le fait que $X_G = c^{-1}(G)$. □

La borne d'erreur (4.76) jouera un rôle essentiel pour démontrer que (QC-R) est une condition suffisante de qualification de contraintes (proposition 4.69).

Les inégalités (4.75) du théorème 4.58 forment donc une famille de bornes d'erreur pour les ensembles $c^{-1}(y+G)$, avec y petit, qui sont de petites perturbations de X_G. Le théorème 4.58 affirme que cette famille de bornes d'erreur est une condition équivalente à (QC-R), alors que son corollaire 4.59 nous informe que l'unique borne d'erreur (4.76) est qu'une conséquence de (QC-R).

La distance à un ensemble vide est infinie, car une distance est définie par un infimum sur cet ensemble (voir (A.4) et (1.1)). Comme la distance à droite dans (4.75) est finie (parce que G est non vide), celle à gauche l'est aussi, ce qui signifie que $c^{-1}(y+G)$ est non vide pour de petites perturbations y de G. C'est ce résultat de *stabilité* (i.e., de non-disparition) pour de petites perturbations qu'affirme le second corollaire, qui ne demande pas d'autre démonstration.

Corollaire 4.60 (stabilité de X_G pour de petites perturbations) *Si c est continûment différentiable dans un voisinage de $x_0 \in X_G$ et si la condition de qualification de Robinson (QC-R)$_{x_0}$ a lieu en x_0, alors pour toute perturbation $y \in \mathbb{F}$ proche de 0, on a*

$$\{x \in \mathbb{E} : c(x) \in y + G\} \neq \varnothing. \tag{4.77}$$

On remarquera que (4.77) ne fait pas de référence au point x_0, qui pourra donc être un point arbitraire dans X_G satisfaisant (QC-R). On verra plus loin que la condition (QC-R) est équivalente à la condition de Mangasarian-Fromovitz (QC-MF) pour le problème (P_{EI}), si bien que l'exemple 4.38 montre que la réciproque de cette proposition est fausse (la stabilité n'implique pas (QC-R)) ; il en faut davantage pour avoir la condition (i) du théorème 4.58, à savoir la régularité métrique de son point (ii).

Le reste de cette section est entièrement consacré à la démonstration du théorème 4.58, dont la généralité suggère la puissance, mais présage aussi de la difficulté de

son analyse. Le lecteur pressé ou non intéressé par la démonstration pourra passer directement à la section 4.5.3 qui suit. Celle-ci n'utilisera de la section présente que le corollaire 4.59.

Voici le schéma de la démonstration du théorème 4.58, qui sera donnée à la page 223. L'analyse utilise la boîte à outils des multifonctions. On introduit les deux multifonctions

$$T : \mathbb{E} \multimap \mathbb{F} : x \mapsto T(x) := c(x) - G, \tag{4.78a}$$

$$T_0 : \mathbb{E} \multimap \mathbb{F} : x \mapsto T_0(x) := c(x_0) + c'(x_0) \cdot (x - x_0) - G. \tag{4.78b}$$

On peut voir T_0 comme une linéarisation de T en x_0. Comme l'image de T_0 est $c(x_0) + c'(x_0)\mathbb{E} - G$, la condition de Robinson $(\text{QC-R})_{x_0}$ s'écrit

$$0 \in \operatorname{int} \mathcal{R}(T_0),$$

ce qui justifie l'introduction de T_0. Pour une multifonction convexe, comme l'est T_0 (mais pas T !), cette dernière condition s'avère être une propriété de *régularité métrique* en $(x_0, 0) \in \mathcal{G}(T_0)$ (théorème 4.63), ce qui veut dire qu'il existe une constante positive μ_0 telle que pour tout (x, y) voisin de $(x_0, 0)$, on a

$$\operatorname{dist}(x, T_0^{-1}(y)) \leqslant \mu_0 \operatorname{dist}(y, T_0(x)).$$

Cette notion est clarifiée ci-dessous, après la définition 4.62. Lorsque x est proche de x_0, $T(x)$ est « proche » de $T_0(x)$ (et réciproquement, dans un sens qui sera précisé), par exemple si l'on suppose la continue différentiabilité de c dans le voisinage de x_0. On utilise alors le fait que la propriété de régularité métrique de T_0 a la faculté de se « diffuser » de T_0 à T (proposition 4.67) : il existe une constante positive μ telle que pour tout (x, y) voisin de $(x_0, 0)$, on a

$$\operatorname{dist}(x, T^{-1}(y)) \leqslant \mu \operatorname{dist}(y, T(x)).$$

Pour obtenir (4.75), il suffit maintenant d'observer que

$$T^{-1}(y) = c^{-1}(y + G) \qquad \text{et} \qquad \operatorname{dist}(y, T(x)) = \operatorname{dist}(c(x), y + G).$$

La première identité provient des équivalences $x' \in T^{-1}(y) \Leftrightarrow y \in T(x') = c(x') - G \Leftrightarrow c(x') \in y + G \Leftrightarrow x' \in c^{-1}(y + G)$; tandis que la seconde découle de $\operatorname{dist}(y, T(x)) = \operatorname{dist}(y, c(x) - G) = \operatorname{dist}(c(x), y + G)$. Ce schéma sera repris dans la démonstration du théorème 4.58 en fin de section (à la page 223). Établissons maintenant les résultats qui permettront de le mettre en œuvre.

Nous commençons par une proposition donnant des expressions équivalentes au fait qu'un point y_0 est intérieur à l'image d'une multifonction convexe. Comme expliqué dans le schéma de démonstration ci-dessus, ce résultat sera appliqué à la multifonction convexe T_0 définie par (4.78b), pas à T définie par (4.78a) qui n'est pas convexe. Nous aurons besoin des deux propriétés de multifonction importantes suivantes.

Définition 4.61 (multifonction ouverte en un point) On dit qu'une multifonction $T : \mathbb{E} \multimap \mathbb{F}$ est *ouverte* en $(x_0, y_0) \in \mathcal{G}(T)$ de ratio $\rho > 0$, s'il existe un voisinage W de (x_0, y_0) et un rayon maximal $r_{\max} > 0$, tels que pour tout $(x, y) \in W \cap \mathcal{G}(T)$ et tout $r \in [0, r_{\max}]$, on ait

$$y + \rho\, r\, \bar{B}_{\mathbb{F}} \subseteq T(x + r\, \bar{B}_{\mathbb{E}}). \tag{4.79}$$

Le ratio ρ donne donc le rapport entre le rayon de la boule $y + \rho\, r\, \bar{B}_{\mathbb{F}}$ de $\mathbb{F}$ s'inscrivant dans l'image par T de la boule $x + r\, \bar{B}_{\mathbb{E}}$ de $\mathbb{E}$. □

En prenant $r = 0$ dans l'inclusion (4.79), on voit que (x, y) doit nécessairement appartenir au graphe de T.

Le dessin de gauche de la figure 4.6 illustre la notion de *multifonction ouverte*. On

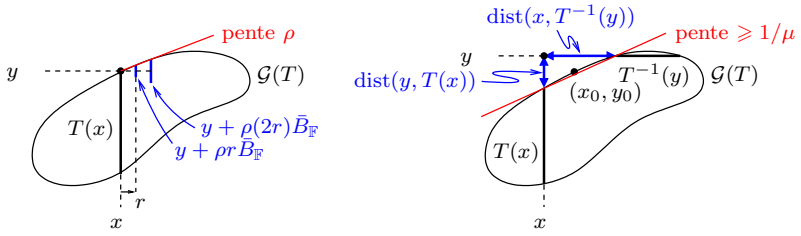


Fig. 4.6. Illustration des notions de multifonction ouverte de ratio ρ (à gauche) et de régularité métrique de module μ (à droite) : les deux notions décrivent la variation de $T(x)$ avec x, la première le fait de l'intérieur au graphe, la seconde le fait de l'extérieur ; ce ne sont pas des notions infinitésimales

y voit les images translatées (pour plus de lisibilité) de $T(x)$ et des boules centrées en y et de rayon maximal dans $T(x + r\bar{B}_{\mathbb{E}})$ et $T(x + 2r\bar{B}_{\mathbb{E}})$, pour un certain rayon $r > 0$. Le taux d'accroissement du rayon de ces boules avec r fournit le ratio ρ. Ce ratio n'est pas une notion infinitésimale ; il ne donne donc pas la variation infinitésimale de la « taille » de $T(x + r\bar{B}_{\mathbb{E}})$ avec r, en $r = 0$, mais en donne une approximation vue de l'intérieur du graphe de T (les couples (x, y) appartiennent à celui-ci).

Si l'« ouverture » de T décrit l'évolution de $T(x)$ avec x, en se plaçant à l'*intérieur* du graphe $\mathcal{G}(T)$, la notion de régularité métrique ci-dessous donne une description similaire en se plaçant à l'*extérieur* du graphe.

Définition 4.62 (multifonction métriquement régulière) On dit qu'une multifonction $T : \mathbb{E} \multimap \mathbb{F}$ est *métriquement régulière* en $(x_0, y_0) \in \mathcal{G}(T)$ de module positif μ, si pour tout (x, y) voisin de (x_0, y_0), on a

$$\mathrm{dist}(x, T^{-1}(y)) \leqslant \mu\ \mathrm{dist}(y, T(x)). \tag{4.80}$$

Cette inégalité suppose que $\mathrm{dist}(\cdot, \varnothing) = +\infty$. □

Le dessin de droite de la figure 4.6 illustre la notion de *régularité métrique* et son lien avec celle de multifonction ouverte. Pour que l'estimation (4.80) apporte de l'information, il faut que (x, y) soit en dehors du graphe (dans le cas contraire les deux distances de (4.80) sont nulles), comme dans la figure. On voit alors que le rapport $\mathrm{dist}(y, T(x))/\,\mathrm{dist}(x, T^{-1}(y)) \simeq 1/\mu$ a un lien avec le ratio de ρ de la figure de gauche.

Il est à noter que (4.79) *et* (4.80) *renferment des résultats d'existence*. Ainsi, lorsque (4.79) a lieu, si $y' \in y + \rho r \bar{B}_{\mathbb{F}}$, alors *il existe* un $x' \in x + r\bar{B}_{\mathbb{E}}$ tel que

$y' \in T(x')$: le problème d'inclusion « étant donné y', trouver x' tel que $y' \in T(x')$ » a une solution. De même, lorsque (4.80) a lieu, $\operatorname{dist}(x, T^{-1}(y))$ est fini (car $\operatorname{dist}(y, T(x))$ est toujours fini), ce qui veut dire que $T^{-1}(y) \neq \varnothing$ ou encore qu'*il existe* un x' tel que $y = T(x')$. Dans la démonstration du théorème ci-dessous, nous utiliserons ces propriétés d'ouverture et de régularité métrique de T de cette manière.

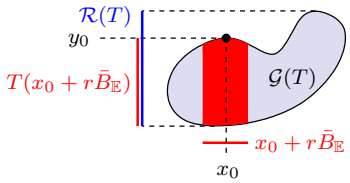
Théorème 4.63 (de la multifonction ouverte) *Soient $\bar{B}_{\mathbb{E}}$ et $\bar{B}_{\mathbb{F}}$ les boules-unités fermées de $\mathbb{E}$ et $\mathbb{F}$ respectivement, $T : \mathbb{E} \multimap \mathbb{F}$ est une multifonction convexe et $(x_0, y_0) \in \mathcal{G}(T)$. Alors les propriétés suivantes sont équivalentes :*

(i) $y_0 \in \operatorname{int} \mathcal{R}(T)$,
(ii) *pour tout rayon $r > 0$, on a $y_0 \in \operatorname{int} T(x_0 + r\bar{B}_{\mathbb{E}})$,*
(iii) *T est ouverte en (x_0, y_0) de ratio $\rho > 0$,*
(iv) *T est métriquement régulière en (x_0, y_0) de module μ.*

On peut prendre $\mu = 1/\rho$ au point (iv) si ρ est donné par le point (iii).

Avant de faire la démonstration de ce théorème, faisons quelques remarques sur ses hypothèses et sa signification.

- Ce théorème généralise aux multifonctions convexes (non linéaires), le *théorème de l'application ouverte* sur les applications linéaires continues, bien connu en analyse fonctionnelle [91] (voir l'exercice B.3 pour la dimension finie). Ce dernier affirme, entre autres, qu'une application linéaire (continue) est surjective si, et seulement si, l'image de la boule unité par cette application contient une boule centrée en zéro dans l'espace d'arrivée. On le retrouve dans l'équivalence $(i) \Leftrightarrow (ii)$ du théorème lorsque T est cette application linéaire (univoque) et $(x_0, y_0) = (0, 0)$ puisqu'alors (i) traduit la surjectivité de T et (ii) la non-vacuité de l'intérieur de $T(B_{\mathbb{E}})$.
- La convexité de T ne peut être omise sans perdre les implications $(i) \Rightarrow (ii)$ et $(i) \Rightarrow (iii)$. En effet, pour la multifonction non convexe T, dont le graphe est donné dans la figure ci-dessous, $y_0 \in \operatorname{int} \mathcal{R}(T)$, mais T ne vérifie certainement pas (ii) et



  n'est certainement pas ouverte en (x_0, y_0).
- Par son implication $(i) \Rightarrow (iv)$, le théorème permet de traduire (QC-R) en une condition de régularité métrique, ce que nous cherchions à faire.
- L'équivalence entre les points (iii) et (iv) explicite le lien entre multifonction ouverte et métriquement régulière de manière rigoureuse. Ainsi, une grande valeur de ρ témoigne d'une variation rapide de $T(x)$ avec x, alors qu'une grande valeur de μ exprime la variation lente de $T(x)$ avec x.

Démonstration de la proposition 4.63. [$(i) \Rightarrow (ii)$] Par la convexité de T et (4.69), $T(x_0 + r\bar{B}_{\mathbb{E}})$ est convexe. Il suffit donc de montrer que y_0 en est un point absorbant pour qu'il lui soit intérieur (corollaire 2.15). Soit $p \in \mathbb{F}$. Comme $y_0 \in \operatorname{int} \mathcal{R}(T)$, on peut trouver un $\alpha > 0$ tel que $y_0 + \alpha p \in \mathcal{R}(T)$. Donc il existe $d_\alpha \in \mathbb{E}$ tel que $y_0 + \alpha p \in T(x_0 + d_\alpha)$. Ce d_α n'est pas nécessairement dans $r\bar{B}_{\mathbb{E}}$. La convexité de T intervient à nouveau ici, pour mettre à l'échelle d_α. Pour $t \in [0,1]$, on a

$$\begin{aligned} y_0 + t\alpha p &= (1-t)y_0 + t(y_0 + \alpha p) \\ &\in (1-t)T(x_0) + tT(x_0 + d_\alpha) \qquad [y_0 \in T(x_0) \text{ et } y_0 + \alpha p \in T(x_0 + d_\alpha)] \\ &\subseteq T((1-t)x_0 + t(x_0 + d_\alpha)) \qquad [\text{convexité de } T] \\ &= T(x_0 + td_\alpha). \end{aligned}$$

Donc $y_0 + t\alpha p \in T(x_0 + r\bar{B}_{\mathbb{E}})$ pour $t > 0$ assez petit.

[$(ii) \Rightarrow (iii)$] Par le point (ii), il existe des rayons $\alpha > 0$ et $\beta > 0$ tels que

$$y_0 + \beta\, \bar{B}_{\mathbb{F}} \subseteq T(x_0 + \alpha\, \bar{B}_{\mathbb{E}}). \tag{4.81}$$

La question est maintenant de savoir si l'on garde une inclusion de ce type en des couples (x, y) différents de (x_0, y_0) mais situés dans $W \cap \mathcal{G}(T)$ et pour des rayons variables. Montrons que l'on peut prendre

$$\begin{cases} W = (x_0 + \alpha\bar{B}_{\mathbb{E}}) \times (y_0 + \frac{\beta}{2}\bar{B}_{\mathbb{F}}) \\ \rho = \frac{\beta}{4\alpha} \\ r_{\max} = 2\alpha. \end{cases}$$

En effet, si $(x, y) \in W \cap \mathcal{G}(T)$ et $t \in [0,1]$ (un paramètre proportionnel à r), on a

$$\begin{aligned} y + t\,\tfrac{\beta}{2}\,\bar{B}_{\mathbb{F}} &= (1-t)y + t(y + \tfrac{\beta}{2}\,\bar{B}_{\mathbb{F}}) \\ &\subseteq (1-t)y + t(y_0 + \beta\,\bar{B}_{\mathbb{F}}) \qquad [y \subseteq y_0 + \tfrac{\beta}{2}\,\bar{B}_{\mathbb{F}}] \\ &\subseteq (1-t)T(x) + tT(x_0 + \alpha\,\bar{B}_{\mathbb{E}}) \qquad [(x,y) \in \mathcal{G}(T) \text{ et } (4.81)] \\ &\subseteq T((1-t)x + t(x_0 + \alpha\,\bar{B}_{\mathbb{E}})) \qquad [T \text{ convexe et } (4.68)] \\ &\subseteq T(x + 2t\alpha\,\bar{B}_{\mathbb{E}}) \qquad [x_0 \in x + \alpha\,\bar{B}_{\mathbb{E}}]. \end{aligned}$$

En faisant le changement de variable $t \curvearrowright r = 2t\alpha$, qui est bien dans $[0, r_{\max}]$, on obtient l'inclusion souhaitée puisqu'alors $t\beta/2 = r\beta/)4\alpha) = \rho r$.

[$(iii) \Rightarrow (iv)$] On peut supposer que le voisinage W de (x_0, y_0) et le rayon maximal $r_{\max}$ de la définition 4.61 vérifient les conditions ci-dessous, avec $\varepsilon_{\mathbb{E}} > 0$ et $\varepsilon_{\mathbb{F}} > 0$:

$$W = (x_0 + \varepsilon_{\mathbb{E}}\bar{B}_{\mathbb{E}}) \times (y_0 + \varepsilon_{\mathbb{F}}\bar{B}_{\mathbb{F}}) \qquad \text{et} \qquad r_{\max} \leqslant \frac{\varepsilon_{\mathbb{F}}}{2\rho}. \tag{4.82}$$

Montrons que la régularité métrique a lieu en (x_0, y_0) avec le module $\mu = 1/\rho$ sur le voisinage W'de (x_0, y_0) vérifiant les conditions ci-dessous :

$$W' = (x_0 + \varepsilon'_{\mathbb{E}}\bar{B}_{\mathbb{E}}) \times (y_0 + \varepsilon'_{\mathbb{F}}\bar{B}_{\mathbb{F}}), \qquad \text{avec } \ \varepsilon'_{\mathbb{E}} \leqslant \varepsilon_{\mathbb{E}} \ \text{ et } \ \rho\varepsilon'_{\mathbb{E}} + \varepsilon'_{\mathbb{F}} \leqslant \rho\, r_{\max}. \tag{4.83}$$

Clairement $W' \subseteq W$, car $\varepsilon'_{\mathbb{E}} \leqslant \varepsilon_{\mathbb{E}}$ et $\varepsilon'_{\mathbb{F}} \leqslant \rho\, r_{\max} \leqslant \varepsilon_{\mathbb{F}}/2 \leqslant \varepsilon_{\mathbb{F}}$. Soit donc $(x, y) \in W'$ et montrons (et cela suffira pour démontrer l'implication) que l'inégalité (4.80) a lieu

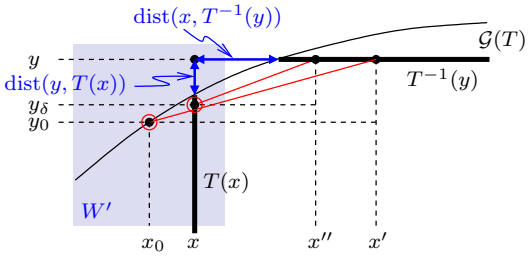


Fig. 4.7. Illustration de la démonstration de l'implication $(iii) \Rightarrow (iv)$ du théorème 4.63

avec $\mu = 1/\rho$ pour ce couple (x, y). La suite de la démonstration est illustrée à la figure 4.7.

Appliquons d'abord (4.79) en $(x_0, y_0) \in \mathcal{G}(T) \cap W$ (premier cercle rouge dans la figure) :

$$\forall\, r \in [0, r_{\max}] : \quad y_0 + \rho\, r\, \bar{B}_{\mathbb{F}} \subseteq T(x_0 + r\, \bar{B}_{\mathbb{E}}).$$

Comme $y \in y_0 + \rho(\|y - y_0\|/\rho)\bar{B}_{\mathbb{F}}$ et que $\|y - y_0\|/\rho \leqslant \varepsilon'_{\mathbb{F}}/\rho \leqslant r_{\max}$, l'inclusion précédente avec $r := \|y - y_0\|/\rho$ affirme l'existence de $x' \in x_0 + (\|y - y_0\|/\rho)\bar{B}_{\mathbb{E}}$ tel que $y = T(x')$. Alors,

$$\begin{aligned}
\operatorname{dist}(x, T^{-1}(y)) &\leqslant \|x - x'\| \qquad [x' \in T^{-1}(y)] \\
&\leqslant \|x - x_0\| + \|x_0 - x'\| \\
&\leqslant \|x - x_0\| + \tfrac{1}{\rho}\|y - y_0\| \qquad [x' \in x_0 + (\|y - y_0\|/\rho)\bar{B}_{\mathbb{E}}] \\
&\leqslant \varepsilon'_{\mathbb{E}} + \tfrac{1}{\rho}\varepsilon'_{\mathbb{F}} \qquad [\text{choix de } x \text{ et } y]. \qquad (4.84)
\end{aligned}$$

On considère deux cas disjoints.

- Si $\varepsilon'_{\mathbb{E}} + \varepsilon'_{\mathbb{F}}/\rho \leqslant \operatorname{dist}(y, T(x))/\rho$, alors (4.84) conduit à $\operatorname{dist}(x, T^{-1}(y)) \leqslant \operatorname{dist}(x, T(x))\rho$, qui est l'inégalité souhaitée.
- Si $\varepsilon'_{\mathbb{E}} + \varepsilon'_{\mathbb{F}}/\rho > \operatorname{dist}(y, T(x))/\rho$, alors, quel que soit $\delta > 0$ assez petit, on peut trouver $y_\delta \in T(x)$ tel que

$$\|y - y_\delta\| \leqslant \operatorname{dist}(y, T(x)) + \delta < \rho\varepsilon'_{\mathbb{E}} + \varepsilon'_{\mathbb{F}} \leqslant \rho r_{\max} \leqslant \frac{1}{2}\varepsilon_{\mathbb{F}}, \qquad (4.85)$$

où les deux dernières inégalités proviennent de (4.83) et (4.82). On cherche à appliquer (4.79) en (x, y_δ) (second cercle rouge dans la figure), qui s'écrit

$$\forall\, r \in [0, r_{\max}] : \quad y_\delta + \rho\, r\, \bar{B}_{\mathbb{F}} \subseteq T(x + r\, \bar{B}_{\mathbb{E}}).$$

  - Montrons que c'est licite. D'abord $(x, y_\delta) \in \mathcal{G}(T)$ parce que $y_\delta \in T(x)$. Ensuite $(x, y_\delta) \in W$ parce que $\|x - x_0\| \leqslant \varepsilon'_{\mathbb{E}} \leqslant \varepsilon_{\mathbb{E}}$ par (4.83) et

$$\|y_\delta - y_0\| \leqslant \|y_\delta - y\| + \|y - y_0\| \leqslant \frac{1}{2}\varepsilon_{\mathbb{F}} + \frac{1}{2}\varepsilon_{\mathbb{F}} = \varepsilon_{\mathbb{F}},$$

    où la dernière inégalité provient de (4.85) ainsi que de $\|y - y_0\| \leqslant \varepsilon'_{\mathbb{F}} \leqslant \rho\, r_{\max} \leqslant \varepsilon_{\mathbb{F}}/2$ par choix de y, (4.83) et (4.82).

- Comme $y \in y_\delta + \rho(\|y-y_\delta\|/\rho)\bar{B}_{\mathbb{F}}$ et que $\|y-y_\delta\|/\rho \leqslant r_{\max}$ par (4.85), l'inclusion précédente avec $r := \|y-y_\delta\|/\rho$ affirme l'existence de $x'' \in x + (\|y-y_\delta\|/\rho)\bar{B}_{\mathbb{E}}$ tel que $y = T(x'')$. On peut maintenant conclure

$$\begin{array}{rcll} \operatorname{dist}(x, T^{-1}(y)) & \leqslant & \|x - x''\| & [x'' \in T^{-1}(y)] \\ & \leqslant & \frac{1}{\rho}\|y - y_\delta\| & [x'' \in x + (\|y-y_\delta\|/\rho)\bar{B}_{\mathbb{E}}] \\ & \leqslant & \frac{1}{\rho}\left(\operatorname{dist}(y, T(x)) + \delta\right) & [(4.85)]. \end{array}$$

  Comme $\delta > 0$ peut-être pris arbitrairement petit, on en déduit l'inégalité recherchée.

[$(iv) \Rightarrow (i)$] On démontre la contraposée. Si $y_0 \notin \operatorname{int}\mathcal{R}(T)$, on peut trouver un $y \notin \mathcal{R}(T)$ aussi proche que l'on veut de y_0. Pour un tel y, $T^{-1}(y)$ est vide, donc $\operatorname{dist}(x_0, T^{-1}(y)) = +\infty$, si bien que l'inégalité (4.80) ne peut avoir lieu pour tout (x, y) proche de (x_0, y_0). □

Le but de la propriété 4.67 ci-dessous est de montrer que la propriété de régularité métrique peut se « diffuser » d'une multifonction à l'autre, en tout cas pour certaines multifonctions, celles de la forme qui nous intéresse, à savoir de la forme des multifonctions T ou T_0 définies en (4.78). On veut dire par là que, sous certaines conditions exprimant la proximité de T_0 et T, si T_0 est métriquement régulière, T l'est aussi (la réciproque a aussi lieu, car ce n'est pas la forme de T_0 et T qui va jouer, mais leur proximité). C'est cette propriété de diffusion qui permet d'obtenir la régularité métrique souhaitée de T à partir de l'hypothèse de régularité métrique de T_0, cette dernière s'exprimant sous la forme de la condition de qualification de Robinson (QC-R), grâce à l'équivalence $(i) \Leftrightarrow (iv)$ du théorème 4.63. Le résultat ne requiert pas la convexité des multifonctions ni la convexité de l'ensemble G.

L'idée de la démonstration proposée (il y en a d'autres !) dérive du changement de perspective suivant [133]. La régularité métrique donne une majoration de $\operatorname{dist}(x, T^{-1}(y))$. Observons que

$$x \in T^{-1}(y) \quad \Longleftrightarrow \quad y \in T(x) \quad \Longleftrightarrow \quad \operatorname{dist}(y, T(x)) = 0,$$

où la dernière équivalence suppose que $T(x)$ est fermé (cette hypothèse est vérifiée si la multifonction T est donnée par (4.78a) avec G fermé). Si l'on introduit la fonction $\varphi_y : \mathbb{E} \to \overline{\mathbb{R}}_+$ définie en $x \in \mathbb{E}$ par

$$\varphi_y(x) := \mu \operatorname{dist}(y, T(x)), \tag{4.86}$$

avec $\mu > 0$, on voit maintenant que

$$T^{-1}(y) \text{ est l'ensemble des zéros de } \varphi_y. \tag{4.87}$$

La propriété de régularité métrique peut alors s'exprimer comme suit

$$\forall\,(x, y) \text{ proche de } (x_0, y_0) : \quad \varphi_y \text{ a un zéro dans } \bar{B}(x, \varphi_y(x)), \tag{4.88}$$

où $\bar{B}(x, r)$ désigne la boule *fermée* de centre x et de rayon r.

On est donc amené à déterminer des conditions assurant qu'une fonction à valeurs dans $\overline{\mathbb{R}}_+$ a un zéro qui n'est pas trop éloigné d'un point x donné. Cela se fait au moyen de deux concepts ad hoc et d'un lemme.

Définition 4.64 (fonction sous-continue) Une fonction $\varphi : \mathbb{E} \to \overline{\mathbb{R}}_+$ est *sous-continue* si pour toute suite $\{x_k\} \to x$ telle que $\varphi(x_k) \to 0$, on a $\varphi(x) = 0$. □

Une fonction continue est sous-continue.

Définition 4.65 (fonction *r*-pentue) Une fonction $\varphi : \mathbb{E} \to \overline{\mathbb{R}}_+$ est *r-pentue*, avec $r \in [0, 1[$ en $x \in \mathbb{E}$, si

$$\forall\, x' \in \bar{B}\left(x, \frac{\varphi(x)}{1-r}\right), \quad \exists\, x'' \in \bar{B}\left(x', \varphi(x')\right) : \quad \varphi(x'') \leqslant r\, \varphi(x').$$

□

On peut exprimer cette propriété comme suit : en tout point x' de la boule $\bar{B}(x, \varphi(x)/(1-r))$, on peut trouver un point x'' qui n'est pas trop éloigné de x' ($x'' \in \bar{B}(x', \varphi(x'))$) et qui fait décroître φ significativement ($\varphi(x'') \leqslant r\, \varphi(x')$).

Une fonction ayant ces deux propriétés a un zéro qui n'est pas trop éloigné du point où elle est r-pentue.

Lemme 4.66 (Lyusternik) *Si $\varphi : \mathbb{E} \to \overline{\mathbb{R}}_+$ est sous-continue et r-pentue en $x \in \operatorname{dom} \varphi$, avec $r \in [0, 1[$, alors φ a un zéro dans $\bar{B}(x, \varphi(x)/(1-r))$.*

Démonstration. On construit une suite $\{x_k\}_{k \geqslant 0} \subseteq \mathbb{E}$ avec les propriétés suivantes :

$$\operatorname{dist}(x_{k+1}, x_k) \leqslant \varphi(x_k) \qquad \text{et} \qquad 0 \leqslant \varphi(x_{k+1}) \leqslant r\varphi(x_k). \tag{4.89}$$

On prend $x_0 = x$. Supposons que $x_0, \ldots, x_k$ soient déterminés et montrons comment calculer x_{k+1}. On a, grâce aux propriétés ci-dessus :

$$\operatorname{dist}(x_k, x) \leqslant \sum_{i=0}^{k-1} \operatorname{dist}(x_{i+1}, x_i) \leqslant \sum_{i=0}^{k-1} \varphi(x_i) \leqslant \left(\sum_{i=0}^{k-1} r^i\right) \varphi(x) \leqslant \frac{\varphi(x)}{1-r}. \tag{4.90}$$

Comme φ est r-pentue en x, on peut alors trouver $x_{k+1} \in \mathbb{E}$ tel que l'on ait (4.89).

Le suite $\{x_k\}$ est de Cauchy, car $\operatorname{dist}(x_{k+1}, x_k) \leqslant r^k \varphi(x)$ par (4.89), si bien qu'elle converge : $x_k \to x \in \bar{B}(x, \varphi(x)/(1-r))$ par (4.90). Par ailleurs, en passant à la limite dans l'inégalité $\varphi(x_{k+1}) \leqslant r\varphi(x_k)$, on voit que $\varphi(x_k) \to 0$. La sous-continuité de φ implique alors que $\varphi(x) = 0$. □

Si c est continue et si la multifonction T est métriquement régulière, alors la fonction φ_y définie en (4.86) est sous-continue et 0-pentue:

- elle est même continue car $x \mapsto \operatorname{dist}(y, T(x)) = \operatorname{dist}(y, c(x) - G) = \operatorname{dist}(c(x) - y, G)$ est la composition des fonctions continues $x \mapsto c(x) - y$ et $\operatorname{dist}(\cdot, G)$,
- elle est 0-pentue par l'expression (4.88) de la régularité métrique.

Il est donc raisonnable de penser qu'une petite perturbation de T gardera ces deux propriétés. C'est l'idée sous-jacente de la démonstration de la proposition suivante.

Proposition 4.67 (diffusion de la régularité métrique) *Soient $c : \mathbb{E} \to \mathbb{F}$ une fonction continue, $\delta : \mathbb{E} \to \mathbb{F}$ une fonction lipschitzienne de module $L > 0$ dans le voisinage d'un point $x_0 \in \mathbb{E}$, $\tilde{c} := c + \delta$ et G un convexe fermé non vide de $\mathbb{F}$. Si la multifonction $T : x \mapsto c(x) - G$ est métriquement régulière en $(x_0, y_0) \in \mathcal{G}(T)$ de module $\mu < 1/L$, alors la multifonction $\tilde{T} : x \mapsto \tilde{c}(x) - G$ est également métriquement régulière en $(x_0, y_0+\delta(x_0)) \in \mathcal{G}(\tilde{T})$ de module $\mu/(1-L\mu)$, c'est-à-dire que pour tout (x, y) voisin de $(x_0, y_0 + \delta(x_0))$, on a*

$$\operatorname{dist}(x, \tilde{T}^{-1}(y)) \leqslant \frac{\mu}{1 - L\mu} \operatorname{dist}(y, \tilde{T}(x)). \tag{4.91}$$

Démonstration. On peut supposer que $\delta(x_0) = 0$ car la régularité métrique est invariante par translation du graphe $\mathcal{G}(T)$ dans $\mathbb{E} \times \mathbb{F}$. Les effets de la continuité se voient mieux si l'on fait apparaître la fonction c qui est dans T pour obtenir, quel que soit $(x, y) \in \mathbb{E} \times \mathbb{F}$:

$$T^{-1}(y) = c^{-1}(y + G), \qquad \tilde{T}^{-1}(y) = \tilde{c}^{-1}(y + G), \tag{4.92a}$$

$$\operatorname{dist}(y, T(x)) = \operatorname{dist}(c(x), y + G), \qquad \operatorname{dist}(y, \tilde{T}(x)) = \operatorname{dist}(\tilde{c}(x), y + G). \tag{4.92b}$$

L'identité (4.92a) provient des équivalences $x \in T^{-1}(y) \Leftrightarrow y \in T(x) \Leftrightarrow y \in c(x) - G$ $\Leftrightarrow c(x) \in y + G \Leftrightarrow x \in c^{-1}(y + G)$. L'identité (4.92b) s'obtient par $\operatorname{dist}(y, T(x)) = \operatorname{dist}(y, c(x) - G) = \operatorname{dist}(c(x), y + G)$.

1) On considère l'application

$$\tilde{\varphi}_y : x \in \mathbb{E} \mapsto \tilde{\varphi}_y(x) := \mu \operatorname{dist}(y, \tilde{T}(x)), \tag{4.93}$$

qui l'on pourra comparer à celle introduite en (4.86). Montrons alors qu'il suffit de démontrer l'affirmation suivante dans laquelle $r := L\mu < 1$:

$$\begin{array}{c} \exists\, U \text{ voisinage de } x_0,\ \exists\, V \text{ voisinage de } y_0,\ \forall\, (x, y) \in U \times V, \\ \text{l'application } \tilde{\varphi}_y \text{ est sous-continue sur } \mathbb{E} \text{ et } r\text{-pentue en } x. \end{array} \tag{4.94}$$

En effet, par le lemme de Lyusternik (lemme 4.66), $\tilde{\varphi}_y$ a alors un zéro dans la boule $\bar{B}(x, \tilde{\varphi}_y(x)/(1 - r))$. Or l'ensemble des zéros de $\tilde{\varphi}_y$ est $\tilde{T}^{-1}(y)$ (par le même raisonnement que celui ayant conduit à (4.87) ; c'est ici que l'on utilise la fermeture de G). Dès lors pour tout $(x, y) \in U \times V$, on obtient (4.91) :

$$\operatorname{dist}(x, \tilde{T}^{-1}(y)) \leqslant \frac{\tilde{\varphi}_y(x)}{1 - r} = \frac{\mu}{1 - L\mu} \operatorname{dist}(y, \tilde{T}(x)).$$

2) Il reste à démontrer (4.94).

Par la continuité de c et celle de la distance à un convexe, on voit que $\tilde{\varphi}_y$ est continue sur $\mathbb{E}$, donc sous-continue sur $\mathbb{E}$.

Déterminons maintenant les voisinages U de x_0 et V de y_0 tels que, pour tout $(x, y) \in U \times V$, $\tilde{\varphi}_y$ soit r-pentue en x.

- Par la μ-régularité métrique de T en (x_0, y_0) et (4.92), on peut trouver des voisinages U_1 de x_0 et V_1 de y_0 tels que pour tout $(x, y) \in U_1 \times V_1$ on a

$$\operatorname{dist}(x, c^{-1}(y + G)) \leqslant \mu \operatorname{dist}(c(x), y + G). \tag{4.95a}$$

- On peut aussi trouver des voisinages $U_2 \subseteq U_1$ de x_0 et $V_2 \subseteq V_1$ de y_0 tels que pour tout $(x, y) \in U_2 \times V_2$, on a

$$\bar{B}\Big(x,\ \mu \operatorname{dist}(\tilde{c}(x), y + G)\Big) \subseteq U_1 \qquad \text{et} \qquad y - \delta(x) \in V_1. \tag{4.95b}$$

  L'inclusion dans U_1 provient de la continuité de $(x, y) \mapsto \operatorname{dist}(\tilde{c}(x), y + G)$ et de sa nullité en $(x_0, y_0) \in \mathcal{G}(T)$; l'appartenance à V_1 provient de la continuité $(x, y) \mapsto y - \delta(x)$ et de sa valeur y_0 en (x_0, y_0).
- Finalement, on prend des voisinages $U \subseteq U_2$ de x_0 et $V \subseteq V_2$ de y_0 tels que pour tout $(x, y) \in U \times V$ on a

$$\bar{B}\left(x,\ \frac{\mu}{1 - r} \operatorname{dist}(\tilde{c}(x), y + G)\right) \subseteq U_2. \tag{4.95c}$$

  C'est également possible par continuité et nullité du rayon $\operatorname{dist}(\tilde{c}(x), y + G)$ en $(x, y) = (x_0, y_0)$ (comme ci-dessus).

Montrons que, quel que soit $(x, y) \in U \times V$, $\tilde{\varphi}_y$ est r-pentue en x. On fixe donc $(x, y) \in U \times V$. Soit $x' \in \bar{B}(x, \tilde{\varphi}_y(x)/(1 - r))$. Il faut maintenant trouver un $x'' \in \bar{B}(x', \tilde{\varphi}_y(x'))$ tel que $\tilde{\varphi}_y(x'') \leqslant r\tilde{\varphi}_y(x')$.

- Comme $x' \in \bar{B}(x, \tilde{\varphi}_y(x)/(1 - r))$, on voit par la définition (4.93) de $\tilde{\varphi}_y$ et (4.95c), que $x' \in U_2$.
- Comme $y \in V \subseteq V_2$, on peut appliquer (4.95b) en $(x, y) \curvearrowright (x', y)$, ce qui montre que

$$\bar{B}\Big(x',\ \mu \operatorname{dist}(\tilde{c}(x'), y + G)\Big) \subseteq U_1 \qquad \text{et} \qquad y' := y - \delta(x') \in V_1. \tag{4.96}$$

- On peut donc utiliser (4.95a) en $(x, y) \curvearrowright (x', y')$, ce qui montre que

$$\operatorname{dist}(x', c^{-1}(y' + G)) \leqslant \mu \operatorname{dist}(c(x'), y' + G).$$

  Comme $c^{-1}(y' + G)$ est fermé, ceci implique qu'il existe

$$x'' \in c^{-1}(y' + G), \tag{4.97}$$

  tel que

$$\begin{aligned}
\operatorname{dist}(x', x'') &\leqslant \mu \operatorname{dist}(c(x'), y' + G) \\
&= \mu \operatorname{dist}(c(x'), y - \delta(x') + G) \qquad [y' = y - \delta(x') \text{ by } (4.96)] \\
&= \mu \operatorname{dist}(\tilde{c}(x'), y + G) \qquad [\tilde{c} = c + \delta] \\
&= \mu \operatorname{dist}(y, \tilde{T}(x')) \qquad [(4.92b)] \\
&= \tilde{\varphi}_y(x') \qquad [(4.93)].
\end{aligned} \tag{4.98}$$

  On a donc trouvé un point x'' a bonne distance de x'. Il reste à montrer qu'il fait bien décroître $\tilde{\varphi}_y$.

- On utilise (4.97) qui se récrit

$$c(x'') \in y' + G \qquad \text{ou} \qquad c(x'') + \delta(x') \in y + G. \tag{4.99}$$

Alors

$$\begin{aligned}
\operatorname{dist}(y, \tilde{T}(x'')) &= \operatorname{dist}(\tilde{c}(x''), y + G) \qquad [(4.92b)] \\
&\leqslant \operatorname{dist}(c(x'') + \delta(x''), c(x'') + \delta(x')) \qquad [(4.99)] \\
&= \operatorname{dist}(\delta(x''), \delta(x')) \\
&\leqslant L \operatorname{dist}(x'', x') \\
&\leqslant L\tilde{\varphi}_y(x') \qquad [(4.98)].
\end{aligned}$$

Si l'on multiplie les membres extrêmes par μ, on obtient grâce à (4.93) :

$$\tilde{\varphi}_y(x'') \leqslant L\mu\tilde{\varphi}_y(x') = r\tilde{\varphi}_y(x'),$$

qui est l'inégalité recherchée. □

Démonstration de la proposition 4.58. Soit T_0 la multifonction définie en (4.78b). Alors

$$(\text{QC-R})_{x_0} \qquad \Longleftrightarrow \qquad 0 \in \operatorname{int} \mathcal{R}(T_0).$$

La multifonction T_0 est convexe, parce que $x \mapsto c(x_0) + c'(x_0) \cdot (x - x_0)$ est affine et G est convexe, et $(x_0, 0) \in \mathcal{G}(T_0)$, parce que $x_0 \in X_G$. Alors on peut utiliser le théorème 4.63, dont l'équivalence $(i) \Leftrightarrow (iv)$ conduit à

$$(\text{QC-R})_{x_0} \qquad \Longleftrightarrow \qquad T_0 \text{ est } \mu_0\text{-métriquement régulière en } (x_0, 0).$$

On applique maintenant la proposition 4.67, avec $T \curvearrowright T_0$, $\tilde{T} \curvearrowright T$ (la multifonction définie en (4.78a)) et $\delta : \mathbb{E} \to \mathbb{F}$ définie en $x \in \mathbb{E}$ par

$$\delta(x) = c(x) - c(x_0) - c'(x_0) \cdot (x - x_0).$$

Comme c est C^1 dans un voisinage de x_0, δ est lipschitzienne dans un voisinage de x_0 (corollaire C.13 du théorème des accroissements finis) :

$$\begin{aligned}
\|\delta(x'') - \delta(x')\| &= \|c(x'') - c(x') - c'(x_0) \cdot (x'' - x')\| \\
&\leqslant \left(\sup_{z \in [x', x'']} \|c'(z) - c'(x_0)\| \right) \|x'' - x'\| \\
&\leqslant L \|x'' - x'\|,
\end{aligned}$$

où $L > 0$ peut être pris aussi petit que l'on veut, pourvu que x' et x'' soient suffisamment proches de x_0 (c' est continue), en particulier $< 1/\mu_0$. Alors, la proposition 4.67 montre que la régularité métrique de T_0 en $(x_0, 0)$ implique celle de T en ce même point. On peut inverser le rôle de T_0 et T, car la convexité d'une des multifonctions n'intervient pas dans la proposition 4.67 et L peut être pris arbitrairement petit en réduisant le voisinage de x_0. Dès lors la régularité métrique de T en $(x_0, 0)$ implique celle de T_0 au même point. Nous avons donc montré que

$$(\text{QC-R})_{x_0} \qquad \Longleftrightarrow \qquad T \text{ est } \mu\text{-métriquement régulière en } (x_0, 0).$$

Il reste à interpréter la partie droite de cette équivalence. Celle-ci signifie que pour tout (x, y) voisin de $(x_0, 0)$, on a

$$\operatorname{dist}(x, T^{-1}(y)) \leqslant \mu \operatorname{dist}(y, T(x)).$$

Comme $\operatorname{dist}(x, T^{-1}(y)) = \operatorname{dist}(x, c^{-1}(y + G))$ et $\operatorname{dist}(y, T(x)) = \operatorname{dist}(c(x), y + G)$, l'équivalence de la proposition 4.58 est démontrée. □

4.5.3 Qualification de Robinson

Rappelons la *condition de qualification de Robinson* (4.100) que l'on écrit ici en $x \in \mathbb{E}$:

$$(\text{QC-R}) \qquad 0 \in \operatorname{int}\{c(x) + c'(x)\mathbb{E} - G\}. \tag{4.100}$$

Avant de montrer que (QC-R) est une condition suffisante de qualification des contraintes de (P_G) au sens de la définition 4.54, donnons quelques formulations équivalentes de cette condition. On rappelle que, pour un convexe C et un point $x \in C$, $\mathrm{T}^a_x C := \mathbb{R}_+(C - x)$ est appelé le cône des directions admissibles de C en x (son adhérence est le cône tangent $\mathrm{T}_x C$).

Proposition 4.68 (autres formulations de (QC-R)) *Si c est dérivable en $x \in X_G$, les propriétés suivantes sont équivalentes :*

$$0 \in \operatorname{int}\{c(x) + c'(x)\mathbb{E} - G\}, \tag{4.101a}$$
$$c'(x)\mathbb{E} - \mathrm{T}^a_{c(x)} G = \mathbb{F}, \tag{4.101b}$$
$$c'(x)\mathbb{E} - \mathrm{T}_{c(x)}\, G = \mathbb{F}, \tag{4.101c}$$
$$\operatorname{adh}(c'(x)\mathbb{E} - \mathrm{T}_{c(x)}\, G) = \mathbb{F}. \tag{4.101d}$$

Démonstration. [(4.101a) ⇒ (4.101b)] Il suffit de montrer l'inclusion "⊇". Soit $y \in \mathbb{F}$. Alors, par (4.101a), $y/t \in c'(x)\mathbb{E} - (G - c(x))$ pour $t > 0$ assez grand, si bien que $y \in c'(x)[t\mathbb{E}] - [t(G - c(x))] \subseteq c'(x)\mathbb{E} - \mathrm{T}^a_{c(x)} G$.

[(4.101b) ⇒ (4.101c) ⇒ (4.101d)] Clair, car les ensembles dans les membres de gauche s'agrandissent.

[(4.101d) ⇒ (4.101b)] On a

$$\begin{aligned}
c'(x)\mathbb{E} - \mathrm{T}_{c(x)}\, G &= c'(x)\mathbb{E} - \operatorname{adh} \mathrm{T}^a_{c(x)} G && [\text{définition de } \mathrm{T}_{c(x)}\, G] \\
&\subseteq \operatorname{adh}(c'(x)\mathbb{E} - \mathrm{T}^a_{c(x)} G) && [\overline{A} + \overline{B} \subseteq \overline{A + B}] \\
&\subseteq \operatorname{adh}(c'(x)\mathbb{E} - \mathrm{T}_{c(x)}\, G) && [\mathrm{T}^a_{c(x)} G \subseteq \mathrm{T}_{c(x)}\, G] \\
&= \mathbb{F} \qquad [(4.101\text{d})].
\end{aligned}$$

En prenant l'adhérence des membres extrêmes et en utilisant (4.101d), on trouve que $\operatorname{adh}(c'(x)\mathbb{E} - \mathrm{T}^a_{c(x)} G) = \mathbb{F}$. Comme $c'(x)\mathbb{E} - \mathrm{T}^a_{c(x)} G$ est convexe (somme de deux convexes), son intérieur relatif est aussi $\mathbb{F}$ (point 3 de la proposition 2.16), si bien que

$$\mathbb{F} = (c'(x)\mathbb{E} - \mathrm{T}^a_{c(x)}G)^\ominus \subseteq c'(x)\mathbb{E} - \mathrm{T}^a_{c(x)}G \subseteq \mathbb{F},$$

où la dernière inclusion a été démontrée ci-dessus. On en déduit (4.101b).

[(4.101b) ⇒ (4.101a)] D'après le corollaire 2.15, il suffit de montrer que 0 est un point absorbant de l'ensemble convexe $C := c(x) + c'(x)\mathbb{E} - G$ de $\mathbb{F}$. Soit $y \in \mathbb{F}$. La condition (4.101b) implique qu'il existe $d \in \mathbb{E}$, $y_1 \in G$ et $\alpha > 0$ tels que $y = c'(x)d - \alpha(y_1 - c(x))$. Pour $t = 1/\alpha$, on a $ty = c(x) + c'(x)(td) - y_1 \in C$; donc 0 est un point absorbant de C. □

Dans le cas du problème (P_{EI}), que l'on retrouve à partir de (P_G) lorsque (4.66) a lieu, on a $\mathrm{T}_{c(x)}\, G = \{0_{\mathbb{R}^{m_E}}\} \times \{v \in \mathbb{R}^{m_I} : v_i \leqslant 0 \text{ si } i \in I^0_x\}$ où $I^0_x := \{i \in I : c_i(x) = 0\}$, si bien que (QC-R) dans sa formulation (4.101c) peut s'exprimer par

$$\forall v \in \mathbb{R}^m, \quad \exists\, d \in \mathbb{E} : \quad c'_E(x)d = v_E \text{ et } c'_{I^0_x}(x)d \leqslant v_{I^0_x}.$$

C'est la formulation des conditions de qualification de Mangasarian-Fromovitz, que l'on trouve au point (ii) de la proposition 4.42. Autrement et brièvement dit, (QC-R) est une généralisation des conditions de Mangasarian-Fromovitz au problème général (P_G), comme nous l'affirmions au début de cette section.

Dans le cas du problème $\inf\{f(x) : x \in G\}$, que l'on retrouve à partir de (P_G), en prenant $\mathbb{F} = \mathbb{E}$ et $c = \mathrm{I}_{\mathbb{E}}$, la formulation (4.101c) de (QC-R) s'exprime par $\mathbb{E} - \mathbb{E} = \mathbb{E}$, qui est toujours vérifiée. C'est normal, (QC-R) ne s'intéresse pas à l'adéquation de G, mais à celle de c qui est ici l'identité, fonction qui ne pose donc pas de difficulté.

On peut aussi noter que (QC-R) est vérifiée si l'une des propriétés suivantes a lieu (exercice 4.22)

- $c'(x)$ est surjective,
- $\exists\, d \in \mathbb{E}$ tel que $c'(x)d \in \operatorname{int} \mathrm{T}_{c(x)}\, G$.

On les rencontre parfois, mais elles sont beaucoup plus restrictives que (QC-R). Par exemple, si $c'(x)$ n'est pas surjective et si $\mathrm{T}_{c(x)}\, G$ est un espace supplémentaire de $\mathcal{R}(c'(x))$, alors (QC-R) a lieu, mais ce n'est pas le cas pour les conditions ci-dessus.

Proposition 4.69 (qualification par (QC-R)) *Si c est continûment différentiable dans le voisinage d'un point $x \in X_G$ et si la condition de Robinson (QC-R) a lieu en ce point, alors la contrainte c est qualifiée pour représenter X_G en x, au sens de la définition 4.54.*

Démonstration. Il faut montrer que (QC-R) entraîne les deux propriétés de (4.71).

[(4.71a)] On utilise la borne d'erreur de Robinson (4.76). D'après la proposition 4.53, il suffit de montrer l'inclusion $\mathrm{T}'_x\, X_G \subseteq \mathrm{T}_x\, X_G$. Soit donc $d \in \mathrm{T}'_x\, X_G$. Pour montrer que $d \in \mathrm{T}_x\, X_G$, il faut construire une suite $\{x_k\} \subseteq X_G$ (a priori une tâche difficile ; on se rappelle de la construction de cette suite, pour le problème (P_E), au moyen du théorème des fonctions implicites) et une suite $\{t_k\} \downarrow 0$ telles que $(x_k - x)/t_k \to d$. La démarche suivie est illustrée à la figure 4.8 : grâce à la borne d'erreur de Robinson (4.76), on peut créer la suite $\{x_k\} \subseteq X_G := c^{-1}(G)$ avec les propriétés souhaitées, grâce à la bonne propriété de la suite $\{c(x + t_k d)\}$ de $\mathbb{F}$ qui est d'être asymptotiquement très proche de la suite $\{y_k\}$. Voici les détails.

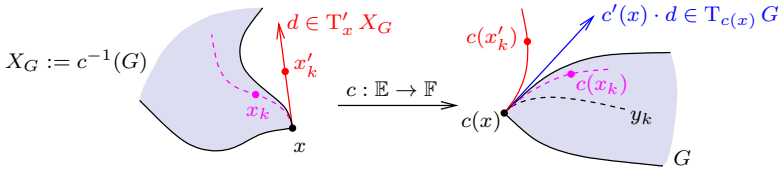


Fig. 4.8. Illustration de la démonstration de (4.71a) pour établir la proposition 4.69

Comme d est dans le cône linéarisant $\mathrm{T}'_x\, X_G$, il vérifie $c'(x) \cdot d \in \mathrm{T}_{c(x)}\, G$, ce qui veut dire qu'il existe une suite de $\{y_k\} \subseteq G$ et une suite $\{t_k\} \downarrow 0$ (c'est la bonne), telles que

$$\frac{y_k - c(x)}{t_k} \to c'(x) \cdot d \qquad \text{ou} \qquad c(x) + c'(x) \cdot (t_k d) - y_k = o(t_k). \tag{4.102}$$

On considère la suite $\{x'_k\} \subseteq \mathbb{E}$ définie par

$$x'_k := x + t_k d.$$

Cette suite n'est pas nécessairement dans X_G et donc la suite $\{c(x'_k)\}$ n'est pas nécessairement dans G, mais elle n'est pas loin de $\{y_k\}$ (qui est dans G) asymptotiquement car on a par la différentiabilité de c :

$$c(x'_k) = c(x) + c'(x) \cdot (t_k d) + o(t_k)$$

et donc

$$c(x'_k) = y_k + o(t_k). \tag{4.103}$$

On construit à présent une suite $\{x_k\} \subseteq X_G$ grâce à la borne d'erreur de Robinson (4.76) en $x_0 \curvearrowright x$. Celle-ci peut s'appliquer en x'_k puisque ce point est proche de x pour k assez grand (et c est C^1 dans un voisinage de x). Elle nous informe donc qu'il existe une constante $\mu \geqslant 0$ telle que, pour k assez grand, on a

$$\begin{aligned} \operatorname{dist}(x'_k, X_G) &\leqslant \mu \operatorname{dist}(c(x'_k), G) && [(4.76)] \\ &\leqslant \mu \operatorname{dist}(c(x'_k), y_k) && [y_k \in G] \\ &= o(t_k) && [(4.103)]. \end{aligned}$$

Ceci veut dire qu'il existe une suite $\{x_k\} \subseteq X_G$ telle que $x_k = x'_k + o(t_k)$. On a construit la suite $\{x_k\}$ souhaitée puisque

$$\frac{x_k - x}{t_k} = \frac{x_k - x'_k}{t_k} + \frac{x'_k - x}{t_k} = \frac{o(t_k)}{t_k} + d \to d.$$

[(4.71b)] On utilise l'expression (4.101c) de (QC-R). On considère une suite convergente du cône $\mathcal{K} := c'(x)^*[(\mathrm{T}_{c(x)}\, G)^+]$, dont on souhaite monter le caractère fermé :

$$c'(x)^* \lambda_k \to y, \tag{4.104}$$

où $\{\lambda_k\} \subseteq (\mathrm{T}_{c(x)}\, G)^+$. Il s'agit de montrer que y est dans $\mathcal{K}$. En dimension finie, il suffit de montrer que la suite $\{\lambda_k\}$ est bornée (on extrait alors une sous-suite convergente,

dont la limite est dans le cône fermé $(\mathrm{T}_{c(x)}\, G)^+$, et on passe à la limite dans (4.104), ce qui montre que $y \in \mathcal{K}$). On raisonne par l'absurde. Si cette suite n'est pas bornée, on peut extraire une sous-suite convergente de $\{\lambda_k/\|\lambda_k\|\}$, qui tend vers un vecteur *non nul*, disons λ qui appartient au cône fermé $(\mathrm{T}_{c(x)}\, G)^+$. On voit sur (4.104) que cette limite vérifie

$$c'(x)^*\lambda = 0 \qquad \text{et} \qquad \lambda \in (\mathrm{T}_{c(x)}\, G)^+.$$

On en déduit que

$$\forall\, (d,p) \in \mathbb{E} \times \mathrm{T}_{c(x)}\, G : \quad \langle \lambda, c'(x)d - p\rangle = \langle c'(x)^*\lambda, d\rangle - \langle \lambda, p\rangle \leqslant 0.$$

On déduit de l'expression (4.101c) de (QC-R), que $\langle \lambda, \mu\rangle \leqslant 0$ pour tout $\mu \in \mathbb{F}$, donc $\lambda = 0$, ce qui contredit sa non-nullité affirmée plus haut. □

Proposition 4.70 (bornitude des multiplicateurs) *Supposons que f et c soient différentiables en $x_* \in X_G$ et qu'il existe un multiplicateur optimal λ_* tel que les conditions d'optimalité (4.73a)-(4.73b) du problème (P_G) aient lieu. Alors le cône asymptotique de l'ensemble multiplicateurs optimaux Λ_* associés à x_* s'écrit*

$$\Lambda_*^\infty = (\mathrm{N}_{c(x_*)}G) \cap \mathcal{N}(c'(x_*)^*) = \left[c'(x_*)\mathbb{E} - T_{c(x_*)}G\right]^+. \tag{4.105}$$

On en déduit que Λ_* est borné si, et seulement si, la condition de qualification de Robinson (QC-R) a lieu en $x = x_*$.

Démonstration. Au vu des conditions d'optimalité (4.73), l'ensemble des multiplicateurs optimaux s'écrit

$$\Lambda_* = \{\lambda_* \in \mathrm{N}_{c(x_*)}G : \nabla f(x_*) + c'(x_*)^*\lambda_* = 0\}.$$

D'après les points 1 et 3 de la proposition 2.9, son cône asymptotique est donné par la première identité de (4.105). Ensuite, en appliquant le point 6 de la proposition 2.43, $\mathrm{N}_{c(x_*)}G = (-\,\mathrm{T}_{c(x_*)}\, G)^+$ et $\mathcal{N}(c'(x_*)^*) = \mathcal{R}(c'(x_*))^\perp = (c'(x_*)\mathbb{E})^+$, on obtient la seconde identité de (4.105).

Ensuite, d'après le corollaire 2.8, Λ_* est borné si, et seulement si, son cône asymptotique est réduit à $\{0\}$. Donc, il reste à montrer que

$$\left[c'(x_*)\mathbb{E} - T_{c(x_*)}G\right]^+ = \{0\} \qquad \Longleftrightarrow \qquad \text{(QC-R)}.$$

L'implication "$\Leftarrow$" est évidente, si l'on considère la forme (4.101c) de (QC-R), c'est-à-dire $c'(x)\mathbb{E} - \mathrm{T}_{c(x)}\, G = \mathbb{F}$. Pour l'implication réciproque "$\Rightarrow$", on introduit le cône convexe fermé $\mathcal{K} := \mathrm{adh}(c'(x_*)\mathbb{E} - T_{c(x_*)}G)$. Par hypothèse et par le point 3 de la proposition 2.43, $\mathcal{K}^+ = \{0\}$. Il résulte alors de la décomposition de Moreau (exercice 2.27) que $\mathcal{K} = \mathbb{F}$ et donc que la formulation (4.101d) de (QC-R) a lieu. □

4.6 Analyse de sensibilité

Dans cette section, nous donnons un certain nombre de résultats concernant la sensibilité des grandeurs optimales du problème (P_{EI}) par rapport à des perturbations. On note X l'ensemble admissible de ce problème. On suppose une perturbation linéaire des contraintes : on se donne un vecteur $p = (p_i)_{i\in E\cup I} \in \mathbb{R}^m$ et on considère le problème perturbé

$$(P^p_{EI}) \quad \begin{cases} \min\ f(x) \\ c_E(x) + p_E = 0 \\ c_I(x) + p_I \leqslant 0. \end{cases}$$

On note X^p l'ensemble admissible de (P^p_{EI}) et X^p_* son ensemble de solutions.

Ce cadre est assez général. Par exemple, on ne gagne rien en généralité en considérant une perturbation non linéaire du type

$$\begin{cases} \min\ \tilde f(x,p) \\ \tilde c_E(x,p) = 0 \\ \tilde c_I(x,p) \leqslant 0, \end{cases}$$

(on suppose que $\tilde f(\cdot,0) \equiv f(\cdot)$, $\tilde c_E(\cdot,0) \equiv c_E(\cdot)$ et $\tilde c_I(\cdot,0) \equiv c_I(\cdot)$) puisque celle-ci peut être obtenue à partir d'une perturbation linéaire d'une autre formulation de (P_{EI}), comme ci-dessous :

$$\begin{cases} \min\ \tilde f(x,y) \\ \tilde c_E(x,y) = 0 \\ \tilde c_I(x,y) \leqslant 0 \\ y = 0 \end{cases} \quad \longrightarrow \quad \begin{cases} \min\ \tilde f(x,y) \\ \tilde c_E(x,y) = 0 \\ \tilde c_I(x,y) \leqslant 0 \\ y = p. \end{cases}$$

Définition 4.71 On appelle *fonction valeur* associée au problème perturbé (P^p_{EI}), la fonction $v : p \in \mathbb{R}^m \to \overline{\mathbb{R}}$ définie par

$$v(p) = \inf_{x\in X^p} f(x), \tag{4.106}$$

où X^p est l'ensemble admissible de (P^p_{EI}) :

$$X^p := \{x \in \mathbb{E} : c_E(x) + p_E = 0,\ c_I(x) + p_I \leqslant 0\}.$$

□

La fonction valeur décrit donc comment varie la valeur optimale du critère lorsque p varie. Elle jouera à nouveau un rôle dans la théorie de la dualité (chapitre 14). Cette fonction peut prendre la valeur $-\infty$ (problème (P^p_{EI}) non borné) ou $+\infty$ (par convention si l'ensemble admissible $X^p = \varnothing$). Observons que la fonction valeur est aussi la fonction marginale de la fonction $\varphi : \mathbb{R}^m \times \mathbb{E} \to \overline{\mathbb{R}}$, définie en $(p,x) \in \mathbb{R}\times\mathbb{E}$ par

$$\varphi(p,x) = f(x) + \mathcal{I}_{X^p}(x), \tag{4.107}$$

où $\mathcal{I}_{X^p}$ désigne l'indicatrice de X^p.

Nous commencerons par établir deux résultats simples qui donnent une signification précieuse des multiplicateurs de Lagrange, objets pour l'instant relativement abstraits qui semblent ne servir qu'à l'écriture des conditions d'optimalité (section 4.6.1).

Nous montrons ensuite comment on peut introduire des chemins à valeurs dans X^{tp}, dont on contrôle soit la pente initiale soit la pente et la courbure initiales (section 4.6.2). Ceux-ci sont alors utilisés pour montrer la continuité directionnelle de la fonction valeur (section 4.6.3). Ces chemins sont aussi utiles pour étudier le comportement des suites convergentes de solutions des problèmes perturbés (section 4.6.4).

4.6.1 Interprétation marginaliste des multiplicateurs optimaux

Cette section rassemble des résultats donnant du sens aux multiplicateurs de Lagrange ou de Karush, Kuhn et Tucker. Le premier (proposition 4.72) est donné sous des hypothèses un peu fortes qui portent sur la régularité de la solution primale-duale de (P^p_{EI}), qui est supposée dépendre univoquement de p. Le second résultat (proposition 4.76) ne suppose ni la régularité des fonctions f et c, ni l'unicité des solutions primales-duales ; le plus souvent, cependant, il requiert la convexité du problème (P_{EI}).

Dans certaines circonstances de régularité que nous allons préciser, le multiplicateur de Lagrange λ_* du problème d'optimisation (P_{EI}) s'interprète comme le gradient de la fonction valeur en $p = 0$. Autrement dit, $(\lambda_*)_i$ donne la variation de la valeur *optimale* de la fonction coût lorsqu'on perturbe la i-ième contrainte tout en gardant les autres contraintes. Ceci est illustré à la figure 4.9, pour un ensemble admissible

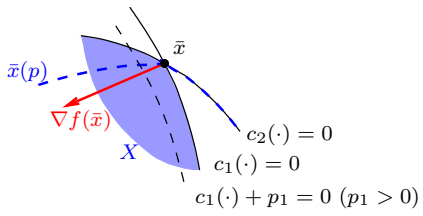


Fig. 4.9. Interpretation marginaliste des multiplicateurs optimaux

défini par deux contraintes d'inégalité : $X = \{x \in \mathbb{R}^2 : c_1(x) \leqslant 0,\ c_2(x) \leqslant 0\}$. Pour une perturbation de la première contrainte $p = (p_1, 0)$, la solution de (P^p_{EI}) pourraît y évoluer comme la courbe en tirets $p \mapsto \bar{x}(p)$ (ici non différentiable). Nous allons voir que les multiplicateurs optimaux donnent la variation au premier ordre du coût f le long de cette courbe en $p = 0$.

Le cas où $(\lambda_*)_i = 0$ est éclairant. Il peut se produire :

- si la contrainte est inactive : on savait déjà qu'alors le multiplicateur correspondant est nul et on comprend bien qu'une petite perturbation de la contrainte n'aura d'incidence ni sur la solution, ni sur la valeur optimale du critère ;
- si la contrainte est active (absence de complémentarité stricte) : ici aussi, on savait que l'on pouvait éliminer la contrainte sans remettre en cause l'optimalité au premier ordre de la solution (le couple (x_*, λ_*) reste solution du système d'optimalité (4.33) modifié) et on apprend avec le résultat ci-dessous que, bien que la solution du problème puisse être affectée par une perturbation de la i-ième contrainte, la valeur optimale du critère ne varie pas *au premier ordre*.

Cette interprétation du multiplicateur optimal est importante pour les applications.

Proposition 4.72 *Supposons que le problème* (P_{EI}) *ait une solution primale-duale unique* (x_*, λ_*) *et que* f *et* c *soient différentiables en* x_*. *On considère le problème perturbé* (P^p_{EI}) *avec* $p = (0, \ldots, 0, p_i, 0, \ldots, 0)$, *pour un indice* $i \in E \cup I$, *et on suppose que pour tout* p_i *voisin de* 0, (P^p_{EI}) *a une solution primale-duale unique* $(\bar{x}(p_i), \bar{\lambda}(p_i))$ *telle que* $p_i \mapsto \bar{x}(p_i)$ *soit différentiable en* 0, $p_i \mapsto \bar{\lambda}(p_i)$ *soit continue en* 0, $\bar{x}(0) = x_*$ *et* $\bar{\lambda}(0) = \lambda_*$. *Alors,*

$$(f \circ \bar{x})'(0) = (\lambda_*)_i.$$

Démonstration. On note comme d'habitude $\ell(x, \lambda) = f(x) + \lambda^\mathsf{T} c(x)$ le lagrangien du problème non perturbé (P_{EI}) et on note $p = (0, \ldots, 0, p_i, 0, \ldots, 0)$ la perturbation des contraintes considérée. En se rappelant que $(\lambda_*)_j = 0$ si $j \in I \setminus I^{0+}_*$, on a

$$\ell(\bar{x}(p_i), \lambda_*) = f(\bar{x}(p_i)) + \sum_{j \in E \cup I^{0+}_*} (\lambda_*)_j c_j(\bar{x}(p_i)).$$

Si $j \in E$, on a $c_j(\bar{x}(p_i)) = -p_j$. Si $j \in I^{0+}_*$, $(\lambda_*)_j > 0$ et par continuité, $(\bar{\lambda}(p_i))_j > 0$ pour une petite perturbation p_i, si bien que par complémentarité on a aussi $c_j(\bar{x}(p_i)) = -p_j$. Dès lors

$$\ell(\bar{x}(p_i), \lambda_*) = f(\bar{x}(p_i)) - \sum_{j \in E \cup I^{0+}_*} (\lambda_*)_j p_j = f(\bar{x}(p_i)) - \lambda_*^\mathsf{T} p.$$

puisque $(\lambda_*)_j = 0$ si $j \notin E \cup I^{0+}_*$. En utilisant le fait que $\nabla_x \ell(x_*, \lambda_*) = 0$, on obtient

$$(f \circ \bar{x})'(0) = \nabla_x \ell(x_*, \lambda_*)^\mathsf{T} \bar{x}'(0) + (\lambda_*)_i = (\lambda_*)_i. \qquad \square$$

En général, les solutions de (P^p_{EI}) ne peuvent pas s'exprimer comme fonction univoque régulière du paramètre de perturbation p et la fonction valeur (4.106) peut ne pas être différentiable en $p = 0$. Nous clarifions ci-dessous le lien entre les multiplicateurs optimaux et la fonciton valeur qui ne suppose pas la différentiabilité des fonctions définissant (P_{EI}), mais qui a besoin de la convexité de la fonction valeur. Cette convexité est assurée si le problème est convexe (proposition 4.75), mais peut l'être pour d'autres raisons. Dans ce cas, nous démontrons à la proposition 4.76 le résultat apothéotique selon lequel l'ensemble des multiplicateurs optimaux n'est autre que $\partial v(0)$, le sous-différentiel de la fonction valeur en zéro. Cette belle identité justifierait à elle seule l'introduction de la notion de sous-différentiel.

Pour y arriver, il nous faut d'abord redéfinir ce que l'on entend par multiplicateur optimal (de Lagrange ou de KKT) si l'on ne suppose pas que les fonctions f et c sont différentiables. Les multiplicateurs optimaux ont en effet été jusqu'à présent définis comme formant le vecteur λ_* permettant d'écrire les conditions d'optimalité de Karush, Kuhn et Tucker (4.33), dans lesquelles les dérivées des fonctions f et c interviennent (dans le gradient du lagrangien essentiellement). Il sera utile d'introduire l'ensemble

$$\Lambda := \{\lambda \in \mathbb{R}^m : \lambda_I \geqslant 0\}$$

dans lequel les mutiplicateurs optimaux résident. Le concept qui va nous permettre ici de nous libérer du besoin de différentiabilité de f et c est celui de point-selle (définition 3.81) du lagrangien ℓ de (P_{EI}) sur $\Omega \times \Lambda$, où Ω est une partie de $\mathbb{E}$, c'est-à-dire de couple $(x_*, \lambda_*) \in \Omega \times \Lambda$ tel que

$$\forall\, (x, \lambda) \in \Omega \times \Lambda : \quad \ell(x_*, \lambda) \leqslant \ell(x_*, \lambda_*) \leqslant \ell(x, \lambda_*). \tag{4.108}$$

Cette notion est plus forte que la stationnarité, car elle exprime aussi la minimalité *globale* sur Ω (voir la proposition 4.74 ci-dessous), mais elle ne requiert pas la différentiabilité de f et c. Par ailleurs, l'utilisation de l'ensemble Ω permet d'obtenir une certaine localisation de la minimalité en x. Nous reviendrons plus longuement sur cette approche utilisant la notion de point-selle au chapitre 14 sur la dualité.

Commençons par établir une expression équivalente de l'inégalité de gauche dans (4.108). Comme $f(x_*)$ est présent dans les deux membres de cette inégalité, elle peut aussi s'écrire $\lambda^T c(x_*) \leqslant \lambda_*^T c(x_*)$, qui est le point (i) de la proposition ci-dessous.

Proposition 4.73 (admissibilité et maximalité de $\ell(x_*, \cdot)$) *Soit $(x_*, \lambda_*) \in \mathbb{E} \times \Lambda$. Alors les propriétés suivantes sont équivalentes :*
(i) $\lambda^T c(x_*) \leqslant \lambda_*^T c(x_*)$ pour tout $\lambda \in \Lambda$,
(ii) $c_E(x_*) = 0$ et $0 \leqslant (\lambda_*)_I \perp c_I(x_*) \leqslant 0$.

Démonstration. $[(i) \Rightarrow (ii)]$ Pour obtenir l'admissibilité de x_*, on prend d'abord $\lambda_E = (\lambda_*)_E + c_E(x_*)$ et $\lambda_I = (\lambda_*)_I \in \mathbb{R}^{m_I}_+$, qui montre que $c_E(x_*) = 0$, et ensuite $\lambda_I = (\lambda_*)_I + c_I(x_*)^+ \in \mathbb{R}^{m_I}_+$, qui montre que $c_I(x_*)^+ = 0$ ou $c_I(x_*) \leqslant 0$. Pour obtenir la complémentarité, on prend $\lambda_I = 0$, qui donne $(\lambda_*)_I c_I(x_*) \geqslant 0$, puis $\lambda_I = 2(\lambda_*)_I$, qui donne $(\lambda_*)_I c_I(x_*) \leqslant 0$ (on peut aussi utiliser le fait que $(\lambda_*)_I \geqslant 0$ et $c_I(x_*) \leqslant 0$), d'où l'égalité $(\lambda_*)_I c_I(x_*) = 0$.

$[(ii) \Rightarrow (i)]$ Soit $\lambda \in \Lambda$. Comme $c_E(x_*) = 0$ par (ii), il suffit de montrer que $\lambda_I c_I(x_*) \leqslant (\lambda_*)_I c_I(x_*)$. Or $(\lambda_*)_I c_I(x_*) = 0$ par la complémentarité dans (ii) et $\lambda_I c_I(x_*) \leqslant 0$ car $\lambda_I \geqslant 0$ pour des λ dans Λ et $c_I(x_*) \leqslant 0$ par (ii). □

Le résultat suivant montre que la « partie primale » x_* d'un point-selle (x_*, λ_*) du lagrangien de (P_{EI}) sur $\Omega \times \Lambda$ est un minimum de f sur $X \cap \Omega$ (rappelons que X désigne l'ensemble admissible de (P_{EI})). Ce résultat est à rapprocher de la condition suffisante CS1 de la proposition 4.33, excepté qu'il n'y a ici aucune hypothèse de différentiabilité ou de convexité. Il est parfois utilisé pour montrer qu'un point est solution globale d'un problème d'optimisation (voir la démonstration du théorème 9.2 et l'exercice 14.11).

Proposition 4.74 (CS0 globale par point-selle) *Soient X l'ensemble admissible de (P_{EI}) et $\Omega \subseteq \mathbb{E}$. Si (x_*, λ_*) est un point-selle du lagrangien de (P_{EI}) sur $\Omega \times \Lambda$, alors x_* minimise f sur $X \cap \Omega$.*

Démonstration. On utilise toutes les conditions exprimant que (x_*, λ_*) est un point-selle du lagrangien sur $\Omega \times \Lambda$. Observons d'abors que $x_* \in X$ (proposition 4.73) et

que $x_* \in \Omega$ (par le fait même que (x_*, λ_*) est un point-selle sur $\Omega \times \Lambda$). Maintenant, si $x \in X \cap \Omega$, on a :

$$\begin{aligned} f(x_*) &= \ell(x_*, \lambda_*) \qquad [c_E(x_*) \text{ et } (\lambda_*)_I^{\mathsf{T}} c_I(x_*) = 0 \text{ (proposition 4.73)}] \\ &\leqslant \ell(x, \lambda_*) \qquad [x_* \text{ minimise } \ell(\cdot, \lambda_*) \text{ sur } \Omega \text{ et } x \in \Omega] \\ &= f(x) + (\lambda_*)_I^{\mathsf{T}} c_I(x) \qquad [c_E(x) = 0] \\ &\leqslant f(x) \qquad [(\lambda_*)_I \geqslant 0 \text{ (proposition 4.73) et } c_I(x) \leqslant 0]. \end{aligned}$$

Ceci montre que x_* minimise f sur $X \cap \Omega$. □

Notre but à présent est d'analyser le sous-différentiel de la fonction valeur en zéro. Celle dont nous parlons tient compte de la localisation par Ω. Nous la noterons

$$v_\Omega : p \in \mathbb{R}^m \to v_\Omega := \inf_{x \in \mathbb{E}} \Big(f(x) + \mathcal{I}_{X^p \cap \Omega}(x) \Big) \in \overline{\mathbb{R}}. \tag{4.109}$$

C'est donc la fonction marginale de la fonction

$$\varphi_\Omega : (p, x) \in \mathbb{R}^m \times \mathbb{E} \mapsto \varphi_\Omega(p, x) := f(x) + \mathcal{I}_{X^p \cap \Omega}(x), \tag{4.110}$$

Pour que le concept de sous-différentiel ait un sens il faut que la fonction valeur soit convexe. La proposition suivante montre qu'il en est bien ainsi si le problème (P_{EI}) est convexe au sens de la définition 4.26 et si Ω est convexe.

Proposition 4.75 (fonction valeur convexe) *Si f et les $\{c_i\}_{i \in I}$ sont convexes, si c_E est affine et si Ω est convexe, la fonction valeur v_Ω définie par (4.109) est convexe.*

Démonstration. Comme fonction marginale de la fonction φ_Ω définie par (4.110), la fonction valeur v_Ω sera convexe si l'on montre que φ_Ω est convexe (proposition 3.35). Le premier terme de la somme définissant φ_Ω étant convexe par hypothèse, il suffit de s'intéresser au second terme et de montrer que pour x, $x' \in \mathbb{E}$, p, $p' \in \mathbb{R}^m$ et $t \in]0, 1[$, on a

$$\mathcal{I}_{X^{(1-t)p+tp'} \cap \Omega}((1-t)x + tx') \leqslant (1-t)\, \mathcal{I}_{X^p \cap \Omega}(x) + t\, \mathcal{I}_{X^{p'} \cap \Omega}(x'),$$

ou encore que

$$x \in X^p \cap \Omega \text{ et } x' \in X^{p'} \cap \Omega \quad \Longrightarrow \quad (1-t)x + tx' \in X^{(1-t)p+tp'} \cap \Omega.$$

Clairement $(1-t)x + tx' \in \Omega$ par convexité de Ω. Par ailleurs, si $x \in X^p$ et $x' \in X^{p'}$, on a

$$c_E(x) + p_E = 0, \quad c_E(x') + p'_E = 0, \quad c_I(x) + p_I \leqslant 0 \quad \text{et} \quad c_I(x') + p'_I \leqslant 0.$$

En utilisant l'affinité de c_E et la convexité des c_i $(i \in I)$, on trouve

$$c_E((1-t)x + tx') + (1-t)p_E + tp'_E = 0 \quad \text{et} \quad c_I((1-t)x + tx') + (1-t)p_I + tp'_I \leqslant 0,$$

ce qui montre l'implication ci-dessus. □

Venons-en maintenant au résultat décrivant $\partial v_\Omega(0)$ comme l'ensemble des multiplicateurs optimaux. Le résultat ne suppose pas la convexité du problème (P_{EI}), ni celle de Ω, mais requiert que v_Ω soir convexe. On rappelle que X est l'ensemble admissible de (P_{EI}).

Proposition 4.76 (sous-différentiel de v_Ω en 0) *Soient $\Omega \subseteq \mathbb{E}$ et x_* est un minimiseur de f sur $X \cap \Omega$. On suppose que la fonction valeur v_Ω déinie par (4.109) est dans* $\mathrm{Conv}(\mathbb{R}^m)$. *Alors*

$$\partial v_\Omega(0) = \{\lambda_* : (x_*, \lambda_*) \text{ est point-selle de } \ell \text{ sur } \Omega \times \Lambda\}.$$

En particulier, $\partial v_\Omega(0) \neq \varnothing$ si, et seulement si, il existe un multiplicateur λ_* tel que (x_*, λ_*) soit point-selle de ℓ sur $\Omega \times \Lambda$.

Démonstration. Comme v_Ω est la fonction marginale de la fonction φ_Ω définie en (4.110), la proposition 3.75 nous apprend que

$$\begin{aligned}
\lambda_* \in \partial v_\Omega(0) &\iff (\lambda_*, 0) \in \partial \varphi_\Omega(0, x_*) \\
&\iff \varphi_\Omega(p, x) \geqslant \varphi_\Omega(0, x_*) + \lambda_*^\mathsf{T} p, \quad \forall p \in \mathbb{R}^m, \quad \forall x \in \mathbb{E} \\
&\iff f(x) \geqslant f(x_*) + \lambda_*^\mathsf{T} p, \quad \forall p \in \mathbb{R}^m, \quad \forall x \in X^p \cap \Omega. \quad (4.111)
\end{aligned}$$

Il reste à montrer que cette dernière condition (4.111) est équivalente au fait que (x_*, λ_*) est un point-selle de ℓ sur $\Omega \times \Lambda$.

Supposons que (4.111) ait lieu. On peut prendre dans (4.111), $x \in \Omega$ et $p = -c(x)$, ce qui donne (pour obtenir la seconde inégalité, on prend $x = x_*$)

$$\ell(x, \lambda_*) \geqslant f(x_*) \qquad \text{et} \qquad \lambda_*^\mathsf{T} c(x_*) \geqslant 0. \tag{4.112}$$

On peut aussi prendre dans (4.111), $p = c(x_*)$ et $x = x_*$, ce qui donne

$$0 \geqslant \lambda_*^\mathsf{T} c(x_*). \tag{4.113}$$

Conclusion : pour tout $x \in \Omega$ et tout $\lambda \in \Lambda$, on a

$$\begin{aligned}
\ell(x_*, \lambda) &\leqslant f(x_*) && [\lambda^\mathsf{T} c(x_*) \leqslant 0, \text{ car } x_* \in X \text{ et } \lambda \in \Lambda] \\
&= \ell(x_*, \lambda_*) && [\lambda_*^\mathsf{T} c(x_*) = 0 \text{ par (4.112) et (4.113)}] \\
&\leqslant \ell(x, \lambda_*) && [(4.112)].
\end{aligned}$$

Inversement, supposons que (x_*, λ_*) soit un point-selle de ℓ sur $\Omega \times \Lambda$. Alors pour tout $x \in \Omega$:

$$\begin{aligned}
f(x_*) &= \ell(x_*, \lambda_*) && [\lambda_*^\mathsf{T} c(x_*) = 0 \text{ par la proposition 4.73}] \\
&\leqslant \ell(x, \lambda_*) && [x_* \in \arg\min\{\ell(x, \lambda_*) : x \in \Omega\}] \\
&= f(x) + \lambda_*^\mathsf{T} c(x).
\end{aligned}$$

Soient $p \in \mathbb{R}^m$ et $x \in X^p \cap \Omega$. On a $(\lambda_*)_E^\mathsf{T} c_E(x) = -(\lambda_*)_E^\mathsf{T} p_E$ et $(\lambda_*)_I^\mathsf{T} c_I(x) \leqslant -(\lambda_*)_I^\mathsf{T} p_I$ (car $(\lambda_*)_I \geqslant 0$ lorsque $\lambda_* \in \Lambda$). Dès lors $\lambda_*^\mathsf{T} c(x) \leqslant -\lambda_*^\mathsf{T} p$ et les membres extrêmes de la relation ci-dessus conduisent à (4.111). □

Terminons cette section par un résultat assurant la non vacuité de $\partial v(0)$, ce qui, par la proposition 4.76, revient à affirmer l'existence d'un point-selle (x_*, λ_*) de ℓ sur $\mathbb{R}^n \times \Lambda$. Il s'agit donc de conditions nécessaires d'optimalité du type de celles conduisant au système d'optimalité de Karush, Kuhn et Tucker (4.33). Comme dans le cas différentiable, l'existence de multiplicateurs optimaux n'est garantie que sous une condition de qualification de contrainte. Nous ferons l'hypothèse de Slater, dans une version moins restrictive que (QC-S) : on ne suppose pas la différentiabilité des contraintes d'inégalité.

Théorème 4.77 (CN0 pour problèmes convexes) *Supposons que f et les $\{c_i\}_{i\in I}$ soient convexes et que c_E soit affine. On accepte un critère $f : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ prenant la valeur $+\infty$, pour autant que son domaine soit non vide. On suppose que les contraintes sont qualifiées au sens de Slater : c'_E est surjective, il existe $\hat{x} \in (\operatorname{dom} f)^\circ$ tel que $c_I(\hat{x}) < 0$ et c est continue en $\hat{x}$. Si x_* est solution de (P_{EI}), alors v est localement lipschitzienne dans un voisinage de 0 et il existe un $\lambda_* \in \mathbb{R}^m$ tel que (x_*, λ_*) soit un point-selle de ℓ sur $\mathbb{R}^n \times \Lambda$.*

Démonstration. D'après la proposition 4.76, il suffit de montrer que $\partial v(0) \neq \varnothing$. D'après la proposition 3.62, cette propriété de non vacuité sera vérifiée si $v \in \operatorname{Conv}(\mathbb{R}^m)$ et $0 \in (\operatorname{dom} v)^\circ$. On sait que v est convexe par la proposition 4.75. D'autre part $v(0) = f(x_*) \in \mathbb{R}$, donc $0 \in \operatorname{dom} v \neq \varnothing$. En montrant que $0 \in (\operatorname{dom} v)^\circ$, on assure du même coup que v est propre (exercice 3.3) et que $\partial v(0) \neq \varnothing$. Soit p voisin de $0 \in \mathbb{R}^m$. Alors, du fait de la surjectivité de $A_E := c'_E$, $h := -A_E^\mathsf{T}(A_E A_E^\mathsf{T})^{-1} p_E$ est voisin de $0 \in \mathbb{E}$ et $c_E(\hat{x} + h) + p_E = 0$; d'autre part, grâce à la continuité de c_I et au fait que $c_I(\hat{x}) < 0$, on a $c_I(\hat{x} + h) + p_I \leqslant 0$. On en déduit que $\hat{x} + h$ est admissible pour (P^p_{EI}), si bien que $v(p) \leqslant f(\hat{x} + h) < +\infty$ ($\hat{x} \in (\operatorname{dom} f)^\circ$).

On notera que v est localement lipschitzienne dans un voisinage de 0 en utilisant le lemme 3.13, qui s'applique puisque v a une minorante affine (proposition 3.7) et que l'on vient de montrer qu'elle est bornée supérieurement dans un voisinage de 0. □

L'exercice 4.24 examine l'allure de la fonction valeur d'un problème sans qualification de contraintes, dans une situation semblable à celle illustrée par le dessin de gauche à la figure 4.5. La fonction valeur est sous-différentiable en zéro si et seulement si le gradient est dans l'espace vectoriel de dimension un engendré par les gradients des contraintes, qui sont colinéaires.

4.6.2 Construction de chemins réguliers dans l'ensemble perturbé

L'utilisation de chemins à valeur dans X a été utile dans l'analyse des conditions du premier ordre (proposition 4.14) et du second ordre (lemme 4.47). Il en est de même ici.

Soit $p \in \mathbb{R}^m$ une perturbation. Nous allons montrer comment on peut construire un chemin $\xi : t \mapsto \xi(t)$ passant par $x \in X$ en $t = 0$ et tel que $\xi(t) \in X^{tp}$ pour $t \geqslant 0$ petit (donc $\xi(t)$ n'est plus nécessairement dans X). Dans un premier temps, on ne contrôle que par la dérivée première $\xi'(0) = d$ de ξ en 0 (proposition 4.78). On verra quelles sont les conditions que $d \in \mathbb{E}$ doit vérifier pour que cela soit possible.

On s'intéresse ensuite à la construction d'un chemin $\xi : t \mapsto \xi(t)$ dont on contrôle les dérivées première $\xi'(0) = d \in \mathbb{E}$ et seconde $\xi''(0) = h \in \mathbb{E}$ en $t = 0$ (proposition 4.79). Nous verrons que d et h doivent vérifier certaines conditions pour qu'un tel chemin puisse être tracé.

Intéressons-nous donc d'abord à la construction de $\xi : t \mapsto \xi(t)$, dont la dérivée $\xi'(0) = d$ en l'origine est donnée. La direction $d \in \mathbb{E}$ ne peut pas être choisie de manière arbitraire. D'une part, on doit avoir $c_E(\xi(t)) + tp_E = 0$, pour tout $|t|$ petit, si bien qu'en dérivant cette identité en $t = 0$, on voit que d doit vérifier

$$c'_E(x) \cdot d + p_E = 0. \tag{4.114}$$

D'autre part, pour $i \in I$, il faut que l'on ait $c_i(\xi(t)) + tp_i \leqslant 0$, pour tout $t \geqslant 0$ petit. Si $c_I \circ \xi$ est continue en x, cette inégalité sera certainement vérifiée pour les indices $i \in I \setminus I^0_x$, car pour ceux-ci $c_i(x) < 0$. Pour $i \in I^0_x$, on a en supposant c_i et ξ différentiable en x et en 0 respectivement :

$$c_i(\xi(t)) + tp_i = t(c'_i(x) \cdot d + p_i) + o(t).$$

Si on se limite à ce développement au premier ordre, il ne suffit pas d'imposer que $c'_i(x) \cdot d + p_i \leqslant 0$ pour assurer la négativité de $c_i(\xi(t)) + tp_i$. C'est pourquoi on choisit d tel que

$$c'_{I^0_x}(x) \cdot d + p_{I^0_x} \leqslant -\epsilon, \tag{4.115}$$

où $\epsilon > 0$ est fixé. Remarquons que, pour $p \in \mathbb{R}^m$ donné, on peut toujours trouver une direction $d \in \mathbb{E}$ vérifiant (4.114) et (4.115) lorsque les conditions de Mangasarian-Fromovitz ont lieu en x (proposition 4.42).

Proposition 4.78 *Soient x un point admissible pour (P_{EI}), $p \in \mathbb{R}^m$ une perturbation et $\epsilon > 0$. On suppose que c_E est C^1 dans un voisinage de x, que $c_{I^0_x}$ est dérivable en x et que $c_{I \setminus I^0_x}$ est continue en x. On suppose également que les conditions de Mangasarian-Fromovitz (QC-MF) ont lieu en x et donc que l'on peut se fixer $d \in \mathbb{E}$ vérifiant (4.114) et (4.115). Alors il existe un chemin $t \mapsto \xi(t) \in \mathbb{E}$ de classe C^1, défini pour $|t|$ suffisamment petit, tel que l'on ait*

$$\xi(0) = x, \quad \xi'(0) = d \quad \text{et} \quad \xi(t) \in X^{tp}, \text{ pour } t \geqslant 0 \text{ petit.}$$

Démonstration. On s'y prend comme dans la proposition 4.14. D'après (QC-MF), $c'_E(x)$ est surjective et on peut donc trouver une matrice Z, de type $(n - m_E) \times n$, telle que

$$\begin{pmatrix} c'_E(x) \\ Z \end{pmatrix} \text{ soit inversible.}$$

On considère l'application $F : \mathbb{E} \times \mathbb{R} \to \mathbb{E}$ définie par

$$F(\xi, t) = \begin{pmatrix} c_E(\xi) + tp_E \\ Z(\xi - x - td) \end{pmatrix}.$$

Cette fonction est de classe C^1 dans un voisinage de $(x, 0)$ et $F'_\xi(x, 0)$ est la matrice inversible ci-dessus. On peut alors appliquer le théorème des fonctions implicites

(théorème C.14) : il existe une fonction $t \in \mathbb{R} \mapsto \xi(t) \in \mathbb{E}$ définie dans un voisinage de 0, de classe C^1, telle que $F(\xi(t), t) = 0$ pour tout $|t|$ petit et $\xi(0) = x$. En dérivant $F(\xi(t), t) = 0$ en $t = 0$, on trouve grâce à (4.114)

$$\begin{pmatrix} c'_E(x) \\ Z \end{pmatrix} \xi'(0) = \begin{pmatrix} -p_E \\ Zd \end{pmatrix} = \begin{pmatrix} c'_E(x) \\ Z \end{pmatrix} d.$$

On en déduit que $\xi'(0) = d$.

Il reste à montrer que $\xi(t) \in X^{tp}$ pour $t \geqslant 0$ petit. On a clairement $c_E(\xi(t)) + tp_E = 0$ pour tout $|t|$ petit. Pour $i \in I \setminus I^0_x$, on a $c_i(x) < 0$ et donc $c_i(\xi(t)) + tp_i \leqslant 0$ pour tout $|t|$ petit, grâce à la continuité de c_i et de ξ. Pour les indices $i \in I^0_x$, on a grâce à la différentiabilité de c_i en x et à (4.115) :

$$c_i(\xi(t)) + tp_i = t(c'_i(x) \cdot d + p_i) + o(t) \leqslant -\epsilon t + o(t) \leqslant 0,$$

pour $t \geqslant 0$ petit. □

Ce résultat a la conséquence importante suivante : pourvu que c soit régulière et que les conditions de Mangasarian-Fromovitz (QC-MF) aient lieu en $x \in X$, l'ensemble perturbé X^{tp} est non vide si $t \geqslant 0$ est suffisamment petit :

$$\forall p \in \mathbb{R}^m,\ \exists\, \bar{t} > 0,\ \forall t \in [0, \bar{t}] :\ X^{tp} \neq \varnothing.$$

Il s'agit d'une propriété de stabilité de l'ensemble admissible de (P_{EI}) sous l'effet de perturbations.

Voyons maintenant comment construire $\xi : t \mapsto \xi(t)$ à valeurs dans X^{tp}, dont les dérivées première $\xi'(0) = d$ et seconde $\xi''(0) = h$ sont données. Les conditions à remplir sont donc

$$\xi(0) = x, \quad \xi'(0) = d, \quad \xi''(0) = h, \quad \xi(t) \in X^{tp}, \text{ pour } t \geqslant 0 \text{ petit.} \tag{4.116}$$

Nous suivons de près la démarche utilisée dans la démonstration du lemme 4.47. Comme ci-dessus, les directions d et $h \in \mathbb{E}$ ne peuvent pas être arbitraires. D'une part, on doit avoir $c_E(\xi(t)) + tp_E = 0$, pour tout $|t|$ petit. En dérivant une première fois, puis une seconde fois cette relation en $t = 0$, on voit qu'il faut que

$$c'_E(x) \cdot d + p_E = 0 \tag{4.117}$$

et

$$c''_E(x) \cdot d^2 + c'_E(x) \cdot h = 0. \tag{4.118}$$

Réaliser $c_o(\xi(t)) + tp_i \leqslant 0$ ne pose pas de difficultés pour les indices $i \in I \setminus I^0_x$, pourvu que $c_i \circ \xi$ soit continue en 0. Pour les indices $i \in I^0_x$, on a $c_i(x) = 0$ et donc

$$c_i(\xi(t)) + tp_i = t(c'_i(x) \cdot d + p_i) + \frac{t^2}{2}\left(c''_i(x) \cdot d^2 + c'_i(x) \cdot h\right) + o(t^2)$$

qui est négatif si

$$c'_{I^0_x}(x_*) \cdot d + p_{I^0_x} \leqslant 0 \tag{4.119}$$

et

$$c''_{I^0_x}(x)\cdot d^2 + c'_{I^0_x}(x)\cdot h \leqslant -\epsilon, \tag{4.120}$$

où $\epsilon > 0$ est fixé. Notons que (4.117)-(4.120) peuvent être vérifiées par des directions d et h, dès lors que les conditions de qualification de Mangasarian-Fromovitz (QC-MF) ont lieu en x : on choisit d'abord d vérifiant (4.117) et (4.119), puis h vérifiant (4.118) et (4.120).

Proposition 4.79 *Soient x un point admissible pour (P_{EI}), $p \in \mathbb{R}^m$ une perturbation et $\epsilon > 0$. On suppose que c_E est C^2 dans un voisinage de x, que $c_{I^0_x}$ est deux fois dérivable en x et que $c_{I\setminus I^0_x}$ est continue en x. On suppose également que les conditions de Mangasarian-Fromovitz (QC-MF) ont lieu en x et donc que l'on peut trouver d et $h \in \mathbb{E}$ vérifiant (4.117), (4.118), (4.119) et (4.120). Alors il existe un chemin $t \mapsto \xi(t) \in \mathbb{E}$ de classe C^2, défini pour $|t|$ suffisamment petit, tel que l'on ait (4.116).*

Démonstration. La technique de démonstration est semblable à celle utilisée dans la démonstration du lemme 4.47. Soit $A_E := c'_E(x)$ la matrice jacobienne de c_E en x. Celle-ci étant surjective, on peut trouver une matrice Z, $(n - m_E) \times n$, telle que

$$\begin{pmatrix} A_E \\ Z \end{pmatrix} \text{ soit inversible.}$$

Considérons alors la fonction $F : \mathbb{E} \times \mathbb{R} \to \mathbb{E}$ définie par

$$F(\xi, t) = \begin{pmatrix} c_E(\xi) + tp_E \\ Z(\xi - x - td - \frac{t^2}{2}h) \end{pmatrix}.$$

Elle est C^2 dans un voisinage de $(x, 0)$, $F(x, 0) = 0$ et $F'_\xi(x, 0)$ est inversible. Par le théorème des fonctions implicites (théorème C.14), il existe alors une fonction $t \mapsto \xi(t)$, définie pour $|t|$ petit, de classe C^2, telle que $F(\xi(t), t) = 0$ pour tout $|t|$ petit et $\xi(0) = x$. En dérivant une première fois $F(\xi(t), t) = 0$ en $t = 0$, on trouve grâce (4.117)

$$\begin{pmatrix} A_E \\ Z \end{pmatrix} \xi'(0) = \begin{pmatrix} -p_E \\ Zd \end{pmatrix} = \begin{pmatrix} A_E \\ Z \end{pmatrix} d.$$

Donc $\xi'(0) = d$. En dérivant deux fois $F(\xi(t), t) = 0$ en $t = 0$, on trouve grâce à (4.118)

$$\begin{pmatrix} A_E \\ Z \end{pmatrix} \xi''(0) = \begin{pmatrix} -c''_E(x)\cdot d^2 \\ Zh \end{pmatrix} = \begin{pmatrix} A_E \\ Z \end{pmatrix} h.$$

Donc $\xi''(0) = h$. Enfin, en suivant le raisonnement fait avant l'énoncé de la proposition, $\xi(t) \in X^{tp}$ pour $t \geqslant 0$ petit, grâce au choix de d et h vérifiant (4.119) et (4.120). □

4.6.3 Continuité directionnelle de la fonction valeur

Nous allons donner des hypothèses assurant la continuité de $t \in \mathbb{R}_+ \mapsto v(tp)$ lorsque $t \downarrow 0$, pour une perturbation p donnée. Il s'agit donc d'une propriété de continuité directionnelle.

Proposition 4.80 *Soit x_* une solution de (P_{EI}). On suppose que c_E est C^1 dans un voisinage de x_*, que $c_{I^0_*}$ est dérivable en x_*, que $c_{I\setminus I^0_*}$ est continue en x_* et que f est continue sur X. On suppose également que les conditions de Mangasarian-Fromovitz (QC-MF) ont lieu en x_*. On se donne $p \in \mathbb{R}^m$ et on suppose que*

$$\text{il existe } r>0, \text{ tel que pour tout } |t| \text{ petit}, \quad \varnothing \neq X^{tp}_* \subseteq B(0,r). \tag{4.121}$$

Alors $t \mapsto v(tp)$ est continue à droite en 0.

Démonstration. Soient $\epsilon > 0$ et $d \in \mathbb{E}$ vérifiant (4.114) et (4.115). Par la proposition 4.78, on peut construire un chemin $t \mapsto \xi(t)$, tel que $\xi(0) = x_*$ et $\xi(t) \in X^{tp}$ pour tout $t \geqslant 0$ petit. Sans utiliser (4.121), on a déjà

$$\limsup_{t\downarrow 0} v(tp) \leqslant \limsup_{t\downarrow 0} f(\xi(t)) = \lim_{t\downarrow 0} f(\xi(t)) = f(x_*) = v(0).$$

Montrons à présent que $\liminf_{t\downarrow 0} v(tp) \geqslant v(0)$. On peut trouver une suite $\{t_k\} \subseteq \mathbb{R}_{++}$ telle que $t_k \downarrow 0$ et

$$\liminf_{t\downarrow 0} v(tp) = \lim_{k\to\infty} v(t_k p).$$

Par (4.121), on a un point $\bar{x}_k \in X^{t_k p}_*$ tel que $v(t_k p) = f(\bar{x}_k)$. De plus $\{\bar{x}_k\} \subseteq B(0,r)$ est bornée et on peut donc en extraire une sous-suite convergente, disons $\bar{x}_k \to \bar{x}$. En passant à la limite dans les équations définissant $X^{t_k p}$, on voit que $\bar{x} \in X$. Par continuité de f, on a pour des sous-suites d'indices k :

$$\liminf_{t\downarrow 0} v(tq) = \lim_{k\to\infty} f(\bar{x}_k) = f(\bar{x}) \geqslant v(0).$$

On a $v(0) \leqslant \liminf_{t\downarrow 0} v(tq) \leqslant \limsup_{t\downarrow 0} v(tq) \leqslant v(0)$; d'où le résultat. □

4.6.4 Étude des sous-suites convergentes de solutions

Considérons une suite de solutions $\{\bar{x}_k\}$ de $(P^{t_k p}_{EI})$ qui converge vers une solution x_* de (P_{EI}), lorsque $t_k \downarrow 0$ (on a vu dans la proposition 4.80 que, sous l'hypothèse (4.121), $\{\bar{x}_k\}$ a des sous-suites qui convergent vers des solutions de (P_{EI})). Dans cette section, nous allons montrer des estimations du type

$$\bar{x}_k - x_* = O(\sqrt{t_k}) \quad \text{ou} \quad \bar{x}_k - x_* = O(t_k).$$

Il faudra des hypothèses plus fortes pour avoir la seconde estimation (c'est bien normal). Les points d'adhérence de $\{(\bar{x}_k - x_*)/\sqrt{t_k}\}$ seront des directions critiques (proposition 4.82), ceux de $\{(\bar{x}_k - x_*)/t_k\}$ seront un peu plus complexes à décrire (proposition 4.83).

Voici une illustration de la proposition 4.82 qui concerne l'estimation en $O(\sqrt{t_k})$.

Exemple 4.81 On considère le problème dans $\mathbb{R}^2$ perturbé par $t_k(0,2) \in \mathbb{R}^2$, avec $t_k \downarrow 0$:

$$\left\{ \begin{array}{l} \min\ x_2 \\ x_2 \geqslant x_1^2 \\ x_2 \geqslant -x_1^2 + 2t_k. \end{array} \right.$$

Clairement $x_* = 0$ et $\bar{x}_k = (\pm\sqrt{t_k}, t_k)$, si bien que $\bar{x}_k - x_* = O(\sqrt{t_k})$ seulement. On a aussi $(\bar{x}_k - x_*)/\sqrt{t_k} \to (\pm 1, 0)$ qui sont des directions critiques.

Si au lieu de prendre x_1^2 dans les contraintes ci-dessus, on prenait $|x_1|^a$, avec $a > 0$, ont aurait $\bar{x}_k = (\pm \sqrt[a]{t_k}, t_k)$ et $\bar{x}_k - x_* = O(\sqrt[a]{t_k})$. Si $a > 2$, ces estimations sont moins bonnes. Si $a < 2$, elles sont meilleures, mais les contraintes ne sont plus de classe C^2. On voit donc que l'hypothèse de régularité de c faite dans la proposition ci-dessous affaiblit le résultat, mais elle est nécessaire car les conditions de d'optimalité du second ordre sont utilisées dans la démonstration. □

Proposition 4.82 *On se donne $p \in \mathbb{R}^m$ et une suite $\{t_k\} \subseteq \mathbb{R}_{++}$ qui converge vers 0 lorsque $k \to \infty$. Pour tout k, on suppose que l'on a une solution $\bar{x}_k$ de $(P_{EI}^{t_k p})$ et que $\{\bar{x}_k\}$ converge vers une solution x_* de (P_{EI}) lorsque $k \to \infty$. On suppose également que c_E est C^1 dans un voisinage de x_* et deux fois dérivable en x_*, que f et $c_{I_*^0}$ sont dérivables dans un voisinage de x_* et deux fois dérivable en x_* et que $c_{I \setminus I_*^0}$ est continue en x_*. On suppose enfin que les conditions de Mangasarian-Fromovitz (QC-MF) ont lieu en x_*, ainsi que la condition suffisante d'optimalité du second ordre faible (4.59). Alors la suite $\{(\bar{x}_k - x_*)/\sqrt{t_k}\}$ est bornée et ses points d'adhérence sont des directions critiques.*

Démonstration. On ne s'intéresse ci-dessous qu'à la sous-suite des $\bar{x}_k$ différents de x_*. S'il n'y a pas de telle sous-suite, le résultat est trivial.

Soient $\epsilon > 0$, $d \in \mathbb{E}$ vérifiant (4.114) et (4.115) en $x = x_*$, et $t \mapsto \xi(t)$ le chemin donné par la proposition 4.78. On a $f(\bar{x}_k) \leqslant f(\xi(t_k))$ et comme $(f \circ \xi)(t_k)$ est différentiable en 0, il existe une constante $C_1 > 0$ telle que

$$f(\bar{x}_k) \leqslant f(x_*) + C_1 t_k. \tag{4.122}$$

En développant $f(\bar{x}_k)$ autour de x_*, cette inégalité devient

$$f'(x_*) \cdot (\bar{x}_k - x_*) + o(\|\bar{x}_k - x_*\|) \leqslant C_1 t_k. \tag{4.123}$$

Faisons également un développement au premier ordre de $c(\bar{x}_k)$ autour de x_*. Les contraintes de $(P_{EI}^{t_k p})$ donnent alors :

$$c_i'(x_*) \cdot (\bar{x}_k - x_*) + o(\|\bar{x}_k - x_*\|) \left\{ \begin{array}{ll} = -t_k p_i, & \text{si } i \in E \\ \leqslant -t_k p_i, & \text{si } i \in I_*^0. \end{array} \right. \tag{4.124}$$

Montrons à présent que $\{(\bar{x}_k - x_*)/\sqrt{t_k}\}$ est bornée. Si ce n'est pas le cas, on peut trouver une sous-suite d'indices k telle que $\|\bar{x}_k - x_*\|/\sqrt{t_k} \to \infty$ et $(\bar{x}_k - x_*)/\|\bar{x}_k - x_*\| \to u \in \mathbb{E}$ (on se rappelle que $\bar{x}_k \neq x_*$). En divisant (4.123) et (4.124) par $\|\bar{x}_k - x_*\|$ et en passant à la limite, on voit que u est une direction critique. Montrons que cela

contredit la CS2 faible (4.59). Soit λ_* le multiplicateur de Lagrange associé à $d = u$ dans (4.59). On note ℓ est lagrangien du problème (P_{EI}) et $L_* = \nabla^2_{xx}\ell(x_*, \lambda_*)$ sa hessienne par rapport à x (pour le produit scalaire euclidien). On a

$$f(\bar{x}_k) = \ell(\bar{x}_k, \lambda_*) - \sum_{i \in E \cup I} (\lambda_*)_i c_i(\bar{x}_k).$$

Comme $c_E(\bar{x}_k) = -t_k p_E$, $c_I(\bar{x}_k) \leqslant -t_k p_I$ et $(\lambda_*)_I \geqslant 0$, la somme dans le second membre de cette égalité est $\leqslant -t_k \lambda_*^{\mathsf{T}} p$. Ensuite, en développant $\ell(\bar{x}_k, \lambda_*)$ autour de x_* au premier ordre et en utilisant $\ell(x_*, \lambda_*) = f(x_*)$ et $\nabla_x \ell(x_*, \lambda_*) = 0$, on trouve

$$f(\bar{x}_k) \geqslant f(x_*) + t_k \lambda_*^{\mathsf{T}} p + \frac{1}{2}(\bar{x}_k - x_*)^{\mathsf{T}} L_* (\bar{x}_k - x_*) + o(\|\bar{x}_k - x_*\|^2). \tag{4.125}$$

Avec (4.122), on obtient pour une constante $C_2 > 0$

$$(\bar{x}_k - x_*)^{\mathsf{T}} L_* (\bar{x}_k - x_*) \leqslant C_2 t_k + o(\|\bar{x}_k - x_*\|^2). \tag{4.126}$$

En divisant par $\|\bar{x}_k - x_*\|^2$ et en passant à la limite, on obtient $u^{\mathsf{T}} L_* u \leqslant 0$, ce qui contredit la CS2 (4.59), puisque $u \neq 0$.

Soit alors $\bar{d}$ un point d'adhérence de la suite bornée $\{(\bar{x}_k - x_*)/\sqrt{t_k}\}$. On divise cette fois (4.123) et (4.124) par $\sqrt{t_k}$ et on passe à la limite dans ces équations en tenant compte de ce que

$$\frac{o(\|\bar{x}_k - x_*\|)}{\sqrt{t_k}} = \frac{o(\|\bar{x}_k - x_*\|)}{\|\bar{x}_k - x_*\|} \frac{\|\bar{x}_k - x_*\|}{\sqrt{t_k}} \to 0.$$

On voit que $\bar{d}$ est une direction critique. □

Si l'on essaye de suivre la même démonstration que ci-dessus pour étudier la suite $\{(\bar{x}_k - x_*)/t_k\}$, on voit qu'on peut la mener jusqu'à l'estimation (4.126), où il faudrait que C_2 soit nul. La difficulté vient de ce que la constante C_1 dans (4.122) peut être supérieure à $\lambda_*^{\mathsf{T}} p$. On surmonte cette difficulté en donnant une estimation plus précise de la constante C_1 dans (4.122) et en choisissant λ_* non plus en fonction de u, mais en fonction de p. Il faut alors que les conditions suffisantes du second ordre *fortes* (4.63) aient lieu pour que l'on puisse conclure.

Proposition 4.83 *On se donne $p \in \mathbb{R}^m$ et une suite $\{t_k\} \subseteq \mathbb{R}_{++}$ qui converge vers 0 lorsque $k \to \infty$. Pour tout k, on suppose que l'on a une solution $\bar{x}_k$ de $(P_{EI}^{t_k p})$ et que $\{\bar{x}_k\}$ converge vers une solution x_* de (P_{EI}) lorsque $k \to \infty$. On suppose également que c_E est C^2 dans un voisinage de x_*, que f et $c_{I^0_*}$ sont dérivables dans un voisinage de x_* et deux fois dérivable en x_* et que $c_{I \setminus I^0_*}$ est continue en x_*. On suppose enfin que les conditions de Mangasarian-Fromovitz (QC-MF) ont lieu en x_*, ainsi que la condition suffisante d'optimalité du second ordre forte (4.63). Alors la suite $\{(\bar{x}_k - x_*)/t_k\}$ est bornée et ses points d'adhérence sont solutions du problème*

$$\begin{cases} \min\limits_{d\in\mathbb{E}} \left(\max\limits_{\lambda_*\in\Lambda_*} \langle L_* d, d\rangle \right) \\ \nabla f(x_*)^\mathsf{T} d = \max\limits_{\lambda_*\in\Lambda_*} \lambda_*^\mathsf{T} p \\ c'_E(x_*)\cdot d + p_E = 0 \\ c'_{I_*^0}(x_*)\cdot d + p_{I_*^0} \leqslant 0. \end{cases} \tag{4.127}$$

Notes

La condition d'optimalité (4.15) des problèmes sans contrainte est parfois appelée *équation de Fermat* pour rappeler une condition similaire trouvée, dans le cas d'un polynôme réel d'une variable réelle, par Pierre de Fermat. Cette découverte, dont il est fait allusion dans l'épigraphe de ce chapitre, daterait de 1629, c'est-à-dire environ quarante ans avant l'invention *officielle* du calcul différentiel par Newton et Leibniz ! Elle ne fut pleinement exposée qu'en 1638 dans une lettre à Roberval. Johannes Kepler avait déjà exprimé cela de manière qualitative en 1615 dans sa *Nouvelle stéréométrie des tonneaux de vins*, en observant qu'une fonction réelle varie très peu au voisinage d'un maximum. Ceci lui permit de résoudre quelques problèmes d'extrémum comme celui du plus grand cylindre inscrit dans une sphère. L'emploi explicite de la dérivée dans ces problèmes est dû à Leibniz (1684)[1], qui utilise déjà les dérivées secondes pour distinguer les minima des maxima. La généralisation au cas des fonctions de deux variables est due à Euler. Celui-ci utilise cette condition en *Calcul des Variations*, où elle porte le nom d'équation d'Euler ou d'Euler-Lagrange. [87, 6, 413]

La célèbre méthode des multiplicateurs (théorème 4.17) est attribuée à Lagrange qui l'énonça dans sa *Méchanique analytique* [372 ; 1788, pages 77-112]. On en trouve toutefois déjà des traces dans des travaux d'Euler sur les problèmes isopérimétriques (1744). Lagrange utilisa d'abord cette méthode pour résoudre un problème de calcul des variations sous contraintes, en dimension infinie donc ! Neuf ans plus tard, dans sa *Théorie des fonctions analytiques* (1797), il applique cette méthode aux problèmes de dimension finie sous contraintes d'égalité de la forme (P_E). [87, 6]

L'étude des problèmes avec contraintes d'inégalité de la forme (P_{EI}) est beaucoup plus récente [1, 207]. La condition nécessaire d'optimalité du premier ordre $\nabla f(x_*) \in (\mathrm{T}_{x_*} X)^+$ du théorème 4.6 pour un ensemble admissible X général, qui nous a servi de point de départ pour établir les conditions d'optimalité du premier ordre de problèmes structurés, était déjà exprimée ainsi par Peano [477, 478] dès 1887, puis par Kantorovitch [344] en 1940, mais ce résultat est passé inaperçu ou a été oublié [485, 172]. La définition 4.30 de condition de qualification des contraintes est

[1] *Nova methodus pro maximis et minimis, itemque tangentibus, qua nec irrationales quantitates moratur.* C'est aussi dans cet ouvrage que Leibniz donne pour la première fois les formules bien connues de dérivation d'un produit, d'un quotient et d'une puissance entière.

parfois associée au nom d'Abadie [208]. Les conditions nécessaires d'optimalité (4.33) ont longtemps été attribuées à Kuhn et Tucker [369 ; 1951]. Après bien des années, on constata que ces conditions avaient déjà été données par Karush [347 ; 1939] dans une thèse qui ne fut jamais publiée, mais qui est décrite dans le compte rendu historique de Kuhn [368]. Une approche différente conduisant au même résultat a été suivie par John [338 ; 1948], également avant les travaux de Kuhn et Tucker. C'est aussi ce point de vue qu'ont pris Mangasarian et Fromovitz [405 ; 1967] pour introduire la condition de qualification (QC-MF) qui porte leur nom. Les conditions de qualification de Slater (QC-S) ont été introduites dans [555 ; 1950]. Sur ces sujets, on lira avec profit la revue de Rockafellar [526 ; 1993] sur les conditions du premier ordre et le petit livre didactique d'Hiriart-Urruty [315]. La monographie de Gauvin [232] est vivifiante, avec une approche qui demande une bonne maîtrise de l'optimisation linéaire. Il donne les conditions nécessaires du second ordre les plus fines, celles des théorèmes 4.48 et 4.49.

Pour obtenir les conditions d'optimalité des problèmes avec contraintes générales de la section 4.5, nous nous sommes mis comme contrainte de garder l'esprit de la démarche suivie pour l'obtention des conditions d'optimalité des problèmes qui précèdent, celle faisant usage du lemme de Farkas. La condition générale de qualification de contrainte de la définition 4.54 s'impose alors naturellement et les conditions du premier ordre en découlent aisément (cette qualification de contrainte se retrouve chez Guignard [294 ; 1969]). La borne d'erreur de Robinson [513] (section 4.5.2) a été obtenue en suivant Bonnans et Shapiro [74 ; 2000] pour le théorème 4.63 (en apportant les simplifications permises par la dimension finie des espaces sous-jacents) et Cominetti [133 ; 1990] pour le lemme 4.66 de Lyusternik et la proposition 4.67 sur la diffusion de la régularité métrique. Ces deux sources contiennent beaucoup d'autres références sur l'origine de ces résultats. La démonstration de la proposition 4.69 a bénéficié de la lecture de [645, 74].

L'étude de l'effet de perturbations sur des problèmes d'optimisation, abordée à la section 4.6, est d'une importance considérable et ne cesse d'être le sujet de publications dans des situations les plus variées. La présentation qui en est donnée s'inspire de [232]. On trouvera un approfondissement de ces questions dans les ouvrages de Levitin [390 ; 1994] et de Bonnans et Shapiro [74 ; 2000], ainsi que chez Rockafellar [525 ; 1981] et beaucoup d'autres.

La théorie en dimension infinie est présentée par Ekeland et Temam [191 ; 1974], Barbu et Precupanu [36 ; 1975], Aubin et Ekeland [25 ; 1984], Bonnans et Shapiro [74 ; 2000], Mordukhovich [436 ; 2006] et Ito et Kunisch [328 ; 2008].

Exercices

4.1. *Propriétés du cône tangent.* Soient $\{X_i\}_{i\in I}$ une famille de parties d'un espace vectoriel $\mathbb{E}$ et $x \in \mathbb{E}$. Alors

1) $\mathrm{T}_x(\cap_{i\in I} X_i) \subseteq \cap_{i\in I}(\mathrm{T}_x\, X_i)$,
2) $\mathrm{T}_x(\cup_{i\in I} X_i) \supseteq \cup_{i\in I}(\mathrm{T}_x\, X_i)$, avec égalité si I est fini,
3) $\mathrm{T}_{(x,y)}(X \times Y) \subseteq (\mathrm{T}_x\, X) \times (\mathrm{T}_y\, Y)$, si X (resp. Y) est une partie d'un espace vectoriel $\mathbb{E}$ (resp. $\mathbb{F}$), avec égalité si X et Y sont convexes.

Montrez par des contre-exemples que l'on n'a pas nécessairement égalité au point 1, ni au point 2 si I est infini dénombrable.

4.2. *Cônes tangent et normal à un convexe.* Lorsque C est un convexe fermé et $x \in C$, les définitions de cône tangent données aux sections 2.5.6 et 4.1.1 coïncident. De même, les définitions de cône normal données aux sections 2.5.2 et 4.1.1 coïncident.

4.3. *Transport affine des cônes tangent et normal à un convexe.* Soient $\mathbb{E}$ et $\mathbb{F}$ deux espaces vectoriels et $a : \mathbb{E} \to \mathbb{F} : x \mapsto Ax + b$ une application affine ($A : \mathbb{E} \to \mathbb{F}$ est linéaire et $b \in \mathbb{F}$). On désigne par $f^{-1}(P)$ l'image réciproque d'un ensemble P par une application f.

1) Si C est un convexe de $\mathbb{E}$ et si $x \in C$, alors

$$\mathrm{T}_{a(x)}\, a(C) = \overline{A(\mathrm{T}_x\, C)}. \tag{4.128}$$

2) Si C est un convexe de $\mathbb{F}$ et si $x \in a^{-1}(C)$, alors

$$\mathrm{T}_x(a^{-1}(C)) = A^{-1}(\mathrm{T}_{a(x)}[C \cap \mathcal{R}(a)]). \tag{4.129}$$

On suppose à présent que $\mathbb{E}$ et $\mathbb{F}$ sont des espaces euclidiens et on note A^* l'adjoint de A.

3) Si C est un convexe de $\mathbb{E}$ et si $x \in C$, alors

$$\mathrm{N}_{a(x)} a(C) = (A^*)^{-1}(\mathrm{N}_x C). \tag{4.130}$$

4) Si C est un convexe de $\mathbb{F}$ et si $x \in a^{-1}(C)$, alors

$$\mathrm{N}_x(a^{-1}(C)) = A^*(\mathrm{N}_{a(x)}[C \cap \mathcal{R}(a)]). \tag{4.131}$$

En particulier, l'ensemble $A^*(\mathrm{N}_{a(x)}[C \cap \mathcal{R}(a)])$ est fermé.

4.4. *CS2 diffuse pour problèmes sans contrainte.* Considérons le problème sans contrainte (4.14). Supposons que f soit deux fois différentiable dans un voisinage de $x_* \in \mathbb{E}$, que $f'(x_*) = 0$ et que $f''(x) \cdot h^2 \geqslant 0$ pour tout $h \in \mathbb{E}$ et tout x dans un voisinage de x_*. Montrez qu'alors x_* est un minimum local de f. Montrez par un contre-exemple que la réciproque est fausse.

4.5. *CS2 pour un maximum sans contrainte.* Supposons que $f : \mathbb{E} \to \mathbb{R}$ soit C^1 dans un voisinage d'un point $x_* \in \mathbb{E}$ et deux fois dérivable en x_*. Supposons également que $\nabla f(x_*) = 0$ et que $\nabla^2 f(x_*)$ soit définie *négative*. Alors x_* est un *maximum* local strict de f.

Remarque. Il y a des CN2 analogues.

4.6. *Optimisation quadratique sans contrainte.* Soient $A \in \mathcal{S}^n$ et $b \in \mathbb{R}^n$. On considère le problème $\min\{f(x) : x \in \mathbb{R}^n\}$, où $f(x) = \frac{1}{2} x^{\mathsf{T}} A x + b^{\mathsf{T}} x$.

1) Montrez que, si $A \succ 0$, le problème a une solution unique.
2) Montrez que les propriétés suivantes sont équivalentes :
   (*a*) le problème a une solution,
   (*b*) $A \succcurlyeq 0$ et $b \in \mathcal{R}(A)$,
   (*c*) f est bornée inférieurement,

   et que dans ce cas, l'ensemble des solutions est de la forme $x_0 + \mathcal{N}(A)$, où x_0 est une solution particulière du problème, et la valeur optimale vaut $-b^{\mathsf{T}} A^{\dagger} b$, où $A^{\dagger}$ est le pseudo-inverse de A.

   Remarque : ce résultat ne peut s'étendre à des polynômes de degré strictement supérieur à 2 ; par exemple, le *polynôme quartique* $x \in \mathbb{R}^2 \mapsto x_1^2 + (1 - x_1 x_2)^2$ est strictement positif sur $\mathbb{R}^2$; il n'atteint donc pas sa borne inférieure qui est nulle (on l'approche en prenant $x_1 \to 0$, $x_1 \neq 0$ et $x_2 = 1/x_1$) [221 ; 1956, annexe (i)].

3) Montrez que la valeur optimale du problème est donnée par

$$\inf_{x\in\mathbb{R}^n}\left(\frac{1}{2}x^{\mathsf{T}}Ax+b^{\mathsf{T}}x\right)=-\frac{1}{2}\inf_{\substack{\alpha\geqslant 0\\ \alpha A\succcurlyeq bb^{\mathsf{T}}}}\alpha.$$

4.7. *Approximation de rang un d'une matrice symétrique.* Soit M une matrice symétrique d'ordre n. On considère le problème

$$\inf_{v\in\mathbb{R}^n}\ \frac{1}{2}\|M-vv^{\mathsf{T}}\|_F^2.$$

qui consiste à trouver une matrice de rang $\leqslant 1$ qui soit la plus proche de M au sens de la norme de Frobenius. Montrez que ce problème a une solution et déterminez ses solutions.

4.8. *CS2 pour un maximum avec contraintes d'égalité.* Supposons que f et c soient C^1 dans un voisinage d'un point $x_*\in\mathbb{E}$ et deux fois dérivables en x_*. Supposons également que $c(x_*)=0$ et qu'il existe $\lambda_*\in\mathbb{R}^m$ tel que l'on ait $\nabla_x\ell(x_*,\lambda_*)=0$ et $\langle L_*d,d\rangle<0$, pour tout $d\in\mathcal{N}(c'(x_*))\setminus\{0\}$. Alors x_* est un *maximum* local strict de f sur $\{x\in\mathbb{E}:c(x)=0\}$.

Remarque. Il y a des CN2 analogues.

Remarque. Les CS2 pour avoir un maximum de f sous contraintes d'inégalité ne s'obtiennent pas directement en changeant le signe de l'inégalité (4.59) dans le théorème 4.49. Par exemple, dans le problème $\min\{x_1^2-x_2^2 : x_1\geqslant 1\}$, l'unique point stationnaire est $x_*=(1,0)$ avec pour multiplicateur $\lambda_*=2$. La hessienne du lagrangien $L_*=\operatorname{diag}(2,-2)$ est tel que $\langle L_*d,d\rangle<0$ pour toute direction non nulle dans le cône critique $C_*=\{d:d_1=0\}$. Pourtant, x_* n'est pas un maximum local strict puisque pour tout $t>0$, $x_*+(t,0)$ est admissible et $f(x_*+(t,0))>f(x_*)$.

4.9. *Élimination hasardeuse de contrainte* [460]. L'exemple suivant montre qu'il faut être très prudent lorsqu'on élimine des contraintes. On considère le problème

$$\begin{cases}\min_{(x_1,x_2)}\ x_1^2+(x_2-1)^2\\ \alpha\,x_1^2=x_2,\end{cases}\tag{4.132}$$

dans lequel $\alpha>0$. On déterminera les solutions de ce problème (4.132) en fonction de $\alpha>0$. On remplace à présent dans le critère de (4.132), x_1^2 par sa valeur donnée par la contrainte, ce qui conduit au problème

$$\min_{x_2}\ \frac{1}{\alpha}\,x_2+(x_2-1)^2.\tag{4.133}$$

Déterminez l'unique solution du problème (4.133). Celle-ci n'est pas nécessairement solution de (4.132). Pourquoi ?

4.10. *Minimisation d'une fonction cubique sur le cercle* [295 ; ex. 9.10]. On considère le problème d'optimisation en $(x_1,x_2)\in\mathbb{R}^2$ suivant

$$\begin{cases}\min\ \frac{1}{3}(x_1^3+x_2^3)\\ \frac{1}{2}(x_1^2+x_2^2-1)=0.\end{cases}\tag{4.134}$$

Ce problème admet les 6 points stationnaires suivants $x=\pm(0,1)$, $x=\pm(1,0)$ et $x=\pm(1/\sqrt{2},1/\sqrt{2})$. En utilisant les conditions d'optimalité du second ordre, déterminez si ces points sont des minima ou des maxima locaux.

4.11. *Minimisation d'une fonction linéaire sur une boule.* Soit $c\in\mathbb{R}^n$ *non nul* (le cas où $c=0$ est trivial). Calculez les solutions du problème $\min\{c^{\mathsf{T}}x : x\in\bar{B}_p\}$, où $\bar{B}_p$ est la

boule unité fermée de $\mathbb{R}^n$ pour la norme ℓ_p définie par (A.5) ; considérez séparément les cas où $p = 1$, $1 < p < \infty$ et $p = \infty$. En déduire l'*inégalité de Hölder* : pour tout x et $y \in \mathbb{R}^n$ et tout $p \in [1, \infty]$, on a

$$|x^\mathsf{T} y| \leqslant \|x\|_p \|y\|_{p'}, \quad \text{avec } \tfrac{1}{p} + \tfrac{1}{p'} = 1.$$

On dit que les nombres p et p' sont *conjugués*. Montrez que, lorsque $1 < p < +\infty$, on a égalité dans celle-ci si, et seulement si, si x est parallèle au vecteur de composante $\operatorname{sgn}(y_i)|y_i|^{p'/p}$ (qui est y si $p = 2$).

4.12. *Quotient de Rayleigh.* Soient $\mathbb{E}$ un espace euclidien, muni du produit scalaire $\langle \cdot, \cdot \rangle$ et de la norme associée $\|\cdot\| := \langle \cdot, \cdot \rangle^{1/2}$ et $A : \mathbb{E} \to \mathbb{E}$ une application linéaire auto-adjointe (pour ce produit scalaire : pour tout x, $y \in \mathbb{E}$, on a $\langle Ax, y \rangle = \langle x, Ay \rangle$). Le *quotient de Rayleigh* $q(x)$ de A en $x \in \mathbb{E} \setminus \{0\}$ est défini par

$$q(x) = \frac{\langle Ax, x \rangle}{\|x\|^2}. \tag{4.135}$$

On s'intéresse au problème de la minimisation de q (et plus généralement de la description de ses points stationnaires) en considérant le problème équivalent

$$\inf_{\substack{x \in \mathbb{E} \\ \|x\|^2 = 1}} \langle Ax, x \rangle, \tag{4.136}$$

qui est plus simple par la quadraticité de son critère et de sa contrainte.

1) Dans quel sens les problèmes (4.135) et (4.136) sont-ils équivalents ? Montrez que ces problèmes ont une solution.
2) Soit $(\bar{x}, \bar{\lambda}) \in \mathbb{E} \times \mathbb{R}$. Montrez que $\bar{x}$ est un vecteur propre unitaire (i.e., de norme 1) de A, de valeur propre $\bar{\lambda}$, si, et seulement si, $\bar{x}$ est un point stationnaire de (4.136) de multiplicateur $\bar{\lambda}$. De plus, $\bar{\lambda}$ est la valeur du critère en ce point stationnaire (ou valeur critique).
3) Montrez que les solutions de (4.136) sont les vecteurs propres unitaires de A de valeur propre minimale $\lambda_{\min}(A)$. En particulier,
$$\forall\, x \in \mathbb{E} : \quad \langle Ax, x \rangle \geqslant \lambda_{\min}(A) \|x\|^2. \tag{4.137}$$
4) Montrez que la condition suffisante du second ordre (4.26) du problème (4.136) est vérifiée en une solution si, et seulement si, la plus petite valeur propre de A est simple (i.e., l'espace vectoriel associé est de dimension 1).
5) On note λ_i les valeurs propres de A. Montrez que
$$\begin{aligned} \max_i |\lambda_i| &= \|A\| \\ \lambda_{\max}(A) &= \|A\| \quad (\text{si } A \succcurlyeq 0) \\ \lambda_{\min}(A) &= \|A^{-1}\|^{-1} \quad (\text{si } A \succ 0). \end{aligned}$$

4.13. *Valeurs singulières d'une matrice.* Rappelons le cadre défini à la section B.5.4. On note $\mathbb{E}$ et $\mathbb{F}$ deux espaces euclidiens sur $\mathbb{R}$ (produits scalaires notés $\langle \cdot, \cdot \rangle$ et normes associées notées $\|\cdot\|$), de dimension respective $n := \dim \mathbb{E}$ et $m := \dim \mathbb{F}$. On considère une application linéaire $A : \mathbb{E} \to \mathbb{F}$ de rang $r \geqslant 1$ et on note $A^* : \mathbb{F} \to \mathbb{E}$ son adjointe. On considère le problème de maximisation en $(x, y) \in \mathbb{E} \times \mathbb{F}$ suivant :

$$\begin{cases} \sup \langle Ax, y \rangle \\ \|x\|^2 = 1 \\ \|y\|^2 = 1. \end{cases} \tag{4.138}$$

On introduit l'application linéaire auto-adjointe $\hat{A} : \mathbb{E} \times \mathbb{F} \to \mathbb{E} \times \mathbb{F} : (x, y) \mapsto (A^* y, Ax)$.

1) Montrez que le problème (4.138) a une solution, que ses contraintes sont qualifiées en tout point admissible et que chaque point stationnaire est un vecteur propre de $\hat{A}$, de valeur propre égale à la valeur du critère en ce point stationnaire.
2) Soit $(\bar{x}, \bar{y})$ une solution de (4.138) et $\bar{\sigma} = \langle A\bar{x}, \bar{y}\rangle$. Montrez que $\bar{\sigma}$ est la valeur propre maximale de $\hat{A}$ (c'est la valeur singulière maximale si $A \neq 0$).

4.14. *Remplacer des contraintes d'égalité par des contraintes d'inégalité.* Si $I \subseteq [1:n]$, on note ci-dessous I^c le complémentaire de I dans $[1:n]$. Soient A une matrice $m \times n$, $b \in \mathbb{R}^m$, $I \subseteq [1:n]$ et $f : \mathbb{E} \to \mathbb{R}$ une fonction convexe différentiable. Si $\bar{x}$ est solution du problème 5

$$\min \{f(x) : Ax = b,\ x_I \geqslant 0,\ x_{I^c} = 0\},$$

alors il existe un sous-ensemble d'indices $J \subseteq [1:n]$, contenant I, tel que $\bar{x}$ est aussi solution du problème

$$\min \{f(x) : Ax = b,\ x_J \geqslant 0,\ x_{J^c} \leqslant 0\}.$$

4.15. *Qualification sans conditions de Mangasarian-Fromovitz.* Soit $X = \{x \in \mathbb{R}^2 : 0 \leqslant x_2 \leqslant x_1^2\}$. Montrez que les contraintes sont qualifiées en $x = 0$, alors que les conditions de qualification de Mangasarian-Fromovitz n'ont pas lieu.

4.16. *Stabilité des conditions de Mangasarian-Fromovitz.* Soit $x \in X_{EI}$ (ensemble défini en (4.29)). Si c est C^1 dans un voisinage de x et si (QC-MF) a lieu en x, (QC-MF) a aussi lieu aux points de X_{EI} voisins de x.

4.17. *Aspect topologique des conditions de Mangasarian-Fromovitz.* Soient A_E et A_J deux matrices de type $m_E \times n$ et $m_J \times n$ vérifiant la condition de régularité (4.39). Montrez, qu'il existe une constante C telle que pour tout $v \in \mathbb{R}^{m_E+m_J}$, on peut trouver $d \in \mathbb{E}$ tel que $A_E d = v_E$, $A_J d \leqslant v_J$ et $\|d\| \leqslant C\|v\|$.

4.18. *Autres expressions de la stationnarité* [128, 401]. On considère le problème (P_{EI}), dont on note X l'ensemble admissible. Soit $x_* \in X$. On suppose que f et $c_{E\cup I^0_*}$ sont dérivables en x_* et que les contraintes sont qualifiées en x_*. Montrez que les propriétés suivantes sont équivalentes :

(*i*) il existe $\lambda_* \in \mathbb{R}^m$ tel que l'on ait les conditions de KKT (4.33) ;
(*ii*) zéro est solution du problème d'optimisation linéaire

$$\begin{cases} \min_v \langle \nabla f(x_*), v\rangle \\ c'_E(x_*)\cdot v = 0 \\ c'_{I^0_*}(x_*)\cdot v \leqslant 0\,; \end{cases}$$

(*iii*) $\forall \alpha \geqslant \|\nabla f(x_*)\|$ et $\forall v \in \mathbb{E}$, on a $\langle \nabla f(x_*), v\rangle + \alpha \operatorname{dist}(v, \mathrm{T}_{x_*} X) \geqslant 0$ ($\operatorname{dist}(\cdot, C)$ désigne la distance à C).

4.19. *Conditions d'optimalité de Fritz-John.* On considère le problème (P_{EI}), avec f et c régulières. Soit $x \in \mathbb{E}$ un point admissible pour ce problème. Si

$$\{d \in \mathbb{E} : f'(x)\cdot d < 0,\ c'_E(x)\cdot d = 0,\ c'_{I^0_x}(x)\cdot d < 0\} = \varnothing,$$

alors il existe $(\lambda_0, \lambda) \in \mathbb{R} \times \mathbb{R}^m$ tel que $(\lambda_0, \lambda) \neq 0$, $\lambda_I \geqslant 0$, $c_I(x)^{\mathsf{T}}\lambda_I = 0$ et

$$\lambda_0 \nabla f(x) + \sum_{i=1}^{m} \lambda_i \nabla c_i(x) = 0.$$

La réciproque a lieu si $c'_E(x)$ est surjective.

4.20. *Un minimum fort est un point stationnaire isolé* [514 ; théorème 2.4]. On considère le problème (P_{EI}) et on suppose que les hypothèses du théorème 4.49 sont vérifiées, avec en plus : f et c deux fois dérivables dans un voisinages de x_*, (4.59) est remplacée par (4.63) (CS2 fortes) et les contraintes sont qualifiées en x_* au sens de (QC-MF). Montrez que, dans ces conditions, x_* est un point stationnaire isolé (il n'y a pas d'autres points stationnaires que x_* dans un voisinage de x_*).

4.21. *CN2 fortes avec (QC-A) ou (QC-IL).* Soit x_* un minimum local du problème (P_{EI}). On suppose que f et c sont deux fois dérivables en x_* et que la condition de qualification (QC-A) ou (QC-IL) a lieu en x_*. Soit Λ_* l'ensemble des multiplicateurs optimaux associés à x_*, que l'on sait être non vide. Pour $\lambda_* \in \Lambda_*$, on définit

$$\tilde{X}_{\lambda_*} := \{x \in \mathbb{E} : c_{E \cup I_*^{0+}}(x) = 0,\ c_{I \setminus I_*^{0+}}(x) \leqslant 0\}.$$

On note C_* le cône critique en x_* et $L_* := \nabla^2_{xx}\ell(x_*, \lambda_*)$ pour un $\lambda_* \in \mathbb{R}^m$ donné.

1) Montrer que $\forall\, \lambda_* \in \Lambda_*$, $\mathrm{T}_{x_*} \tilde{X}_{\lambda_*} = C_*$,
2) Montrer que $\forall\, \lambda_* \in \Lambda_*$, $\forall\, d \in C_*$, $\langle L_* d, d\rangle \geqslant 0$,
3) Dans l'exemple (4.48), (QC-MF) a lieu en x_* mais pas (QC-A) ni (QC-IL). Montrer que l'égalité du point 1 n'est pas vérifiée, quel que soit $\lambda_* \in \Lambda_*$.

4.22. *Conditions suffisantes de qualification (QC-R)* [645]. Montrez que la condition de qualification (QC-R) est vérifiée si l'une des propriétés suivantes a lieu

(i) $c'(x)$ est surjective,
(ii) $\exists\, d \in \mathbb{E}$ tel que $c'(x)d \in \operatorname{int} \mathrm{T}_{c(x)}\, K$.

4.23. Soient $\mathbb{E}$ et $\mathbb{F}$ deux espaces euclidiens. On considère le problème d'optimisation (P_G) auquel on ajoute la contrainte d'appartenance à un ensemble "$x \in Q$", ce qui conduit au problème suivant

$$(P_{QG}) \quad \begin{cases} \min\ f(x) \\ x \in Q \\ c(x) \in G, \end{cases}$$

où $f : \mathbb{E} \to \mathbb{R}$ et $c : \mathbb{E} \to \mathbb{F}$ sont des fonctions différentiables, Q est un convexe fermé non vide de $\mathbb{E}$ et G est un convexe fermé non vide de $\mathbb{F}$. L'ensemble admissible de ce problème est noté $X_{QG} := \{x \in Q : c(x) \in G\} = Q \cap c^{-1}(G)$.

1. Soit $x \in X_{QG}$. Montrez que

$$\mathrm{T}_x\, X \subseteq \mathrm{T}'_x\, X := \{d \in \mathrm{T}_x\, Q : c'(x)d \in \mathrm{T}_{c(x)}\, G\}. \tag{4.139}$$

On dit que c *est qualifiée en* $x \in X_{QG}$ *pour représenter* X_{QG} si

$$\mathrm{T}_x\, X = \mathrm{T}'_x\, X \qquad \text{et} \qquad (\mathrm{T}_x\, Q)^+ + c'(x)^* \left[(\mathrm{T}_{c(x)}\, G)^+\right] \text{ est fermé.} \tag{4.140}$$

2. Montrez que si x_* est un minimum local de (P_{QG}) et si c est qualifiée en x_* pour représenter X_{QG}, alors il existe $\lambda_* \in \mathbb{F}$ tel que

$$\nabla f(x_*) + c(x_*)^* \lambda_* \in (\mathrm{T}_{x_*}\, Q)^+ \qquad \text{et} \qquad \lambda_* \in \mathrm{N}_{c(x_*)} G. \tag{4.141}$$

3. Montrez que c est qualifiée en $x_0 \in X_{QG}$ pour représenter X_{QG} si

$$0 \in \operatorname{int}\Big(c(x_0) + c'(x_0)(Q - x_0) - G\Big). \tag{4.142}$$

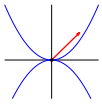
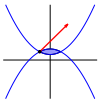
4. Montrer que la condition (4.142) est equivalente au fait qu'il existe une constante $\mu \geqslant 0$, telle que pour tout $(x, q, g) \in \mathbb{E} \times \mathbb{E} \times \mathbb{F}$ proche de $(x_0, 0, 0)$, on a

$$\operatorname{dist}(x, X_{QG}^{qg}) \leqslant \mu\Big(\operatorname{dist}(c(x), g+G) + \operatorname{dist}(x, q+Q)\Big), \tag{4.143}$$

où $X_{QG}^{qg} := \{x \in q+Q : c(x) \in g+G\}$.

4.24. *Fonction valeur d'un problème sans qualification de contraintes.* On considère le problème d'optimisation dans $\mathbb{R}^2$ à gauche ci-dessous, ainsi que le problème perturbé par $p \in \mathbb{R}^2$ à droite :

$$\begin{cases} \inf\ \alpha x_1 + x_2 \\ x_1^2 - x_2 \leqslant 0 \\ x_1^2 + x_2 \leqslant 0 \end{cases} \qquad \text{et} \qquad \begin{cases} \inf\ \alpha x_1 + x_2 \\ x_1^2 - x_2 + p_1 \leqslant 0 \\ x_1^2 + x_2 + p_2 \leqslant 0, \end{cases}$$

où $\alpha \in \mathbb{R}$ est un paramètre. On note $v(p)$ la valeur optimale du problème de droite (v est donc la fonction valeur du problème de gauche, voir la section 4.6.1).

1) Montrez que v est convexe et prend pour valeur

$$v(p) = \begin{cases} p_1 - \alpha^2/4 & \text{si } p_1 + p_2 \leqslant -\alpha^2/2 \\ (p_1 - p_2)/2 - |\alpha|\sqrt{|p_1 + p_2|/2} & \text{si } p_1 + p_2 \in\,]-\alpha^2/2, 0[\\ p_1 & \text{si } p_1 + p_2 = 0 \\ +\infty & \text{si } p_1 + p_2 > 0. \end{cases}$$

2) En déduire que v est sous-différentiable en zéro si, et seulement si, $\alpha = 0$.
3) On suppose que $\alpha = 0$. Montrez que $\partial v(0) = \{p^* \in \mathbb{R}^2 : 1 \leqslant p_1^* = p_2^* + 1\}$ (pour le produit scalaire euclidien), et que cet ensemble est bien l'ensemble des multiplicateurs optimaux de KKT du problème non perturbé.

4.25. *Minimisation d'un maximum de fonctions* [530 ; exemple 3.31]. Soit $\{f_i\}_{i\in[1:m]}$ une famille formée de $m \in \mathbb{N}^*$ fonctions $f_i : \mathbb{E} \to \mathbb{R}$ différentiables définies sur une espace vectoriel $\mathbb{E}$. On considère le problème de minimiser la fonction $f : \mathbb{E} \to \mathbb{R}$ définie en $x \in \mathbb{E}$ par

$$f(x) = \max_{i\in[1:m]} f_i(x).$$

Montrez que si $\bar{x}$ est un minimum local de f et $I := \{i \in [1:m] : f_i(\bar{x}) = \inf_{x\in\mathbb{E}} f(x)\}$, alors il existe $\alpha_I \in \mathbb{R}^{|I|}$ tels que

$$\alpha_I \geqslant 0, \qquad \sum_{i\in I} \alpha_i = 1 \qquad \text{et} \qquad \sum_{i\in I} \alpha_i \nabla f_i(\bar{x}) = 0.$$

4.26. *Exemple d'utilisation des conditions du second ordre.* On considère le problème en $x \in \mathbb{R}^2$:

$$\begin{cases} \min\ -\frac{1}{2}(x_1^2 + x_2^2) \\ x_2 \geqslant x_1^2 - 1 \\ x_1 \geqslant 0. \end{cases}$$

Déterminez analytiquement les points stationnaires de ce problème et, parmi eux, ceux qui sont des minima/maxima locaux/globaux (stricts ou pas).

5 Prolégomènes à l'algorithmique

Les quelques nouveaux qui réussissent à passer le font tout au moins en partie parce qu'ils ont de la chance, mais surtout parce qu'ils font ce qu'il faut : ils excellent à construire des machines à survivre.

R. Dawkins (1976), *Le Gène Égoïste*. [157]

Avec ce chapitre nous commençons l'étude des algorithmes de résolution des problèmes d'optimisation. Quelques réponses aux questions communes à tout algorithme, à leur implémentation ou utilisation, ont été rassemblées ici. Celles relatives à la comparaison des algorithmes en fonction de la vitesse de convergence des suites qu'ils génèrent ou du nombre d'opérations qu'ils requièrent sont respectivement abordées aux sections 5.1 et 5.2. La section 5.3 fait quelques remarques sur le (pré-)conditionnement des problèmes d'optimisation, c'est-à-dire sur les moyens de les transformer analytiquement de manière à les rendre plus faciles à résoudre numériquement et moins sensibles aux erreurs d'arrondi, lesquelles sont inévitablement commises en calcul flottant. Le calcul d'un gradient est une opération fondamentale, mécanique et fastidieuse, de l'optimisation numérique. La méthode de l'état adjoint, qui est souvent la bonne méthode à utiliser dans les problèmes de commande optimale, est exposée à la section 5.4. Il ne faudra jamais perdre de vue qu'il existe une méthode générale, théoriquement efficace, pour évaluer un gradient : la différentiation automatique. Nous détaillons les modes de différentiation direct et inverse à la section 5.5. Nous concluons avec la section 5.6 qui discute de deux aspects des codes d'optimisation : des différentes structures que peut avoir un code utilisant un module d'optimisation et des profils de performance qui apportent une image graphique de l'efficacité relative de solveurs sur un banc d'essai de problèmes-tests.

5.1 Vitesse de convergence des suites

In those days, we needed faster convergence to get results in the few hours between expected failures of the Avidac's large roomful of a few thousand bytes of temperamental electrostatic memory.

W.C. Davidon, dans la nouvelle introduction de son article fondateur des méthodes de quasi-Newton, écrit en 1959, mais qui ne fut publié qu'en 1991, lors de la parution du premier numéro de la revue *SIAM Journal on Optimization* [155, 156].

Les algorithmes d'optimisation génèrent des suites $\{x_k\}_{k\geqslant 1}$ convergeant vers une solution x_* du problème (dans les bons cas !). Ils procèdent ainsi parce que, dans la plupart des problèmes que nous allons rencontrer, il n'est pas possible de calculer une solution en un nombre fini d'opérations arithmétiques. Les éléments x_k de la suite générée sont appelés des *itérés*. Étant donné un itéré x_k, on calcule l'itéré suivant x_{k+1} de telle sorte que celui-ci soit, si possible, plus proche de la solution cherchée x_* que ne l'est x_k.

Se pose alors le problème de comparer l'efficacité des algorithmes par l'examen des suites qu'ils génèrent, dans le pire des cas. La notion de vitesse de convergence permet de qualifier le comportement asymptotique (pour x_k proche de x_*) d'une suite. On distingue deux classes de vitesses de convergence : les vitesses de convergence en quotient (section 5.1.1) et en racine (section 5.1.2). Ces notions sont motivées par des considérations pratiques : on peut en effet démontrer que les suites générées par les algorithmes étudiés dans cet ouvrage ont de telles vitesses de convergence.

5.1.1 Vitesses de convergence en quotient

Dans cette section, on cherche à qualifier la vitesse de convergence d'une suite $\{x_k\}$ d'un espace normé $\mathbb{E}$, de norme notée $\|\cdot\|$, en comparant la norme de l'*erreur* $x_k - x_*$ de deux itérés successifs, c'est-à-dire en comparant $\|x_k - x_*\|$ et $\|x_{k+1} - x_*\|$. On supposera toujours que l'erreur ne s'annule pas ($x_k \neq x_*$, pour tout indice k). Cette hypothèse est raisonnable, car dans les algorithmes bien conçus, dès que $x_k = x_*$, la suite devient stationnaire après x_k (tous les itérés suivants sont égaux à x_*) et il n'y a plus de sens à parler de vitesse de convergence. On s'intéresse donc au *quotient*

$$\frac{\|x_{k+1} - x_*\|}{\|x_k - x_*\|^\alpha}, \tag{5.1}$$

où α est un entier non nul. L'intérêt pour ce quotient provient du fait qu'on peut souvent l'estimer en faisant un développement de Taylor autour de x_* des fonctions définissant le problème que l'on cherche à résoudre et dont x_* est solution.

Les noms des vitesses de convergence de cette section seront préfixés par « q- » pour rappeler qu'il s'agit de vitesse de convergence en quotient.

Numériquement, plus rapide est la convergence, plus vite augmente le *nombre de chiffres significatifs corrects* de x_k, c'est-à-dire de chiffres significatifs identiques à ceux de x_*. Donnons une définition plus précise de cette notion. Si x_k est un vecteur, on ne peut pas définir par un scalaire la correction des chiffres significatifs de toutes

ses composantes, mais on peut le faire *en moyenne* au sens de la norme sur $\mathbb{E}$. On suppose que $x_* \neq 0$ car on ne peut définir ce que sont les chiffres significatifs de zéro. Si $\|x_k - x_*\|/\|x_*\|$ vaut 10^{-4}, on dira que x_k a 4 chiffres significatifs corrects. Ceci conduit à la définition suivante.

Définition 5.1 (nombre de chiffres significatifs corrects) Le *nombre de chiffres significatifs corrects* de x_k par rapport à $x_* \neq 0$ est le nombre réel défini par

$$\sigma_k := -\log_{10} \frac{\|x_k - x_*\|}{\|x_*\|}. \qquad \square$$

Lorsque $x_* \neq 0$, on peut exprimer les vitesses de convergence en quotient en utilisant σ_k plutôt que le quotient (5.1), ce que nous ferons.

Il est parfois intéressant de vérifier numériquement que les suites générées par un algorithme ont bien la vitesse de convergence attendue, celle démontrée théoriquement pour l'algorithme considéré. Bien sûr, c'est une manière de vérifier que l'algorithme est bien implémenté, mais il y a une autre motivation. Par exemple, sous certaines hypothèses de régularité, on verra que l'algorithme de Newton converge q-quadratiquement (théorème 10.2) ; cet algorithme procède par linéarisation de la fonction qu'il cherche à annuler ; vérifier que la convergence des suites générées est bien q-quadratique est alors une indication sur la correction du calcul des dérivées.

Comme on ne connaît pas la solution, on ne peut vérifier la vitesse de convergence en quotient attendue par l'examen du quotient (5.1), qu'en résolvant deux fois le problème ; la première fois pour calculer une approximation précise de la solution x_*, la seconde pour faire l'examen des quotients susmentionnés ; c'est dommage. On pourrait aussi mémoriser tous les itérés et considérer que le dernier itéré est la solution, mais en grande dimension, cette une opération est bien trop gourmande en place mémoire. On peut éviter cette double résolution ou la mémorisation des itérés si l'on arrive à exprimer la vitesse de convergence en termes d'une quantité dont la limite est connue, typiquement nulle. Il en est ainsi si l'algorithme cherche à annuler une fonction

$$F : \mathbb{E} \to \mathbb{E},$$

pourvu que

$$F(x_*) = 0 \qquad \text{et} \qquad F(x) \sim (x - x_*). \tag{5.2}$$

L'écriture $F(x) \sim (x - x_*)$ signifie ici que, pour une norme $|||\cdot|||$ sur $\mathbb{E}$,

$$\exists\, C \geqslant 1,\ \forall\, x \text{ voisin de } x_* :\ C^{-1}|||F(x)||| \leqslant |||x - x_*||| \leqslant C|||F(x)|||. \tag{5.3}$$

En dimension finie, cette propriété ne dépend pas du choix de la norme $|||\cdot|||$ et est vérifiée dans les codnitions énoncées dans la proposition suivante.

Proposition 5.2 (équivalence asymptotique) *Si F est différentiable en x_*, si $F(x_*) = 0$ et si $F'(x_*)$ est inversible, alors (5.3) a lieu.*

Démonstration. Par la différentiablité de F en x_* et la nullité de $F(x_*)$, on a

$$F(x) = F'(x_*) \cdot (x - x_*) + o(\|x - x_*\|).$$

On en déduit directement que $\|F(x)\| = O(\|x - x_*\|)$ pour x voisin de x_*. Inversement, grâce à l'inversibilité de $F'(x_*)$, on a $x - x_* = F'(x_*)^{-1}F(x) + o(\|x - x_*\|)$, d'où l'on déduit que $\|x - x_*\| = O(\|F(x)\|)$. □

Les vitesses de convergence d'une suite $\{x_k\}$ présentées ci-dessous seront également exprimées en termes du logarithme de $\|F(x_k)\|$ pour une fonction F vérifiant (5.2) et une autre norme $\|\cdot\|$, de manière à permettre une vérification numérique de cette convergence.

La mise en évidence de l'équivalence (5.2) n'est pas toujours aussi aisée que dans la démonstration de la proposition 5.2, en particulier lorsque F n'est pas différentiable en x_* dans le sens classique de Fréchet. C'est le cas, par exemple, pour la fonction que l'on cherche à annuler par l'algorithme newtonien en optimisation avec contraintes (chapitre 15). Pour certaines vitesses de convergence de $\{x_k\}$ vers x_*, on dispose d'un autre moyen de les mettre en évidence théoriquement ou numériquement sans la connnaissance de la limte x_*, à savoir par l'examen de la vitesse de convergence vers zéro de la suite $\{s_k\}$ des déplacements

$$s_k := x_{k+1} - x_k. \tag{5.4}$$

L'intérêt de cette suite est, en l'occurrence, de converger vers une limite connue, zéro, lorsque $x_k \to x_*$. L'estimation de la vitesse de convergence peut alors se faire au cours des itérations.

Nous rencontrerons essentiellement trois vitesses de convergence en quotient : les vitesses de convergence q-linéaire, q-superlinéaire et q-quadratique. Le préfixe « q » utilisé dans ces appellations, qui rappelle le mot « quotient », on l'a dit, est parfois omis.

Convergence q-linéaire

Définition 5.3 (convergence q-linéaire) On dit qu'une suite $\{x_k\} \subseteq \mathbb{E}$ converge *q-linéairement* vers x_* s'il existe une norme $\|\cdot\|$, un scalaire $\tau \in [0, 1[$ et un indice $k_1 \geqslant 1$, tels que pour tout $k \geqslant k_1$ on ait

$$\|x_{k+1} - x_*\| \leqslant \tau \|x_k - x_*\|. \tag{5.5}$$

Le paramètre τ est appelé le *taux de convergence linéaire*. □

Il faut donc que la norme de l'erreur décroisse strictement à chaque itération à partir d'une certaine itération, avec un taux de convergence τ strictement plus petit que 1. Cette propriété dépend du choix de la norme que l'on utilise pour mesurer l'erreur, car l'estimation (5.5) peut être vraie pour une norme et, malgré l'équivalence des normes, peut ne plus être vérifiée avec $\tau < 1$ pour une autre norme.

Le résultat suivant fait le lien entre la convergence q-linéaire et le nombre σ_k de chiffres significatifs corrects des itérés (définition 5.1). Sa démonstration est proposée à l'exercice 5.2.

Proposition 5.4 (convergence q-linéaire en termes de σ_k) *La suite $\{x_k\}_{k\geqslant 1}$ converge q-linéairement vers $x_* \neq 0$ pour une norme $\|\cdot\|$ si, et seulement si, il existe une constante $\sigma > 0$ et un indice $k_1 \geqslant 1$ tels que pour tout $k \geqslant k_1$ on ait*

$$\sigma_{k+1} \geqslant \sigma_k + \sigma,$$

où σ_k est défini avec la norme $\|\cdot\|$.

Il est difficile d'établir un lien entre la convergence linéaire de la suite $\{x_k\}$ et celle de la suite $\{F(x_k)\}$ où la fonction F vérifie (5.2), à cause de la constante C intervenant dans (5.3) et du taux de convergence τ qui doit être strictement inférieur à 1. On trouve la même difficulté pour établir une équivalence entre la convergence linéaire de $\{x_k\}$ et celle des déplacements $s_k := x_{k+1} - x_k$.

En général, on s'attend à ce qu'un algorithme d'optimisation différentiable calcule des suites convergeant plus rapidement que q-linéairement.

Exemple d'algorithme générant des suites q-linéairement convergentes

- L'algorithme du gradient pour minimiser une fonction quadratique strictement convexe (proposition 7.2).

Convergence q-superlinéaire

Définition 5.5 (convergence q-superlinéaire) On dit qu'une suite $\{x_k\} \subseteq \mathbb{E}$ converge *q-superlinéairement* vers x_* si pour tout $\tau > 0$, il existe un indice $k_\tau \geqslant 1$, tels que pour tout $k \geqslant k_\tau$ on ait

$$\|x_{k+1} - x_*\| \leqslant \tau \|x_k - x_*\|.$$

Il revient au même de dire que $\|x_{k+1} - x_*\|/\|x_k - x_*\| \to 0$ ou encore que

$$\|x_{k+1} - x_*\| = o(\|x_k - x_*\|). \qquad \square$$

Cette propriété est indépendante du choix de la norme. Clairement, une suite convergeant q-superlinéairement converge q-linéairement.

Le résultat suivant fait le lien entre la convergence q-superlinéaire et le nombre σ_k de chiffres significatifs corrects des itérés (définition 5.1). Sa démonstration est proposée à l'exercice 5.2.

Proposition 5.6 (convergence q-superlinéaire en termes de σ_k) *La suite $\{x_k\}_{k\geqslant 1}$ converge q-superlinéairement vers $x_* \neq 0$ pour une norme $\|\cdot\|$ si, et seulement si,*

$$\sigma_{k+1} - \sigma_k \to +\infty,$$

où σ_k est défini avec la norme $\|\cdot\|$.

Voici une manière de vérifier numériquement la convergence q-superlinéaire d'une suite par l'intermédiaire d'une fonction s'annulant au point limite. La démonstration de la proposition est proposée à l'exercice 5.3.

Proposition 5.7 (convergence q-superlinéaire en termes *F*) *Soit* $F : \mathbb{E} \to \mathbb{E}$ *une fonction vérifiant* (5.2)*. La suite* $\{x_k\}$ *converge q-superlinéairement vers* x_* *si, et seulement si,*

$$\log \|F(x_{k+1})\| - \log \|F(x_k)\| \to -\infty.$$

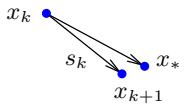
On peut le dire autrement : la suite $\{x_k\}$ converge q-superlinéairement si, et seulement si, le tracé $k \mapsto \log \|F(x_k)\|$ a une « pente » estimée par la différence $\log \|F(x_{k+1})\| - \log \|F(x_k)\|$ qui tend vers $-\infty$.

Pour des suites $\{z_k\}$ et $\{z'_k\}$ dans un espace normé, convergeant vers zéro, on utilisera occasionnellement la notation

$$z_k \sim z'_k \tag{5.6}$$

pour dire qu'il existe une constante $C \geqslant 1$ telle que, pour tout indice k, on a $C^{-1}\|z_k\| \leqslant \|z'_k\| \leqslant C\|z_k\|$, ce que l'on peut aussi écrire $z_k = O(\|z'_k\|)$ et $z'_k = O(\|z_k\|)$. En mots simples : $\{z_k\}$ et $\{z'_k\}$ convergent vers zéro à la même « vitesse ».

Si $\{x_k\}$ converge superlinéairement vers x_*, asymptotiquement, la norme de l'erreur à l'itération suivante $\|x_{k+1} - x_*\|$ est très petite devant celle de l'itération courante $\|x_k - x_*\|$, ce qui implique que le déplacement $s_k := x_{k+1} - x_k$ est proche de $x_* - x_k$ (voir le dessin ci-contre). Ceci a pour conséquence que $(x_k - x_*) \sim s_k$, comme le montre le lemme ci-dessous.



Lemme 5.8 (suites équivalentes) *Si la suite* $\{x_k\}$ *converge superlinéairement vers* x_**, alors* $(x_k - x_*) \sim s_k$*.*

Démonstration. On a

$$s_k = (x_{k+1} - x_*) - (x_k - x_*) = -(x_k - x_*) + o(\|x_k - x_*\|).$$

On en déduit le résultat. □

Cette équivalence entre la suite des erreurs $\{x_k - x_*\}$ et celle des déplacements $\{s_k\}$ permet de montrer que celles-ci convergent superlinéairement vers leur limite respective simultanément.

Proposition 5.9 (convergence q-superlinéaire en termes de s_k)
1) *Si une suite $\{x_k\}$ converge superlinéairement vers x_*, alors la suite des déplacements $\{s_k\}$, définie par (5.4), converge superlinéairement vers zéro.*
2) *Inversement, si, pour une suite donnée $\{x_k\}$, la suite des déplacements $\{s_k\}$ converge superlinéairement vers zéro, alors $\{x_k\}$ est une suite de Cauchy qui converge superlinéairement vers sa limite.*

DÉMONSTRATION. 1) D'après le lemme 5.8, la convergence superlinéaire de x_k vers x_* implique que $s_k \sim (x_k - x_*)$. On déduit alors de $x_{k+1} - x_* = o(\|x_k - x_*\|)$ que $s_{k+1} = o(\|s_k\|)$, qui est la marque de la convergence superlinéaire de s_k vers zéro.

2) On peut supposer que $s_k \neq 0$ pour tout $k \geqslant 1$ (si $s_k = 0$ avec k assez grand, la convergence superlinéaire de $\{s_k\}$ implique alors que $s_i = 0$ pour tout $i \geqslant k$ et donc que la suite $\{x_k\}$ est stationnaire, ce qui implique immédiatement le résultat). Du fait de la convergence superlinéaire de $\{s_k\}$ vers zéro et du fait que l'on s'intéresse au comportement asymptotique des suites, on peut aussi supposer que, pour tout $k \geqslant 1$, on a

$$\varepsilon_k := \sup_{i \geqslant k} \frac{\|s_{i+1}\|}{\|s_i\|} < 1. \tag{5.7}$$

Observons que la suite $\{\varepsilon_k\}$ tend vers zéro en décroissant.

Commençons par montrer que, pour $k \geqslant 1$ et $i \geqslant 1$, on a

$$\|s_{k+i}\| \leqslant \varepsilon_k^i \|s_k\|. \tag{5.8}$$

Fixons $k \geqslant 1$ et montrons l'inégalité par récurrence sur i. Par (5.7), l'inégalité a lieu pour $i = 1$. Supposons maintenant qu'elle ait lieu pour un $i \geqslant 1$ et montrons la pour $i + 1$. Par (5.7), la décroissance de $\{\varepsilon_k\}$ et l'hypothèse de récurrence, nous pouvons conclure

$$\|s_{k+i+1}\| \leqslant \varepsilon_{k+i}\|s_{k+i}\| \leqslant \varepsilon_k \|s_{k+i}\| \leqslant \varepsilon_k^{i+1}\|s_k\|.$$

Montrons à présent que $\{x_k\}$ est une suite de Cauchy, qui converge donc vers une limite que l'on note x_*, et que, pour tout $\varepsilon > 0$, il existe un k_ε tel que, pour tout $k \geqslant k_\varepsilon$, on a

$$\|x_k - x_*\| \leqslant (1 + \varepsilon)\|s_k\|. \tag{5.9}$$

Soit $\varepsilon > 0$. Pour $l > k$ et k suffisamment grand pour que $1/(1 - \varepsilon_k) \leqslant 1 + \varepsilon$, on a

$$\begin{aligned}
\|x_k - x_l\| &= \|\textstyle\sum_{i=0}^{l-k-1} s_{k+i}\| \\
&\leqslant \textstyle\sum_{i=0}^{l-k-1} \|s_{k+i}\| \\
&\leqslant (\textstyle\sum_{i=0}^{l-k-1} \varepsilon_k^i)\|s_k\| \qquad [(5.8)] \\
&\leqslant \tfrac{1}{1-\varepsilon_k}\|s_k\| \qquad [\varepsilon_k < 1] \\
&\leqslant (1 + \varepsilon)\|s_k\| \qquad [1/(1 - \varepsilon_k) \leqslant 1 + \varepsilon].
\end{aligned}$$

Ceci montre que $\{x_k\}$ est une suite de Cauchy, dont la limite est notée x_*. En faisant tendre $l \to \infty$ dans l'inégalité ci-dessus, on obtient (5.9).

Montrons à présent que, pour tout $\varepsilon > 0$, il existe un k_ε tel que, pour tout $k \geqslant k_\varepsilon$, on a

$$\|x_k - x_*\| \geqslant (1-\varepsilon)\|s_k\|. \tag{5.10}$$

Soit $\varepsilon > 0$. Pour $l > k$ et k suffisamment grand pour que $\varepsilon_k/(1-\varepsilon_k) \leqslant \varepsilon$, on a

$$\begin{aligned}
\|x_k - x_l\| &= \|\textstyle\sum_{i=0}^{l-k-1} s_{k+i}\| \\
&\geqslant \|s_k\| - \|\textstyle\sum_{i=1}^{l-k-1} s_{k+i}\| \\
&\geqslant \|s_k\| - \textstyle\sum_{i=1}^{l-k-1} \|s_{k+i}\| \\
&\geqslant (1 - \textstyle\sum_{i=1}^{l-k-1} \varepsilon_k^i)\|s_k\| \qquad [(5.8)] \\
&\geqslant (1 - \tfrac{\varepsilon_k}{1-\varepsilon_k})\|s_k\| \qquad [\varepsilon_k < 1] \\
&\geqslant (1-\varepsilon)\|s_k\| \qquad [\varepsilon_k/(1-\varepsilon_k) \leqslant \varepsilon].
\end{aligned}$$

En faisant tendre $l \to \infty$, on obtient (5.10).

Les inégalités (5.9) et (5.10), avec $\varepsilon \in]0,1[$, montrent que $s_k \sim (x_k - x_*)$. Alors la convergence superlinéaire de $\{s_k\}$ vers zéro, $\|s_{k+1}\| = o(\|s_k\|)$, implique celle de $\{x_k\}$ vers x_*. □

Exemple d'algorithme générant des suites q-superlinéairement convergentes

- Les algorithmes de quasi-Newton en optimisation (proposition **??**) ou pour résoudre un système d'équations non linéaires.

Convergence q-quadratique

Définition 5.10 (convergence q-quadratique) On dit qu'une suite $\{x_k\} \subseteq \mathbb{E}$ converge *q-quadratiquement* vers x_* s'il existe une constante $C \geqslant 0$, telle que pour tout $k \geqslant 1$ on ait

$$\|x_{k+1} - x_*\| \leqslant C\|x_k - x_*\|^2. \qquad \square$$

Le quotient des erreurs successives $\|x_{k+1} - x_*\|/\|x_k - x_*\| \leqslant C\|x_k - x_*\|$ d'une suite quadratiquement convergente tend vers zéro ; une telle suite converge donc q-superlinéairement.

Le résultat suivant fait le lien entre la convergence q-quadratique et le nombre σ_k de chiffres significatifs corrects des itérés (définition 5.1). Sa démonstration est proposée à l'exercice 5.2.

Proposition 5.11 (convergence q-quadratique en termes de σ_k) *La suite $\{x_k\}_{k\geqslant 1}$ converge q-quadratiquement vers $x_* \neq 0$ pour une norme $\|\cdot\|$ si, et seulement si, il existe une constante $C \in \mathbb{R}$ telle que*

$$\sigma_{k+1} \geqslant 2\sigma_k + C,$$

où σ_k est défini avec la norme $\|\cdot\|$. Dans ce cas

$$\liminf_{k\to\infty} \frac{\sigma_{k+1}}{\sigma_k} \geqslant 2. \tag{5.11}$$

Remarque 5.12 (odublement du nombre de chiffres significatifs corrects à chaque itération) De manière imagée, on peut exprimer verbalement l'inégalité (5.11) en disant qu'« une suite $\{x_k\}$ convergeant q-quadratiquement a des éléments x_k dont le nombre de chiffres significatifs corrects double à chaque itération *asymptotiquement* ». C'est une convergence très rapide puisque l'on atteint alors très vite le nombre maximal de chiffres significatifs qu'un ordinateur donné peut représenter (15..16 pour des nombres en double précision pour la *norme IEEE 754*). □

Voici à présent deux manières de vérifier numériquement la convergence q-quadratique d'une suite, sans connaître le point limite x_*. La première méthode (proposition 5.13) utilise une fonction s'annulant au point limite et vérifiant (5.2), ce qui requiert l'existence d'une telle fonction (voir la proposition 5.2). La seconde méthode (proposition 5.14) utilise la suite des déplacements $\{s_k\}$, où $s_k := x_{k+1} - x_k$. Les démonstrations de ces propositions sont proposées aux exercices 5.3 et 5.4.

Proposition 5.13 (convergence q-quadratique en termes F) *Soit $F : \mathbb{E} \to \mathbb{E}$ une fonction vérifiant (5.2). La suite $\{x_k\}$ converge q-quadratiquement vers x_* si, et seulement si, il existe une constante $C \in \mathbb{R}$ telle que*

$$\log \|F(x_{k+1}\| \leqslant 2 \log \|F(x_k)\| + C.$$

Dans ce cas

$$\liminf_{k\to\infty} \frac{\log \|F(x_{k+1})\|}{\log \|F(x_k)\|} \geqslant 2.$$

Proposition 5.14 (convergence q-quadratique en termes de s_k)
1) *Si une suite $\{x_k\}$ converge quadratiquement vers x_*, alors la suite des déplacements $\{s_k\}$, définie par (5.4), converge quadratiquement vers zéro.*
2) *Inversement, si, pour une suite donnée $\{x_k\}$, la suite des déplacements $\{s_k\}$ converge quadratiquement vers zéro, alors $\{x_k\}$ est une suite de Cauchy qui converge quadratiquement vers sa limite.*

Exemple d'algorithme générant des suites q-quadratiquement convergentes

- L'algorithme de Newton en optimisation ou pour résoudre un système d'équations non linéaires (théorèmes 10.2 et 10.3).

Typiquement, ce sont donc les algorithmes qui procèdent par linéarisation des équations à résoudre qui génèrent des suites convergeant q-quadratiquement.

Convergence avec q-ordre plus élevé

Définition 5.15 (convergence avec q-ordre) On dit qu'une suite $\{x_k\} \subseteq \mathbb{E}$ converge avec un *q-ordre* $\alpha \in \mathbb{R}_{++}$ vers x_* s'il existe une constante $C \geqslant 0$, telle que pour tout $k \geqslant 1$ on ait

$$\|x_{k+1} - x_*\| \leqslant C\|x_k - x_*\|^\alpha.$$

On dit que la convergence est *q-cubique* si $\alpha = 3$, qu'elle est *q-quartique* si $\alpha = 4$ et qu'elle est *q-quintique* si $\alpha = 5$. □

Clairement, la convergence avec un q-ordre α implique la convergence avec un q-ordre $\alpha' \in]0, \alpha]$.

Une vitesse de convergence avec un q-ordre $\alpha > 2$ est garantie par certains algorithmes, mais ceux-ci ne se rencontrent pas souvent. D'ailleurs, on cherche rarement à construire un algorithme assurant une telle vitesse de convergence à ses suites, parce que la convergence q-quadratique ($\alpha = 2$) permet d'avoir une solution très précise en très peu d'itérations (remarque 5.12) et requiert déjà une itération qui peut être couteuse en place mémoire et temps de calcul lorsque la dimension du problème est grande.

5.1.2 Vitesses de convergence en racine ▲

On dit que $\{x_k\}$ converge *r-linéairement* vers x_* s'il existe un scalaire $\tau \in [0, 1[$ tel que

$$\limsup_{k\to\infty} \|x_k - x_*\|^{\frac{1}{k}} \leqslant \tau. \tag{5.12}$$

Le scalaire τ est appelé le *taux de convergence* r-linéaire.

Contrairement à la convergence q-linéaire, la notion de convergence r-linéaire ne dépend pas de la norme utilisée pour mesurer l'erreur. En effet, si $|\cdot|$ est une autre norme, il existe une constante $C > 0$ telle que $C|\cdot| \leqslant \|\cdot\|$ (équivalence des normes). Comme $C^{1/k} \to 1$ lorsque $k \to \infty$, (5.12) implique que l'on a aussi $\limsup_{k\to\infty} |x_k - x_*|^{1/k} \leqslant \tau$.

D'autre part, une suite convergeant q-linéairement converge r-linéairement, comme cela se voit sur l'estimation $\|x_k - x_*\| \leqslant r^{k-k_1}\|x_{k_1} - x_*\|$ que l'on a pour les indices assez grand ($> k_1$) d'une suite q-linéairement convergente. Nous avons donc introduit une notion plus faible que la convergence q-linéaire. En fait la convergence r-linéaire est liée à la convergence q-linéaire d'une suite majorant la norme de l'erreur $\|x_k - x_*\|$. C'est parce que l'on peut parfois montrer la convergence q-linéaire d'un majorant de l'erreur que cette notion est utile.

Proposition 5.16 *Soit $\{x_k\}$ une suite convergeant vers x_*. La convergence est r-linéaire si, et seulement si, il existe une suite $\{\beta_k\}$ convergeant q-linéairement vers 0 telle que $\|x_k - x_*\| \leqslant \beta_k$ pour tout $k \geqslant 1$.*

Démonstration. Supposons que $\{x_k\}$ vérifie (5.12) et choisissons $r_1 \in]r, 1[$. Alors il existe un indice k_1 tel que pour tout $k \geqslant k_1$: $\|x_k - x_*\|^{1/k} \leqslant r_1$. En définissant $\beta_k := \|x_k - x_*\|$ si $k < k_1$ et $\beta_k := r_1^k$ si $k \geqslant k_1$, on a $\|x_k - x_*\| \leqslant \beta_k$ et la suite $\{\beta_k\}$ converge q-linéairement vers 0.

Inversement, si $\|x_k - x_*\| \leqslant \beta_k$, avec $\{\beta_k\}$ convergeant q-linéairement vers 0, on a pour k plus grand qu'un indice k_1 et pour un nombre $r \in [0, 1[$:

$$\|x_k - x_*\|^{\frac{1}{k}} \leqslant \beta_k^{\frac{1}{k}} \leqslant r^{\frac{k-k_1}{k}} \beta_{k_1}^{\frac{1}{k}} \to r < 1.$$

□

5.2 Notions de complexité ▲

On peut trouver la solution de certains problèmes d'optimisation, ou une solution à une précision $\varepsilon > 0$ donnée, en un nombre fini d'étapes. Il en va ainsi des problèmes d'optimisation linéaire ou plus généralement d'optimisation quadratique convexe et même de certains problèmes d'optimisation convexe ou non convexe. Dans ce cas, la notion de vitesse de convergence introduite à la section 5.1 n'est plus très pertinente, puisque la suite générée est stationnaire à partir d'un certain indice. Pour apprécier la qualité des algorithmes de résolution de tels problèmes on préfère compter le nombre d'étapes qu'ils requièrent. Une étape de l'algorithme peut être une itération (concept un peu flou), une opération arithmétique, une opération sur un bit, *etc.* Il faudra toujours bien préciser cet unité élémentaire du calcul. La *théorie de la complexité* s'intéresse à ces questions. Dans cette section, nous passons en revue les notions de cette théorie les plus souvent rencontrées en optimisation. Des présentations allant plus en profondeur sont données par Garey et Johnson [229] et Papadimitriou et Steiglitz [473]. Nous nous plaçons ici dans l'esprit du livre de Vavasis [606] qui présente cette théorie en vue de son application à l'optimisation.

5.2.1 Famille de problèmes, machine de Turing

Dans la théorie de la complexité, on regroupe les problèmes en familles dans le but de pouvoir comparer l'efficacité des algorithmes en fonction de la « taille » des problèmes d'une même famille. On parlera par exemple de la *famille* OL *des problèmes d'optimisation linéaire* (ensemble des problèmes qui consistent à minimiser une fonction linéaire sur un polyèdre convexe, voir les chapitres 17 et 18), de la *famille* OQ *des problèmes d'optimisation quadratique* (ensemble des problèmes qui consistent à minimiser une fonction quadratique sur un polyèdre convexe, voir le chapitre **??**), *etc.* Un problème d'une famille est spécifié par un jeu de données, dont la structure dépend de la famille considérée et dont la représentation ou l'*encodage* doit être précisé.

Étant donné la diversité des problèmes et la variété des questions que l'on peut se poser sur un problème donné, les spécialistes de la théorie de la complexité ont ressenti le besoin de décrire plus formellement ce qu'est une famille de problèmes. Un problème d'une famille doit pouvoir être décrit par un nombre fini de caractères d'un alphabet donné. Un problème ainsi décrit est appelé une *instance* de la famille de problèmes. De manière formelle, une *famille de problèmes* est alors une application

$$F : I \to S$$

entre la collection I des instances de la famille et un certain ensemble S qui pourra dépendre de la famille considérée et de ce que l'on considère comme une solution du problème. En réalité, « évaluer F en i » signifie le plus souvent « résoudre le problème i de la famille ». Considérons par exemple, la famille OL des problèmes d'optimisation linéaire.

Famille 5.17 (OL) On se donne deux entiers $n \geqslant 1$ et $m \geqslant 0$, $c \in \mathbb{Q}^n$, $A \in \mathbb{Q}^{m \times n}$ et $b \in \mathbb{Q}^m$. On considère le problème d'optimisation

$$\begin{cases} \inf\ c^{\mathsf{T}}x \\ Ax = b \\ x \geqslant 0. \end{cases}$$

Une instance i de la famille OL pourra prendre la forme d'un quintuplet $(n, m, c, A, b) \in \mathbb{N}^* \times \mathbb{N} \times \mathbb{Q}^n \times \mathbb{Q}^{mn} \times \mathbb{Q}^m$. L'ensemble S pourra être $\mathbb{Q} \cup \{-\infty, +\infty\}$, si l'on s'intéresse à la valeur optimale $F(i)$ du problème et au fait qu'il peut ne pas être réalisable ($F(i) = +\infty$) ou ne pas être borné ($F(i) = -\infty$). Il pourra aussi être $\mathbb{Q}^n \cup \{-\infty, +\infty\}$, si l'on s'intéresse à la solution x du problème et au fait que ce problème peut ne pas être réalisable ou borné. □

Dans la famille OL, on a choisi de décrire les problèmes en prenant l'ensemble des nombres rationnels $\mathbb{Q}$ comme alphabet. On aurait aussi pu choisir l'ensemble $\{0, 1\}$ des bits représentant ces rationnels. Les nombres réels $\mathbb{R}$ sont parfois utilisés [67], mais il n'est pas clair que l'on puisse alors bien faire la distinction entre la minimisation de problèmes d'optimisation quadratiques convexes et non convexes [606]. Quant aux nombres flottants, utilisés par les ordinateurs, ils semblent peu adaptés à l'étude de la complexité pour des raisons qui sont discutées dans [606 ; page 33 et chapitre 3].

Plus un problème est de grande dimension, plus il faut de caractères pour le décrire. Comme on est particulièrement intéressé par l'influence de la dimension d'un problème sur le nombre d'opérations à effectuer pour le résoudre, la longueur de l'instance jouera un rôle important. On la note

$$L(i) \quad \text{ou} \quad L,$$

selon qu'il est important ou pas de préciser l'instance considérée.

On dira qu'une famille $F : I \to S$ est formée de *problèmes de décision* si S est formé de deux éléments, traditionnellement dénommés `vrai` et `faux`. La théorie de la complexité traite principalement de ces familles de problèmes. Un problème d'optimisation « $\inf\{f(x) : x \in X\}$ » n'est cependant pas un problème de décision. Pour ce ramener à cette notion, on considère parfois la *version décisionnelle* d'un problème d'optimisation : étant donnés f, X et un seuil σ, il s'agit de déterminer si $\inf\{f(x) : x \in X\} \leqslant \sigma$.

Pour parler de la complexité d'un problème calculatoire et donc compter le nombre d'opérations à réaliser pour le résoudre, il faut spécifier la machine effectuant les calculs en précisant les opérations qui peuvent y être exécutées. Le nombre d'opérations à effectuer pour évaluer $F(i)$ sera d'autant plus grand que le nombre d'opérations licites est réduit. Par exemple, on peut considérer que l'addition de deux entiers est une unique opération ou que ce sont les opérations sur les bits représentant ces entiers qui sont les opérations élémentaires. Dans le premier cas, le nombre d'opérations élémentaires pour réaliser une addition est égal à un; dans le second cas, il dépendra de la grandeur ou longueur des entiers.

On a imaginé différents modèles de machines, mais celui qui semble le mieux refléter les capacités des ordinateurs actuels pour résoudre les problèmes que l'on se pose est, selon les spécialistes de ces questions, celui de la *Machine de Turing* [610 ; 2009]. Dans cette brève introduction, nous n'aurons jamais besoin de savoir ce que peut faire exactement une telle machine, ni quelles sont ses opérations licites, si bien que nous renvoyons le lecteur aux ouvrages cités au début de cette section pour une description précise de ce calculateur. On peut toutefois le décrire approximativement.

Il prend en entrée la description de l'instance i du problème à résoudre et il fournit en sortie la réponse $F(i)$. Le calculateur dispose d'une *mémoire linéaire* et d'un *pointeur* adressant une case-mémoire. Il agit en fonction de son *état*, qui fait partie d'une liste *finie*, et du *symbole* de la case-mémoire adressée par le pointeur. Il y a une *liste finie de règles*, définissant ce que le calculateur doit faire en fonction de chaque couple état-symbole possible, ce qui signifie que la machine est *déterministe*. Les actions du calculateur comprennent la modification du contenu de la case-mémoire sous le pointeur, le déplacement du pointeur d'un nombre fini de positions, la modification de ses états, *etc.*

5.2.2 Classes P et NP

On dit qu'une famille de problèmes $F : I \to S$ peut être *évaluée en temps polynomial* pour une machine de Turing de référence, si celle-ci peut évaluer cette fonction en toute instance $i \in I$, et que cette évaluation se fait en un nombre fini d'opérations qui est majoré par une fonction polynomiale de $L(i)$.

Définition 5.18 (classe P, problème polynomial) On dit qu'une famille de problèmes de décision $F : I \to \{\texttt{vrai}, \texttt{faux}\}$ est dans P, on dit aussi que ses problèmes sont *polynomiaux*, si F peut être évaluée en temps polynomial (pour une machine de Turing de référence). □

Beaucoup de familles de problèmes sont dans P et on peut voir là une première raison d'introduire cette classe. Très souvent le degré du polynôme est faible, si bien que ces problèmes peuvent être résolus pour des dimensions (ou plus précisément, des longueurs d'instance L) assez élevées. Les propriétés des polynômes rendent aussi cette notion très souple. Ainsi, parce qu'une composition de polynômes est un polynôme, un algorithme sera polynomial s'il se décompose en un nombre polynomial de modules pouvant chacun être exécuté en temps polynomial de degré constant.

Exemples 5.19 *Familles de problèmes* P : OL (optimisation linéaire, voir le chapitre 18), OQC (optimisation quadratique convexe). □

Pour un alphabet $\mathcal{A}$, on note $\mathcal{A}^*$ l'ensemble de toutes les chaînes finies formées de caractères de $\mathcal{A}$.

Définition 5.20 (classe NP) On dit qu'une famille de problèmes de décision $F : I \to \{\texttt{vrai}, \texttt{faux}\}$ est dans NP, on dit aussi que ses problèmes sont *NP*, s'il existe un polynôme p, un alphabet fini $\mathcal{A}$ et une fonction $G : I \times \mathcal{A}^* \to \{\texttt{vrai}, \texttt{faux}\}$ qui peut être évaluée en temps polynomial, tels que

(NP$_1$) $\forall i \in I$ tel que $F(i) = \texttt{vrai}$, $\exists j \in \mathcal{A}^*$ tel que $L(j) \leqslant p(L(i))$ et $G(i, j) = \texttt{vrai}$,
(NP$_2$) $\forall i \in I$ tel que $F(i) = \texttt{faux}$, $\forall j \in \mathcal{A}^*$, on a $G(i, j) = \texttt{faux}$. □

Cette définition compliquée mérite quelques éclaircissements.

1. On ne demande pas ici que F soit polynomial (F serait alors dans la classe P), mais que G le soit.

2. On peut interpréter l'évaluation de $F(i)$ comme la recherche d'une solution de l'instance $i \in I$, alors que l'évaluation de $G(i,j)$ peut être interprétée comme une vérification que $j \in \mathcal{A}^*$ est une solution de i : si i a une solution (c.-à-d., $F(i) = \texttt{vrai}$), disons $j \in \mathcal{A}^*$ (avec la condition $L(j) \leqslant p(L(i))$, de manière à ce que l'on ne puisse pas prendre pour j, hormis cas triviaux, la réunion de tous les candidats-solutions possibles), on peut vérifier en temps polynomial qu'il en est ainsi (c.-à-d., $G(i,j) = \texttt{vrai}$); si i n'a pas de solution (c.-à-d., $F(i) = \texttt{faux}$), on peut vérifier en temps polynomial qu'un candidat-solution arbitraire $j \in \mathcal{A}^*$ n'est en fait pas une solution (c.-à-d., $G(i,j) = \texttt{faux}$). De manière approximative (en négligeant (NP_2) en particulier), on peut dire qu'on ne demande pas que les instances des familles de la classe NP soient résolubles en temps polynomial, mais que l'on puisse vérifier en temps polynomial qu'une solution en est effectivement une (c'est (NP_1), cette description passe aussi sous silence la condition $L(j) \leqslant p(L(i))$).
3. On a
$$\mathsf{P} \subseteq \mathsf{NP}.$$
En effet, si F est polynomial, on peut prendre pour $\mathcal{A}$ un alphabet fini quelconque et définir G par $G(i,j) = F(i)$ pour tout $(i,j) \in I \times \mathcal{A}^*$.
4. On ne sait pas si $\mathsf{P} = \mathsf{NP}$ (beaucoup pense aujourd'hui que $\mathsf{P} \neq \mathsf{NP}$). Beaucoup de familles de problèmes sont dans NP, sans que l'on sache si elles sont dans P. Cela justifie l'introduction de cette notion.

La définition de la classe NP traite des familles de problèmes de décision. Il ne semble pas simple de l'étendre aux problèmes qui ne le soient pas.

5.2.3 Problèmes NP-complets et NP-ardus

On dit qu'une famille de problèmes de décision $F_1 : I_1 \to \{\texttt{vrai}, \texttt{faux}\}$ est *polynomialement réductible* en une autre famille de problèmes de décision $F_2 : I_2 \to \{\texttt{vrai}, \texttt{faux}\}$, s'il existe une fonction $\rho : I_1 \to I_2$ transformant *toutes* les instances de I_1 en instances de I_2 avec les deux propriétés suivantes :

(PR_1) $\forall i_i \in I_1$, $\rho(i_1)$ peut être évaluée en temps polynomial,
(PR_2) $\forall i_i \in I_1$, $F_1(i_1) = \texttt{vrai}$ si, et seulement si, $F_2(\rho(i_1)) = \texttt{vrai}$.

L'opérateur ρ est appelé une *réduction*. On peut montrer que cette relation est transitive : si F_1 est polynomialement réductible en F_2 et si F_2 est polynomialement réductible en F_3, alors F_1 est polynomialement réductible en F_3. D'autre part, il est clair que si F_1 est polynomialement réductible en F_2 et si $F_2 \in \mathsf{P}$, alors $F_1 \in \mathsf{P}$.

Définition 5.21 (classe NPC, problème NP-complet) On dit qu'une famille de problèmes de décision $F : I \to \{\texttt{vrai}, \texttt{faux}\}$ est dans NPC, on dit aussi que ses problèmes sont *NP-complets*, si elle est dans NP et si toute famille de problèmes dans NP est polynomialement réductible en F. □

En quelque sorte, c'est la sous-classe des familles de problèmes « les plus difficiles » de NP. Si on pouvait montrer qu'une famille de problèmes de NPC est dans P, alors on aurait $\mathsf{P} = \mathsf{NP}$. D'autre part, grâce à la relation de transitivité énoncée ci-dessus,

$$\left.\begin{array}{r} F_1 \in \mathsf{NPC} \\ F_2 \in \mathsf{NP} \\ F_1 \text{ polynomialement réductible en } F_2 \end{array}\right\} \quad \Longrightarrow \quad F_2 \in \mathsf{NPC}.$$

Pour montrer qu'une famille F est dans NPC, on utilise souvent cette dernière implication : on montre que $F \in \mathsf{NP}$ et on montre qu'une famille de NPC est polynomialement réductible en F. Il faut pour cela connaître des problèmes NPC. En voici quelques-uns.

Exemples 5.22 *Familles de problèmes* NP*-complets* : une version décisionnelle de OQ (proposition **??**), CLIQUE, SOUS-SOMME (famille 5.24). □

Famille 5.23 (SAT) Un *littéral* est une variable booléenne v (pouvant prendre la valeur 0 ou 1) ou sa négation (notée $\neg v := 1 - v$). Une *clause* est un OU (noté $\vee$; $v_1 \vee v_2 = 1$ sauf si $v_1 = v_2 = 0$) entre un nombre fini de littéraux (par exemple $v_1 \vee \neg v_2$). Enfin, une *forme normale conjonctive* est un ET (noté $\wedge$; $v_1 \wedge v_2 = 1$ si, et seulement si, $v_1 = v_2 = 1$) entre un nombre fini de clauses; par exemple

$$(v_1) \wedge (v_1 \vee \neg v_2 \vee v_3) \wedge (\neg v_1 \vee \neg v_3).$$

Un problème SAT consiste à déterminer s'il existe une affectation des variables booléennes v_i d'une forme normale conjonctive qui lui donne la valeur 1. Il s'agit d'une famille de problèmes NP*-complets*. □

Famille 5.24 (SOUS-SOMME) On se donne un nombre fini $n \geqslant 1$ d'entiers $a_1, \ldots, a_n$, un entier b et on se pose la question de savoir si l'on peut trouver un sous-ensemble d'indices $I \subseteq \{1, \ldots, n\}$ tel que

$$\sum_{i \in I} a_i = b.$$

Il s'agit donc d'un problème de décision. □

Beaucoup de problèmes de calcul scientifique, en particulier les problèmes d'optimisation, ne sont pas des problèmes de décision. Afin d'éviter de passer par une version décisionnelle de ces problèmes, qui est parfois éloignée du problème que l'on se pose au départ, on introduit des notions de complexité qui leurs sont propres, comme la suivante.

Définition 5.25 (classe NP-ardu, problème NP-ardu) On dit qu'une famille de problèmes $F : I \to S$ est dans NP-ardu, on dit aussi que ses problèmes sont *NP-ardus*, si une (ou n'importe quelle) famille de problèmes de NPC peut être évaluée en temps polynomial par une machine de Turing utilisant F comme un oracle (c'est-à-dire que cette machine est supposée évaluer F en une étape). □

Par conséquent, une famille F de NP-ardu doit être formée de problèmes suffisamment difficiles à résoudre pour que, lorsque évaluer $F(i)$ devient une opération élémentaire du calculateur, les problèmes de NPC deviennent polynomiaux pour de tels calculateurs. Évidemment, on a

$$\mathsf{NPC} \subseteq \mathsf{NP\text{-}ardu}$$

puisque si $F \in \mathsf{NPC}$, on peut évaluer F (un problème de NPC) en une opération (donc en temps polynomial) en utilisant F comme oracle.

Exemple 5.26 *Famille de problèmes* NP-*ardus* : OQ (proposition **??**). □

Nous avons représenté à la figure 5.1 les liens entre les différentes classes de familles

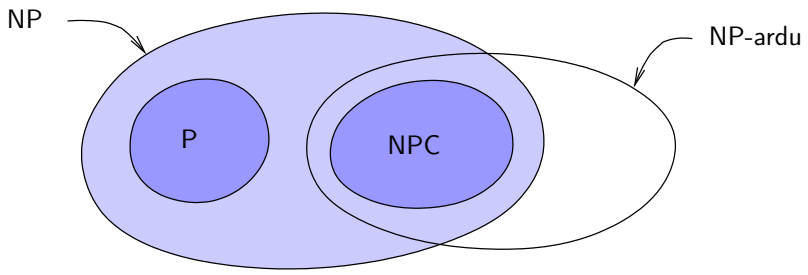


Fig. 5.1. Quelques classes de familles de problèmes si P ≠ NP

de problèmes présentées dans cette section, en supposant que P ≠ NP (ce qui peut être remis en question dans l'avenir). Si P = NP, les trois classes P, NP et NPC sont confondues.

5.3 Conditionnement d'un problème d'optimisation ▲

5.3.1 Notions de conditionnement

Le calcul numérique des solutions d'un problème (d'optimisation ou pas) se fait sur des calculateurs qui utilisent des nombres flottants, avec une précision finie donc. Le *conditionnement* d'un problème est un indicateur de l'effet de cette précision finie sur le résultat. On cherche par cette mesure à savoir si une petite perturbation des données (celle due à cette présicion finie par exemple) peut avoir une grande incidence sur la ou les solutions. Cette incidence dépendra du problème, mais peut aussi dépendre de l'algorithme qui le résout. On comprend alors pourquoi il n'y a pas une unique définition du conditionnement, ou que l'on fait parfois allusion à celui-ci de manière peu précise, ou encore que celui-ci sert de bouc émissaire permettant de justifier le mauvais comportement d'une approche algorithmique !

Résoudre un problème d'optimisation quadratique convexe sans contrainte revient à résoudre sa condition d'optimalité qui est un système linéaire (chapitre 8). Le conditionnement d'un tel problème d'optimisation est, par définition, celui défini pour un système linéaire $Ax = b$. L'effet d'une perturbation de b sur la solution x est estimé par la proposition suivante.

Proposition 5.27 (conditionnement d'un système linéaire) *Soit* $\|\cdot\|$ *une norme matricielle subordonnée à une norme vectorielle notée de la même maniè-*

re. Si A est inversible, $b \neq 0$, $x = A^{-1}b$ et $x' = A^{-1}b'$, alors

$$\frac{\|x' - x\|}{\|x\|} \leqslant \|A\| \, \|A^{-1}\| \, \frac{\|b' - b\|}{\|b\|}.$$

DÉMONSTRATION. Clairement $x' - x = A^{-1}(b' - b)$, si bien que $\|x' - x\| \leqslant \|A^{-1}\| \, \|b' - b\|$. On obtient le résultat en utilisant l'inégalité $\|b\| \leqslant \|A\| \, \|x\|$. □

Le scalaire $\|A\| \, \|A^{-1}\|$ permet donc d'avoir une estimation de l'erreur *relative* commise sur x à partir de l'erreur relative sur b. On le prend comme définition du conditionnement de A.

Définition 5.28 (conditionnement d'une matrice) *Le conditionnement d'une matrice inversible A pour une norme matricielle $\|\cdot\|$ est le nombre*

$$\kappa(A) := \|A\| \, \|A^{-1}\|.$$

On note $\kappa_p(A)$ le conditionnement de A pour la norme matricielle ℓ_p.

Si $\|\cdot\|$ est subordonnée à une norme vectorielle, $\kappa(A) \geqslant 1$ (en effet $1 = \|I\| = \|AA^{-1}\| \leqslant \|A\| \, \|A^{-1}\|$).

C'est la même notion que pour la résolution d'un système linéaire. La définir comme distance à l'ensemble des problèmes « singuliers ». Par exemple, Kahan [340 ; 1966] montre que le conditionnement ℓ_p mesure la distance relative en norme ℓ_p à l'ensemble des matrices singulières :

$$\kappa_p(A) = \min_{X \text{ singular}} \frac{\|X - A\|_p}{\|A\|_p}.$$

On appelle *préconditionnement*, toute méthode destinée à améliorer le conditionnement du problème. En général, on transforme le problème en un problème équivalent dont le conditionnement est meilleur.

Noter ce qu'en dit Moré [438 ; 1983; Section 2: scaling and preconditioning]. Dans N et qN, noter que ces algorithmes sont invariants par changement de variables, localement.

5.4 Calcul de dérivées par état adjoint

La méthode de l'état adjoint est une technique souvent utilisée pour calculer les dérivées d'une fonction dépendant *implicitement* des variables par rapport auxquelles on veut la dériver. Elle permet d'éviter le calcul coûteux de la dérivée de la fonction implicite et est d'une utilisation constante dans les problèmes de commande optimale [384, 398]. Nous allons dans un premier temps donner un cadre formel au problème que nous nous posons, ainsi que des exemples de problèmes d'optimisation

où cette technique peut être utile (section 5.4.1). À la section 5.4.2, nous montrons comment l'état adjoint peut être utilisé pour calculer un gradient et ceci est illustré sur des problèmes de commande optimale à la section 5.4.3. Enfin, le calcul de produits hessienne-vecteur par état adjoint est abordé à la section 5.4.4.

5.4.1 Position du problème

Soient $\mathbb{Y}$, $\mathbb{U}$ et $\mathbb{Z}$ trois espaces vectoriels et

$$F : \mathbb{Y} \times \mathbb{U} \to \mathbb{Z} : (y, u) \mapsto F(y, u)$$

une fonction régulière utilisée pour décrire un système par ce que l'on appelle son *équation d'état*

$$F(y, u) = 0. \tag{5.13}$$

Cela peut être l'équation d'équilibre d'un système statique ou l'équation d'évolution d'un modèle dynamique. Dans le cadre des problèmes de commande optimale, les variables u sont appelées *variables de commande* du système et les variables y sont appelées *variables d'état* du système.

Ce qui distingue y et u est le rôle que jouent ces variables dans l'équation d'état. Pour une valeur donnée aux variables de commande u, on suppose que l'équation (5.13) détermine l'état y du système. Ceci sera certainement vrai, localement, dans le voisinage d'un couple état-commande (y_0, u_0) vérifiant (5.13), si F est C^1 dans un voisinage de (y_0, u_0) et si la jacobienne de F en (y_0, u_0) par rapport à y

$$F'_y(y_0, u_0) \text{ est inversible.} \tag{5.14}$$

En effet, on sait alors, en vertu du théorème des fonctions implicites (théorème C.14), qu'il existe une unique *fonction implicite*

$$u \in U_0 \mapsto y(u) \in Y_0 \tag{5.15}$$

définie dans un voisinage U_0 de u_0 et à valeurs dans un voisinage Y_0 de y_0. Celle-ci vérifie

$$\begin{cases} y(u_0) = y_0 \\ F(y(u), u) = 0, \quad \forall u \in U_0. \end{cases} \tag{5.16}$$

Cette fonction implicite décrit l'état $y(u)$ du système lorsque la commande u varie dans U_0.

Supposons à présent que l'on donne une fonction scalaire φ dépendant de y et de u :

$$\varphi : \mathbb{Y} \times \mathbb{U} \to \mathbb{R} : (y, u) \mapsto \varphi(y, u).$$

Si y est fonction implicite de u pour l'équation d'état (5.13), on peut considérer φ comme une fonction de u seul, ou plus précisément considérer la fonction $\psi : U_0 \to \mathbb{R}$ définie par

$$\psi(u) = \varphi(y(u), u).$$

Le problème que l'on se pose est de calculer, *de manière efficace*, le gradient de ψ par rapport à u.

Commençons par calculer les dérivées directionnelles de ψ. Si $v \in \mathbb{U}$, on a

$$\psi'(u)\cdot v = \varphi'_y(y(u),u)\cdot y'(u)\cdot v + \varphi'_u(y(u),u)\cdot v. \tag{5.17}$$

Donnons nous un produit scalaire $\langle\cdot,\cdot\rangle$ sur $\mathbb{U}$ pour pouvoir calculer le gradient de ψ. Nous aurons également besoin d'un produit scalaire sur $\mathbb{Y}$ que l'on notera de la même manière. On note alors A^* l'adjoint d'un opérateur linéaire $A \in \mathcal{L}(\mathbb{U},\mathbb{Y})$ pour ces produits scalaires. Avec $y = y(u)$, on a

$$\nabla\psi(u) = y'(u)^*\nabla_y\varphi(y,u) + \nabla_u\varphi(y,u).$$

Si cette formule contenait $y'(u)$ au lieu de $y'(u)^*$, il s'agirait simplement d'évaluer une dérivée directionnelle de y, ce qui n'est pas une opération coûteuse : si la direction est w, $y'(u)\cdot w$ s'obtient en résolvant un seul système linéaire

$$F'_y(y,u)\,(y'(u)\cdot w) = -F'_u(y,u)\cdot w.$$

Mais avec $y'(u)^*$, il semble bien qu'il faille évaluer $y'(u)$ complètement, ce qui demande a priori le calcul de $\dim\mathbb{U}$ dérivées directionnelles ou encore de $\dim\mathbb{U}$ systèmes linéaires :

$$F'_y(y,u)y'(u) = -F'_u(y,u). \tag{5.18}$$

La méthode de l'état adjoint que nous présentons ci-après permet d'éviter ces $\dim\mathbb{U}$ résolutions. La présence de l'adjoint d'un opérateur de dérivation montre en fait qu'il s'agit d'une « dérivation cotangente », situation à laquelle la méthode de l'état adjoint est bien adaptée.

Avant cela donnons deux exemples dans le domaine de l'optimisation où cette technique peut être utilisée.

Optimisation avec contraintes d'égalité

Supposons que l'on cherche à résoudre le problème d'optimisation sous contrainte d'égalité suivant

$$\begin{cases} \min f(x) \\ F(x) = 0, \end{cases}$$

où x se partitionne en $x = (y,u)$ et $F'_y(y,u)$ est inversible. L'approche est particulièrement efficace lorsque F est linéaire en y, puisque dans ce cas l'évaluation de la fonction implicite $u \mapsto y(u)$ peut se faire en résolvant un unique système linéaire.

Dans ce cas, il est possible et souvent préférable de remplacer le problème avec contraintes d'égalité en x ci-dessus par le problème en u sans contrainte équivalent

$$\min_u f(y(u),u),$$

où $y(u)$ est la fonction implicite donnée par l'équation d'état $F(y,u) = 0$. Ce problème est en effet d'une dimension ($\dim\mathbb{U}$) qui peut être beaucoup plus petite que celle du problème original ($\dim\mathbb{Y} + \dim\mathbb{U}$).

Lors de la résolution numérique de ce problème, le calcul du gradient de $u \mapsto \psi(u) = f(y(u),u)$ se fera sans difficulté par la technique de l'état adjoint (avec $\varphi = f$) si le calcul de l'état $y(u)$ associé à une commande u est aisé, ce qui est souvent le cas lorsque F est linéaire en y. Si F n'est pas linéaire en y, il sera souvent nécessaire d'utiliser un processus itératif pour calculer $y(u)$, si bien que la technique de l'état adjoint s'impose moins.

Optimisation avec contraintes d'égalité et d'inégalité

La même idée peut être utilisée si l'on a en plus des contraintes d'inégalité *en petit nombre* :

$$\begin{cases} \min f(x) \\ F(x) = 0 \\ c(x) \leqslant 0, \end{cases}$$

où c est à valeurs vectorielles. On suppose toujours que x se partitionne en $x = (y, u)$ et que F'_y est inversible. Ici aussi on aura intérêt à ce que F soit linéaire en y.

Dans ce cas également, il pourra être préférable de remplacer le problème original par le problème équivalent en u seulement :

$$\begin{cases} \min f(y(u), u) \\ c(y(u), u) \leqslant 0. \end{cases}$$

Les gradients de $u \mapsto f(y(u), u)$ et des $u \mapsto c_i(y(u), u)$ se calculeront par la technique de l'état adjoint (avec $\varphi = f$ ou un des c_i), d'où la nécessité de ne pas avoir trop de contraintes. Il n'est pas nécessaire que c soit linéaire en y.

5.4.2 Calcul de gradients

Étant donné l'importance pratique de cette méthode, très simple, nous donnons deux approches permettant de déduire les formules. La première est la plus rapide dans ce cadre formel, mais fait intervenir l'expression de la dérivée de la fonction implicite donnée par (5.18), qu'il n'est pas toujours aisé d'écrire. Celle-ci n'est pas nécessaire dans la seconde approche, qui s'avère souvent la plus utile dans les applications.

Calcul direct

On remplace dans (5.17) la valeur de $y'(u)$ donnée par (5.18), à savoir

$$y'(u) = -F'_y(y, u)^{-1} F'_u(y, u).$$

Ceci donne

$$\psi'(u) \cdot v = \varphi'_y(y(u), u) \cdot \left(-F'_y(y, u)^{-1} F'_u(y, u) \cdot v \right) + \varphi'_u(y(u), u) \cdot v.$$

Avec le produit scalaire $\langle \cdot, \cdot \rangle$ supposé donné et en simplifiant les notations par $y \equiv y(u)$, cette relation s'écrit aussi

$$\begin{aligned} \langle \nabla\psi(u), v \rangle &= -\langle \nabla_y \varphi(y, u), F'_y(y, u)^{-1} F'_u(y, u) \cdot v \rangle + \langle \nabla_u \varphi(y, u), v \rangle \\ &= -\langle F'_u(y, u)^* \left(F'_y(y, u)^{-1} \right)^* \nabla_y \varphi(y, u), v \rangle + \langle \nabla_u \varphi(y, u), v \rangle \\ &= \langle F'_u(y, u)^* p, v \rangle + \langle \nabla_u \varphi(y, u), v \rangle, \end{aligned}$$

si p est solution du système linéaire

$$F'_y(y, u)^* p = -\nabla_y \varphi(y, u). \tag{5.19}$$

Cette équation est appelée l'*équation de l'état adjoint* ou plus simplement l'*équation adjointe* (elle fait intervenir l'opérateur adjoint de $F'_y(y, u)$). Elle est linéaire en p.

Pour u donné et $y = y(u)$ calculé, elle permet de déterminer $p = p(u)$, appelé *état adjoint*. Alors, d'après ce qui précède, le gradient de ψ en u s'écrit comme fonction de u, $y = y(u)$ et de $p = p(u)$:

$$\nabla\psi(u) = F'_u(y,u)^* p + \nabla_u \varphi(y,u). \tag{5.20}$$

Utilisation d'un lagrangien

Pour $p \in \mathbb{Z}$, on introduit le *lagrangien*

$$\ell(y,u,p) = \varphi(y,u) + \langle p, F(y,u)\rangle.$$

La valeur du vecteur p sera choisie plus tard. Elle dépendra du point u_0 où l'on veut calculer le gradient de ψ, mais dans la démarche qui suit on considère p indépendant de u. Si $u \mapsto y(u)$ est une fonction implicite pour (5.13), on a grâce à (5.16)

$$\psi(u) = \ell(y(u),u,p), \quad \forall u \in U_0,$$

si bien que le gradient de ψ peut aussi s'obtenir en calculant celui de $u \mapsto \ell(y(u),u,p)$. On choisira p de manière à ce que ce dernier calcul soit simple.

Pour $v \in \mathbb{U}$, on a en $y = y(u)$

$$\begin{aligned}\psi'(u)\cdot v &= \varphi'_y(y,u)\cdot y'(u)\cdot v + \varphi'_u(y,u)\cdot v + \langle p, F'_y(y,u)\cdot y'(u)\cdot v + F'_u(y,u)\cdot v\rangle \\ &= \langle \nabla_u\varphi(y,u), v\rangle + \langle F'_u(y,u)^* p, v\rangle + \langle \nabla_y\varphi(y,u) + F'_y(y,u)^* p, y'(u)\cdot v\rangle,\end{aligned}$$

où on a regroupé les termes dépendant de $y'(u)\cdot v$. L'idée maîtresse de cette approche est maintenant de faire disparaître les termes où $y'(u).v$ intervient (on se rappelle qu'on essaye d'éviter le calcul de $y'(u)$!), en *choisissant* p de telle sorte que le dernier terme s'annule :

$$F'_y(y,u)^* p = -\nabla_y \varphi(y,u).$$

On retrouve l'équation adjointe (5.19). Alors, d'après ce qui précède le gradient de ψ en u s'écrit comme fonction de u, $y = y(u)$ et de $p = p(u)$:

$$\nabla\psi(u) = \nabla_u\varphi(y,u) + F'_u(y,u)^* p.$$

On retrouve (5.20).

Résumé des opérations

Pour résumer, la méthode de l'état adjoint pour le calcul de $\nabla\psi(u_0)$ se déroule en trois étapes. On suppose $u_0 \in \mathbb{U}$ donné. Ensuite, on on procède de la manière suivante.

Schéma 5.29 (calcul d'un gradient par état adjoint)

1. On calcule l'état y_0, solution de l'équation d'état (5.13), avec $u = u_0$, qui peut être non linéaire.

2. On calcule l'état adjoint p_0, solution de l'équation adjointe (5.19), avec $u = u_0$ et $y = y_0$, qui est toujours linéaire en p.
3. On calcule le gradient de ψ en u_0 par la formule (5.20), avec $u = u_0$, $y = y_0$ et $p = p_0$.

Comme annoncé, ce calcul ne fait pas intervenir $y'(u_0)$.

On se rappelle souvent les formules (5.19) et (5.20), en remarquant que l'équation adjointe s'écrit également

$$\nabla_y \ell(y_0, u_0, p) = 0$$

et que le gradient est aussi donné par

$$\nabla\psi(u_0) = \nabla_u \ell(y_0, u_0, p_0).$$

Cependant, dans certains cas, il est préférable de reprendre la démarche suivie pour arriver à ces formules plutôt que de les appliquer directement. En particulier, si les équations d'état sont des équations aux dérivées partielles ou des équations d'évolution, les conditions aux limites ou initiales se trouvent plus facilement en reprenant la démarche que nous avons exposée, plutôt qu'en appliquant les formules de dérivation du lagrangien (voir la section 5.4.3).

Une interprétation de l'état adjoint

Comme un multiplicateur de Lagrange, l'état adjoint p a une interprétation marginaliste, que nous énonçons dans la proposition 5.30 ci-dessous. Pour cela, considérons le système perturbé par le vecteur λ :

$$F(y, u) + \lambda = 0.$$

Si $F'_y(y_0, u_0)$ est inversible, on peut encore exprimer la solution de ce système perturbé dans un voisinage de $(u, \lambda) = (u_0, 0)$ par une fonction implicite

$$(u, \lambda) \mapsto \tilde{y}(u, \lambda).$$

Comme $F(\tilde{y}(u, 0), u) = 0$ pour tout u voisin de u_0, l'unicité de la fonction implicite implique que $\tilde{y}(u, 0) = y(u)$ est la valeur en u de la fonction implicite du problème non perturbé. On peut aussi considérer la fonction

$$\tilde{\psi}(u, \lambda) = \varphi(\tilde{y}(u, \lambda), u).$$

Proposition 5.30 *L'état adjoint p associé à (y, u), solution du système non perturbé, donne la variation de la valeur de $\tilde{\psi}$ en $(u, 0)$ par rapport à une perturbation des équations d'état :*

$$p = \nabla_\lambda \tilde{\psi}(u, 0).$$

Démonstration. Pour montrer cela, on calcule $\nabla_\lambda \tilde{\psi}(u, 0)$ par la méthode de l'état adjoint. On introduit le lagrangien du problème perturbé

$$\tilde{\ell}(y,u,\lambda,p) = \varphi(y,u) + \langle p, F(y,u) + \lambda \rangle.$$

Soit u donné. L'état du système perturbé correspondant à ce u et à $\lambda = 0$ est l'état y du système non perturbé correspondant à la commande u. L'état adjoint du système perturbé en $(y,u,0)$ s'obtient par

$$\nabla_y \tilde{\ell}(y,u,0,p) \equiv \nabla_y \varphi(y,u) + F'_y(y,u)^* p = 0.$$

Il est donc identique à l'état adjoint correspondant à (y,u) dans le système non perturbé. Enfin, le gradient de $\tilde{\psi}$ par rapport à λ en $(u,0)$ s'obtient par

$$\nabla_\lambda \tilde{\psi}(u,0) = \nabla_\lambda \tilde{\ell}(y,u,0,p) = p.$$

D'où le résultat. □

5.4.3 Exemples

Nous donnons ci-dessous deux exemples illustrant l'utilisation de la méthode de l'état adjoint. Ils sont paradigmatiques. Dans le premier exemple, l'équation d'état est une équation différentielle ordinaire. On est donc dans une situation plus générale que dans le cadre abstrait ci-dessus, puisque les objets manipulés appartiennent à un espace vectoriel de dimension infinie, mais les principes à appliquer sont identiques. Dans le second exemple, l'équation d'état est une équation différentielle ordinaire discrétisée. On pourrait alors appliquer les formules de la section 5.4.2 précédente pour obtenir le gradient, mais, comme nous l'avons déjà dit, il est plus efficace d'appliquer le principe ayant conduit à ces formules, de manière à en dégager la structure. Dans certains cas, il faut un peu de doigté pour aboutir à une formulation opérationnelle. La démarche exposée peut aussi s'appliquer lorsque les équations d'état sont des équations aux dérivées partielles. Ce sujet sort du cadre de cette monographie. Il est traité en détail dans [398].

Cas où l'équation d'état est une équation différentielle

On suppose que l'état du système est une fonction dépendant du temps, définie sur un intervalle $[0,T]$ de $\mathbb{R}$ $(T > 0)$, à valeurs dans $\mathbb{R}^n$:

$$y : t \in [0,T] \to y(t) \in \mathbb{R}^n.$$

Cet état est supposé déterminé par une équation différentielle ordinaire et sa condition initiale :

$$\begin{cases} \dot{y}(t) = \varphi(y(t),t), & \text{pour tout } t \in]0,T[\\ y(0) = Au, \end{cases} \tag{5.21}$$

appelée *équation d'état* du système. Dans (5.21), $\dot{y}$ désigne la dérivée de y par rapport à t, $\varphi : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ est une application régulière, A est une matrice $n \times m$ et $u \in \mathbb{R}^m$ est la variable de commande. Dans certains problèmes de commande optimale, on cherche à déterminer la commande u, qui agit ici sur la condition initiale du système, de manière à minimiser un critère de la forme

$$f(u) = \int_0^T J(y(t,u),t)\,\mathrm{d}t + J^u(u) + J^T(y(T,u)),$$

où J est une fonction de $\mathbb{R}^n \times \mathbb{R}$ dans $\mathbb{R}$ représentant un coût distribué dans le temps, J^u est une fonction de $\mathbb{R}^m$ dans $\mathbb{R}$ représentant un coût sur la commande du système et J^T est une fonction de $\mathbb{R}^n$ dans $\mathbb{R}$ représentant un coût sur l'état final du système. Nous avons écrit $y(t) \equiv y(t,u)$ et $y(T) \equiv y(T,u)$ pour expliciter la dépendance de y en u, qui provient de l'équation d'état (5.21).

On s'intéresse ici au calcul de $\nabla f(u)$, le gradient de f en u. Il se déduira aisément de $f'(u)\cdot v$, la dérivée directionnelle de f en u dans la direction $v \in \mathbb{R}^m$. La difficulté est de prendre en compte la dépendance de y en u. Le calcul proposé est formel, car il ne précise pas les espaces fonctionnels, dont le choix sort du cadre de cette monographie.

Comme dans le cadre abstrait ci-dessus, on écrit $f(u)$ comme un lagrangien au moyen d'un état adjoint, qui est ici une fonction $p : t \in [0,T] \to p(t) \in \mathbb{R}^n$:

$$\begin{aligned} f(u) &= \int_0^T J(y(t,u),t)\,\mathrm{d}t + J^u(u) + J^T(y(T,u)) \\ &\quad + \int_0^T p(t)^{\mathsf{T}}\left(\dot{y}(t,u) - \varphi(y(t,u),t)\right)\,\mathrm{d}t. \end{aligned}$$

Cette identité a lieu quel que soit p, pourvu que y vérifie l'équation d'état (5.21). On choisira p ultérieurement, de manière à éliminer la dérivée de la fonction implicite $u \mapsto y(\cdot,u)$. On a

$$\begin{aligned} f'(u)\cdot v &= \int_0^T J'_y(y,t)\cdot(y'(t,u)\cdot v)\,\mathrm{d}t \\ &\quad + (J^u)'(u)\cdot v + (J^T)'(y(T,u))\cdot(y'(T,u)\cdot v) \\ &\quad + \int_0^T p(t)^{\mathsf{T}}\left(\dot{y}'(t,u)\cdot v - \varphi'_y(y,t)\cdot(y'(t,u)\cdot v)\right)\,\mathrm{d}t. \end{aligned}$$

Pour faire disparaître la dérivée en temps de $y'(t,u)$, on intègre par parties :

$$\begin{aligned} &\int_0^T p(t)^{\mathsf{T}}\left(\dot{y}'(t,u)\cdot v\right)\,\mathrm{d}t \\ &= -\int_0^T \dot{p}(t)^{\mathsf{T}}\left(y'(t,u)\cdot v\right)\,\mathrm{d}t + p(t)^{\mathsf{T}}\left(y'(t,u)\cdot v\right)\Big|_0^T \\ &= -\int_0^T \dot{p}(t)^{\mathsf{T}}\left(y'(t,u)\cdot v\right)\,\mathrm{d}t + p(T)^{\mathsf{T}}\left(y'(T,u)\cdot v\right) - p(0)^{\mathsf{T}}Av, \end{aligned}$$

où on a utilisé le fait que $y(0,u) = Au$ pour tout $u \in \mathbb{R}^n$ et donc $y'(0,u)\cdot v = Av$. On aurait pu faire cette intégration par parties dans l'expression de $f(u)$, avant dérivation. En regroupant les termes contenant $y'(t,u)\cdot v$ et $y'(T,u)\cdot v$ dans $f'(u)\cdot v$, on obtient

$$\begin{aligned} f'(u)\cdot v &= (J^u)'(u)\cdot v \\ &\quad + \int_0^T \left(-\dot{p}(t) - \varphi'_y(y,t)^{\mathsf{T}}p(t) + \nabla_y J(y,t)\right)^{\mathsf{T}}\left(y'(t,u)\cdot v\right)\,\mathrm{d}t \\ &\quad + \Big(p(T) + \nabla J^T(y(T))\Big)^{\mathsf{T}}\Big(y'(T,u)\cdot v\Big) - p(0)^{\mathsf{T}}Av. \end{aligned}$$

On a désigné de la même façon la dérivée $\varphi'_y(y,t)$ et la matrice jacobienne associée. L'étape importante de ce calcul est maintenant d'éliminer $y'(t,u)\cdot v$ et $y'(T,u)\cdot v$, qui sont coûteux à calculer, en annulant leur facteur par un choix adéquat de p. Ceci conduit à l'*équation adjointe*, qui est une équation différentielle ordinaire linéaire en p, avec une condition *finale*

$$\left\{\begin{array}{l} -\dot p(t) = \varphi'_y(y(t),t)^{\mathsf T} p(t) - \nabla_y J(y(t),t), \quad \text{pour tout } t\in\,]0,T[\\ p(T) = -\nabla J^T(y(T)). \end{array}\right. \tag{5.22}$$

On obtient alors

$$f'(u)\cdot v = (J^u)'(u)\cdot v - p(0)^{\mathsf T} A v$$

et donc

$$\nabla f(u) = \nabla J^u(u) - A^{\mathsf T} p(0). \tag{5.23}$$

Pour résumer, le gradient $\nabla f(u)$ de f en u s'obtient en déterminant d'abord y comme solution de l'équation d'état (5.21), qui est souvent non linéaire, ensuite p comme solution de l'équation adjointe (5.22), toujours linéaire, et enfin le gradient par (5.23). Ce calcul permet d'éviter la détermination des m trajectoires $t\mapsto y'(t,u)\cdot e_i$, pour $i=1,\dots,m$, formant la dérivée de la fonction implicite $u\mapsto y(\cdot,u)$.

On observera que, si l'équation d'état est progressive (c.-à-d., elle s'intègre à partir d'une condition initiale), l'équation adjointe est rétrograde (c.-à-d., elle s'intègre à partir d'une condition finale). Notons également que, pour intégrer l'équation adjointe, il faut connaître l'état y à chaque instant. En pratique (sur les modèles discrétisés), il faut donc mémoriser cet état, ce qui peut demander beaucoup d'espace-mémoire : l'efficacité du calcul est donc contrebalancée par un besoin de mémoire. On retrouvera cette propriété dans le mode inverse de différentiation (section 5.5.3).

Cas où l'équation d'état est une équation différentielle discrétisée

Supposons à présent que l'état $y : t\in[0,T]\to y(t)\in\mathbb{R}^n$ du système considéré soit décrit par l'équation différentielle

$$\left\{\begin{array}{l} \dot y(t) = \varphi(y(t),u(t),t), \quad \text{pour tout } t\in\,]0,T[\\ y(0) = y_0, \end{array}\right. \tag{5.24}$$

dans laquelle $y_0\in\mathbb{R}^n$ est la valeur donnée de l'état à l'instant initial et la commande est à présent une fonction $u : t\in[0,T]\to u(t)\in\mathbb{R}^m$ agissant à chaque instant sur le système par l'intermédiaire de $\varphi : \mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}\to\mathbb{R}^n$. Le critère que l'on cherche à minimiser par rapport à u est le suivant :

$$\int_0^T J(y(t,u),u(t),t)\,\mathrm{d}t + J^T(y(T,u)), \tag{5.25}$$

où $J : \mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}\to\mathbb{R}$ représente le coût instantané et $J^T : \mathbb{R}^n\to\mathbb{R}$ représente le coût associé à l'état final du système. Nous avons encore écrit $y(t)\equiv y(t,u)$ pour insister sur le fait que y dépend de u par l'intermédiaire de l'équation d'état (5.24).

On s'intéresse cette fois à la forme discrétisée de ce modèle. Pour cela, on décompose l'intervalle d'intégration $[0,T]$ en N pas de temps $\Delta t_i > 0$, $1\leqslant i\leqslant N$, tels que

$T = \sum_{i=1}^{N} \Delta t_i$. On note $t_0 = 0$, puis $t_i = t_{i-1} + \Delta t_i$, pour $1 \leqslant i \leqslant N$, (donc $t_N = T$) et $u_i = u(t_i) \in \mathbb{R}^m$. Pour simplifier, on supposera que l'approximation $y_i \in \mathbb{R}^n$ de $y(t_i)$ est obtenue par le *schéma d'Euler implicite* (pour d'autres schémas d'intégration, on pourra consulter [230 ; 1997] par exemple) :

$$y_{i+1} = y_i + \Delta t_i \, \varphi(y_{i+1}, u_{i+1}, t_i), \quad \text{pour } 0 \leqslant i \leqslant N-1, \tag{5.26}$$

démarrant sur la condition initiale y_0. On prendra comme approximation du critère continu, le critère discret suivant

$$f(\mathbf{u}) = \sum_{i=1}^{N} J(y_i(\mathbf{u}), u_i, t_i) \, \Delta t_i + J^T(y_N(\mathbf{u})),$$

où $\mathbf{u} = \{u_i\}_{i=1}^N \subseteq \mathbb{R}^m$ et où la notation $y_i \equiv y_i(\mathbf{u})$ met en évidence la dépendance de y_i par rapport à $\mathbf{u}$. On cherche à calculer le gradient de f.

Comme dans le cadre abstrait et le cas continu ci-dessus, on écrit $f(\mathbf{u})$ comme un lagrangien au moyen d'un état adjoint, qui est ici un ensemble de $N+1$ vecteurs $\mathbf{p} = \{p_i\}_{i=0}^N \subseteq \mathbb{R}^n$ (le vecteur $p_N \in \mathbb{R}^n$ apparaîtra plus loin) :

$$\begin{aligned} f(\mathbf{u}) \;=\; & \sum_{i=1}^{N} J(y_i(\mathbf{u}), u_i, t_i) \, \Delta t_i + J^T(y_N(\mathbf{u})) \\ & + \sum_{i=0}^{N-1} p_i^{\mathsf{T}} \Big(y_{i+1} - y_i - \Delta t_i \, \varphi(y_{i+1}, u_{i+1}, t_i) \Big). \end{aligned}$$

Cette identité a lieu quel que soit $\mathbf{p}$, pourvu que $\mathbf{y} = \{y_i\}_{i=0}^N$ vérifie l'équation d'état (5.26). On choisira $\mathbf{p}$ ultérieurement, de manière à éliminer la dérivée de la fonction implicite $\mathbf{u} \mapsto \mathbf{y}(\mathbf{u})$. Pour varier, commençons par faire une « somme par parties » (équivalent de l'intégration par parties de l'exemple précédent), avant dérivation. On a

$$\sum_{i=0}^{N-1} p_i^{\mathsf{T}}(y_{i+1} - y_i) = \sum_{i=1}^{N} p_{i-1}^{\mathsf{T}} y_i - \sum_{i=0}^{N-1} p_i^{\mathsf{T}} y_i = \sum_{i=1}^{N} (p_{i-1} - p_i)^{\mathsf{T}} y_i + p_N^{\mathsf{T}} y_N - p_0^{\mathsf{T}} y_0.$$

Il en résulte

$$\begin{aligned} f(\mathbf{u}) \;=\; & \sum_{i=1}^{N} J(y_i(\mathbf{u}), u_i, t_i) \, \Delta t_i + J^T(y_N(\mathbf{u})) \\ & + \sum_{i=1}^{N} \Big((p_{i-1} - p_i)^{\mathsf{T}} y_i - \Delta t_{i-1} \, p_{i-1}^{\mathsf{T}} \varphi(y_i, u_i, t_{i-1}) \Big) \\ & + p_N^{\mathsf{T}} y_N - p_0^{\mathsf{T}} y_0. \end{aligned}$$

On peut à présent calculer la dérivée directionnelle de f en $\mathbf{u}$ dans une direction $\mathbf{v} = \{v_i\}_{i=1}^N \subseteq \mathbb{R}^m$:

$$
\begin{aligned}
f'(\mathbf{u})\cdot\mathbf{v} \;=\; & \sum_{i=1}^{N}\Big(J'_y(y_i,u_i,t_i)\cdot(y'_i(\mathbf{u})\cdot\mathbf{v})+J'_u(y_i,u_i,t_i)\cdot v_i\Big)\Delta t_i\\
& +(J^T)'(y_N)\cdot(y'_N(\mathbf{u})\cdot\mathbf{v})\\
& +\sum_{i=1}^{N}\Big((p_{i-1}-p_i)^{\mathsf{T}}(y'_i(\mathbf{u})\cdot\mathbf{v})\\
& \qquad -\Delta t_{i-1}\,p_{i-1}^{\mathsf{T}}\varphi'_y(y_i,u_i,t_{i-1})\cdot(y'_i(\mathbf{u})\cdot\mathbf{v})\\
& \qquad -\Delta t_{i-1}\,p_{i-1}^{\mathsf{T}}\varphi'_u(y_i,u_i,t_{i-1})\cdot v_i\Big)\\
& +p_N^{\mathsf{T}}(y'_N(\mathbf{u})\cdot\mathbf{v}).
\end{aligned}
$$

Comme précédemment, l'étape importante du calcul consiste à éliminer les dérivées $y'_i(\mathbf{u})\cdot\mathbf{v}$ $(1\leqslant i\leqslant N)$, qui sont coûteuses à évaluer, en annulant ses facteurs par un choix approprié de $\mathbf{p}$. Ceci conduit à l'*équation adjointe* suivante

$$
\left\{\begin{array}{l}
\Big(I-\Delta t_{i-1}\,\varphi'_y(y_i,u_i,t_{i-1})^{\mathsf{T}}\Big)p_{i-1}=p_i-\Delta t_i\,\nabla_y J(y_i,u_i,t_i),\quad \text{pour } i=N,\dots,1\\
p_N=-\nabla J^T(y_N).
\end{array}\right. \tag{5.27}
$$

On obtient alors

$$
\nabla f(\mathbf{u})=\Big\{\Delta t_i\,\nabla_u J(y_i,u_i,t_i)-\Delta t_{i-1}\,\varphi'_u(y_i,u_i,t_{i-1})^{\mathsf{T}}p_{i-1}\Big\}_{i=1}^{N}, \tag{5.28}
$$

qui est un vecteur de $\mathbb{R}^{Nm}$.

Résumons les opérations. Le gradient $\nabla f(\mathbf{u})$ de f en $\mathbf{u}=\{u_i\}_{i=1}^N\in\mathbb{R}^{Nm}$ s'obtient en déterminant d'abord $\mathbf{y}=\{y_i\}_{i=1}^N\in\mathbb{R}^{Nn}$ à partir de sa condition initiale y_0 comme solution de l'équation d'état (5.26), qui est souvent non linéaire. Ensuite $\mathbf{p}=\{p_i\}_{i=0}^{N-1}\in\mathbb{R}^{Nn}$ est déterminé à partir de sa condition finale $p_N=-\nabla J^T(y_N)$ comme solution de l'équation adjointe (5.27), toujours linéaire. Enfin le gradient est obtenu par (5.28). Comme dans le cas continu analysé précédemment, l'équation d'état est progressive et l'équation adjointe est rétrograde.

5.4.4 Calcul de produits hessienne-vecteur

On se place à nouveau dans le cadre de la section 5.4.1, où des variables y et u sont reliées par une *équation d'état*

$$F(y,u)=0,$$

permettant d'écrire localement y comme une fonction de u et où l'on cherche à dériver la fonction

$$u\mapsto\psi(u):=\varphi(y(u),u).$$

On s'intéresse à présent au produit de la hessienne de ψ en u (pour un produit scalaire $\langle\cdot,\cdot\rangle$ donné) par un vecteur v.

Le schéma de calcul 5.29 peut être vu comme évaluant successivement, à partir de u donné, les variables y, puis p, et enfin $\nabla\psi$, toutes vues comme des fonctions de u.

Il suffit donc de calculer la dérivée directionnelle de ces quantités dans la direction v, pour obtenir finalement $\nabla^2\psi(u)v$. On note $\dot{y} = y'(u)$ et $\dot{p} = p'(u)$. La dérivation de l'équation d'état donne une équation linéaire permettant de calculer $\dot{y}$:

$$F'_y(y,u)\,\dot{y} = -F'_u(y,u).$$

La dérivation de l'équation adjointe donne une équation linéaire permettant de calculer $\dot{p}$:

$$F'_y(y,u)^*\,\dot{p} = -\left(F''_{yy}(y,u)\cdot\dot{y} + F''_{yu}(y,u)\cdot v\right)^*\,p - \left(\varphi''_{yy}(y,u)\cdot\dot{y} + \varphi''_{yu}(y,u)\cdot v\right).$$

Il reste à dériver l'équation donnant le gradient de ψ pour obtenir le produit hessienne-vecteur :

$$\psi''(u)\cdot v = \left(F''_{yu}(y,u)\cdot\dot{y} + F''_{uu}(y,u)\cdot v\right)^*\,p + \left(\varphi''_{yu}(y,u)\cdot\dot{y} + \varphi''_{uu}(y,u)\cdot v\right).$$

On observera que les systèmes linéaires intervenant dans le calcul de $\nabla\psi(u)$ ou $\nabla^2\psi(u)v$, utilisent tous la matrice $F'_y(y,u)$ ou son adjointe. Si une factorisation de cette matrice s'impose, celle-ci ne devra être faite qu'une seule fois par valeur de u.

5.5 Différentiation automatique

En optimisation, l'objet numérique de base est le gradient. Les méthodes numériques qui peuvent fonctionner sans calcul de gradients sont lentes. On peut alors se demander si les n quantités formant le gradient d'une fonction scalaire définie sur $\mathbb{R}^n$ peuvent se calculer en un temps raisonnable, en particulier si n est très grand. On pourrait en effet penser que cela prend n fois plus de temps que le calcul de la fonction elle-même. S'il en était ainsi, les techniques numériques utilisant le gradient devraient être cantonnées à l'optimisation de problèmes de taille petite ou moyenne. Nous montrerons dans cette section (proposition 5.32) que, par le mode inverse de différentiation automatique, on peut toujours calculer le gradient d'une fonction en un temps qui est du même ordre que celui nécessaire au calcul de la fonction elle-même (sans facteur proportionnel à n). Il s'agit donc d'un résultat fondamental pour l'optimisation numérique.

La *différentiation automatique*[1] est un ensemble de techniques permettant d'obtenir les dérivées exactes (aux erreurs d'arrondi près) d'une fonction représentée par un programme informatique, écrit par exemple en Fortran ou en C. Si c'est un programme qui représente la fonction à dériver, c'est aussi un programme qui est généré par le différentiateur. Il faudra l'exécuter pour obtenir la valeur de la dérivée de la fonction en un point donné.

La méthode se distingue donc de la différentiation symbolique, qui suppose que la fonction à dériver admet une représentation symbolique au moyen de formules mathématiques et qui génère une représentation, également symbolique, de la fonction dérivée. On la trouve dans des codes comme Macsyma [607 ; 1985] ou Maple [116 ;

[1] En anglais, le vocable *Automatic Differentiation* tend à être remplacé par *Computational Differentiation* ou *Algorithmic Differentiation* [288, 291]. En français, on rencontre parfois l'expression *Différentiation par Programme*.

1988]. Elle se distingue aussi de la différentiation par différences finies, méthode par laquelle les dérivées partielles d'une fonction $f : \mathbb{R}^n \to \mathbb{R}^m$ sont *approchées* en faisant varier insensiblement et successivement chacune des variables, tout en gardant les autres fixées :

$$\frac{\partial f}{\partial x_i}(x) \simeq \frac{f(x + t_i e_i) - f(x)}{t_i}, \quad i = 1, \ldots, n,$$

où e_i est le i-ième vecteur de la base canonique de $\mathbb{R}^n$ et $t_i > 0$ est « bien choisi » (ni trop grand ni trop petit !). Avec cette dernière technique les dérivées ne peuvent pas être calculées avec précision. De plus, par cette technique le calcul d'un gradient demande l'évaluation de f aux n points $x + t_i e_i$, $1 \leqslant i \leqslant n$, ce qui peut demander beaucoup de temps de calcul.

Un des attraits de la différentiation automatique est de pouvoir générer un programme calculant le gradient d'une fonction à valeurs scalaires, dont le temps d'exécution relatif à celui du programme original est indépendant du nombre de variables par rapport auxquelles on dérive. Il s'agit en fait de génération automatique de codes adjoints. Cette technique s'est concrétisée dans des codes de différentiation tels que ADOL-C [], ADIFOR [] ou TAPENADE [].

Dans cette section, nous présentons quelques aspects théoriques de la différentiation automatique. Pour plus de détails, on consultera les ouvrages et monographies cités en fin de chapitre.

5.5.1 Modèle de programme

Pour obtenir les formules permettant de générer de manière automatique un code calculant les dérivées d'une fonction représentée par un programme, on commence par considérer le cas d'un programme formé d'une suite d'instructions d'affectation. Un tel programme a l'avantage d'avoir une représentation mathématique simple avec laquelle on peut travailler. Les règles que l'on déduit de ce modèle permettent alors de comprendre les règles de transformation de codes à mettre en œuvre sur des programmes plus complexes, voire généraux, tels que ceux écrits en Fortran ou C.

On note $v_1, v_2, \ldots, v_N$ les N variables du code informatique (ou cases-mémoire) sur lesquelles travaille le programme modèle. Il n'y a aucun ordre sur ces variables. En particulier, il n'est pas nécessaire qu'elles soient évaluées dans l'ordre de leur indice. C'est simplement une manière commode de les désigner toutes. On supposera que les n *variables d'entrée* ($n \leqslant N$), celles par rapport auxquelles on dérive, encore appelée *variables indépendantes*, sont les variables $v_1, \ldots, v_n$. Les m *variables de sortie* ($m \leqslant N$), celles que l'on veut dériver, sont supposées être les variables $v_{N-m+1}, \ldots, v_N$. Des variables peuvent être à la fois d'entrée et de sortie. On dira qu'une variable v_i devient *active* lorsqu'on lui affecte une valeur qui dépend de la valeur donnée aux variables indépendantes.

Le programme modèle que nous considérerons est donc supposé être formé d'une suite de K instructions d'affectation exécutées l'une après l'autre, ce que l'on peut écrire :

$$v_{\mu_k} := \varphi_k(v_{D_k}), \quad k = 1, \ldots, K. \tag{5.29}$$

À chaque instruction k, le programme modifie la variable v_{μ_k} au moyen d'une fonction φ_k, en utilisant les variables v_i, $i \in D_k$, où D_k est une partie de $[1 : N]$. Dans ce modèle, on ne se donne aucune restriction sur $\mu_k \in [1 : N]$, qui peut en particulier faire partie

de D_k. Si on note x les variables d'entrée et f les variables de sorties, on cherche donc à différentier une fonction

$$f : x \in \mathbb{R}^n \mapsto f(x) \in \mathbb{R}^m,$$

qui est représentée par le programme modèle (5.29) et dont la valeur en un point x peut être obtenue en exécutant ce programme.

Pour fixer les idées, considérons l'exemple suivant écrit en Fortran, dans lequel on calcule une variable $f = \texttt{v5}$ à partir de la donnée d'un couple de variables $x = (\texttt{v1}, \texttt{v2})$:

```
v3 = v1 + v2**2
v4 = v1**2 * sin(v3)
v4 = v4/v3
v5 = exp(v4)
```

La correspondance entre ce programme et le modèle (5.29) est détaillée au tableau 5.1.

indice k de l'instruction	instruction	indice de la variable modifiée	fonction φ_k	indices de dépendance
1	`v3=v1+v2**2`	$\mu_1 = 3$	$\varphi_1(a,b) = a + b^2$	$D_1 = \{1, 2\}$
2	`v4=v1**2*sin(v3)`	$\mu_2 = 4$	$\varphi_2(a,b) = a^2 \sin(b)$	$D_2 = \{1, 3\}$
3	`v4=v4/v3`	$\mu_3 = 4$	$\varphi_3(a,b) = b/a$	$D_3 = \{3, 4\}$
4	`v5=exp(v4)`	$\mu_4 = 5$	$\varphi_4(a) = \exp(a)$	$D_4 = \{4\}$

Tableau 5.1. Correspondance entre l'exemple de programme et le modèle (5.29)

On désigne par $\mathcal{F}$ l'ensemble des *fonctions intermédiaires*, c'est-à-dire les fonctions φ_k utilisées dans le programme modèle (5.29). Il n'y a pas de restriction sur ces fonctions, si ce n'est qu'elles doivent être différentiables et à valeurs scalaires (le cas où elles sont à valeurs vectorielles se traite de manière analogue, voir l'exercice 5.7). Elles peuvent représenter les opérations élémentaires (`+`, `-`, `*`, `/`), les fonctions intrinsèques du langage informatique utilisé, des compositions de ces fonctions ou encore des ensembles d'instructions (sous-routines, procédures, fonctions). On notera $\mathcal{F}_{_F} = \{$`+`, `-`, `*`, `/`, `sqrt`, `exp`, `log`, `log10`, `sin`, `cos`, `tan`, `asin`, `acos`, `atan`, `sinh`, `cosh`, `tanh`$\}$ la classe des fonctions utilisables en Fortran (langage choisi pour estimer la complexité des calculs ci-dessous).

5.5.2 Différentiation en mode direct

Le *mode direct* de différentiation est le plus simple et le plus intuitif. Il est bien adapté au calcul des dérivées directionnelles de la fonction f représentée par le programme.

Soit $d \in \mathbb{R}^n$ la direction dans laquelle on veut différentier f. On cherche donc à calculer $f'(x) \cdot d$. Pour cela, il est bon de voir chaque variable du code comme une fonction des variables d'entrée $x = (v_1, \ldots, v_n)$. Donc à chaque v_i $(1 \leqslant i \leqslant N)$

correspond une dérivée directionnelle $\dot{v}_i := v_i'(x)\cdot d$. Alors, en différentiant l'instruction k du programme (5.29), on obtient

$$\dot{v}_{\mu_k} = \sum_{i\in D_k} \frac{\partial \varphi_k}{\partial v_i}(v_{D_k})\dot{v}_i.$$

On a donc une formule permettant de calculer $\dot{v}_{\mu_k} := v'_{\mu_k}(x)\cdot d$ à partir des $\dot{v}_i$, $i \in D_k$. Grâce à celle-ci, on peut « propager » le calcul des dérivées directionnelles des variables évaluées dans le code, parallèlement à leur évaluation. C'est l'idée utilisée dans le mode direct de différentiation.

Remarquons que $\dot{v}_i = d_i$, pour $1 \leqslant i \leqslant n$. Pour calculer $f'(x)\cdot d$, il suffit donc d'initialiser les dérivées directionnelles des variables indépendantes comme suit :

$$\dot{v}_i := d_i, \quad \text{pour } 1 \leqslant i \leqslant n,$$

et ensuite de propager les dérivées directionnelles dans le code par les instructions suivantes :

$$\left.\begin{array}{r} \dot{v}_{\mu_k} := \sum_{i\in D_k} \frac{\partial \varphi_k}{\partial v_i}(v_{D_k})\dot{v}_i \\ v_{\mu_k} := \varphi_k(v_{D_k}) \end{array}\right\}, \quad k = 1,\ldots,K.$$

On calcule $\dot{v}_{\mu_k}$ avant v_{μ_k} au cas où φ_k dépendrait de v_{μ_k}. Il faut en effet évaluer les dérivées partielles $\frac{\partial \varphi_k}{\partial v_i}$ avec la valeur non modifiée de v_{μ_k}. En fin d'exécution, on récupère dans $\dot{v}_{N-m+1}, \ldots, \dot{v}_N$, la dérivée $f'(x)\cdot d$:

$$f'(x)\cdot d := (\dot{v}_{N-m+1},\ldots,\dot{v}_N).$$

Ce mode de différentiation porte le nom de *mode direct* et le code réalisant ces opérations est appelé *code linéaire tangent* de (5.29). Celui-ci peut être résumé ainsi :

$$\boxed{\begin{array}{l} \text{pour } i = 1,\ldots,n \\ \quad \dot{v}_i := d_i \ ; \\ \text{pour } k = 1,\ldots,K \\ \quad \dot{v}_{\mu_k} := \sum_{i\in D_k} \frac{\partial \varphi_k}{\partial v_i}(v_{D_k})\dot{v}_i \ ; \\ \quad v_{\mu_k} := \varphi_k(v_{D_k}) \ ; \\ f'(x)\cdot d := (\dot{v}_{N-m+1},\ldots,\dot{v}_N) \ . \end{array}} \tag{5.30}$$

On voit que l'on a associé à chaque variable v_i du code, une variable $\dot{v}_i$ contenant sa dérivée directionnelle. Si toutes les variables sont actives, l'espace-mémoire requis est donc exactement le double de celui utilisé par le programme original. En ce qui concerne le temps de calcul $T(f, f'\cdot d)$ nécessaire à l'exécution de (5.30), on peut le comparer au temps $T(f)$ de l'exécution de (5.29).

Proposition 5.31 *Le temps $T(f, f'\cdot d)$ nécessaire au calcul de f et de sa dérivée directionnelle $f'\cdot d$ dans la direction d au moyen de l'algorithme (5.30) vérifie*

l'estimation suivante:

$$\frac{T(f, f'\cdot d)}{T(f)} \leqslant C_{\mathcal{F}},$$

où $T(f)$ est le temps nécessaire au calcul de f par l'algorithme (5.29) et $C_{\mathcal{F}}$ est une constante ne dépendant que des fonctions intermédiaires φ_k de la bibliothèque $\mathcal{F}$.

Démonstration. D'après l'algorithme (5.30), on a

$$T(f, f'\cdot d) = \sum_{k=1}^{K} T(\varphi_k, \varphi'_k).$$

Ensuite, en utilisant l'inégalité de Cauchy-Schwarz généralisée (A.9), on obtient

$$\begin{aligned} T(f, f'\cdot d) &\leqslant \max_{1\leqslant k\leqslant K} \left(\frac{T(\varphi_k, \varphi'_k)}{T(\varphi_k)}\right) \sum_{k=1}^{K} T(\varphi_k) \\ &\leqslant \max_{\varphi\in\mathcal{F}} \left(\frac{T(\varphi, \varphi')}{T(\varphi)}\right) T(f) \\ &= C_{\mathcal{F}}\, T(f), \end{aligned}$$

où la constante

$$C_{\mathcal{F}} = \max_{\varphi\in\mathcal{F}} \left(\frac{T(\varphi, \varphi')}{T(\varphi)}\right)$$

ne dépend que de la classe $\mathcal{F}$ des fonctions intermédiaires. □

Par exemple si toutes les fonctions intermédiaires φ_k sont prises dans $\mathcal{F}_{_F}$ (famille des fonctions Fortran), on montre, sous des hypothèses raisonnables sur le temps respectif de l'évaluation des fonctions de $\mathcal{F}_{_F}$, que

$$\frac{T(f, f'\cdot d)}{T(f)} \leqslant 4,$$

et ceci quel que soit le nombre m de variables dérivées (voir [243 ; 1991]).

À titre d'illustration, voyons comment s'écrit le code linéaire tangent dans le cas de l'exemple de la page 278. Supposons que les deux composantes de la direction d soient données dans les variables `d1` et `d2`. En utilisant le suffixe `tl` pour désigner les variables $\dot{v}_i$ associées aux variables v_i du code et contenant leur dérivée directionnelle, le code linéaire tangent s'écrit :

```
v1tl = d1
v2tl = d2
v3tl = v1tl + 2*v2*v2tl
v3 = v1 + v2**2
v4tl = 2*v1*v1tl*sin(v3) + v1**2*cos(v3)*v3tl
v4 = v1**2 * sin(v3)
```

```
aux = v4/v3
v4tl = (v4tl - aux*v3tl)/v3
v4 = aux
aux = exp(v4)
v5tl = aux * v4tl
v5 = aux
```

On a utilisé une variable auxiliaire `aux`, de manière à éviter de calculer plusieurs fois certaines quantités.

Ce qu'il faut retenir sur le mode direct :

1. Le mode direct de différentiation est le bon mode pour calculer la dérivée d'un grand nombre de variables (ou variables de sortie) par rapport à un petit nombre de variables indépendantes (ou variables d'entrée).
2. C'est donc aussi le bon mode pour calculer une dérivée directionnelle d'une application $f : \mathbb{R}^n \to \mathbb{R}^m$.
3. À chaque variable v_i du code, on associe une variable $\dot{v}_i$ qui contient la dérivée directionnelle de v_i dans une direction d donnée. Sa valeur est nulle (et alors $\dot{v}_i$ peut ne pas exister dans le code linéaire tangent) si v_i n'est pas influencée par les variables par rapport auxquelles on dérive (c'est-à-dire si elle n'est pas active).
4. Le code linéaire tangent peut s'obtenir en « dérivant » le code original ligne par ligne. Si dans ce dernier une instruction s'interprète comme une application $\{v_i : i \in D_k\} \mapsto v_{\mu_k}$, les instructions à ajouter pour obtenir le code linéaire tangent forment une application entre variables $\dot{v}_i$ de la forme $\{\dot{v}_i : i \in D_k\} \mapsto \dot{v}_{\mu_k}$.

5.5.3 Différentiation en mode inverse

Le *mode inverse* de différentiation est moins intuitif et plus complexe à mettre en œuvre que le mode direct. Il trouve son utilité lorsqu'on veut calculer des dérivées par rapport à un grand nombre de variables. Ce mode est une méthode de génération automatique de codes adjoints, encore appelés codes linéaires cotangents, concept familier en commande optimale.

Il y a plusieurs manières d'introduire le mode inverse. Dans l'état de l'art [243 ; 1991], quatre approches ont été recensées : l'approche du graphe de calcul [535 ; 1984] [326 ; 1984] [327 ; 1984], la méthode des substitutions rétrogrades [565 ; 1980] [353 ; 1984], la méthode de dualité et l'approche de Speelpenning [565 ; 1980]. Chacune de ces approches apporte un éclairage différent sur ce mode de différentiation, toujours surprenant, mais c'est la dernière qui est la plus simple à exposer et à adapter à diverses situations. C'est aussi celle que nous allons présenter.

D'abord, il est intéressant de voir l'instruction k de (5.29) comme une transformation agissant sur l'ensemble des variables $v_1, \ldots, v_N$ du code et qui laisse inchangées toutes ces variables sauf v_{μ_k}. On peut donc lui associer une transformation $\Phi_k : \mathbb{R}^N \to \mathbb{R}^N$, définie par

$$\left(\Phi_k(v)\right)_i = \begin{cases} v_i & \text{si } i \neq \mu_k \\ \varphi_k(v_{D_k}) & \text{si } i = \mu_k. \end{cases}$$

À présent, on peut voir le programme (5.29) comme une composition de K fonctions :

$$(\Phi_K \circ \cdots \circ \Phi_1)(v).$$

Afin de lever certaines ambiguïtés, on note $v^0 = v$ et

$$v^k = (\Phi_k \circ \cdots \circ \Phi_1)(v)$$

la valeur des variables du code après exécution des k premières instructions.

Soit alors d dans $\mathbb{R}^m$, espace d'arrivée de f. Avec le mode inverse, on cherche à calculer le gradient de l'application $x \mapsto d^{\mathsf{T}} f(x)$ pour le produit scalaire euclidien, c'est-à-dire le vecteur des dérivées partielles

$$\nabla(d^{\mathsf{T}} f)(x) = \left(\frac{\partial(d^{\mathsf{T}} f)}{\partial x_i}(x)\right)_{1 \leqslant i \leqslant n}.$$

Il s'agit donc d'une dérivation cotangente. On a en notant 0_p le vecteur nul de $\mathbb{R}^p$,

$$d^{\mathsf{T}} f(x) = \begin{pmatrix} 0_{N-m} \\ d \end{pmatrix}^{\mathsf{T}} (\Phi_K \circ \cdots \circ \Phi_2 \circ \Phi_1)(x, 0_{N-n}),$$

si bien que

$$(d^{\mathsf{T}} f)'(x) \cdot y = \begin{pmatrix} 0_{N-m} \\ d \end{pmatrix}^{\mathsf{T}} \Phi_K'(v^{K-1}) \cdots \Phi_2'(v^1) \Phi_1'(v) \begin{pmatrix} I_n \\ 0_{(N-n)\times n} \end{pmatrix} y,$$

où I_n est la matrice unité d'ordre n et $0_{p\times q}$ est la matrice nulle de type $p \times q$. On en déduit

$$\nabla(d^{\mathsf{T}} f)(x) = \begin{pmatrix} I_n & 0_{n\times(N-n)} \end{pmatrix} \Phi_1'(v)^{\mathsf{T}} \Phi_2'(v^1)^{\mathsf{T}} \cdots \Phi_K'(v^{K-1})^{\mathsf{T}} \begin{pmatrix} 0_{N-m} \\ d \end{pmatrix}.$$

Dans le produit de matrices du membre de droite, on a un vecteur à droite et une matrice à gauche. Le nombre d'opérations pour l'évaluer sera donc moindre si l'on effectue les multiplications matricielles de droite à gauche. Pour effectuer ces produits, on introduit des *variables duales* $\overline{v} \in \mathbb{R}^N$, que l'on initialise au vecteur à droite de l'expression ci-dessus

$$\overline{v} := \begin{pmatrix} 0_{N-m} \\ d \end{pmatrix}.$$

Ensuite, ces variables duales sont mises à jour en effectuant les produits matriciels sus-mentionnés, c'est-à-dire

$$\overline{v} := \Phi_k(v^{k-1})^{\mathsf{T}} \overline{v}, \quad \text{pour } k = K, \ldots, 1.$$

Le dernier produit matriciel nous apprend que le gradient $\nabla(d^{\mathsf{T}} f)(x)$ se trouve dans les n premières composantes du vecteur $\overline{v}$ ainsi obtenu.

Pour écrire cette procédure de façon précise, il reste à observer que la jacobienne de $\Phi_k(v^{k-1})$ est la matrice $N \times N$

$$\Phi'_k(v^{k-1}) = \begin{pmatrix} 1 & & & & & & & \\ & \ddots & & & & & & \\ & & 1 & & & & & \\ \frac{\partial\varphi_k}{\partial v_1}(v^{k-1}) & \cdots & \cdots & \frac{\partial\varphi_k}{\partial v_{\mu_k}}(v^{k-1}) & \cdots & \cdots & \frac{\partial\varphi_k}{\partial v_N}(v^{k-1}) & \\ & & & & 1 & & & \\ & & & & & \ddots & & \\ & & & & & & 1 & \end{pmatrix},$$

où seule sa μ_k-ième ligne peut ne pas être formée de vecteurs de base de $\mathbb{R}^N$. Alors la mise à jour $\overline{v} := \Phi_k(v^{k-1})^{\mathsf{T}}\overline{v}$ s'écrit

$$\left\{ \begin{array}{ll} \overline{v}_i := \overline{v}_i & \text{si } i \notin D_k \cup \{\mu_k\} \\ \overline{v}_i := \overline{v}_i + \dfrac{\partial\varphi_k}{\partial v_i}(v^{k-1})\overline{v}_{\mu_k} & \text{si } i \in D_k \setminus \{\mu_k\} \\ \overline{v}_{\mu_k} := \dfrac{\partial\varphi_k}{\partial v_{\mu_k}}(v^{k-1})\overline{v}_{\mu_k}. & \end{array} \right. \tag{5.31}$$

Il faut d'abord mettre à jour les $\overline{v}_i$ pour $i \neq \mu_k$ avant $\overline{v}_{\mu_k}$, car il est clair que c'est l'ancienne valeur de $\overline{v}_{\mu_k}$ qu'il faut utiliser dans le calcul des $\overline{v}_i$ ($i \in D_k \setminus \{\mu_k\}$).

Ce mode de différentiation porte le nom de *mode inverse* et le code réalisant ces opérations est appelé *code linéaire cotangent*. Il calcule $\nabla(d^{\mathsf{T}}f)(x)$ par le programme suivant

$$\boxed{\begin{array}{l} \text{pour } i = 1, \ldots, N-m \\ \quad \overline{v}_i := 0 \ ; \\ \text{pour } i = N-m+1, \ldots, N \\ \quad \overline{v}_i := d_i \ ; \\ \text{pour } k = K, \ldots, 1 \\ \quad \text{pour } i \in D_k \setminus \{\mu_k\} \\ \qquad \overline{v}_i := \overline{v}_i + \frac{\partial\varphi_k}{\partial v_i}(v^{k-1})\overline{v}_{\mu_k} \ ; \\ \quad \overline{v}_{\mu_k} := \frac{\partial\varphi_k}{\partial v_{\mu_k}}(v^{k-1})\overline{v}_{\mu_k} \ ; \\ \nabla(d^{\mathsf{T}}f)(x) := (\overline{v}_1, \ldots, \overline{v}_n) \ . \end{array}} \tag{5.32}$$

On voit que cette procédure s'adapte bien à la situation où on veut dériver une seule variable ($m = 1$ et d est le scalaire 1) par rapport à beaucoup de variables d'entrée (n grand). Dans ce cas, il suffit d'initialiser toutes les variables duales à 0 sauf $\overline{v}_N$ que l'on initialise à 1.

Illustrons cela en appliquant le mode inverse sur l'exemple de la page 278, lorsqu'on veut dériver la variable `v5` par rapport aux variables `v1` et `v2`. On commence par exécuter le code original en sauvegardant certaines quantités. Le choix de l'information à sauvegarder est délicat. Pour coller à la présentation faite ci-dessus, nous avons choisi de sauvegarder les dérivées partielles $\frac{\partial\varphi_k}{\partial v_i}(v^{k-1})$ dans une pile `p` munie du pointeur `ip`, sans essayer d'optimiser le code de manière à garder une certaine lisibilité.

```
ip = 1
p(ip) = 2*v2
```

```
v3 = v1 + v2**2
ip = ip + 1
p(ip) = 2*v1*sin(v3)
ip = ip + 1
p(ip) = v1**2 * cos(v3)
v4 = v1**2 * sin(v3)
ip = ip + 1
p(ip) = 1/v3
ip = ip + 1
p(ip) = -v4/v3**2
v4 = v4/v3
ip = ip + 1
p(ip) = exp(v4)
v5 = exp(v4)
```

Ensuite vient l'exécution du code linéaire cotangent. On a utilisé le suffixe `ad` pour désigner les variables duales $\overline{v}_i$ associées aux variables v_i.

```
v1ad = 0
v2ad = 0
v3ad = 0
v4ad = 0
v5ad = 1
v4ad = v4ad + p(ip)*v5ad
v5ad = 0
ip = ip - 1
v3ad = v3ad + p(ip)*v4ad
ip = ip - 1
v4ad = p(ip)*v4ad
ip = ip - 1
v3ad = v3ad + p(ip)*v4ad
ip = ip - 1
v1ad = v1ad + p(ip)*v4ad
v4ad = 0
ip = ip - 1
v2ad = v2ad + p(ip)*v3ad
v1ad = v1ad + v3ad
v3ad = 0
```

En fin d'exécution, le gradient de `v5` par rapport à `v1` et `v2` se trouve dans (`v1ad`, `v2ad`).

On peut à présent comprendre la difficulté de la mise en œuvre du mode inverse. Il faut d'abord exécuter le code direct (5.29) afin de mémoriser les « quantités » permettant de reconstituer les dérivées partielles $\frac{\partial \varphi_k}{\partial v_i}(v^{k-1})$ au moment où celles-ci sont utilisées dans le code linéaire cotangent (5.32). Remarquons que cette utilisation se fait en ordre inverse (de $k = K, \ldots, 1$) de l'ordre de mémorisation. Il y a donc un problème de gestion de la mémoire qui ne se posait pas avec le mode direct. Une stratégie extrême consiste à mémoriser, à chaque itération k, toute l'information nécessaire à l'évaluation des dérivées partielles $\frac{\partial \varphi_k}{\partial v_i}(v^{k-1})$, $i \in D_k$. Ceci conduit en général à un besoin en place-mémoire à peu près proportionnel au nombre d'instructions où les variables actives interviennent de façon non linéaire soit, en première approximation,

proportionnel au temps d'exécution du code original (en fait cela dépend beaucoup de ce que l'on mémorise). Cette stratégie ne peut donc fonctionner que pour les petits problèmes. Pour les grands problèmes, il est préférable de mémoriser une partie des variables du code à certains instants de l'exécution de (5.29) (voir [287 ; 1992] pour une technique indépendante de la structure du code).). Dans un problème d'évolution, par exemple, on pourra ne mémoriser que les variables d'état du problème à chaque pas de temps. Ensuite les dérivées partielles $\frac{\partial \varphi_k}{\partial v_i}(v^{k-1})$ sont recalculées à partir des informations mémorisées. On parvient ainsi à obtenir des codes moins gourmands en place-mémoire, au prix d'une augmentation du temps de calcul.

En ce qui concerne, le temps d'exécution du couple (code original, code linéaire cotangent) par rapport au temps d'exécution du code original, on a le résultat suivant.

Proposition 5.32 *Le temps $T(f, \nabla(d^{\mathsf{T}} f))$ nécessaire au calcul de f et de $\nabla(d^{\mathsf{T}} f)$ par (5.29) et (5.32) vérifie l'estimation suivante :*

$$\frac{T(f, \nabla(d^{\mathsf{T}} f))}{T(f)} \leqslant C'_{\mathcal{F}}, \tag{5.33}$$

où $T(f)$ est le temps nécessaire au calcul de f par (5.29) et $C'_{\mathcal{F}}$ est une constante ne dépendant que des fonctions intermédiaires φ_k de la bibliothèque $\mathcal{F}$.

Démonstration. L'examen des algorithmes (5.29) et (5.32) permet d'écrire :

$$T(f, \nabla(d^{\mathsf{T}} f)) = \sum_{k=1}^{K} T(\varphi_k, \nabla \varphi_k).$$

On ne compte pas dans $T(f, \nabla(d^{\mathsf{T}} f))$ le temps nécessaire à l'initialisation des variables p_i. Ensuite, en utilisant l'inégalité de Cauchy-Schwarz généralisée (A.9), on obtient

$$T(f, \nabla(d^{\mathsf{T}} f)) \leqslant \max_{1 \leqslant k \leqslant K} \left(\frac{T(\varphi_k, \nabla \varphi_k)}{T(\varphi_k)} \right) \sum_{k=1}^{K} T(\varphi_k) = C'_{\mathcal{F}}\, T(f),$$

où la constante

$$C'_{\mathcal{F}} := \max_{1 \leqslant k \leqslant K} \left(\frac{T(\varphi_k, \nabla \varphi_k)}{T(\varphi_k)} \right)$$

ne dépend que de la classe $\mathcal{F}$ des fonctions intermédiaires. □

On peut donner une estimation de $C'_{\mathcal{F}}$ lorsque les fonctions φ_k sont prises dans $\mathcal{F}_{\text{F}}$ (famille des fonctions Fortran). Sous des hypothèses raisonnables concernant les temps d'exécution relatifs des fonctions Fortran, on a (voir [286 ; 1989] [243 ; 1991])

$$\frac{T(f, \nabla(d^{\mathsf{T}} f))}{T(f)} \leqslant 5.$$

Par conséquent, le calcul de $f(x)$ et de $\nabla(d^{\mathsf{T}} f(x))$ par la méthode décrite demande un temps de calcul au plus égal à 5 fois celui nécessaire au calcul de $f(x)$ et cela quel que

soit le nombre de variables et quelle que soit la complexité du programme original. Comme cela a été montré dans la démonstration ci-dessus, cette inégalité ne tient pas compte du surcoût lié aux accès-mémoire qui peuvent être importants en mode inverse. En pratique, cette estimation n'est donc pas toujours vérifiée. Cependant, même pour de grands codes, il est courant d'obtenir une borne $C'_{\mathcal{F}}$ voisine de 3.

Ce qu'il faut retenir sur le mode inverse :

1. Le mode inverse de différentiation est le bon mode pour calculer les dérivées partielles d'une seule variable (ou variable de sortie) par rapport à un grand nombre de variables (ou variables d'entrée).
2. C'est donc le bon mode pour calculer le gradient d'une application à valeurs scalaires $f : \mathbb{R}^n \to \mathbb{R}$ ou le gradient de $d^{\mathsf{T}} f$ si f est à valeurs vectorielles.
3. À chaque variable v_i du code, on associe une variable duale $\overline{v}_i$ qui contient la variation de la variable de sortie (ou de $d^{\mathsf{T}} f(x)$, pour un vecteur d donné, en cas de sortie vectorielle) par rapport à une perturbation de la variable v_i. Sa valeur est nulle si v_i n'a pas d'influence sur la variable de sortie.
4. Le code linéaire cotangent peut s'obtenir en « dualisant » le code original ligne par ligne. Cette dualisation doit se faire en ordre inverse de l'ordre d'exécution dans le code original. Si dans ce dernier une instruction s'interprète comme une application $\{v_i : i \in D_k\} \mapsto v_{\mu_k}$, les instructions correspondantes dans le code linéaire cotangent forment une application entre variables duales de la forme $(\overline{v}_i : i \in D_k \cup \{\mu_k\}) \mapsto \{\overline{v}_i : i \in D_k \cup \{\mu_k\}\}$.
5. La difficulté principale dans l'écriture de codes linéaires cotangents est la gestion-mémoire permettant de transmettre du code original au code linéaire cotangent, l'information nécessaire à l'évaluation des dérivées partielles $\frac{\partial \varphi_k}{\partial v_i}(v^{k-1})$.

5.6 Développement de codes d'optimisation

5.6.1 Communication directe et inverse ▲

La communication est bien organisée permettant à des groupes différents d'acteurs d'écrire des programmes indépendamment.

Nous appelons ci-dessous *optimiseur*, un code informatique dans lequel est implémenté un algorithme destiné à résoudre un problème d'optimisation ayant une structure bien précise. On dit que l'optimiseur est écrit en *communication directe* (CD), s'il est destiné à être utilisé comme à la figure 5.2. Dans cette figure, nous avons représenté en bleu et en pointillés la partie qui doit être écrite par l'utilisateur du code d'optimisation et en rouge et en lignes continues la partie du programme qui doit être écrite par l'auteur du code d'optimisation. L'optimiseur sera donc appelé par un programme écrit par l'utilisateur. L'optimiseur est écrit quant à lui indépendamment du problème à résoudre et lorsqu'il a besoin d'information sur ce problème, il appelle un *simulateur*, contenant le code qui évalue les objets définissant le problème et qui ont un sens pour l'optimiseur. Ainsi, en général, le simulateur évalue le critère $f(x)$ et/ou les contraintes $c(x)$ en l'itéré x fourni par l'optimiseur. C'est le fanion

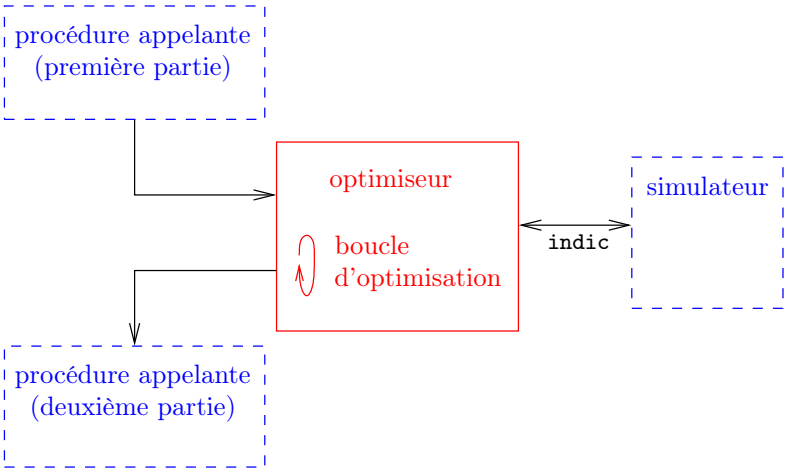


Fig. 5.2. Shéma d'un code avec un optimiseur utilisant la communication directe

`indic`, positionné par l'optimiseur, qui spécifiera ce qui lui est nécessaire. Le rôle de l'optimiseur est de mettre à jour x. Pour le dire autrement, l'optimiseur ne connaît pas le problème qu'il est occupé à résoudre et ne connaît que les objets abstraits qui décrivent la structure du problème; c'est en fait un code généraliste travaillant sur les objets mathématiques, comme nous allons le faire dans cet ouvrage. De son côté, le simulateur ne connaît pas les subtilités propres à la technique algorithmique mise en œuvre pour résoudre le problème (multiplicateur de Lagrange, fonction duale, facteur de pénalisation lui sont étrangers) et est décrit par les grandeurs physiques, chimiques, économiques ou mathématiques..., qui ont du sens pour lui.

Un optimiseur peut aussi être écrit en *communication inverse* (CI). L'organisation d'un tel code est représentée dans le schéma de la figure 5.3.

Le tableau 5.2 compare les deux types de communication. Le signe $\ominus$ signale que le mode en rapport avec ce signe est plus faible sur la caractéristique considérée que l'autre mode de communication. Le signe $\oplus$ signale le contraire : en ce qui concerne la caractéristique considérée, le mode en rapport avec ce signe est avantagé par rapport à l'autre mode. Passons en revue ces caractéristiques. L'inconvénient principal de

	communication directe (CD)	communication inverse (CI)
écriture du simulateur	$\ominus$	$\oplus$
environnement multi-langage	$\ominus$	$\oplus$
écriture de l'optimiseur	$\oplus$	$\ominus$

Tableau 5.2. Comparaison des modes de communication directe et inverse

la CD, selon nous, est de demander à l'utilisateur du code d'optimisation de devoir

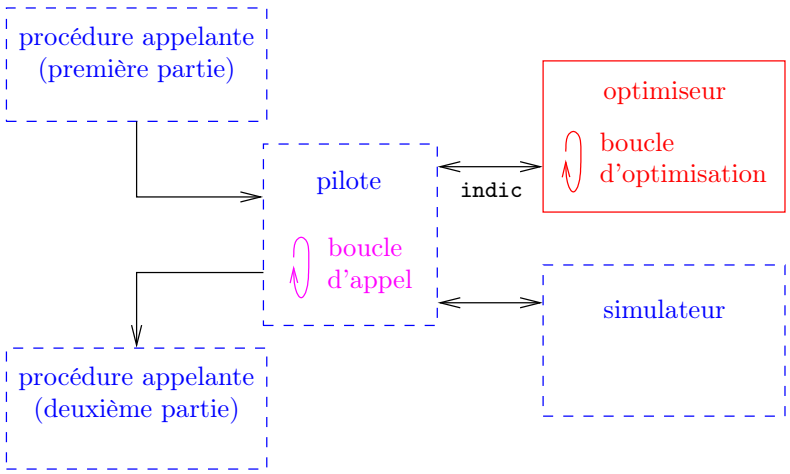


Fig. 5.3. Shéma d'un code avec un optimiseur utilisant la communication inverse

écrire un *simulateur*, c'est-à-dire une procédure à la séquence d'appel stricte réalisant les tâches demandées par l'optimiseur (calcul des fonctions définissant le problème d'optimisation et éventuellement de leurs dérivées). De plus, l'instruction d'appel de ce simulateur a sa structure figée par ce que le code d'optimisation a prévu. Cette rigidité est parfois gênante. Si des langages de programmation différents sont utilisés dans le code utilisateur et dans l'optimiseur, un seul interfaçage devra être introduit en CI et deux en CD. Il faut en effet que le programme appelant prévoit un appel de l'optimiseur dans son propre langage dans les deux modes de communication, mais en CD une interface doit aussi être prévue entre l'optimiseur et le simulateur supposés avoir été écrits dans des langages différents. Ainsi il peut être délicat d'utiliser un optimiseur écrit en Fortran et en CD dans un code Matlab.

5.6.2 Profils de performance

Comment comparer deux codes d'optimisation ? Comment savoir si une idée apporte une amélioration à un algorithme ? Ces questions ne se posent pas seulement en optimisation, mais dans tout domaine du calcul scientifique dans lequel on cherche à améliorer des solveurs généralistes. Il est coutumier de comparer les résultats des solveurs en question sur des bancs d'essai de problèmes-tests. Cela conduit à des tableaux de résultats qu'il n'est pas aisé d'analyser. Il est en effet rare qu'un solveur se comporte mieux qu'un autre sur tous les problèmes-tests de la collection considérée. Les profils de performance, que nous allons introduire, améliorent la situation en remplaçant les tableaux par des courbes dont l'interprétation est plus rapide pourvu que l'on en ait les clés de lecture.

Soient $\mathcal{S}$ un ensemble de solveurs et $\mathcal{P}$ un ensemble de problèmes-tests. Les profils de performance sont utilisés pour comparer l'efficacité *relative* des solveurs de $\mathcal{S}$ sur les problèmes-tests de $\mathcal{P}$. Soit

$$\tau_{p,s} := \text{performance du solveur } s \text{ sur le problème-test } p,$$

où la *performance* est un critère de comparaison, tel que le temps CPU ou le nombre d'évaluations de fonction ou de dérivée. Pour que ce qui suit ait un sens, il faut qu'une telle performance ait une valeur plus faible lorsque le solveur est meilleur. La *performance relative* d'un solveur s (par rapport aux autres solveurs de $\mathcal{S}$) sur un problème-test p est le rapport

$$\rho_{p,s} = \frac{\tau_{p,s}}{\tau_{p,\min}}, \quad \text{où } \tau_{p,\min} := \min\{\tau_{p,s'} : s' \in \mathcal{S}\}.$$

Bien sûr $\rho_{p,s} \geqslant 1$. Par ailleurs, un solveur s ne réussissant pas à résoudre un problème-test p devrait normalement avoir une performance $\tau_{p,s}$ infinie, ce qui ne se traite pas bien numériquement. Pour cette raison, on décide de limiter les performances relatives $\rho_{p,s}$ à une valeur maximale $\bar{\rho} > 1$ (ou de limiter les performances $\tau_{p,s}$ à $\bar{\rho}\,\tau_{p,\min}$). Dès lors, on aura $\rho_{p,s} = \bar{\rho}$, soit parce que le solveur s a en réalité sur le problème-test p une performance supérieure à $\bar{\rho}\,\tau_{p,\min}$, soit parce qu'il ne réussit pas à résoudre le problème p.

Le *profil de performance* du solveur s (par rapport aux autres solveurs de $\mathcal{S}$) est la fonction

$$t \in [1, \bar{\rho}] \mapsto \wp_s(t) := \frac{|\{p \in \mathcal{P} : \rho_{p,s} \leqslant t\}|}{|\mathcal{P}|} \in [0, 1],$$

où $|\cdot|$ désigne le cardinal d'un ensemble (le nombre de ses éléments). Cette fonction est croissante, constante par morceaux et semi-continue supérieurement. Il y a un profil pour chaque solveur de $\mathcal{S}$. La figure 5.4 donne un exemple de profils de performance,

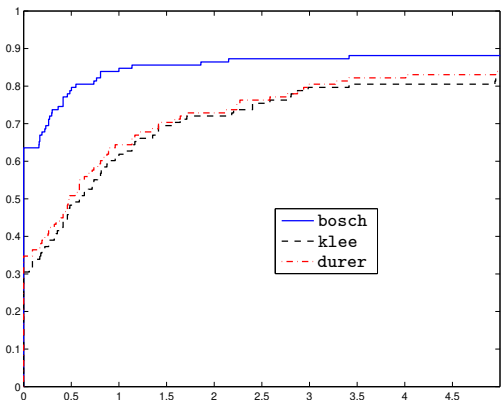


Fig. 5.4. Profils de performance typiques pour trois solveurs

correspondant à trois solveurs appelés `bosch`, `durer` et `klee`. Le solveur `bosch` apparaît plus performant que les deux autres, pour des raisons qui devraient être claires après avoir pris connaissance des clés de lecture données ci-après.

Seuls trois faits doivent être gardés à l'esprit pour avoir une interprétation correcte des profils de performance :

- $\wp_s(1)$ donne la fraction de problèmes-tests pour lesquels le solveur s est le meilleur; comme deux solveurs peuvent avoir le même score et que certains solveurs peuvent ne pas réussir à résoudre un problème-test, il n'est pas garanti que l'on ait $\sum_{s\in\mathcal{S}} \wp_s(1) = 1$;
- par définition de $\bar{\rho}$, $\wp_s(\bar{\rho}) = 1$; par ailleurs, pour $\varepsilon > 0$ petit, $\wp_s(\bar{\rho} - \varepsilon)$ donne la fraction de problèmes-tests que le solveur s peut résoudre; cette valeur est indépendante de la performance choisie pour la comparaison;
- on peut donner l'interprétation suivante de la valeur $\wp_s(t)$, obtenue en inversant l'application $t \mapsto \wp_s(t)$: pour la fraction $\wp_s(t)$ de problèmes-tests de $\mathcal{P}$, la performance du solveur s n'est jamais pire que t fois celle du meilleur solveur (celui-ci varie en général d'un problème-test à l'autre); de ce point de vue, l'argument auquel $\wp_s$ atteint sa valeur « presque » maximale $\wp_s(\bar{\rho} - \varepsilon)$ est significative.

Avec les profils de performance, l'efficacité relative des solveurs s de $\mathcal{S}$ apparaît en un coup d'œil : plus haut est le graphe de $\wp_s$, meilleur est le solveur s.

5.7 Pourquoi étudier l'optimisation numérique ?

Pourquoi est-il utile d'étudier les algorithmes d'optimisation ? La question mérite d'être posée. On pourrait en effet penser que ceci n'est plus nécessaire puisqu'il existe un certain nombre de bibliothèques de programmes d'optimisation (voir la section **??**) dans lesquels on pourra souvent trouver le code d'optimisation convenant au problème que l'on veut résoudre. Notons que ce point de vue pourrait être tenu à propos d'autres domaines de l'analyse numérique, tels que l'algèbre linéaire numérique, l'étude des schémas d'intégration des équations différentielles, les techniques de résolution des équations aux dérivées partielles, *etc.* Les remarques suivantes ne font pas le tour de la question, mais devraient permettre de nuancer cette position.

1. Face à un problème d'optimisation à résoudre, il faut faire un choix d'algorithmes. Si le problème a une structure classique (voir la section **??** pour une classification simplifiée) et ne demande pas trop de temps de calcul, ce choix pourra se faire facilement et un code de bibliothèque pourra être satisfaisant. Il est cependant nécessaire de bien comprendre les possibilités de ces codes et donc des algorithmes qui y sont implémentés pour faire un bon choix. Il faut aussi pouvoir saisir les subtilités de leur documentation pour pouvoir adapter avec pertinence leurs options au problème à résoudre, ce qui requiert également une bonne maîtrise des algorithmes implémentés.
2. Les codes d'optimisation fonctionnent rarement en boîte noire. Ceci est dû au fait qu'une solution d'un problème d'optimisation s'obtient grâce à un processus qui doit converger. Il ne s'agit donc pas d'un processus ayant un nombre déterminé d'opérations. Diagnostiquer un comportement inattendu est souvent délicat. Les codes d'optimisation bien conçus fournissent parfois ces diagnostics, mais il faut pouvoir les comprendre. Ici aussi une bonne connaissance des algorithmes est essentielle.
3. Les codes que l'on trouve dans les bibliothèques d'optimisation sont conçus pour résoudre des problèmes à la structure classique. Même si ceux-ci ont été écrits de

manière à ce qu'ils puissent être utilisés pour résoudre des problèmes variés, on gagnera souvent à les adapter au problème qui nous intéresse. Ceci est d'autant plus vrai que le problème est de grande taille, lorsque les temps de calcul sont important. Modifier les algorithmes de manière à exploiter au mieux la structure d'un problème demande de solides connaissances en algorithmique.

4. Le développement des algorithmes n'est pas terminé. Il est donc nécessaire d'enseigner ces matières aux futurs chercheurs ou ingénieurs, aux futurs numériciens qui feront progresser la discipline.

Cet ouvrage est donc destiné aux étudiants désirant se familiariser avec les techniques numériques en optimisation. Il s'adresse également aux praticiens souhaitant implémenter un algorithme ou comprendre le comportement d'un code d'optimisation sur un problème particulier. Il sera également utile aux chercheurs voulant connaître les techniques permettant d'analyser les algorithmes, d'étudier et de prévoir leur comportement. L'ouvrage contient en effet de nombreux résultats de convergence que l'on ne trouve que dans les revues spécialisées.

Notes

L'importance de la classe P a été soulignée par Edmonds [189 ; 1965] et Cobham [130 ; 1965]. Comme autres ouvrages sur la théorie de la complexité, citons [66, 475, 476].

L'utilisation de l'état adjoint pour calculer des produits hessienne-vecteur a été proposé en météorologie par Le Dimet, Navon et Daescu [383]. Le livre de Griewank (et Walther) [288, 291] et sa revue [289] constituent d'excellents états de l'art sur les techniques de différentiation automatique. Ils ont été écrits par l'un des spécialistes les plus actifs dans ce domaine. On pourra aussi se référer aux monographies [290] et [243]. Le comportement de la différentiation automatique sur des programmes contenant des processus itératifs est analysé dans [238 ; 1992]. Pour des résultats de complexité en différentiation automatique, nous renvoyons le lecteur à [449 ; 2008] et sa bibliographie.

La notion de profil de performance a été introduite par Dolan et Moré [171 ; 2002] et est maintenant couramment utilisée pour comparer l'efficacité des codes d'optimisation. Divers environnements de comparaison de codes permettent de les générer automatiquement, citons LIBOPT [242].

Exercices

5.1. Soient $\mathbb{E}$ un espace normé, dont la norme est notée $\|\cdot\|$, et $F : \mathbb{E} \to \mathbb{E}$ une application différentiable en $x_* \in \mathbb{E}$ dont la dérivée $F'(x_*)$ est inversible. Alors $F(x) \sim (x - x_*)$ au sens de (5.3).

5.2. *Vitesse de convergence en termes du nombre de chiffres significatifs corrects.* Démontrer les propositions 5.4, 5.6 et 5.11.

5.3. *Vitesse de convergence en termes d'une fonction s'annulant au point limite.* Démontrer les propositions 5.7 et 5.13.

5.4. *Convergence q-quadratique en termes de déplacements.* Démontrer la proposition 5.14.

5.5. *Suite de Fejér* [539]. Soit $\mathbb{E}$ un espace euclidien (produit scalaire $\langle\cdot,\cdot\rangle$ et norme associée $\|\cdot\|$). La *cible de Fejér* d'une suite $\{x_k\}\subseteq\mathbb{E}$ est l'ensemble

$$C(\{x_k\}) := \{x\in\mathbb{E} : \text{pour tout } k\in\mathbb{N}, \text{ on a } \|x_{k+1}-x\|\leqslant\|x_k-x\|\}.$$

On dit qu'une suite $\{x_k\}$ de $\mathbb{E}$ est une *suite de Fejér* si $C(\{x_k\})\neq\varnothing$.

1) $C(\{x_k\})$ est un convexe fermé (éventuellement vide).
2) Une suite de Fejér est bornée.
3) Si une suite $\{x_k\}$ a un point d'adhérence $\bar{x}\in C(\{x_k\})$, alors elle converge vers $\bar{x}$.
4) Si $\{x_k\}$ est suite de Fejér, alors

$$x_{k+1}\in\bigcap_{x\in C(\{x_k\})}\bar{B}(x,\|x_k-x\|).$$

Note. *Lipót Fejér* est un mathématicien hongrois (1880-1959). L'étude systématique des suites de Fejér et de leur lien avec l'optimisation a commencé avec le mathématicien russe Eremin [192, 193, 194 ; 1965-1979].

5.6. *Dérivée du critère du problème* (5.24)–(5.25). On considère le problème de commande optimale (5.24)–(5.25). Soit $f(u)$ le critère (5.25). Montrez que, pour $v:[0,T]\to\mathbb{R}^m$,

$$f'(u)\cdot v=\int_0^T\Big(J'_u(y(t),u(t),t)\cdot v(t)-p(t)^{\mathsf{T}}\varphi'_u(y(t),u(t),t)\cdot v(t)\Big)\,\mathrm{d}t,$$

où y est la solution de (5.24) et p est l'état adjoint défini comme solution de l'équation adjointe

$$\begin{cases}\dot{p}(t)=-\varphi'_y(y(t),u(t),t)^{\mathsf{T}}p(t)+\nabla_y J(y(t),u(t),t), & \text{pour tout } t\in\,]0,T[\\ p(T)=-\nabla J^T(y(T)).\end{cases}$$

5.7. *DA d'instructions à valeurs vectorielles.* Supposons que l'on ait dans le code original une instruction à valeurs vectorielles (par exemple par l'intermédiaire d'une procédure) de la forme

$$v_M:=\varphi(v_D),$$

où M et D sont des parties de $[1:N]$, pouvant être disjointes ou non et v_M désigne le vecteur de $\mathbb{R}^{|M|}$ dont les composantes sont les variables v_i du code avec indices $i\in M$ (désignation analogue pour v_D). On indice par $i\in M$ les composantes de φ. Montrez que les instructions du code linéaire tangent s'écrivent

$$\dot{v}_M:=\varphi'(v_D)\dot{v}_D,\quad v_M:=\varphi(v_D)$$

et celles du code adjoint s'écrivent

$$\overline{v}_D:=\overline{v}_{D\setminus M}+\varphi'(v_D)^{\mathsf{T}}\overline{v}_M,\quad \overline{v}_{M\setminus D}:=0. \tag{5.34}$$

5.8. *DA de solveurs linéaires.* Supposons qu'une partie d'un programme consiste à résoudre le système linéaire, ce que l'on peut écrire

$$x:=A^{-1}b,$$

où la matrice carrée inversible A et le vecteur $b\in\mathbb{R}^n$ peuvent dépendre des variables indépendantes, qu'il n'est pas nécessaire de spécifier ici. Montrez que les instructions du code linéaire tangent s'écrivent

$$x := A^{-1}b, \quad \dot{x} := A^{-1}(\dot{b} - \dot{A}x)$$

et celles du code adjoint s'écrivent

$$x := A^{-1}b, \quad \overline{x} := A^{-\mathsf{T}}\overline{x}, \quad \overline{b} := \overline{b} + \overline{x}, \quad \overline{A} := \overline{A} - \overline{x}x^{\mathsf{T}}, \quad \overline{x} := 0.$$

Que deviennent les instructions des codes tangent et adjoint si le code original s'écrit $b := A^{-1}b$ (la solution du système linéaire est placée dans b).

Remarque. Dans le cas où la matrice A a une *structure creuse* et où seuls les éléments d'indices $(i,j) \in N$ de A sont mémorisés, il n'est pas nécessaire de s'intéresser aux $\overline{A}_{ij}$ pour $(i,j) \notin N$ et on pourra ne mettre à jour que les éléments de $\overline{A}$ avec indices (i,j) dans N par $\overline{A}_{ij} := \overline{A}_{ij} - \overline{x}_i x_j$. Autrement dit, il ne faut pas introduire de variables duales associées aux éléments de A qui sont toujours nuls.

5.9. *DA d'une fonction quadratique.* Donner les codes direct et adjoint d'un code qui calcule la fonction quadratique

$$f := \frac{1}{2}\, x^{\mathsf{T}}Ax + b^{\mathsf{T}}x.$$

Partie II

Méthodes de l'optimisation sans contrainte

6 Méthodes à directions de descente

Il est aisé d'en conclure que la valeur de $f(x-\alpha\nabla f(x))$ *deviendra inférieure à* $f(x)$ *si* α *est suffisamment petit. Si, maintenant,* α *vient à croître, et si, comme nous l'avons supposé, la fonction* f *est continue, la valeur de* $f(x-\alpha\nabla f(x))$ *décroîtra jusqu'à ce qu'elle* [...] *coïncide avec une valeur minimum, déterminée par l'équation à une inconnue* $D_\alpha f(x-\alpha\nabla f(x)) = 0$. *Il suffira donc, ou de résoudre cette dernière équation, ou du moins d'attribuer à* α *une valeur suffisamment petite, pour obtenir une nouvelle valeur de* f *inférieure à* $f(x)$. *Si la nouvelle valeur de* f *n'est pas un minimum, on pourra en déduire, en opérant toujours de la même manière, une troisième valeur plus petite encore ; et, en continuant ainsi, on trouvera successivement des valeurs de* f *de plus en plus petites, qui convergeront vers une valeur minimum de* f.

L.-A. Cauchy, extrait de [109 ; 1847, page 537], récrit avec les notations du présent ouvrage.

Ce chapitre introduit une classe importante d'algorithmes de résolution des problèmes d'optimisation sans contrainte. Le concept central est celui de direction de descente (section 6.1). On le retrouvera dans des contextes variés, également pour résoudre des problèmes avec contraintes. Tous les algorithmes d'optimisation n'entrent pas dans ce cadre. Une autre classe importante de méthodes se fonde sur la notion de région de confiance qui sera vue au chapitre 9.

Après avoir décrit le fonctionnement d'un algorithme à directions de descente (section 6.1), nous donnons quelques exemples d'algorithmes de ce type (section 6.2), qui seront étudiés plus en détail dans d'autres chapitres. Nous décrivons ensuite les principales règles de recherche linéaire (section 6.3) et étudions la contribution de la recherche linéaire à la convergence et à la complexité itérative des algorithmes à directions de descente (section 6.4). Nous concluons ce chapitre par sa section 6.5, où sont énoncés des critères permettant d'estimer la qualité de la direction de descente proche d'une solution : celui de l'admissibilité asymptotique du pas unité (section 6.5.1) et celui de la convergence superlinéaire (section 6.5.2).

6.1 Principes généraux

Considérons le problème d'optimisation sans contrainte

$$\min_{x\in\mathbb{R}^n} f(x), \tag{6.1}$$

où $f : \mathbb{R}^n \to \mathbb{R}$ est supposée régulière. On supposera donné un produit scalaire $\langle\cdot,\cdot\rangle$ sur $\mathbb{R}^n$ et on notera $\|\cdot\| = \langle\cdot,\cdot\rangle^{1/2}$ la norme associée. On note aussi $\nabla f(x)$ et $\nabla^2 f(x)$ le gradient et la hessienne de f en x pour ce produit scalaire.

On s'intéresse ici à une classe d'algorithmes qui sont fondés sur la notion de direction de descente. On dit que d est une *direction de descente* de f en $x \in \mathbb{R}^n$ si

$$f'(x) \cdot d < 0.$$

Par définition du gradient (section 7.1), il revient au même de dire que $\langle \nabla f(x), d\rangle < 0$ ou encore que d fait avec l'opposé du gradient $-\nabla f(x)$ un angle θ, appelé *angle de descente*, qui est strictement plus petit que $90°$:

$$\theta := \arccos \frac{\langle -\nabla f(x), d\rangle}{\|\nabla f(x)\|\,\|d\|} \in \left[0, \frac{\pi}{2}\right[. \tag{6.2}$$

La notion d'angle définie ci-dessus dépend du produit scalaire et n'est pas invariante par rotation des vecteurs ! L'ensemble des directions de descente de f en x,

$$\{d \in \mathbb{R}^n : \langle \nabla f(x), d\rangle < 0\},$$

forme un demi-espace ouvert de $\mathbb{R}^n$ (illustration à la figure 6.1).

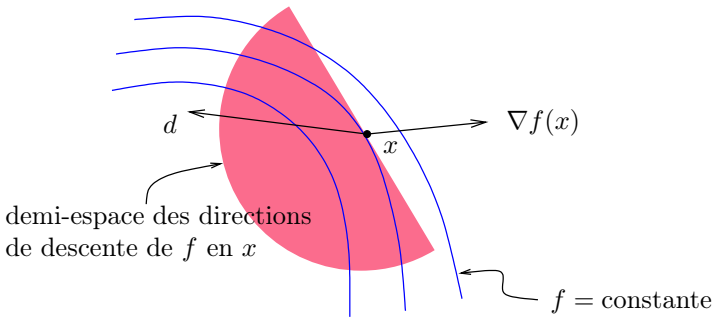


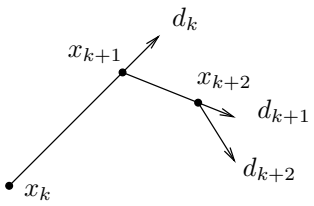
Fig. 6.1. Demi-espace (translaté) des directions de descente d de f en x.

Par définition de la dérivée, on voit que si d est une direction de descente,

$$f(x + \alpha d) < f(x), \text{ pour tout } \alpha > 0 \text{ suffisamment petit}$$

et donc que f décroît strictement dans la direction d. De telles directions sont intéressantes en optimisation car, pour faire décroître f, il suffit de faire un déplacement le

long de d. Les méthodes à directions de descente utilisent cette idée pour minimiser une fonction (voir ce qu'en disait Cauchy dans le texte en épigraphe de ce chapitre, lorsque $d = -\nabla f(x)$). Elles construisent la suite des itérés $\{x_k\}_{k\geqslant 1}$ approchant une solution x_* de (6.1) par la récurrence



$$x_{k+1} = x_k + \alpha_k d_k, \quad \text{pour } k \geqslant 1, \tag{6.3}$$

où $\alpha_k > 0$ est appelé le *pas* et d_k est une direction de descente de f en x_k. Pour définir une méthode à directions de descente il faut donc spécifier deux choses :

- dire comment la direction d_k est calculée; la manière de procéder donne le nom à l'algorithme;
- dire comment on détermine le pas α_k; c'est ce que l'on appelle la *recherche linéaire*; la section 6.3 sera consacrée à la description des principales techniques de recherche linéaire.

Décrivons cette classe d'algorithmes de manière précise.

Algorithme 6.1 (méthode à directions de descente — une itération)

On suppose qu'au début de l'itération k, on dispose d'un itéré $x_k \in \mathbb{R}^n$.

1. *Test d'arrêt* : Si $\nabla f(x_k) \simeq 0$, arrêt de l'algorithme.
2. Choix d'une direction de descente $d_k \in \mathbb{R}^n$.
3. *Recherche linéaire* : déterminer un pas $\alpha_k > 0$ le long de d_k de manière à « faire décroître f suffisamment ».
4. $x_{k+1} := x_k + \alpha_k d_k$.

Le test de l'étape 1 est discuté dans les paragraphes qui suivent. Le sens à donner à l'expression « faire décroître f suffisamment » à l'étape 3 sera précisé dans la section 6.3.

Dans les problèmes sans contrainte, il est normal que le test d'arrêt (étape 1) porte sur la petitesse du gradient : $\nabla f(x_k) \simeq 0$. C'est en effet ce que suggère la condition nécessaire d'optimalité du premier ordre $\nabla f(x_*) = 0$ (voir (4.15)). Comme x_k n'est jamais exactement égal à x_*, ce test ne pourra fonctionner que si $\nabla f(x)$ est faible en norme pour x voisin de x_*, ce qui revient pratiquement à supposer que f est de classe $\mathcal{C}^1$. Par ailleurs, un tel test d'arrêt suggère qu'un algorithme à directions de descente ne peut pas trouver mieux qu'un point stationnaire de f. C'est en effet souvent le cas, mais ce point faible est rarement rédhibitoire en pratique. On peut noter qu'il existe une version élaborée des méthodes à régions de confiance qui permet de trouver un minimum local, évitant ainsi les points stationnaires qui n'ont pas cette propriété de minimalité locale. Nous étudierons cette approche au chapitre 9.

On est parfois tenté d'arrêter l'algorithme si le critère f ne décroît presque plus. Ceci n'est pas sans risque et il vaut mieux ne pas utiliser un tel test d'arrêt, car une faible variation du critère peut se produire loin d'une solution. En effet, au premier ordre, $f(x_{k+1}) \simeq f(x_k)$ revient à $\alpha_k \langle \nabla f(x_k), d_k \rangle \simeq 0$, ce qui peut arriver si le pas α_k est petit (c'est en général très suspect) ou si la direction de descente fait avec l'opposé du gradient un angle proche de 90 degrés, une situation qui se rencontre fréquemment (si l'algorithme est bien conçu, cela traduit un mauvais conditionnement du problème).

Même si le test d'arrêt de l'étape 1 est suggéré par la théorie, on peut s'interroger sur sa pertinence, du point de vue suivant : peut-on préciser dans quelle mesure le fait d'avoir un petit gradient implique que l'itéré est proche d'un point stationnaire de f ? Le cas où f est quadratique strictement convexe est instructif :

$$f(x) = \frac{1}{2} x^{\mathsf{T}} A x - b^{\mathsf{T}} x, \qquad \text{avec } A \succ 0.$$

Minimiser f revient alors à déterminer l'unique solution x_* du système linéaire $Ax = b$. Par ailleurs, le gradient de f (pour le produit scalaire euclidien) est le résidu du système linéaire : $\nabla f(x) = Ax - b$. Or on sait bien que, si le conditionnement de A est élevé, on peut très bien avoir $\|Ax - b\|_2$ petit et une erreur $\|x - x_*\|_2$ importante. Le test d'arrêt portant sur la petitesse du gradient doit donc être interprété avec précaution. Le lemme suivant apportera un nouvel éclairage sur cette question, toujours dans le cas où la fonction est quadratique.

Lemme 6.2 (Wilkinson) *Soient $A \in \mathbb{R}^{n\times n}$ non nul, $b \in \mathbb{R}^n$ et $x \in \mathbb{R}^n$ non nul. Alors*

$$\frac{\|Ax - b\|_2}{\|A\|_2 \, \|x\|_2} = \min \left\{ \frac{\|\Delta A\|_2}{\|A\|_2} : (A + \Delta A)x = b \right\}.$$

Démonstration. Notons $r := Ax - b$. Si ΔA vérifie $(A + \Delta A)x = b$, on a $r = -(\Delta A)x$, dont on déduit $\|r\|_2 \leqslant \|\Delta A\|_2 \, \|x\|_2$ ou encore

$$\frac{\|r\|_2}{\|A\|_2 \, \|x\|_2} \leqslant \frac{\|\Delta A\|_2}{\|A\|_2}.$$

On obtient l'égalité ci-dessus en prenant $\Delta A = -r x^{\mathsf{T}} / \|x\|_2^2$, qui vérifie bien $(A + \Delta A)x = b$. □

Ce lemme nous apprend que si le gradient $\nabla f(x) = Ax - b$ est petit devant $\|A\|_2 \, \|x\|_2$, alors x est solution d'un système linéaire $(A + \Delta A)x = b$ (ou minimiseur de la fonction quadratique précédente avec A remplacé par $A + \Delta A$), avec ΔA petit devant A. On dit que l'on a interprété la précision de x en termes d'*erreur amont* (*backward error* [313]).

Certains algorithmes convergent lorsque $\alpha_k = 1$ pour tout indice k, donc sans faire de recherche linéaire. Il en est ainsi de l'algorithme proximal de la section 7.2 ou de l'algorithme de Newton étudié au chapitre 10. Mais le plus souvent, ces algorithmes ne sont définis et ne convergent que si le premier itéré x_1 est suffisamment proche d'une solution (c'est le cas pour l'algorithme de Newton) ou si la fonction possède des propriétés particulières (c'est la convexité du critère qui joue un rôle important

dans l'algorithme proximal). Pour ces algorithmes, on peut voir l'introduction du pas α_k calculé par recherche linéaire comme une technique de *globalisation de la convergence*, c'est-à-dire une technique permettant de forcer la convergence de la suite des itérés même lorsque le premier itéré est loin d'une solution. Nous verrons à la section 6.4 comment la recherche linéaire contribue en effet de manière significative à la convergence des itérés.

L'introduction d'un algorithme se fait d'ailleurs souvent comme suit. On le conçoit à partir de considérations locales (dans l'algorithme de Newton, on linéarise les équations qui déterminent le type de solution recherchée), pour qu'il ait une bonne vitesse de convergence proche d'une solution. On utilise ensuite une technique de globalisation de la convergence, comme la recherche linéaire (section 6.3) ou les régions de confiance (chapitre 9), pour forcer la convergence. Des conditions pour que la recherche linéaire n'empêche pas de retrouver l'algorithme initial local (avec $\alpha_k = 1$ donc) seront mises en évidence à la section 6.5.1.

6.2 Exemples de méthodes à directions de descente

Oublions un instant la recherche linéaire et concentrons nous sur quelques exemples de directions de descente. On note

$$g_k := \nabla f(x_k)$$

le gradient de f en x_k pour un produit scalaire $\langle \cdot, \cdot \rangle$.

6.2.1 Algorithme du gradient (ou de la plus profonde descente)

Dans cet algorithme, on prend pour direction de recherche

$$\boxed{d_k = -g_k,}$$

appelée *direction du gradient* ou *de la plus profonde descente*. Cette dernière appellation vient du fait que, si g_k est non nul, la direction est parallèle à la solution du problème en $d \in \mathbb{R}^n$ ci-dessous dans lequel on minimise le modèle affine de f (développement au premier ordre) sur une boule de rayon $\Delta > 0$ quelconque :

$$\begin{cases} \min\ f(x_k) + \langle g_k, d \rangle \\ \|d\| \leqslant \Delta. \end{cases}$$

Lorsque $\Delta > 0$ et $g_k \neq 0$, la solution de ce problème est en effet $d = -(\Delta/\|g_k\|)g_k$ (exercice 4.11).

La direction du gradient est évidemment une direction de descente si x_k n'est pas un point stationnaire ($g_k \neq 0$) puisque

$$f'(x_k) \cdot (-g_k) = \langle g_k, -g_k \rangle = -\|g_k\|^2 < 0.$$

L'algorithme qui utilise cette direction de descente porte le nom d'*algorithme du gradient* ou d'*algorithme de la plus profonde descente*.

Par son utilisation de directions de plus profonde descente et par sa simplicité de mise en œuvre, l'algorithme du gradient semble séduisant. Cependant, si le produit scalaire utilisé pour calculer le gradient n'est pas bien choisi, cet algorithme convergera très lentement. Si l'on n'a pas d'idées sur ce que doit être le bon produit scalaire, il vaudra donc mieux éviter cette méthode. On notera que, pour minimiser une fonction quadratique strictement convexe de deux variables (ce qui correspond à résoudre un système linéaire de deux équations linéaires à deux inconnues), l'algorithme demande en général un nombre infini d'itérations, alors que la solution est évidente et aisément calculable à la main ou par d'autres algorithmes en un nombre fini d'opérations. En pratique, on observe souvent que $-g_k$ est une bonne direction de descente loin d'une solution mais qu'elle est à éviter dès que l'on entre dans le voisinage d'une solution x_*, là où les termes du second ordre d'un développement de Taylor de f autour de x_* jouent un grand rôle. En fait, comme nous le verrons, le défaut de cet algorithme est d'ignorer la courbure de f, qui est décrite par sa hessienne.

Malgré ses piètres performances numériques, cet algorithme mérite d'être étudié. Les techniques utilisées pour l'analyser servent en effet souvent de guide dans l'étude d'algorithmes plus complexes. C'est à ce titre que nous en en dirons davantage à la section 7.1.

6.2.2 Algorithme du gradient conjugué

L'*algorithme du gradient conjugué* peut être vu comme une légère modification de l'algorithme du gradient puisque la direction le long de laquelle le pas α_k sera déterminé s'écrit ($k = 1$ est l'indice du premier itéré) :

$$\boxed{d_k = \begin{cases} -g_1 & \text{si} \quad k = 1 \\ -g_k + \beta_k d_{k-1} & \text{si} \quad k \geqslant 2. \end{cases}}$$

Cette direction est appelée *direction du gradient conjugué*. Le scalaire $\beta_k \in \mathbb{R}$ peut prendre différentes valeurs, ce qui donne à l'algorithme des propriétés différentes.

La forme de cette direction sera justifiée au chapitre 8 dans lequel l'algorithme est étudié en détail. Remarquons déjà que si l'on choisit α_{k-1} de manière à minimiser le critère le long de la direction précédente (c'est-à-dire si α_{k-1} minimise la fonction $\alpha \mapsto f(x_{k-1} + \alpha d_{k-1})$, ce qui implique que $\langle g_k, d_{k-1} \rangle = 0$), la direction d_k est bien de descente en un point non stationnaire ($g_k \neq 0$), puisque

$$f'(x_k) \cdot d_k = \langle g_k, d_k \rangle = -\|g_k\|^2 < 0.$$

Cette manière de déterminer le pas n'est pas acceptable en pratique lorsque le critère est non linéaire, sans propriétés particulières (voir la section 6.3.2). Nous renvoyons le lecteur à la section 8.2.6 pour une discussion sur ce sujet.

6.2.3 Algorithme de Newton

Dans l'*algorithme de Newton* pour l'optimisation sans contrainte, on détermine une direction d_k par la formule suivante :

$$\boxed{d_k = -\nabla^2 f(x_k)^{-1} g_k.}$$

Cette direction est appelée *direction de Newton*. Il faut évidemment que la hessienne de f en l'itéré courant soit inversible pour que cette définition ait un sens. Cet algorithme sera étudié en détail au chapitre 10.

Remarquons que si x_* est un minimum vérifiant les conditions d'optimalité du second ordre (CS2), $\nabla^2 f(x_*)$ est définie positive ($\langle \nabla^2 f(x_*)v, v\rangle > 0$, pour tout $v \neq 0$), et donc $\nabla^2 f(x)$ est également définie positive lorsque x est proche de x_*. Dans le voisinage d'une telle solution, d_k est bien définie et est une direction de descente puisque (on suppose aussi que $g_k \neq 0$)

$$f'(x_k) \cdot d_k = -\langle g_k, \nabla^2 f(x_k)^{-1} g_k\rangle = -f''(x_k) \cdot (\nabla^2 f(x_k)^{-1} g_k)^2 < 0.$$

6.2.4 Algorithmes de quasi-Newton

Les *algorithmes de quasi-Newton* s'inspirent de la méthode de Newton pour définir la direction de recherche. Celle-ci s'écrit :

$$\boxed{d_k = -M_k^{-1} g_k,}$$

où M_k est une matrice d'ordre n auto-adjointe (pour le produit scalaire $\langle \cdot, \cdot \rangle$). Cette matrice M_k est générée par des formules de mise à jour qui seront étudiées au chapitre 11. La direction d_k ci-dessus est appelée *direction de quasi-Newton*.

En optimisation, on s'arrangera souvent pour que M_k soit également définie positive ($\langle M_k v, v\rangle > 0$, pour tout $v \neq 0$). Dans ce cas, d_k est une direction de descente de f puisqu'avec $v_k = M_k^{-1} g_k \neq 0$, on a

$$f'(x_k) \cdot d_k = -\langle g_k, M_k^{-1} g_k\rangle = -\langle M_k v_k, v_k\rangle < 0.$$

On observera qu'alors l'*angle de descente* θ_k défini en (6.2) est contrôlé par le *conditionnement* de la matrice M_k, noté $\kappa(M_k)$. D'après l'exercice 4.12, on a en effet

$$\cos\theta_k = \frac{\langle -g_k, d_k\rangle}{\|g_k\|\,\|d_k\|} = \frac{\langle g_k, M_k^{-1} g_k\rangle}{\|g_k\|\,\|M_k^{-1} g_k\|} \geqslant \frac{\lambda_{\min}(M_k^{-1})}{\lambda_{\max}(M_k^{-1})} = \frac{1}{\kappa(M_k)}. \tag{6.4}$$

6.2.5 Algorithme de Gauss-Newton

On s'intéresse ici à un problème d'optimisation sans contrainte particulier, celui de minimiser la norme ℓ_2 d'une fonction $r : \mathbb{R}^n \to \mathbb{R}^m$ (en général $m \gg n$), dont les composantes r_i sont appelées les *résidus* :

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|r(x)\|_2^2. \tag{6.5}$$

C'est ce qu'on appelle un *problème de moindres-carrés non linéaire*.

On note $J(x) = r'(x)$ la jacobienne $m \times n$ de r en x. Alors le gradient et la hessienne du critère f de (6.5) pour le produit scalaire euclidien s'écrivent

$$\nabla f(x) = J(x)^{\mathsf{T}} r(x) \qquad \text{et} \qquad \nabla^2 f(x) = J(x)^{\mathsf{T}} J(x) + \sum_{i=1}^{m} r_i(x) \nabla^2 r_i(x).$$

Dans l'*algorithme de Gauss-Newton*, on détermine d_k comme solution particulière (il peut y en avoir plusieurs) du système linéaire

$$\left(J(x_k)^\mathsf{T} J(x_k)\right) d_k = -J(x_k)^\mathsf{T} r(x_k). \tag{6.6}$$

Si $J(x_k)$ est injective, on obtient

$$\boxed{d_k = -\left(J(x_k)^\mathsf{T} J(x_k)\right)^{-1} J(x_k)^\mathsf{T} r(x_k).}$$

Cette direction est appelée *direction de Gauss-Newton*. Comparée à la direction de Newton sur (6.5), elle n'utilise qu'une partie de la hessienne de f, de manière à éviter le calcul des dérivées secondes des résidus, qui sont souvent coûteuses à évaluer. Cet algorithme sera étudié en détail au chapitre 19.

La direction de Gauss-Newton d_k est de descente lorsque x_k n'est pas stationnaire ($J(x_k)^\mathsf{T} r(x_k) \neq 0$), puisque

$$\begin{aligned} f'(x_k) \cdot d_k = \nabla f(x_k)^\mathsf{T} d_k &= -r(x_k)^\mathsf{T} J(x_k) d_k \\ &= -d_k^\mathsf{T} \left(J(x_k)^\mathsf{T} J(x_k)\right) d_k = -\|J(x_k) d_k\|_2^2 < 0. \end{aligned}$$

La stricte négativité vient du fait que $J(x_k)d_k = 0$ impliquerait par (6.6) que $J(x_k)^\mathsf{T} r(x_k) = 0$, ce que l'on a supposé ne pas avoir lieu.

6.3 La recherche linéaire

6.3.1 Vue d'ensemble

Dans cette section, nous décrivons les techniques les plus fréquemment rencontrées pour déterminer un pas $\alpha_k > 0$ le long d'une direction de descente d_k. C'est ce que l'on appelle *faire de la recherche linéaire*. Il s'agit de réaliser deux objectifs.

Le premier objectif est de *faire décroître f suffisamment*. Cela se traduit le plus souvent par la réalisation d'une inégalité de la forme

$$f(x_k + \alpha_k d_k) \leqslant f(x_k) + \text{« un terme négatif »}. \tag{6.7}$$

Le terme négatif, disons ν_k, joue un rôle-clé dans la convergence de l'algorithme utilisant cette recherche linéaire. L'argument est le suivant. Si $f(x_k)$ est minorée (il existe une constante C telle que $f(x_k) \geqslant C$ pour tout k), alors ce terme négatif tend nécessairement vers zéro : $\nu_k \to 0$. C'est souvent à partir de la convergence vers zéro de cette suite que l'on parvient à montrer que le gradient lui-même doit tendre vers zéro. Le terme négatif devra prendre une forme bien particulière si l'on veut pouvoir en tirer de l'information. En particulier, il ne suffit pas d'imposer $f(x_k + \alpha_k d_k) < f(x_k)$.

Le second objectif de la recherche linéaire est d'*empêcher le pas $\alpha_k > 0$ d'être trop petit*, trop proche de zéro. Le premier objectif n'est en effet pas suffisant car l'inégalité (6.7) est en général satisfaite par des pas $\alpha_k > 0$ arbitrairement petit. Or ceci peut entraîner une « fausse convergence », c'est-à-dire la convergence des itérés vers un point non stationnaire, comme le montre l'observation suivante. Si l'on prend

$$0 < \alpha_k \leqslant \frac{\epsilon}{2^k \|d_k\|},$$

la suite $\{x_k\}$ générée par (6.3) est de Cauchy, puisque pour $1 \leqslant l < k$ on a

$$\|x_k - x_l\| = \left\| \sum_{i=l}^{k-1} \alpha_i d_i \right\| \leqslant \sum_{i=l}^{k-1} \frac{\epsilon}{2^i} \to 0, \quad \text{lorsque } l \to \infty.$$

Donc $\{x_k\}$ converge, disons vers un point $\bar{x}$. En prenant $l = 1$ et $k \to \infty$ dans l'estimation ci-dessus, on voit que $\bar{x} \in \bar{B}(x_1, \epsilon)$ et donc $\bar{x}$ ne saurait être solution s'il n'y a pas de solution dans $B(x_1, \epsilon)$. On a donc arbitrairement forcé la convergence de $\{x_k\}$ en prenant des pas très petits.

Pour simplifier les notations, on définit la restriction de f à la droite $\{x_k + \alpha d_k : \alpha \in \mathbb{R}\}$ comme la fonction

$$h_k : \alpha \mapsto h_k(\alpha) = f(x_k + \alpha d_k). \tag{6.8}$$

6.3.2 Recherches linéaires « exactes »

Comme on cherche à minimiser f, il semble naturel de chercher à minimiser le critère le long de d_k et donc de déterminer le pas α_k comme solution du problème

$$\min_{\alpha \geqslant 0} h_k(\alpha).$$

C'est ce que l'on appelle la *règle de Cauchy* et le pas déterminé par cette règle est appelé *pas de Cauchy* ou pas *optimal* (voir figure 6.2). Dans certains cas, on préférera le plus petit point stationnaire de h_k qui fait décroître cette fonction :

$$\alpha_k = \inf\{\alpha \geqslant 0 : h_k'(\alpha) = 0,\ h_k(\alpha) < h_k(0)\}.$$

On parle alors de *règle de Curry* et le pas déterminé par cette règle est appelé *pas de Curry* (voir figure 6.2). De manière un peu imprécise, ces deux règles sont parfois qualifiées de *recherche linéaire exacte* alors que les règles présentées plus loin sont qualifiées de *recherche linéaire inexacte*. Comme souvent en optimization, ces appellations sont trompeuses, car, comme on le verra, la recherche linéaire exacte n'est pas avantageuse et il faut lui préférer des recherches linéaires inexactes.

Ces deux règles ne sont utilisées que dans des cas particuliers, par exemple lorsque h_k est quadratique. En effet, pour une fonction non linéaire arbitraire,

- il peut ne pas exister de pas de Cauchy ou de Curry,
- la détermination de ces pas demande en général beaucoup de temps de calcul et ne peut de toute façon pas être faite avec une précision infinie,
- l'efficacité supplémentaire éventuellement apportée à un algorithme par une recherche linéaire exacte ne permet pas, en général, de compenser le temps perdu à déterminer un tel pas,
- les résultats de convergence autorisent d'autres types de règles, moins gourmandes en temps de calcul.

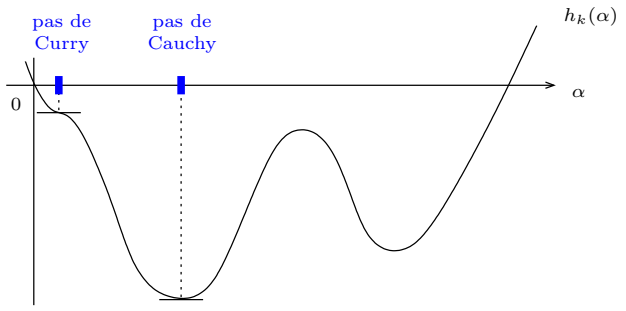


Fig. 6.2. Règles de Cauchy et de Curry.

Au lieu de demander que α_k minimise h_k, on préfère imposer des conditions moins restrictives, plus facilement vérifiées, qui permettent toutefois de contribuer à la convergence des algorithmes. En particulier, il n'y aura plus un unique pas (ou quelques pas) vérifiant ces conditions mais tout un intervalle de pas (ou plusieurs intervalles), ce qui rendra d'ailleurs leur recherche plus aisée. C'est ce que l'on fait avec les règles d'Armijo, de Goldstein et de Wolfe décrites ci-dessous.

6.3.3 Règles d'Armijo et de Goldstein

Une condition naturelle est de demander que f décroisse autant qu'une portion $\omega_1 \in]0,1[$ de ce que ferait le modèle linéaire de f en x_k. Cela conduit à l'inégalité suivante, parfois appelée *condition d'Armijo* ou *condition de décroissance linéaire*:

$$f(x_k+\alpha_k d_k) \leqslant f(x_k) + \omega_1 \alpha_k \langle g_k, d_k\rangle. \tag{6.9}$$

Elle est de la forme (6.7), car ω_1 devra être choisi dans $]0,1[$. On voit bien à la figure 6.3 ce que signifie cette condition. Il faut qu'en α_k, la fonction h_k prenne une

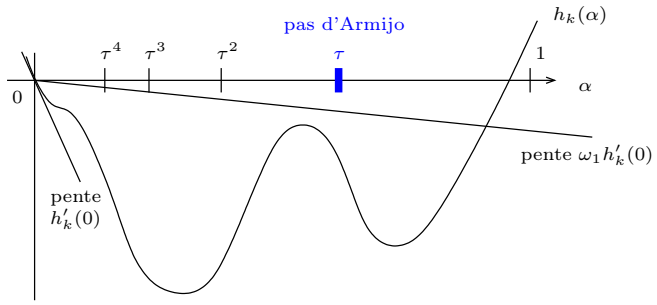


Fig. 6.3. Règle d'Armijo (version simplifiée).

valeur plus petite que celle prise par la fonction affine $\alpha \mapsto f(x_k) + \omega_1 \alpha \langle g_k, d_k\rangle$. En pratique, la constante ω_1 est prise très petite, de manière à satisfaire (6.9) le plus

facilement possible. Typiquement, $\omega_1 = 10^{-4}$. Notons que cette constante ne doit pas être adaptée aux données du problème et donc que l'on ne se trouve pas devant un choix de valeur délicat. On montrera toutefois que, dans certains algorithmes, il est important de prendre $\omega_1 < \frac{1}{2}$ pour que le pas unité ($\alpha_k = 1$) soit accepté lorsque x_k est proche d'une solution (voir par exemple la proposition 6.15).

Il est clair d'après la figure 6.3 que l'inégalité (6.9) est toujours vérifiée si $\alpha_k > 0$ est suffisamment petit. Cela se démontre aussi facilement. En effet dans le cas contraire, on aurait une suite de pas strictement positifs $\{\alpha_{k,i}\}_{i\geqslant 1}$ convergeant vers 0 lorsque $i \to \infty$ et tels que (6.9) n'ait pas lieu pour $\alpha_k = \alpha_{k,i}$. En retranchant $f(x_k)$ dans les deux membres, en divisant par $\alpha_{k,i}$ et en passant à la limite quand $i \to \infty$, on trouverait $\langle g_k, d_k\rangle \geqslant \omega_1 \langle g_k, d_k\rangle$, ce qui contredirait le fait que d_k est une direction de descente ($\omega_1 < 1$).

D'autre part, on a vu qu'il était dangereux d'accepter des pas trop petits, cela pouvait conduire à une fausse convergence. Il faut donc un mécanisme supplémentaire qui empêche le pas d'être trop petit. On utilise souvent la technique de rebroussement due à Armijo [21 ; 1966] ou celle de Goldstein [263 ; 1965].

Dans sa version la plus simple, la *technique de rebroussement* consiste à prendre $\alpha_k = \tau^{i_k}$, où $\tau \in]0,1[$ est une constante et i_k est le plus petit entier tel que l'on ait (6.9) (voir figure 6.3). C'est le fait de prendre pour α_k le plus grand réel dans $\{1, \tau, \tau^2, \ldots\}$ permettant de vérifier (6.9) qui garantit que ce pas ne sera pas trop petit. On voit bien pourquoi cette technique porte le nom de rebroussement : on essaye d'abord $\alpha_k = 1$ et si ce pas n'est pas acceptable, on rebrousse chemin en essayant des pas plus petits τ, τ^2, *etc.*

Cette version simplifiée de la technique de rebroussement est souvent améliorée dans les codes soignés. D'abord, s'il est opportun d'essayer le pas unité en premier lieu dans les algorithmes fondés sur la méthode de Newton, ce n'est pas toujours le cas pour d'autres algorithmes. Ensuite les pas intermédiaires τ^i sont imposés a priori, sans que l'on puisse tenir compte des valeurs de f calculées aux points $x_k + \tau^i d_k$. Ces valeurs peuvent en effet servir à estimer un pas α_k vérifiant (6.9). On a plus de liberté en utilisant l'algorithme ci-dessous.

Algorithme 6.3 (règle d'Armijo)

1. Choisir $\alpha_k^1 > 0$ et $\tau \in]0, \frac{1}{2}[$; $i = 1$;
2. Tant que (6.9) n'est pas vérifiée avec $\alpha_k = \alpha_k^i$:
   2.1. Choisir $\alpha_k^{i+1} \in [\tau\alpha_k^i, (1-\tau)\alpha_k^i]$;
   2.2. Accroître i de 1;
3. $\alpha_k = \alpha_k^i$.

Typiquement, on prend $\tau \in [10^{-2}, 10^{-1}]$ et le pas α_k^{i+1} est déterminé à l'étape 2.1 par interpolation (voir la section 6.3.5). Le pas déterminé par cette règle de recherche linéaire est appelé *pas d'Armijo*. Cette règle est très souvent utilisée dans l'algorithme de Newton.

La technique de rebroussement a la faiblesse de choisir le pas α_k plus petit que le pas-candidat α_k^1, qui peut être arbitraire et peut s'avérer trop petit dans certains

cas. La *règle de Goldstein* remédie à cet inconvénient. Dans celle-ci, le pas $\alpha_k > 0$ est déterminé de manière à vérifier

$$f(x_k+\alpha_k d_k) \leqslant f(x_k) + \omega_1 \alpha_k \langle g_k, d_k\rangle. \tag{6.10a}$$
$$f(x_k+\alpha_k d_k) \geqslant f(x_k) + \omega_2 \alpha_k \langle g_k, d_k\rangle, \tag{6.10b}$$

où les constantes ω_1 et ω_2 sont choisies telles que

$$0 < \omega_1 < \omega_2 < 1.$$

Typiquement, on prend $\omega_1 = 10^{-4}$ et $\omega_2 = 0.99$. On reconnait (6.9) dans (6.10a),qui assure une décroissance suffisante de f. C'est l'inégalité (6.10b) qui empêche le pas d'être trop petit. Le pas déterminé par cette règle est appelé *pas de Goldstein*. La règle de Goldstein est illustrée à la figure 6.4.

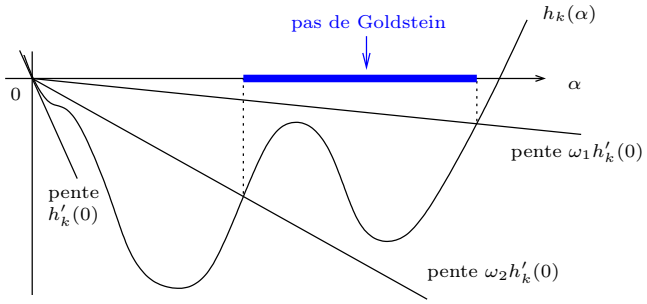


Fig. 6.4. Règle de Goldstein.

On déduit facilement du lemme ci-dessous, que l'on peut trouver un pas vérifiant (6.10) dès que h_k est continue et bornée inférieurement (le pas $\bar\alpha_k$ dont on montre l'existence convient).

Lemme 6.4 *Si $h_k : \mathbb{R}_+ \to \mathbb{R}$ définie par (6.8) est continue et bornée inférieurement, si d_k est une direction de descente de f en x_k et si $\omega_1 \in]0,1[$, alors il existe un pas $\bar\alpha_k > 0$ tel que $h_k(\bar\alpha_k) = f(x_k) + \omega_1 \bar\alpha_k \langle g_k, d_k\rangle$.*

Démonstration. Soit $A = \{\alpha > 0 : (6.9) \text{ est vérifiée pour tout } \alpha_k \in [0,\alpha]\}$. Comme $\omega_1 < 1$ et $h_k'(0) < 0$, A est non vide. D'autre part, $\bar\alpha_k = \sup A$ est fini, car h_k est bornée inférieurement et $\omega_1 > 0$. Par continuité de h_k, le résultat est vérifié avec ce pas $\bar\alpha_k$. □

La détermination d'un pas d'Armijo par rebroussement ne présente pas de difficulté particulière : l'algorithme 6.3 est explicite. À l'inverse, il n'est pas aisé de voir comment on peut trouver un pas de Goldstein en un nombre fini d'étapes. L'exercice 6.3 aborde cette question.

6.3.4 Règle de Wolfe

Dans la *règle de Wolfe*, le pas α_k est déterminé de manière à satisfaire les deux inégalités suivantes, appelées *conditions de Wolfe* (voir figure 6.5) :

$$f(x_k+\alpha_k d_k) \leqslant f(x_k) + \omega_1 \alpha_k \langle g_k, d_k\rangle, \tag{6.11a}$$

$$\langle \nabla f(x_k+\alpha_k d_k), d_k\rangle \geqslant \omega_2 \langle g_k, d_k\rangle, \tag{6.11b}$$

où les constantes ω_1 et ω_2 sont choisies telles que

$$0 < \omega_1 < \omega_2 < 1.$$

Typiquement, on prend $\omega_1 = 10^{-4}$ et $\omega_2 = 0.99$. La première inégalité n'est autre que la condition de décroissance linéaire (6.9), tandis que le rôle de (6.11b) est d'empêcher le pas d'être trop petit. Cette dernière inégalité s'écrit en effet aussi $h'_k(\alpha_k) \geqslant \omega_2 h'_k(0)$, qui n'est pas vérifiée pour $\alpha_k = 0$ (car $\omega_2 < 1$ et $h'_k(0) < 0$) et, par la continuité supposée de h'_k, n'est pas non plus vérifiée pour de petits pas $\alpha_k > 0$. Comme nous le verrons au chapitre 11, cette règle de recherche linéaire est bien adaptée aux algorithmes de quasi-Newton. Le pas déterminé par cette règle est appelé *pas de Wolfe*.

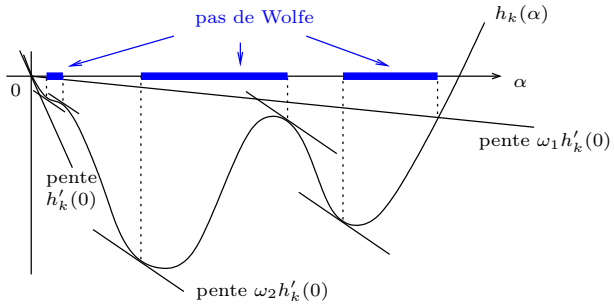


Fig. 6.5. Règle de Wolfe.

Proposition 6.5 *Si d_k est une direction de descente de f en x_k, si $h_k : \mathbb{R}_+ \to \mathbb{R}$ définie par (6.8) est dérivable et bornée inférieurement et si $0 < \omega_1 < \omega_2 < 1$, alors il existe un pas $\alpha_k > 0$ vérifiant les conditions de Wolfe (6.11a)–(6.11b).*

Démonstration. Le pas $\alpha_k = \bar{\alpha}_k$ construit dans la démonstration du lemme 6.4 convient. Il vérifie en effet (6.11a). D'autre part, par définition de $\bar{\alpha}_k$ (voir la démonstration du lemme 6.4)

$$h_k(\alpha_{k,i}) > f(x_k) + \omega_1 \alpha_{k,i} \langle g_k, d_k\rangle,$$

pour une suite de pas $\alpha_{k,i} \downarrow \bar{\alpha}_k$ quand $i \to \infty$. En retranchant $h_k(\bar{\alpha}_k) = f(x_k) + \omega_1 \bar{\alpha}_k \langle g_k, d_k\rangle$ dans les deux membres de cette inégalité, en divisant par $(\alpha_{k,i} - \bar{\alpha}_k)$ et

en passant à la limite quand $i \to \infty$, on obtient $h'_k(\bar{\alpha}_k) \geqslant \omega_1 \langle g_k, d_k \rangle \geqslant \omega_2 \langle g_k, d_k \rangle$ (car $\omega_1 \leqslant \omega_2$ et $\langle g_k, d_k \rangle < 0$). Donc (6.11b) est vérifiée avec $\alpha_k = \bar{\alpha}_k$. □

La démonstration précédente n'est pas constructive. En pratique, on utilise des algorithmes spécifiques pour trouver un pas de Wolfe. En voici un, particulièrement simple, dont on peut montrer qu'il trouve un pas de Wolfe en un nombre fini d'étapes (proposition 6.7). Il génère une suite d'intervalles $[\underline{\alpha}, \overline{\alpha}]$ emboîtés, dans lesquels il trouve finalement un pas α_k vérifiant les deux conditions de Wolfe (6.11a) et (6.11b). Les constantes τ_i et τ_e sont utilisées respectivement dans les phases d'interpolation (on cherche un nouveau pas dans l'intervalle, étapes 3 et 4.3.4) et d'extrapolation (on cherche un nouveau pas dans un intervalle non borné à droite, étape 4.3.3).

Algorithme 6.6 (Fletcher-Lemaréchal)

1. Soient $\underline{\alpha} := 0$, $\overline{\alpha} := +\infty$, $\tau_i \in]0, \frac{1}{2}[$ et $\tau_e > 1$;
   On se donne un premier pas $\alpha > 0$;
2. Répéter :
   2.1. Si (6.11a) n'est pas vérifiée avec $\alpha_k = \alpha$;
   2.2. Alors $\overline{\alpha} = \alpha$ et on choisit un nouveau pas α dans l'intervalle

$$[(1-\tau_i)\underline{\alpha} + \tau_i \overline{\alpha}, \tau_i \underline{\alpha} + (1-\tau_i)\overline{\alpha}]; \tag{6.12}$$

   2.3. Sinon
      2.3.1. Si (6.11b) est vérifiée avec $\alpha_k = \alpha$,
      2.3.2. Alors on sort avec $\alpha_k = \alpha$;
      2.3.3. Sinon
         2.3.3.1. $\underline{\alpha} = \alpha$;
         2.3.3.2. Si $\overline{\alpha} = +\infty$;
         2.3.3.3. Alors choisir un nouveau pas $\alpha \in [\tau_e \underline{\alpha}, \infty[$;
         2.3.3.4. Sinon choisir un nouveau pas α dans l'intervalle (6.12).

Lorsqu'on sort à l'étape 2.3.2, le pas α_k vérifie clairement (6.11a) et (6.11b).

Proposition 6.7 *Si d_k est une direction de descente de f en x_k, si $h_k : \mathbb{R}_+ \to \mathbb{R}$ définie par (6.8) est dérivable et bornée inférieurement et si $0 < \omega_1 < \omega_2 < 1$, alors l'algorithme de Fletcher-Lemaréchal trouve un pas $\alpha_k > 0$ vérifiant les conditions de Wolfe (6.11a)–(6.11b) en un nombre fini d'étapes.*

Démonstration. La finitude de l'algorithme se démontre par l'absurde, c.-à-d., en supposant que la boucle *Répéter* (instruction 2) est parcourue indéfiniment ou encore que l'on n'exécute jamais l'instruction 2.3.2.

Observons d'abord que l'instruction 2.3.3.3 n'est exécutée qu'un nombre fini de fois. En effet, dans le cas contraire, on aurait $\underline{\alpha} \to +\infty$ et

$$h_k(\underline{\alpha}) \leqslant h_k(0) + \omega_1 \underline{\alpha} h'_k(0). \tag{6.13}$$

Mais alors h_k ne serait pas bornée inférieurement sur $\mathbb{R}_+$ (car $\omega_1 h'_k(0) < 0$ par hypothèse), ce qui est contraire aux hypothèses. Par conséquent, après un nombre fini d'étapes, $\overline{\alpha} < +\infty$. Les nouveaux pas-candidats α sont alors toujours choisis dans l'intervalle (6.12). Comme à chaque tour de boucle, on a soit $\underline{\alpha} = \alpha$ soit $\overline{\alpha} = \alpha$, la longueur de l'intervalle $[\underline{\alpha}, \overline{\alpha}]$ tend vers zéro ($0 < \tau_i < 1/2$) et il existe un $\hat{\alpha} \geqslant 0$ tel que $\underline{\alpha} \to \hat{\alpha}$ et $\overline{\alpha} \to \hat{\alpha}$. Nous allons montrer que les propriétés de h_k en $\hat{\alpha}$ sont contradictoires.

Par continuité de h_k, en passant à la limite dans (6.13), on trouve $h_k(\hat{\alpha}) \leqslant h_k(0) + \omega_1 \hat{\alpha} h'_k(0)$. D'autre part, par les instructions 2.1 et 2.2, (6.11a) n'est pas vérifiée en $\alpha = \overline{\alpha}$:

$$h_k(\overline{\alpha}) > h_k(0) + \omega_1 \overline{\alpha} h'_k(0). \tag{6.14}$$

A la limite, on trouve $h_k(\hat{\alpha}) \geqslant h_k(0) + \omega_1 \hat{\alpha} h'_k(0)$. Avec l'inégalité précédente vérifiée par $\hat{\alpha}$, on obtient

$$h_k(\hat{\alpha}) = h_k(0) + \omega_1 \hat{\alpha} h'_k(0), \tag{6.15}$$

En $\alpha = \underline{\alpha}$, (6.11b) n'est pas vérifiée : $h'_k(\underline{\alpha}) < \omega_2 h'_k(0)$. A la limite, on trouve

$$h'_k(\hat{\alpha}) \leqslant \omega_2 h'_k(0), \tag{6.16}$$

En retranchant (6.15) de (6.14), en divisant le résultat par $\overline{\alpha} - \hat{\alpha} > 0$ et en passant à la limite, on trouve

$$h'_k(\hat{\alpha}) \geqslant \omega_1 h'_k(0). \tag{6.17}$$

Les inégalités (6.16), (6.17), $h'_k(0) < 0$ et $\omega_1 < \omega_2$ sont contradictoires. □

Pour certains algorithmes (par exemple le gradient conjugué non linéaire, voir la section 8.2.6), il est parfois nécessaire d'avoir une condition plus restrictive que (6.11b). Dans la *règle de Wolfe forte*, on cherche un pas $\alpha_k > 0$ tel que l'on ait :

$$\begin{aligned} f(x_k + \alpha d_k) &\leqslant f(x_k) + \omega_1 \alpha_k \langle g_k, d_k \rangle \\ |\langle \nabla f(x_k + \alpha d_k), d_k \rangle| &\leqslant \omega_2 \, |\langle g_k, d_k \rangle|. \end{aligned}$$

6.3.5 Mise en œuvre

Choix du premier pas

Certaines directions de descente ont un « pas naturel », un pas qui a de grandes chances d'être accepté par les différentes conditions définissant les règles de recherche linéaire vues aux sections 6.3.3 et 6.3.4. Il en est ainsi des directions newtoniennes (sections 6.2.3 et 6.2.4) pour lesquelles le pas naturel est 1. C'est ce que nous verrons à la section 6.5.1. Ce pas sera alors essayé en premier lieu et, puisqu'il est souvent accepté, permettra de faire de la recherche linéaire sans trop d'évaluations de fonction.

D'autres directions n'ont pas de pas naturel évident. Il en est ainsi des directions du gradient conjugué non linéaire et de Gauss-Newton (sections 6.2.2 et 6.2.5). Pour de telles directions de descente d_k, on détermine parfois le premier pas à partir d'un modèle quadratique

$$\alpha \mapsto \varphi_k(\alpha) = a_{0,k} + a_{1,k}\, \alpha + \frac{a_{2,k}}{2}\, \alpha^2,$$

interpolant $\alpha \mapsto f(x_k + \alpha d_k)$ à partir de la donnée de $f_k := f(x_k)$, de $p_k := f'(x_k) \cdot d_k$ et de la *décroissance attendue* du critère à l'itération k, notée Δ_k. On préfère cette dernière quantité à la dérivée directionnelle seconde $f''(x_k) \cdot d_k^2$, de manière à éviter le calcul généralement coûteux de cette dernière et à assurer la stricte convexité du modèle quadratique ($f''(x_k) \cdot d_k^2$ peut en effet ne pas être strictement positif). Les conditions naturelles d'interpolation $\varphi_k(0) = f_k$ et $\varphi_k'(0) = p_k$ permettent de donner une valeur aux deux premiers coefficients :

$$a_{0,k} = f_k \qquad \text{et} \qquad a_{1,k} = p_k < 0.$$

La courbure $a_{2,k}$ est déterminée en imposant que la décroissance maximale de φ_k, qui vaut $\varphi_k(0) - \inf \varphi_k = p_k^2/(2a_{2,k})$, soit égale à Δ_k :

$$a_{2,k} = \frac{p_k^2}{2\Delta_k} > 0.$$

On prend alors comme premier pas à essayer dans la recherche linéaire, celui qui donne la décroissance maximale du modèle quadratique φ_k de f ainsi obtenu, à savoir

$$\alpha_k := \frac{-2\Delta_k}{f'(x_k) \cdot d_k}. \tag{6.18}$$

Ce pas est appelé le *pas de Fletcher*. Il est aussi parfois utilisé à la première itération des algorithmes quasi-newtoniens (chapitre 11), car le pas unité le long de la première direction $-\nabla f(x_1)$ n'est pas approprié.

Le pas de Fletcher reporte la détermination du premier pas sur celle de la décroissance attendue Δ_k du critère et on peut légitimement se demander si l'on a progressé. C'est le cas si l'on a une estimation $f_{\min}$ de la valeur minimale de f (par exemple, on peut prendre $f_{\min} \simeq 0$ dans les problèmes de moindres-carrés non linéaires à résidu optimal presque nul). Dans ce cas, il est raisonnable de déterminer α_k par (6.18) avec

$$\Delta_k = \gamma\Big(f(x_k) - f_{\min}\Big),$$

où γ est de l'ordre de 10^{-2} ou 10^{-1}. À défaut d'information venant du problème à résoudre, certains développeurs choisissent parfois $\Delta_k = f(x_{k-1}) - f(x_k)$, qui est la décroissance de f réalisée à l'itération précédente (on suppose que $k \geqslant 2$). Cette valeur n'est pas recommandée pour les algorithmes à la convergence superlinéaire, car alors $f(x_k) - f(x_{k+1})$ est asymptotiquement beaucoup plus petit que $f(x_{k-1}) - f(x_k)$.

Interpolation et extrapolation ▲

Contrôle du nombre d'essais de pas ▲

Il n'est pas judicieux de se donner a priori un nombre maximal d'essais de pas, car la recherche d'un pas satisfaisant le long d'une direction de descente arbitraire peut requérir de nombreux essais à certaine itération particulièrement difficile. Heureusement, c'est rarement le cas pour les directions de descente classiques. Toutefois, la situation peut se présenter; par exemple à la première itération des algorithmes

quasi-newtoniens (chapitre 11) parce que l'algorithme n'a pas eu le temps de mettre à l'échelle la direction de recherche ou encore proche de la convergence lorsque les erreurs d'arrondi prévalent.

Lorsque la recherche linéaire détermine le pas par encadrement comme dans l'algorithme de Fletcher-Lemaréchal, il semble préférable de s'arrêter lorsque l'intervalle où l'on recherche le pas devient trop petit. Cependant, ce n'est pas la longueur minimale de l'intervalle de recherche en α qui importe, mais le déplacement en x correspondant; cela permet en effet de ne pas faire dépendre la longueur minimale de l'intervalle de la grandeur de la direction de descente. Ce point de vue revient en fait à se donner une *résolution en* x, c'est-à-dire une grosseur aux « pixels » discrétisant l'espace des paramètres x : on considérera alors que l'on ne peut pas distinguer deux points appartenant à un même pixel. En pratique, cette résolution sera donnée par l'utilisateur du code, qui devrait bien connaître l'ordre de grandeur des *variations* des paramètres à optimiser x, au moyen d'un vecteur

$$\delta \in \mathbb{R}^n_{++},$$

dont les composantes sont petites et strictement positives. Un *pixel* est alors un parallélipipède rectangle dont les côtés ont pour longueur les $\delta_i > 0$. On décide que la recherche linéaire ne peut pas distinguer deux points x et $y \in \mathbb{R}^n$ tels que, pour tout i, $|x_i - y_i| \leqslant \delta_i$. La longueur minimale de l'intervalle de recherche du pas dans la direction d sera alors déterminé par

$$\min_{1 \leqslant i \leqslant n} \frac{\delta_i}{|d_i|}.$$

La donnée des δ_i permet aussi de faire une mise à l'échelle du problème, de telle sorte que les nouvelles variables $\tilde{x}_i := t\, x_i/\delta_i$ aient des *variations* nominales identiques. Dans la transformation précédente, la constante $t > 0$ peut être choisie de manière à avoir la plupart des $\tilde{x}_i$ de l'ordre de l'unité.

6.4 Convergence des méthodes à directions de descente

6.4.1 Condition de Zoutendijk

Dans cette section, on étudie la *contribution* de la recherche linéaire à la convergence des algorithmes à directions de descente et à leur complexité itérative. Ce n'est qu'une contribution, parce que la recherche linéaire ne peut à elle seule assurer la convergence des itérés. On comprend bien que le choix de la direction de descente joue aussi un rôle. Cela se traduit par une condition, dite de Zoutendijk, dont on peut tirer quelques informations qualitatives intéressantes.

On dit qu'une règle de recherche linéaire satisfait la *condition de Zoutendijk* s'il existe une constante $C_z > 0$ telle que pour tout indice $k \geqslant 1$ on ait

$$f(x_{k+1}) \leqslant f(x_k) - C_z \|g_k\|^2 \cos^2 \theta_k, \tag{6.19}$$

où θ_k est l'*angle de descente*, celui que fait d_k avec $-g_k$, que nous avons défini en (6.2) par

$$\cos\theta_k = \frac{\langle -g_k, d_k\rangle}{\|g_k\|\,\|d_k\|}. \tag{6.20}$$

La proposition suivante et les commentaires qui suivent montrent comment on se sert de la condition de Zoutendijk pour établir la convergence d'algorithmes à direction de descente et leur complexité intérative.

Proposition 6.8 (utilité de la condition de Zoutendijk) *Si la suite $\{x_k\}$ générée par un algorithme d'optimisation vérifie la condition de Zoutendijk (6.19) et si $f_* := \inf_k f(x_k)$ est fini, alors*

$$\sum_{k\geqslant 1} \|g_k\|^2 \cos^2\theta_k \leqslant \frac{f(x_1) - f_*}{C_z} < \infty. \tag{6.21}$$

En particulier, s'il existe une constante strictement positive γ minorant $\{\cos\theta_k\}$, alors $g_k \to 0$ et, pour tout $\varepsilon > 0$, on obtient $\|g_k\| \leqslant \varepsilon$ pour un indice k inférieur à

$$K_\varepsilon := \left\lceil \frac{f(x_1) - f_*}{\gamma^2 C_z}\,\varepsilon^{-2} \right\rceil. \tag{6.22}$$

Démonstration. En sommant les inégalités (6.19), on a

$$\sum_{k=1}^{K} \|g_k\|^2 \cos^2\theta_k \leqslant \frac{f(x_1) - f(x_{K+1})}{C_z} \leqslant \frac{f(x_1) - f_*}{C_z}, \tag{6.23}$$

car $f(x_{K+1}) \geqslant f_*$. On a obtenu (6.21).

Supposons à présent qu'il existe une constant strictement positive γ telle que $\cos\theta_k \geqslant \gamma$ pour tout $k \geqslant 1$. Alors (6.21) montre que la série $\sum_{k\geqslant 1} \|g_k\|^2$ est convergente, ce qui implique que son terme générique $\|g_k\|^2$ (et donc g_k) tend vers zéro. De plus, par (6.23), on a

$$\sum_{k=1}^{K} \|g_k\|^2 \leqslant \frac{f(x_1) - f_*}{\gamma^2 C_z}$$

ou encore

$$\min_{k\in[1\,:\,K]} \|g_k\|^2 \leqslant \frac{f(x_1) - f_*}{\gamma^2 C_z K}.$$

Si le membre de droite de cette inégalité est inférieur à ε^2, ce qui revient à dire que K est supérieur au K_ε donné dans l'énoncé, le membre de gauche l'est aussi ; autrement dit un des $\|g_k\|$ pour $k \in [1\,:\,K_\varepsilon]$ est inférieur à ε. □

Considérons la méthode du gradient. Dans celle-ci, $\cos\theta_k = 1$ et donc sous les hypothèses de la proposition précédente, $g_k \to 0$. On obtient donc directement la « convergence » de la méthode du gradient (on ne peut montrer la convergence des itérés eux-mêmes que dans de rares cas). Plus généralement, si d_k fait avec $-g_k$ un angle θ_k qui ne se rapproche pas de $\frac{\pi}{2}$ (on veut dire par là que le $\cos\theta_k$ reste uniformément positif), l'algorithme est convergent ($g_k \to 0$). On est dans ce dernier

cas, lorsque $d_k = -M_k^{-1} g_k$ avec une matrice M_k définie positive de conditionnement borné (utiliser la minoration (6.4)).

L'inégalité (6.22) exprime la *complexité itérative* des algorithmes à directions de descente : le nombre d'itérations requis pour obtenir un gradient de norme inférieure à un $\varepsilon > 0$ donné est en $O(\varepsilon^{-2})$. Ce n'est pas une estimation époustouflante, puisque pour atteindre une précision de 10^{-10} sur la norme du gradient, elle nous dit qu'il faudra de l'ordre de 10^{20} itérations, de quoi dissuader tout utilisateur potentiel de ces méthodes. En réalité, c'est une estimation très pessimiste, car pour la plupart des problèmes rencontrés, la convergence est obtenue beaucoup plus rapidement que cela. L'estimation (6.22) est remarquable toutefois, car elle ne fait pas intervenir la dimension $n := \dim \mathbb{E}$ de l'espace sur lequel f est définie.

La démarche que l'on suit pour montrer la convergence d'un algorithme avec recherche linéaire peut à présent être schématisée. Si $\{f(x_k)\}$ est minorée (ce qui est une hypothèse très raisonnable lorsqu'on cherche à minimiser une fonction), le « terme négatif » dans (6.7) doit tendre vers zéro (c'est l'argument utilisé dans la démonstration de la proposition 6.8). Grâce à la recherche linéaire, on parvient à majorer ce terme par une constante positive fois $(-\|g_k\|^2 \cos^2 \theta_k)$ (c'est ce que montrerons les propositions 6.10, 6.11 et 6.13, ci-dessous). Dès lors $\|g_k\| \cos \theta_k \to 0$. Si le cosinus de θ_k reste uniformément positif, $g_k \to 0$ et l'algorithme « converge ». A posteriori, ceci explique aussi pourquoi on a besoin de trouver un pas tel que l'on ait un peu plus que l'inégalité $f(x_{k+1}) < f(x_k)$: c'est la convergence vers zéro du terme négatif faisant décroître $f(x_k)$ qui contribue à la convergence. S'il n'y a pas de terme négatif, on ne peut en général rien dire sur la convergence de l'algorithme.

Il faut se garder de déduire de cette observation, que l'on dispose à présent d'un moyen efficace pour forcer la convergence d'un algorithme à directions de descente, à savoir celui qui empêcherait les directions d_k de faire avec l'opposé du gradient un angle trop proche de $\frac{\pi}{2}$. Ce serait une bien mauvaise idée. En effet, les directions générées par un algorithme d'optimisation sont en général obtenues à partir de considérations cherchant à ce que l'on ait une convergence rapide vers la solution. Pour obtenir cette propriété, l'algorithme doit parfois construire des directions avec un $\cos \theta_k$ proche de zéro. On sent donc bien qu'il est délicat de donner un seuil à partir duquel la direction d_k devrait être « redressée ». Suivre cette approche pourrait brider un algorithme efficace. Il faut aussi nuancer ce dernier propos, car les méthodes à régions de confiance (qui seront étudiées au chapitre 9) ont un mécanisme qui redresse la direction de déplacement vers l'opposé du gradient dans les situations critiques, mais cette opération est la conséquence d'un autre principe algorithmique qui a tout son sens.

Les propositions 6.10, 6.11, 6.12 et 6.13 ci-dessous précisent les circonstances dans lesquelles la condition de Zoutendijk (6.19) est vérifiée avec les règles de Cauchy, de Curry, d'Armijo, de Goldstein et de Wolfe. On y utilisera souvent l'estimation suivante de l'erreur commise par le modèle linéaire d'une fonction $\mathcal{C}_L^{1,1}$.

Lemme 6.9 (erreur quadratique du modèle linéaire) *Soient Ω un ouvert de $\mathbb{E}$ et $f : \Omega \to \mathbb{R}$ une fonction $\mathcal{C}_L^{1,1}$. Alors pour tout x et $y \in \Omega$, on a*

$$|f(y) - f(x) - f'(x) \cdot (y-x)| \leqslant \frac{L}{2}\|y-x\|^2. \tag{6.24}$$

DÉMONSTRATION. Comme $f \in \mathcal{C}_L^{1,1}(\Omega)$, on a

$$f(y) - f(x) - f'(x)(y-x) = \int_0^1 [f'(x+t(y-x)) - f'(x)](y-x)\,\mathrm{d}t.$$

Dès lors

$$\begin{aligned}
|f(y) - f(x) - f'(x)(y-x)| &= \left|\int_0^1 [f'(x+t(y-x)) - f'(x)](y-x)\,\mathrm{d}t\right| \\
&\leqslant \int_0^1 \|[f'(x+t(y-x)) - f'(x)](y-x)\|\,\mathrm{d}t \\
&\leqslant \int_0^1 \|f'(x+t(y-x)) - f'(x)\|\,\|y-x\|\,\mathrm{d}t \\
&\leqslant \int_0^1 Lt\|y-x\|^2\,\mathrm{d}t \\
&= \frac{L}{2}\|y-x\|^2. \qquad \square
\end{aligned}$$

Proposition 6.10 (condition de Zoutendijk pour la règle de Curry) *Soit $f : \mathbb{R}^n \to \mathbb{R}$ une fonction différentiable sur un voisinage de $\mathcal{N}_1 := \{x \in \mathbb{R}^n : f(x) \leqslant f(x_1)\}$ et $\mathcal{C}_L^{1,1}$ sur $\mathcal{N}_1$. On considère un algorithme à directions de descente d_k dans lequel le pas α_k est déterminé de manière à avoir*

$$f(x_k+\alpha_k d_k) \leqslant f(x_k+\hat{\alpha}_k d_k),$$

où $\hat{\alpha}_k$ est le pas de Curry (que l'on suppose donc exister). Alors, pour tout $k \geqslant 1$, la condition de Zoutendijk (6.19) est vérifiée avec la constante

$$C_z = \frac{1}{2L}.$$

DÉMONSTRATION. Par la règle de Curry, $f(x_k+\hat{\alpha}_k d_k) \leqslant f(x_k+\alpha d_k)$, pour tout $\alpha \in [0, \hat{\alpha}_k]$. On utilisant l'hypothèse $f(x_k+\alpha_k d_k) \leqslant f(x_k+\hat{\alpha}_k d_k)$ et (6.24), on obtient alors pour tout $\alpha \in [0, \hat{\alpha}_k]$:

$$f(x_{k+1}) \leqslant f(x_k) + \alpha\langle g_k, d_k\rangle + \frac{\alpha^2 L}{2}\|d_k\|^2. \tag{6.25}$$

Le membre de droite atteint un minimum lorsque α prend la valeur

$$\tilde{\alpha}_k = \frac{-\langle g_k, d_k\rangle}{L\|d_k\|^2}.$$

D'autre part, $\tilde{\alpha}_k \leqslant \hat{\alpha}_k$, car par la condition de Lipschitz, on a

$$0 = f'(x_k+\hat{\alpha}_k d_k)\cdot d_k \leqslant \langle g_k, d_k\rangle + \hat{\alpha}_k L\|d_k\|^2.$$

On peut donc prendre $\alpha = \tilde{\alpha}_k$ dans l'inégalité (6.25). Ceci donne

$$f(x_{k+1}) \leqslant f(x_k) - \frac{\langle g_k, d_k\rangle^2}{2L\|d_k\|^2} = f(x_k) - \frac{1}{2L}\|g_k\|^2\cos^2\theta_k,$$

qui est l'inégalité recherchée. □

Comme la valeur prise par f au pas de Cauchy est inférieure à celle prise au pas de Curry, la conclusion de cette proposition reste vraie pour la règle de Cauchy. On pourrait également étendre les résultats ci-dessous en supposant qu'en $x_k+\alpha_k d_k$, f prend une valeur inférieure à celle prise avec le pas d'Armijo ou de Wolfe. Un tel raisonnement nous sera utile dans la démonstration du lemme 6.14.

Proposition 6.11 (condition de Zoutendijk pour la règle d'Armijo) *Soit $f : \mathbb{R}^n \to \mathbb{R}$ une fonction différentiable sur un voisinage de $\mathcal{N}_1 := \{x \in \mathbb{R}^n : f(x) \leqslant f(x_1)\}$ et $\mathcal{C}^{1,1}_L$ sur $\mathcal{N}_1$. On considère un algorithme à directions de descente d_k, qui génère une suite $\{x_k\}$ en utilisant la recherche linéaire d'Armijo avec α^1_k uniformément positif. Alors, il existe une constante $C > 0$ telle que, pour tout $k \geqslant 1$, l'une des conditions suivantes est vérifiées*

$$f(x_{k+1}) \leqslant f(x_k) - C|\langle g_k, d_k\rangle| \tag{6.26a}$$
$$f(x_{k+1}) \leqslant f(x_k) - C_z\|g_k\|^2\cos^2\theta_k, \tag{6.26b}$$

avec $C_z = 2\tau\omega_1(1-\omega_1)/L$.

Démonstration. Si le pas $\alpha_k = \alpha^1_k$ est accepté, on a (6.26a), car α^1_k est uniformément positif. Dans le cas contraire, (6.9) n'est pas vérifiée avec un pas $\bar{\alpha}_k \leqslant \alpha_k/\tau$, c'est-à-dire

$$f(x_k+\bar{\alpha}_k d_k) > f(x_k) + \omega_1\bar{\alpha}_k\langle g_k, d_k\rangle.$$

En utilisant (6.24), on obtient

$$f(x_k+\bar{\alpha}_k d_k) \leqslant f(x_k) + \bar{\alpha}_k\langle g_k, d_k\rangle + \frac{L}{2}\bar{\alpha}_k^2\|d_k\|^2.$$

En combinant les deux inégalités, en retranchant $f(x_k)+\bar{\alpha}_k\langle g_k, d_k\rangle$ des deux membres, en utilisant la définition (6.20) de θ_k et en simplifiant par $\bar{\alpha}_k\|d_k\| > 0$, on obtient

$$(1-\omega_1)\|g_k\|\cos\theta_k \leqslant \frac{L}{2}\bar{\alpha}_k\|d_k\| \leqslant \frac{L}{2\tau}\alpha_k\|d_k\|,$$

puisque $\bar{\alpha}_k \leqslant \alpha_k/\tau$. Nous avons obtenu une minoration (compliquée) du pas α_k. On exprime ensuite que f décroît suffisamment par (6.9), ce qui conduit à

$$\begin{aligned} f(x_{k+1}) &\leqslant f(x_k) - \omega_1 \alpha_k \|g_k\| \, \|d_k\| \cos\theta_k \\ &\leqslant f(x_k) - \frac{2\tau\omega_1(1-\omega_1)}{L} \|g_k\|^2 \cos^2\theta_k. \end{aligned}$$

On en déduit (6.19) avec la constante C_z donnée dans l'énoncé. □

Proposition 6.12 (condition de Zoutendijk pour la règle de Goldstein) *Soit $f : \mathbb{R}^n \to \mathbb{R}$ une fonction différentiable sur un voisinage de $\mathcal{N}_1 := \{x \in \mathbb{R}^n : f(x) \leqslant f(x_1)\}$ et $\mathcal{C}^{1,1}_L$ sur $\mathcal{N}_1$. On considère un algorithme à directions de descente d_k, qui génère une suite $\{x_k\}$ en utilisant la recherche linéaire de Goldstein (6.10). Alors, pour tout $k \geqslant 1$, la condition de Zoutendijk (6.19) est vérifiée avec la constante*

$$C_z = \frac{2\omega_1(1-\omega_2)}{L}.$$

Démonstration. Exprimons d'abord que le pas α_k n'est pas trop petit en utilisant (6.10b) et (6.24) :

$$f(x_k) + \omega_2 \alpha_k \langle g_k, d_k\rangle \leqslant f(x_{k+1}) \leqslant f(x_k) + \alpha_k \langle g_k, d_k\rangle + \frac{L}{2}\alpha_k^2 \|d_k\|^2.$$

En retranchant $f(x_k) + \alpha_k \langle g_k, d_k\rangle$ des deux membres extrêmes, en utilisant la définition (6.20) de θ_k et en simplifiant par $\alpha_k \|d_k\| > 0$, on obtient que

$$(1-\omega_2)\|g_k\| \cos\theta_k \leqslant \frac{L}{2}\alpha_k \|d_k\|.$$

qui est bien une minoration (compliquée) du pas α_k. On exprime ensuite que f décroît suffisamment par (6.10a), ce qui conduit à

$$\begin{aligned} f(x_{k+1}) &\leqslant f(x_k) - \omega_1 \alpha_k \|g_k\| \, \|d_k\| \cos\theta_k \\ &\leqslant f(x_k) - \frac{2\omega_1(1-\omega_2)}{L} \|g_k\|^2 \cos^2\theta_k. \end{aligned}$$

On en déduit (6.19) avec la constante C_z donnée dans l'énoncé. □

Proposition 6.13 (condition de Zoutendijk pour la règle de Wolfe) *Soit $f : \mathbb{R}^n \to \mathbb{R}$ une fonction différentiable sur un voisinage de $\mathcal{N}_1 := \{x \in \mathbb{R}^n : f(x) \leqslant f(x_1)\}$ et $\mathcal{C}^{1,1}_L$ sur $\mathcal{N}_1$. On considère un algorithme à directions de descente d_k, qui génère une suite $\{x_k\}$ en utilisant la recherche linéaire de Wolfe (6.11). Alors, pour tout $k \geqslant 1$, la condition de Zoutendijk (6.19) est vérifiée avec la constante*

$$C_z = \frac{\omega_1(1-\omega_2)}{L}.$$

DÉMONSTRATION. Exprimons d'abord que le pas α_k n'est pas trop petit en utilisant (6.11b), qui permet d'écrire

$$(1-\omega_2)\,|\langle g_k, d_k\rangle| \leqslant \langle g_{k+1} - g_k, d_k\rangle.$$

Comme f est $\mathcal{C}^{1,1}_L$, on en déduit

$$(1-\omega_2)\|g_k\|\cos\theta_k \leqslant L\alpha_k\|d_k\|,$$

qui est bien une minoration (compliquée) du pas α_k. On exprime ensuite que f décroît suffisamment par (6.11a), ce qui conduit à

$$\begin{aligned} f(x_{k+1}) &\leqslant f(x_k) - \omega_1\alpha_k\|g_k\|\,\|d_k\|\cos\theta_k \\ &\leqslant f(x_k) - \frac{\omega_1(1-\omega_2)}{L}\|g_k\|^2\cos^2\theta_k. \end{aligned}$$

On en déduit (6.19) avec la constante C_z donnée dans l'énoncé. □

6.4.2 Suites minimisantes spéciales

Grâce à la condition de Zoutendijk, en appliquant une des règles de recherche linéaire décrites à la section 6.3 à l'algorithme du gradient, on peut obtenir des suites minimisantes ayant la propriété supplémentaire d'avoir un gradient qui tend vers zéro.

Lemme 6.14 (suite minimisante spéciale) *Soit $f : \mathbb{R}^n \to \mathbb{R}$ une fonction bornée inférieurement, dérivable dans le voisinage d'un de ses ensembles de sous-niveau $\mathcal{N}$ et de dérivée lipschitzienne sur $\mathcal{N}$. Alors, on peut trouver une suite minimisante $\{x_k\} \subseteq \mathcal{N}$ de f telle que $f'(x_k) \to 0$.*

DÉMONSTRATION. Soit $\{x'_k\}$ une suite minimisante de f incluse dans $\mathcal{N}$. On construit $\{x_k\} \subseteq \mathcal{N}$ comme suit. On prend $x_1 = x'_1$. Soit alors x_k donné ($k \geqslant 1$). Si $\nabla f(x_k) = 0$, on prend $x_{k+1} = x'_{k+1}$. Sinon on calcule d'abord $x''_{k+1} = x_k - \alpha_k\nabla f(x_k)$, où le pas $\alpha_k > 0$ est déterminé par recherche linéaire de Wolfe (par exemple); et on prend pour x_{k+1} le point x'_{k+1} ou x''_{k+1} donnant la plus petite valeur de f. Montrons qu'une sous-suite de la suite $\{x_k\}$ ainsi construite convient.

Observons que, comme $f(x_k) \leqslant f(x'_k)$ pour tout k, la suite $\{x_k\}$ est aussi minimisante. Par ailleurs, si, pour une sous-suite d'indices k, on a $\nabla f(x_k) = 0$, le résultat est démontré en sélectionnant cette sous-suite. Sinon, pour des indices assez grands, on a $f(x_{k+1}) \leqslant f(x''_{k+1}) \leqslant f(x_k) - \omega_1\alpha_k\|\nabla f(x_k)\|^2$ et $\langle\nabla f(x''_{k+1}), -\nabla f(x_k)\rangle \geqslant \omega_2\langle\nabla f(x_k), -\nabla f(x_k)\rangle$. En raisonnant comme dans la démonstration de la proposition 6.13, on montre que $\alpha_k \geqslant (1-\omega_2)/L > 0$, où L est la constante de Lipschitz de f'. Dès lors $\nabla f(x_k) \to 0$. □

6.5 Propriétés asymptotiques

6.5.1 Admissibilité asymptotique du pas unité

On s'intéresse dans cette section à la mise en évidence de conditions assurant que le pas unité $\alpha_k = 1$ est accepté asymptotiquement par les différentes règles de recherche linéaire présentées dans la section 6.3. Par exemple, lorsqu'on considère la règle de décroissance suffisante (6.9), on cherche à savoir quand est-ce que l'on peut garantir que l'on a

$$f(x_k + d_k) \leqslant f(x_k) + \omega_1 \langle \nabla f(x_k), d_k \rangle,$$

pour k suffisamment grand et lorsque x_k converge vers un minimum x_* de f. On cherche des conditions sur le type de minimum x_*, sur d_k et sur les constantes intervenant dans l'inégalité à vérifier (ici ω_1). Cette propriété est très souhaitable, car elle permet de dire que le pas unité a quelques chances d'être accepté et que c'est donc le premier pas à essayer dans la recherche linéaire. En pratique, essayer en premier lieu le pas unité permet d'éviter de faire trop d'évaluations de fonction pendant la recherche linéaire. Nous allons montrer qu'il en est ainsi dans le voisinage d'un minimum fort (c.-à-d., vérifiant les conditions d'optimalité du second ordre de la proposition 4.11).

Sans grande perte de généralité, on peut supposer que la direction de descente est obtenue comme solution du système linéaire

$$M_k d_k = -\nabla f(x_k), \tag{6.27}$$

où M_k est une matrice d'ordre n auto-adjointe inversible. La propriété d'acceptation asymptotique du pas unité est d'ailleurs typique des algorithmes de Newton (chapitre 10), pour lesquels $M_k = \nabla f(x_k)$, et de quasi-Newton (chapitre 11). Une partie des conditions d'acceptation asymptotique du pas unité porte sur la matrice M_k qui doit satisfaire l'estimation (6.28) ci-dessous. Celle-ci est du type $t_k \geqslant o(\|u_k\|)$, ce qui veut dire qu'il doit exister une suite de réels $\{s_k\}$ tels que $t_k \geqslant s_k$ et $s_k/\|u_k\| \to 0$ lorsque $k \to \infty$. Dès lors, (6.28) est équivalente à $\langle M_k d_k, d_k \rangle \geqslant \langle \nabla^2 f(x_*) d_k, d_k \rangle + o(\|d_k\|^2)$, qui est vérifiée si les matrices M_k sont assez « grosses ». Cela n'a rien d'étonnant puisqu'alors les directions d_k sont petites (d'après (6.27)) et le pas unité le long de ces directions a en effet plus de chance d'être accepté. D'autre part, (6.28) est clairement vérifiée pour $M_k = \nabla^2 f(x_k)$, donc dans l'algorithme de Newton.

Proposition 6.15 (admissibilité asymptotique du pas unité par l'inégalité de décroissance suffisante) *Soit f une fonction de classe $\mathcal{C}^2$ dans le voisinage d'un point x_*, minimum local fort du problème (6.1). On note $\nabla f(x)$ et $\nabla^2 f(x)$ le gradient et la hessienne de f pour un même produit scalaire. Soit $\{x_k\}$ une suite générée par la récurrence $x_{k+1} = x_k + d_k$, où d_k est solution de (6.27). On suppose que $\{x_k\}$ converge vers x_* et que les matrices M_k sont auto-adjointes inversibles et vérifient la condition*

$$\langle (M_k - \nabla^2 f(x_*)) d_k, d_k \rangle \geqslant o(\|d_k\|^2), \quad \textit{lorsque } k \to \infty. \tag{6.28}$$

Alors, pour k suffisamment grand, on a

$$f(x_k + d_k) \leqslant f(x_k) + \omega_1 \langle \nabla f(x_k), d_k \rangle,$$

pourvu que $\omega_1 \in]0, \frac{1}{2}[$.

DÉMONSTRATION. On commence par développer $f(x_k + d_k)$ autour de x_k au second ordre :

$$f(x_k + d_k) = f(x_k) + \langle \nabla f(x_k), d_k \rangle + \frac{1}{2} \langle \nabla^2 f(x_k) d_k, d_k \rangle + o(\|d_k\|^2).$$

Alors, en utilisant (6.27) et (6.28), on trouve

$$\begin{aligned} & f(x_k + d_k) - f(x_k) - \omega_1 \langle \nabla f(x_k), d_k \rangle \\ &= (1 - \omega_1) \langle \nabla f(x_k), d_k \rangle + \frac{1}{2} \langle \nabla^2 f(x_k) d_k, d_k \rangle + o(\|d_k\|^2) \\ &= \left(\frac{1}{2} - \omega_1 \right) \langle \nabla f(x_k), d_k \rangle - \frac{1}{2} \langle (M_k - \nabla^2 f(x_k)) d_k, d_k \rangle + o(\|d_k\|^2) \\ &\leqslant \left(\frac{1}{2} - \omega_1 \right) \langle \nabla f(x_k), d_k \rangle + o(\|d_k\|^2). \end{aligned}$$

Le résultat sera obtenu si l'on montre qu'il existe une constante $C > 0$ telle que $\langle \nabla f(x_k), d_k \rangle = -\langle M_k d_k, d_k \rangle \leqslant -C\|d_k\|^2$ pour k grand (alors le dernier membre ci-dessus est bien négatif pour k grand). Or ceci découle de (6.28) qui entraîne

$$\langle M_k d_k, d_k \rangle \geqslant \langle \nabla^2 f(x_*) d_k, d_k \rangle + o(\|d_k\|^2) \geqslant C\|d_k\|^2,$$

puisque les conditions du deuxième ordre sont supposées être vérifiées en x_*. □

6.5.2 Conditions de convergence superlinéaire ▲

Proposition 6.16 (Dennis et Moré [166]) *Soit $F : \mathbb{R}^n \to \mathbb{R}^n$ une fonction différentiable en un point $x_* \in \mathbb{R}^n$ tel que $F(x_*) = 0$ et $F'(x_*)$ est inversible. Soit $\{x_k\}$ une suite convergeant vers x_* et vérifiant pour tout indice k :*

$$F(x_k) + M_k(x_{k+1} - x_k) = 0, \tag{6.29}$$

où $\{M_k\}$ est une suite de matrices. Alors $\{x_k\}$ converge q-superlinéairement si, et seulement si,

$$(M_k - F'(x_*))(x_{k+1} - x_k) = o(\|x_{k+1} - x_k\|). \tag{6.30}$$

DÉMONSTRATION. Par la différentiabilité de F en x_* et $F(x_*) = 0$, on a $F(x_k) = F'(x_*)(x_k - x_*) + o(\|x_k - x_*\|)$, si bien que

$$\begin{aligned}(M_k - F'(x_*))(x_{k+1} - x_k) &= -F(x_k) - F'(x_*)(x_{k+1} - x_k) \qquad [(6.29)] \\ &= -F'(x_*)(x_{k+1} - x_*) + o(\|x_k - x_*\|).\end{aligned}$$

Si $\{x_k\}$ converge superlinéairement, on en déduit directement (6.30). Réciproquement, si (6.30) a lieu, on en déduit que $F'(x_*)(x_{k+1} - x_*) = o(\|x_k - x_*\|)$ et donc la convergence superlinéaire de $\{x_k\}$ grâce à l'inversibilité de $F'(x_*)$. □

Notes

Pour les règles de recherche linéaire présentées dans ce chapitre, on pourra consulter les travaux originaux de Curry [149 ; 1944], d'Armijo [21 ; 1966], de Wolfe [623, 624 ; 1969-1971], de Zoutendijk [644 ; 1970], de Fletcher [210 ; 1980] et de Lemaréchal [388 ; 1981]. On trouvera une présentation générale des techniques de détermination de pas par encadrement au chapitre 3 de [73].

Si l'on parvient à faire tendre le gradient vers zéro sans trop de difficulté, en général, on n'a pas pour autant nécessairement la convergence *des itérés* vers un minimum de la fonction, même avec l'algorithme du gradient et une recherche linéaire raisonnable. Un contre-exemple est donné par Gonzaga [269 ; 2000]. La dernière partie de l'énoncé de la proposition 6.8 et son estimation du nombre d'itérations en $O(\varepsilon^{-2})$ pour obtenir un gradient plus petit que $\varepsilon > 0$ en norme est reprise de [452 ; 2004, p. 29].

La suite minimisante spéciale du lemme 6.14 peut aussi être obtenue en utilisant le principe variationnel d'Ekeland [190 ; 1974], au lieu de la recherche linéaire comme dans la démonstration proposée ici (voir Borwein et Zhu [80 ; 2010, lemme 2.4.2]).

De manière à accepter plus rapidement le pas le long de la direction de recherche d_k, il est parfois intéressant de faire ce que l'on appelle de la *recherche linéaire non monotone* : on remplace la condition de descente suffisante (6.9) par

$$f(x_{k+1}) \leqslant \left(\max_{0 \leqslant j \leqslant m(k)} f(x_{k-j}) \right) + \omega_1 \alpha_k \langle g_k, d_k \rangle,$$

qui est donc plus facilement vérifiée que (6.9). On doit imposer des conditions sur le décalage d'indice maximal $m(k)$ pour préserver la convergence : $m(1) = 0$, $0 \leqslant m(k) \leqslant \min[m(k-1) + 1, M]$. En pratique, $M \in \mathbb{N}$ est une petite constante entière, souvent prise égale à 3. Donc $\{f(x_k)\}$ n'est plus nécessairement décroissante (c'est dans ce sens que la recherche linéaire est dite « non monotone »). Grippo, Lampariello et Lucidi [292 ; 1986] utilisent cette RL avec l'algorithme de Newton, Lucidi et Roma [400 ; 1994] avec le gradient conjugué non linéaire.

Certains algorithmes, à l'utilité pratique discutable, déterminent un pas le long d'une direction de descente, sans chercher à faire décroître le critère f, mais en utilisant des formules déterminées par des considérations diverses. Ainsi, la règle de Barzilai-Borwein [38] utilise une estimation de la courbure de f suivant la direction de descente; les règles de [365, 3, 498, 499] utilisent des propriétés de minimalité le long de d_k (minimalité de la norme du gradient ou d'une combinaison convexe de celle-ci et du critère) et permettent d'étendre les résultats de convergence de la méthode du gradient

de la section 7.1. On ne peut démontrer des résultats de convergence que sous des hypothèses restrictives.

L'idée d'utiliser la régularisation de Moreau-Yosida, comme dans l'algorithme proximal, pour la résolution de problèmes d'optimisation mal conditionnés remonte au moins à Bellman, Kalaba et Lockett [42 ; 1966, chapitre V]. Ce principe fut généralisé par Martinet [409 ; 1970] à la minimisation de fonctions convexes et par Rockafellar [523, 524 ; 1976] à la recherche de zéros d'opérateurs monotones maximaux ainsi qu'à l'optimisation convexe. L'influence des *préconditionneurs* R_k est analysée, par exemple, par Qi et Chen [502 ; 1997], Chen et Fukushima [118 ; 1999]. On pourra aussi consulter [131]. La section 7.2.2 s'inspire largement de [317 ; section XV.4.2].

Exercices

6.1. *Directions de descente en optimisation différentiable.*

1) Les dessins de la figure 6.6 représentent une courbe de niveau (ou iso-valeur) d'une fonction quadratique $f : \mathbb{R}^2 \to \mathbb{R}$ (f est constante sur cette courbe). Dans quels dessins (A ou B ou les deux) la direction d est-elle de descente au point x ?

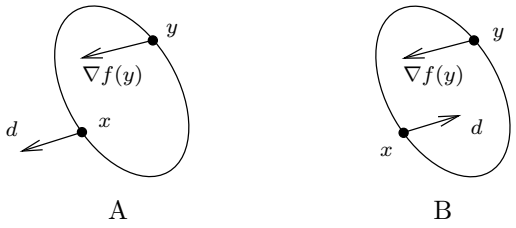


Fig. 6.6. Directions de descente ?

2) Soient $f : \mathbb{R}^n \to \mathbb{R}$ une fonction différentiable, $x \in \mathbb{R}^n$ et $\nabla f(x)$ le gradient de f en x pour un produit scalaire $\langle \cdot, \cdot \rangle$ (on note $\| \cdot \|$ la norme associée). Soit enfin $d \in \mathbb{R}^n$ une direction *non nulle* telle que

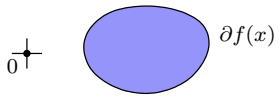
$$\|\nabla f(x) + \alpha d\| \leqslant \|\nabla f(x)\|,$$

où $\alpha > 0$. Montrez que d est une direction de descente de f en x.

6.2. *Directions de descente en optimisation non différentiable.*

1) Soient $f : \mathbb{R}^n \to \mathbb{R}$ une fonction convexe différentiable et x et y deux points de $\mathbb{R}^n$ tels que $f(y) < f(x)$. Montrez que $y - x$ est une direction de descente de f en x.

2) Soient $f : \mathbb{R}^2 \to \mathbb{R}$ une fonction convexe et $D := \{d \in \mathbb{R}^2 : f'(x; d) \leqslant 0\}$ l'ensemble des directions de descente au sens large (il n'y a pas d'inégalité stricte) de f en un point $x \in \mathbb{R}^2$. On a représenté ci-dessous le sous-différentiel $\partial f(x)$ de f en x pour le produit scalaire euclidien et l'origine 0.



a) Déterminez graphiquement l'ensemble D.
b) Montrez par un contre-exemple que $-\partial f(x)$ n'est pas nécessairement contenu dans D (autrement dit, l'opposé d'un sous-gradient n'est pas nécessairement une direction de descente).
c) Montrez que, si p est la projection de 0 sur $\partial f(x)$, alors $f'(x;-p) \leqslant -\|p\|^2$ (donc, si $0 \notin \partial f(x)$, $-p$ est une direction de descente de f en x).

3) Soient $F : \mathbb{R}^n \to \mathbb{R}^m$ une fonction différentiable et $f : \mathbb{R}^n \to \mathbb{R}$ la fonction définie par $f(x) = \|F(x)\|$, où $\|\cdot\|$ est une norme sur $\mathbb{R}^m$. Soit $d \in \mathbb{R}^n$ une direction telle que

$$\|F(x) + F'(x) \cdot d\| < \|F(x)\|.$$

Montrez que f a des dérivées directionnelles et que d est une direction de descente de f en x, dans le sens où $f'(x;d) < 0$.

6.3. *Détermination d'un pas de Goldstein.* Donnez un algorithme déterminant un pas de Goldstein (vérifiant donc (6.10)) en un nombre fini d'étapes sous des conditions semblables à celles de la proposition 6.7 (dans ce cas, on n'a besoin de la dérivabilité de f qu'en zéro et de sa continuité).

6.4. *Toute direction de descente est un gradient.* Toute direction de descente d'une fonction $f : \mathbb{E} \to \mathbb{R}$ en x, définie sur un espace vectoriel $\mathbb{E}$, est l'opposé du gradient de f en x pour un certain produit scalaire sur $\mathbb{E}$.

7 Algorithmes du premier ordre

Il n'y pas d'algorithme meilleur que tout autre quel que soit le critère de performance que l'on adopte, même dans le champ restreint de l'optimisation sans contrainte. Les critères d'appréciation sont en effet multiples : vitesse de convergence des suites générées ou de la valeur optimale, coût interne (nombre d'opérations) ou externe (nombre d'évaluations de fonctions) de l'itération, complexité itérative de l'algorithme (nombre total d'itérations pour atteindre un seuil donné), espace mémoire requis, *etc.* On trouve plutôt tout un éventail de méthodes plus ou moins bien adaptées à des problèmes particuliers. Il faut donc bien connaître les caractéristiques de ces algorithmes pour qu'en présence d'un problème donné, l'on puisse faire un choix à bon escient. Ce chapitre présente un premier ensemble de techniques qui sont toutes utiles à des degrés divers et selon le contexte ; cela sera précisé aux cours des sections qui leur sont consacrées. Elles ont la caractéristique d'être du *premier ordre*, c'est-à-dire de n'utiliser de la fonction à minimiser, que sa valeur et sa dérivée première.

L'algorithme du gradient est une bien mauvaise méthode (elle converge trop lentement), mais son analyse sert de référence à l'étude d'autres algorithmes plus complexes; nous l'aborderons à la section 7.1. Le chapitre étudie l'algorithme proximal pour la minimisation de fonctions convexes (section 7.2). Celui-ci, bien que non implémentable en pratique, sera utilisé pour interpréter certains algorithmes de dualité au chapitre 14.

Connaissances supposées. La section 7.2 sur l'algorithme proximal suppose connues les notions de sous-différentiabilité de fonctions convexes, de point proximal et de régularisée de Moreau-Yosida (sections 3.6 et 3.7).

7.1 Algorithme du gradient

On ne le répétera sans doute jamais assez : *l'algorithme du gradient est une très mauvaise méthode d'optimisation* ! La preuve en est qu'elle requiert génériquement un nombre infini d'itérations pour minimiser une fonction quadratique strictement convexe de deux variables. Sachant que ce problème est équivalent à la résolution d'un système linéaire de deux équations à deux inconnues, dont la solution peut s'obtenir manuellement par élimination d'une des deux variables, on comprend l'ampleur de l'inefficacité. Pire, si le rapport des valeurs propres du hessien de la fonction est grand (mauvais conditionnement du problème), une précision raisonnable sur la solution ne peut jamais être atteinte en un temps de calcul supportable. L'opinion la plus couramment rencontrée est qu'il est toujours préférable d'utiliser l'algorithme de BFGS à mémoire limitée (section 11.2.5), qui a un champ d'application à peu près identique

– oracle (ou simulateur) et espace mémoire semblables – et n'est guère plus difficile à implémenter. Cette opinion reste vraie selon nous, mais ne condamne pas pour autant l'algorithme du gradient. D'abord, on ne peut se passer de son étude, car elle sert de base à celle d'algorithmes plus complexes. Ensuite, il permet d'interpréter d'autres algorithmes qui ne se présentent pas dans leur conception comme un algorithme de gradient, mais qui y sont apparentés (nous pensons à la relaxation lagrangienne et à la méthode des multiplicateurs). Enfin, l'utilisation de l'optimisation dans de nouvelles applications (traitement du signal, analyse des données massives, apprentissage, *etc* [470, 568]) a remis cet algorithme et ses variantes au goût du jour.

Dans cette section, $\mathbb{E}$ est un espace euclidien, dont le produit scalaire est noté $\langle \cdot, \cdot \rangle$ et la norme associée est notée $\| \cdot \|$. L'espace vectoriel $\mathbb{E}$ étant supposé de dimension finie, le nombre de variables des problèmes d'optimisation posés sur $\mathbb{E}$ en est la dimension

$$n := \dim \mathbb{E}.$$

7.1.1 Définition

On s'intéresse dans cette section au problème de minimisation sans contrainte

$$f_* := \inf_{x \in \mathbb{E}} f(x),$$

où $f : \Omega \to \mathbb{R}$ est une fonction définie et différentiable sur un ouvert $\Omega \subseteq \mathbb{E}$.

L'*algorithme du gradient* (ou *de la plus profonde descente*) a déjà été introduit à la section 6.2.1. Il génère une suite $\{x_k\} \subseteq \mathbb{E}$ par la récurrence

$$\boxed{x_{k+1} = x_k - \alpha_k g_k,} \tag{7.1}$$

où $g_k = \nabla f(x_k)$ est le gradient de f pour le produit scalaire de $\mathbb{E}$ et $\alpha_k > 0$ est un pas déterminé par recherche linéaire (section 6.3).

Il s'agit donc d'un algorithme à directions de descente (chapitre 6), dont les directions sont les *antigradients* $-g_k$, des directions opposées au gradient. Ces directions ont la particularité d'avoir un *angle de descente* θ_k défini en (6.2) qui est nul, si bien que son cosinus vaut un :

$$\forall\, k \geqslant 1 : \qquad \cos \theta_k = 1. \tag{7.2}$$

Cette particularité implique des résultats de convergence et de complexité itérative aisé à déterminer.

Le lemme suivant est utilisé pour étudier la vitesse de convergence de la méthode du gradient.

Lemme 7.1 (inégalité de Kantorovitch [345]) *Soient M une matrice d'ordre n symétrique définie positive et $v \in \mathbb{R}^n$ de norme 1. Alors*

$$(v^\mathsf{T} M v)(v^\mathsf{T} M^{-1} v) \leqslant \frac{(\lambda_{\max} + \lambda_{\min})^2}{4 \lambda_{\max} \lambda_{\min}},$$

où $\lambda_{\max}$ et $\lambda_{\min}$ sont les valeurs propres maximale et minimale de M.

DÉMONSTRATION. □

Proposition 7.2 *Soit f une fonction quadratique strictement convexe de hessien H (défini positif). Utilisée pour minimiser cette fonction, la méthode du gradient à pas optimal génère une suite $\{x_k\}$ convergeant q-linéairement vers l'unique minimum x_* de f. Plus précisément,*

$$\frac{\|x_{k+1} - x_*\|_H}{\|x_k - x_*\|_H} \leqslant \frac{\kappa - 1}{\kappa + 1}, \quad \text{pour } k \geqslant 1,$$

où $\|\cdot\|_H = (\cdot^\mathsf{T} H \cdot)^{1/2}$ est la norme associée à H et κ est le conditionnement de H.

DÉMONSTRATION. □

7.2 Algorithme proximal

7.2.1 Définition

Soit $\mathbb{E}$ un espace euclidien (produit scalaire $\langle\cdot,\cdot\rangle$ et norme associée $\|\cdot\|$). À un opérateur auto-adjoint défini positif M, on associe le produit scalaire

$$\langle\cdot,\cdot\rangle_M : (u,v) \in \mathbb{E}\times\mathbb{E} \mapsto \langle u,v\rangle_M := \langle Mu, v\rangle$$

et la norme

$$\|\cdot\|_M : u \in \mathbb{E} \mapsto \|u\|_M := \langle u,u\rangle_M^{1/2}.$$

On s'intéresse dans cette section à la minimisation d'une fonction $f \in \overline{\mathrm{Conv}}(\mathbb{E})$ (c.-à-d., propre, convexe et fermée), qui n'est pas nécessairement différentiable. Soit $\{M_k\}$ une suite d'opérateurs auto-adjoints définis positifs, qui seront utilisés comme *préconditionneurs* de l'algorithme proximal présenté ci-après, et peuvent être générés au fur et à mesure que l'algorithme progresse. Rappelons que le point proximal d'un point $x \in \mathbb{E}$, associé à la fonction f et à l'opérateur M_k, est le point noté et défini par

$$P_k(x) := \operatorname*{arg\,min}_{y\in\mathbb{E}} \left(f(y) + \frac{1}{2}\|y - x\|^2_{R_k^{-1}} \right). \tag{7.3}$$

Cette notion a été introduite à la section 3.7.1. Nous ferons évidemment souvent référence aux résultats de cette section. Par la condition nécessaire et suffisante d'optimalité du problème dans (7.3), on a

$$x_p = P_k(x) \qquad \Longleftrightarrow \qquad \exists\, g_p \in \partial f(x_p) : \quad x_p = x - R_k g_p.$$

où le sous-différentiel $\partial f(x)$ est calculé pour le produit scalaire originel $\langle\cdot,\cdot\rangle$.

L'*algorithme proximal* génère une suite $\{x_k\}$ par la formule

$$\boxed{x_{k+1} := P_k(x_k) = x_k - R_k g_{k+1},} \qquad \text{où} \qquad g_{k+1} \in \partial f(x_{k+1}). \tag{7.4}$$

On peut interpréter cet algorithme de diverses manières.

1) *C'est une méthode de sous-gradient implicite pour minimiser f.*
Le sous-gradient g_{k+1} est en effet évalué en x_{k+1} (qui est inconnu), plutôt qu'en x_k (dans l'algorithme du gradient de la section 7.1). Donc l'équation (7.4) est non linéaire en x_{k+1}, si bien que le nouvel itéré x_{k+1} s'obtiendra en général par un procédé itératif, par exemple celui qui consiste à résoudre itérativement le problème d'optimisation dans (7.3).

2) *C'est un algorithme à directions de descente sur f avec pas unité.*
La direction $-R_k g_{k+1}$ est en effet une direction de descente de f en x_k parce que $R_k g_{k+1}$ est une direction de montée en x_{k+1} et que f est convexe. On a en effet

$$f'(x_{k+1}; R_k g_{k+1}) \geqslant \langle g_{k+1}, R_k g_{k+1}\rangle \geqslant 0, \tag{7.5}$$

où la première inégalité vient du fait que $g_{k+1} \in \partial f(x_{k+1})$ (point (i) de la proposition 3.55). Puis, par le point (iv) de la proposition 3.19 et (7.4) :

$$f'(x_k; x_{k+1}-x_k) \leqslant -f'(x_{k+1}; x_k - x_{k+1}) = -f'(x_{k+1}; R_k g_{k+1}) \leqslant 0.$$

On observe aussi que l'algorithme s'impose systématiquement un pas unité le long de la direction $-R_k g_{k+1}$. On constate aussi que ce pas fait décroître f à chaque itération, puisqu'en prenant $y = x_k$ dans (7.3), on obtient

$$f(x_{k+1}) \leqslant f(x_{k+1}) + \frac{1}{2}\|x_{k+1} - x_k\|^2_{R_k^{-1}} \leqslant f(x_k). \tag{7.6}$$

3) *C'est l'algorithme du gradient, pour le produit scalaire $\langle\cdot,\cdot\rangle_{R_k^{-1}}$, avec pas unité, pour minimiser la régularisée de Moreau-Yosida de f*, à savoir la fonction

$$\tilde f : x \in \mathbb{E} \mapsto \tilde f(x) := \inf_{y\in\mathbb{E}} \left(f(y) + \frac{1}{2}\|y - x\|^2_{R_k^{-1}} \right).$$

En effet, d'après la proposition 3.80, le gradient de $\tilde f$ pour le produit scalaire $\langle\cdot,\cdot\rangle$ s'écrit

$$\nabla \tilde f(x_k) = R_k^{-1}(x_k - x_{k+1}) = g_{k+1}.$$

Dès lors $R_k g_{k+1}$ est le gradient de $\tilde f$ en x_{k+1} pour le produit scalaire $\langle\cdot,\cdot\rangle_{R_k^{-1}}$.
Notons que les itérés générés font aussi décroître $\tilde f$ de façon monotone. En effet, (7.6) se récrit $f(x_{k+1}) \leqslant \tilde f(x_k) \leqslant f(x_k)$, si bien que l'on a

$$\tilde f(x_{k+1}) \leqslant f(x_{k+1}) \leqslant \tilde f(x_k).$$

Voilà un algorithme bien étrange, puisque pour minimiser f, il faut qu'à chaque itération, l'algorithme proximal minimise la fonction

$$f^k : y \in \mathbb{E} \mapsto f^k(y) := f(y) + \frac{1}{2}\|y - x_k\|^2_{R_k^{-1}},$$

qui semble bien être aussi compliquée que f. Ce point de vue doit être relativisé au vu des remarques suivantes.

1) La fonction f^k à minimiser (à chaque itération, ne l'oublions pas) peut être plus attrayante que f, du fait de sa *forte convexité*. Pour certains algorithmes, cette propriété est une aubaine, permettant d'accélérer leur convergence et de mieux la contrôler. Cette fonction a aussi un minimum unique, ce qui n'est pas nécessairement le cas de f.

2) Comme nous le verrons au chapitre 14, certains algorithmes de dualité peuvent être *interprétés* comme des algorithmes proximaux. Ces algorithmes de dualité s'appliquent en fait à des fonctions f dont l'évaluation résulte de la minimisation d'une fonction (le lagrangien) et il n'est pas plus coûteux d'évaluer f que de minimiser f^k (pourvu que $R_k \succ 0$), qui revient à minimiser une autre fonction (le lagrangien augmenté). Ces algorithmes de dualité ont donc tout leur sens. Leur interprétation en termes d'algorithme proximal permet alors d'en obtenir des propriétés difficiles à mettre en évidence autrement.

3) Observons que si f est *séparable*, c'est-à-dire si elle s'écrit comme suit

$$f(x) = \sum_{j=1}^{N} f_j(x_{D_j}),$$

où les D_j sont de *petites* parties disjointes de l'ensemble des indices $[1:n]$, il en est de même de f^k (pourvu que R_k soit diagonale par blocs). C'est une propriété intéressante lorsqu'on cherche à résoudre de grands problèmes par des techniques de décomposition.

4) L'algorithme proximal a un *effet stabilisant*. Lorsque f a plusieurs minimiseurs, l'algorithme génère une suite convergeant vers l'un d'entre eux. L'algorithme proximal est parfois utilisé pour stabiliser des algorithmes qui ne convergeraient pas sans modification lorsque le problème considéré devient singulier.

7.2.2 Convergence

Les propositions 7.3 et 7.4 ci-desous explorent quelques propriétés de convergence de l'algorithme proximal, dans des situations de plus en plus restrictives.

Un des rôles de l'opérateur R_k dans l'itération de l'algorithme proximal peut se comprendre en examinant l'influence de son ordre de grandeur dans le problème (7.3) dont x_{k+1} est la solution. Si R_k est *très petit*, la pénalité introduite par le terme quadratique est *très grande*, si bien que le déplacement $x_{k+1} - x_k$ peut devenir *très petit* au point d'empêcher le progrès vers une solution (de ce point de vue, l'opérateur $R_k = +\infty I$ conviendrait parfaitement, puisqu'il permettrait de résoudre le problème de minimisation en une seule itération). Comme nous le verrons dans la démonstration de la proposition 7.3 ci-dessous, la condition suivante

$$\sum_{k\geqslant 0} \lambda_{\min}(R_k) = +\infty \tag{7.7}$$

est suffisante.

L'algorithme proximal à métrique variable R_k est difficile à analyser du fait de la modification de R_k à chaque itération, qui empêche d'utiliser les relations de monotonie issues de la formule (3.67), permettant de comparer deux itérations successives ou les itérés de l'itération courante à une solution du problème (un argument classique du cas où $R_k = r_k I$). Le résultat de convergence de la proposition 7.3 repose alors sur la monotonie suivante, observée au point 2 de la page 328 : l'algorithme proximal fait décroître la fonction f à chaque itération.

On note $\tilde{f}(x_k)$ la valeur optimale dans (7.3), qui est la valeur de la régularisée de Moreau-Yosida en x_k. Comme x_k minimise f si, et seulement si, $\tilde{f}(x_k) = f(x_k)$, il est naturel de s'intéresser à l'écart entre ces deux valeurs, que l'on note

$$\begin{aligned}
\delta_k &:= f(x_k) - \tilde{f}(x_k) \\
&= f(x_k) - f(x_{k+1}) - \frac{1}{2}\|x_{k+1} - x_k\|^2_{R_k^{-1}} \qquad [\text{définition de } \tilde{f}(x_k)] \\
&= f(x_k) - f(x_{k+1}) - \frac{1}{2}\langle R_k g_{k+1}, g_{k+1}\rangle \qquad [(7.4)]. \\
&\geqslant 0 \qquad [(7.6)].
\end{aligned} \tag{7.8}$$

On peut obtenir mieux que cette inégalité en utilisant l'inégalité de convexité $f(x_{k+1}) + f'(x_{k+1}; x_k - x_{k+1}) \leqslant f(x_k)$ et (7.5), à savoir

$$f(x_{k+1}) + \langle R_k g_{k+1}, g_{k+1}\rangle \leqslant f(x_k). \tag{7.9}$$

Proposition 7.3 (algorithme proximal, R_k général) *Soient $\{x_k\}$ la suite générée par l'algorithme proximal et $\{g_k\}$ la suite des sous-gradients qui interviennent dans (7.4).*

1) *La suite $\{f(x_k)\}$ décroît vers une valeur $f_* \in \mathbb{R} \cup \{-\infty\}$.*
2) *Si $f_* > -\infty$, alors $\sum_k \delta_k < +\infty$.*

On suppose désormais que (7.7) a lieu.

3) *Si $f_* > -\infty$, alors zéro est point d'adhérence de la suite $\{g_k\}$.*
4) *Si $\{x_k\}$ est bornée, alors les points d'adhérence de $\{x_k\}$ sont des minima de f (en particulier, f a un minimum).*

Démonstration. 1) Le fait que $\{f(x_k)\}$ décroisse a été observé en (7.6) ou (7.9). Comme les $f(x_k) \in \mathbb{R}$, $f(x_k)$ décroît vers une limite f_* dans $\mathbb{R} \cup \{-\infty\}$.

2) En sommant les égalités (7.8), on obtient

$$\sum_{k\geqslant 0} \delta_k + \frac{1}{2}\sum_{k\geqslant 0} \langle R_k g_{k+1}, g_{k+1}\rangle = f(x_0) - f_* < \infty.$$

Comme la seconde série est positive, on en déduit le résultat.

3) Grâce à l'identité précédente, on a aussi

$$\frac{1}{2}\sum_{k\geqslant 0}\lambda_{\min}(R_k)\|g_{k+1}\|^2 < +\infty.$$

On en déduit qu'il existe une sous-suite de $\{g_k\}$ qui tend vers zéro, sinon il existerait une constante $\gamma > 0$ tel que $\|g_k\| \geqslant \gamma$ et l'inégalité ci-dessus contredirait (7.7).

4) Si $\{x_k\}$ est bornée, alors $f_* > -\infty$, car $f(x_k) \to f_*$ et f a une minorante affine (proposition 3.7). Il suffit de montrer que f_* est la valeur minimale de f, puisque $f(x_k) \to f_*$ (point 1) et qu'alors un point d'adhérence x_* de $\{x_k\}$ sera tel que $f(x_*) = f_*$ (f est fermée), ce qui implique que x_* minimise f. Par le point 3, il existe une sous-suite d'indices $\mathcal{K}$ telle que $\{g_k\}_{k\in\mathcal{K}} \to 0$. Comme $\{x_k\}$ est bornée, il existe une sous-suite d'indices $\mathcal{K}' \subseteq \mathcal{K}$ telle que $\{x_k\}_{k\in\mathcal{K}'} \to x_*$. On peut alors passer à la limite dans $g_k \in \partial f(x_k)$ (résultat de (7.4)) pour obtenir $0 \in \partial f(x_*)$ (point (iii) de la proposition 3.67), c'est-à-dire que x_* minimise f et donc que $f_* = \inf f$. □

L'algorithme proximal a davantage de proprétés lorsque $R_k = r_k R$, où r_k est un scalaire strictement positif et R est un opérateur auto-adjoint défini positif fixé. Le cas où $R_k = r_k I$ est souvent rencontré. Dans ce cas, la condition de convergence (7.7) devient la condition sur les r_k suivante :

$$\sum_{k\geqslant 0} r_k = +\infty. \tag{7.10}$$

Comme exemple de propriété supplémentaire : la suite $\{x_k\}$ est nécessairement minimisante, même si elle n'est pas bornée (voir le point 1 ci-dessous).

Proposition 7.4 (algorithme proximal, $R_k = r_k R$) *Soient $\{x_k\}$ la suite générée par l'algorithme proximal avec $R_k = r_k R$, où r_k est un scalaire strictement positif et R est un opérateur auto-adjoint défini positif, et $\{g_k\}$ la suite des sous-gradients qui interviennent dans (7.4). On suppose que (7.10) a lieu.*

1) *La suite $\{f(x_k)\}$ décroît vers la valeur minimale de f (qui peut être $-\infty$).*
2) *Si, de plus, $\{r_k\}$ est bornée et f a un minimiseur, alors $\{x_k\}$ converge vers un minimiseur de f.*

Démonstration. 1) L'argument consiste à examiner comment $\{x_k\}$ dévie d'une suite de Fejér, la déviation étant mesurée au moyen de la fonction f. On regarde donc comment évolue l'écart $x_k - x$ en norme R^{-1}, pour un $x \in \mathbb{E}$ pour l'instant arbitraire. La récurrence $x_{k+1} = x_k - r_r R g_{k+1}$ conduit à $x_{k+1} - x = x_k - x - r_r R g_{k+1}$, puis à l'identité suivante en prenant le carré des normes R^{-1} :

$$\|x_{k+1} - x\|^2_{R^{-1}} = \|x_{k+1} - x_k\|^2_{R^{-1}} + 2\langle x_{k+1} - x_k, x_k - x\rangle_{R^{-1}} + \|x_k - x\|^2_{R^{-1}}. \tag{7.11}$$

L'idée-clé est d'estimer les deux premiers termes du membre de droite (qui font éventuellement dévier $\{x_k\}$ d'une suite de Fejér) par une variation de f. On a

$$\begin{aligned} f(x) &\geqslant f(x_{k+1}) + \langle g_{k+1}, x - x_{k+1}\rangle \qquad \text{[inégalité de convexité]} \\ &= f(x_{k+1}) + \langle R_k^{-1}(x_k - x_{k+1}), x - x_{k+1}\rangle \qquad \text{[(7.4)]} \\ &= f(x_{k+1}) + \langle R_k^{-1}(x_k - x_{k+1}), x - x_k\rangle + \langle R_k^{-1}(x_k - x_{k+1}), x_k - x_{k+1}\rangle. \end{aligned}$$

On fait ensuite apparaître δ_k donné par (7.8) :

$$f(x) \geqslant f(x_k) + \frac{1}{r_k}\,\langle x_k - x_{k+1}, x - x_k\rangle_{R^{-1}} + \frac{1}{2r_k}\,\|x_k - x_{k+1}\|_{R^{-1}}^2 - \delta_k.$$

En multipliant par $2r_k$, on obtient une majoration des deux premiers termes du membre de droite de (7.11), qui devient

$$\|x_{k+1} - x\|_{R^{-1}}^2 \leqslant \|x_k - x\|_{R^{-1}}^2 + 2r_k\Big[f(x) - f(x_k) + \delta_k\Big]. \tag{7.12}$$

Choisissons maintenant $x \in \mathbb{E}$. On sait d'après le point 1 de la propoisiton 7.3, que $f(x_k)$ décroît vers une limite f_*. Si $f_* = -\infty$, le résultat est démontré. Supposons désormais que $f_* > -\infty$. Si $f_* \neq \inf f$, on peut trouver un $x \in \mathbb{E}$ et un $\eta > 0$ tel que $f(x) + \eta \leqslant f(x_k)$ pour tout indice $k \geqslant 0$. Alors l'inégalité précédente devient

$$\forall\, k \geqslant 0 : \quad \|x_{k+1} - x\|_{R^{-1}}^2 \leqslant \|x_k - x\|_{R^{-1}}^2 + 2r_k(\delta_k - \eta).$$

Mais $\delta_k \to 0$ par le point 2 de la proposition 7.3, si bien que pour k_1 assez grand :

$$\forall\, k \geqslant k_1 : \quad \|x_{k+1} - x\|_{R^{-1}}^2 \leqslant \|x_k - x\|_{R^{-1}}^2 - \eta r_k.$$

En sommant ces inégalités de k_1 à $k_2 \geqslant k_1$, on obtient

$$0 \leqslant \|x_{k_2+1} - x\|_{R^{-1}}^2 \leqslant \|x_{k_1} - x\|_{R^{-1}}^2 - \eta \sum_{k=k_1}^{k_2} r_k,$$

si bien que la série $\sum_k r_k$ serait convergente, en contradiction avec l'hypothèse (7.10).

2) Soit $\bar{x}$ un minimiseur de f. En prenant $x = \bar{x}$ dans (7.12), on obtient

$$\forall\, k \geqslant 0 : \quad \|x_{k+1} - \bar{x}\|_{R^{-1}}^2 \leqslant \|x_k - \bar{x}\|_{R^{-1}}^2 + 2r_k\delta_k. \tag{7.13}$$

Si $\{r_k\}$ est bornée, la série $\sum_k r_k\delta_k$ converge et, en sommant les inégalités précédentes, on voit que $\{x_k\}$ est bornée.

Soit $\bar{x}$ un point d'adhérence de $\{x_k\}$, qui est nécessairement un minimiseur de f (car $f(x_k) \to \inf f$, par le point 1, et f est s.c.i.). Montrons que toute la suite $\{x_k\}$ convergence vers $\bar{x}$. Soit $\varepsilon > 0$. Il suffit de montrer que $\|x_k - \bar{x}\|_{R^{-1}}^2 \leqslant \varepsilon$ pour k assez grand. Par définition de $\bar{x}$, on peut trouver un indice k_1 tel que

$$\|x_{k_1} - \bar{x}\|_{R^{-1}}^2 \leqslant \frac{\varepsilon}{2} \qquad \text{et} \qquad 2\sum_{k \geqslant k_1} r_k\delta_k \leqslant \frac{\varepsilon}{2}.$$

En sommant les inégalités (7.13), on trouve que $\|x_k - \bar{x}\|_{R^{-1}}^2 \leqslant \varepsilon$ pour tout $k \geqslant k_1$. □

7.3 Méthode de Gauss-Seidel

La méthode de Gauss-Seidel est initialement une méthode itérative de résolution d'un système d'équations linéaires (de dimension finie) de la forme $Ax = b$ (section 7.3.1), ce qui signifie qu'elle génère une suite qui converge vers une solution de cette équation, lorsque celle-ci en a une et lorsque des conditions de convergence sont satisfaites (par exemple lorsque A est symétrique définie positive). L'algorithme suppose que la diagonale de A est formée d'éléments non nuls. Elle se décline aussi en une version « par blocs ».

Le principe de la méthode peut s'étendre à la résolution de systèmes d'équations non linéaires (section 7.3.2) et à l'optimisation (section 7.3.3), mais avec des conditions d'efficacité moins claires. En optimisation, l'utilité de cette approche dépendra beaucoup de la structure du problème. Le principe gauss-seidelien permet aussi d'interpréter d'autres algorithmes.

7.3.1 En algèbre linéaire

Version élément par élément

Rappelons d'abord ce qu'est la méthode de Gauss-Seidel pour résoudre en $x \in \mathbb{R}^n$ le système linéaire

$$Ax = b,$$

dans lequel $A \in \mathbb{R}^{n\times n}$ et $b \in \mathbb{R}^n$. Il s'agit d'un algorithme itératif, générant donc une suite $\{x_k\} \subseteq \mathbb{R}^n$. On interrompt le calcul de la suite lorsque l'itéré courant, disons x_k, est jugé suffisamment proche d'une solution, par exemple parce que la norme du *résidu* $\|Ax_k - b\|$ est petite.

Soit $x_k = ((x_k)_1, \ldots, (x_k)_n) \in \mathbb{R}^n$ l'itéré courant. L'itéré suivant $x_{k+1} = ((x_{k+1})_1, \ldots, (x_{k+1})_n) \in \mathbb{R}^n$ se calcule en n étapes, comme suit.

- *Étape* 1. Si l'on suppose que $a_{11} \neq 0$ et connaissant $((x_k)_2, \ldots, (x_k)_n)$, on peut calculer $(x_{k+1})_1$ au moyen de la première équation du système linéaire $Ax = b$. De manière plus précise, $(x_{k+1})_1 \in \mathbb{R}$ est pris comme l'unique solution de

$$a_{11}\boxed{(x_{k+1})_1} + a_{12}(x_k)_2 + \cdots + a_{1n}(x_k)_n = b_1.$$

- *Étape* 2. Si l'on suppose que $a_{22} \neq 0$ et connaissant $((x_{k+1})_1, (x_k)_3, \ldots, (x_k)_n)$, on peut calculer $(x_{k+1})_2$ au moyen de la deuxième équation du système linéaire $Ax = b$. De manière plus précise, $(x_{k+1})_2 \in \mathbb{R}$ est pris comme l'unique solution de

$$a_{21}(x_{k+1})_1 + a_{22}\boxed{(x_{k+1})_2} + a_{23}(x_k)_3 + \cdots + a_{2n}(x_k)_n = b_2.$$

- *Étape* $i \in [1\,{:}\,n]$ (cas général). Si l'on suppose que $a_{ii} \neq 0$ et connaissant $((x_{k+1})_1, \ldots, (x_{k+1})_{i-1}, (x_k)_{i+1}, \ldots, (x_k)_n)$, on peut calculer $(x_{k+1})_i$ au moyen de la i-ième équation du système linéaire $Ax = b$. De manière plus précise, $(x_{k+1})_i \in \mathbb{R}$ est pris comme l'unique solution de

$$a_{i1}(x_{k+1})_1 + \cdots + a_{i,i-1}(x_{k+1})_{i-1} + a_{ii}\boxed{(x_{k+1})_i} + a_{i,i+1}(x_k)_{i+1} + \cdots + a_{in}(x_k)_n = b_i.$$

En résumé, on calcule les composantes $(x_{k+1})_i$ de x_{k+1} de manière séquentielle pour $i = 1, \ldots, n$ par

$$\boxed{(x_{k+1})_i = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij}(x_{k+1})_j - \sum_{j=i+1}^{n} a_{ij}(x_k)_j \right).}$$

La formule fait intervenir les éléments $(x_{k+1})_j$ $(j = 1, \ldots, i-1)$ calculés dans les étapes précédentes.

L'expression matricielle de l'algorithme suppose que la matrice A se décompose comme suit

$$A = L + D + U,$$

où D est la partie diagonale de A, L sa partie triangulaire inférieure stricte et U sa partie triangulaire supérieure stricte. Une itération de la méthode de Gauss-Seidel, celle passant de x_k à x_{k+1}, consiste alors à résoudre le système triangulaire inférieur

$$\boxed{(L + D)x_{k+1} = b - Ux_k,}$$

de « haut en bas », c'est-à-dire en déterminant successivement $(x_{k+1})_1$, $(x_{k+1})_2$, …, $(x_{k+1})_n$.

Version par blocs

La méthode de Gauss-Seidel peut se décliner en une « *version par blocs* ». Celle-ci procède de manière similaire à la méthode élément par élément décrite ci-dessus, mais en remplaçant l'utilisation des éléments de A par des sous-matrices de A, appelées ici des *blocs*. On suppose que l'ensemble des indices $[1:n]$ est partitionné en p sous-intervalles (non vides et deux-à-deux disjoints) :

$$[1:n] = I_1 \cup I_2 \cup \cdots \cup I_p.$$

La matrice A et le vecteur b sont alors décomposés comme suit

$$A = \begin{pmatrix} A_{I_1 I_1} & A_{I_1 I_2} & \cdots & A_{I_1 I_p} \\ A_{I_2 I_1} & A_{I_2 I_2} & \cdots & A_{I_2 I_p} \\ \vdots & \vdots & \ddots & \vdots \\ A_{I_p I_1} & A_{I_p I_2} & \cdots & A_{I_p I_p} \end{pmatrix} \qquad \text{et} \qquad b = \begin{pmatrix} b_{I_1} \\ b_{I_2} \\ \vdots \\ b_{I_p} \end{pmatrix},$$

où A_{IJ} est la sous-matrice de A obtenue en sélectionnant les éléments avec indices de ligne dans I et indices de colonnes dans J, tandis que b_I est le sous-vecteur de b obtenu en sélectionnant les éléments avec indices dans I.

La méthode de Gauss-Seidel par blocs suppose que les sous-matrices principales $A_{I_i I_i}$, avec $i \in [1:p]$, sont inversibles.

Une itération de la méthode de Gauss-Seidel par blocs, celle passant de x_k à x_{k+1}, s'écrit de la même manière que la méthode élément par élément, à savoir

$$\boxed{(L + D)x_{k+1} = b - Ux_k,}$$

mais avec des définitions différentes de L, D et U :

$$L = \begin{pmatrix} 0 & \cdots & \cdots & 0 \\ A_{I_2 I_1} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ A_{I_p I_1} & \cdots & A_{I_p I_{p-1}} & 0 \end{pmatrix}, \qquad D = \begin{pmatrix} A_{I_1 I_1} & 0 & \cdots & 0 \\ 0 & A_{I_2 I_2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & A_{I_p I_p} \end{pmatrix}$$

et $U = A - L - D$. La résolution du système triangulaire par blocs ci-dessus, se fait également de « haut en bas », c'est-à-dire en déterminant successivement $(x_{k+1})_{I_1}$, $(x_{k+1})_{I_2}$, …, $(x_{k+1})_{I_p}$.

7.3.2 Pour les systèmes non linéaires

Le principe de la méthode de Gauss-Seidel peut également s'appliquer à la résolution d'un système d'équations non linéaires $F(x) = 0$, où $F : \mathbb{R}^n \to \mathbb{R}^n$. Ce système s'écrit donc sous la forme de n équations non linéaires à n inconnues :

$$\begin{cases} F_1(x_1, x_2, \ldots, x_n) = 0 \\ F_2(x_1, x_2, \ldots, x_n) = 0 \\ \cdots \\ F_n(x_1, x_2, \ldots, x_n) = 0. \end{cases}$$

La *méthode de Gauss-Seidel* résout ce système de manière itérative, en générant donc une suite $\{x_k\} \subseteq \mathbb{R}^n$. On interrompt le calcul de la suite lorsque l'itéré courant, disons x_k, est jugé suffisamment proche d'une solution, par exemple parce que la norme du *résidu* $\|F(x_k)\|$ est petite.

Soit $x_k = ((x_k)_1, \ldots, (x_k)_n) \in \mathbb{R}^n$ l'itéré courant. L'itéré suivant $x_{k+1} = ((x_{k+1})_1, \ldots, (x_{k+1})_n) \in \mathbb{R}^n$ se calcule en n étapes, comme suit.

- *Étape* 1. Connaissant $((x_k)_2, \ldots, (x_k)_n)$, on calcule $(x_{k+1})_1$ comme solution de l'équation non linéaire (cette solution est supposée exister) :

$$F_1(\boxed{(x_{k+1})_1}, (x_k)_2, \ldots, (x_k)_n) = 0.$$

- *Étape* 2. Connaissant $((x_{k+1})_1, (x_k)_3, \ldots, (x_k)_n)$, on calcule $(x_{k+1})_2$ comme solution de l'équation non linéaire (cette solution est supposée exister) :

$$F_1((x_{k+1})_1, \boxed{(x_{k+1})_2}, (x_k)_2, \ldots, (x_k)_n) = 0.$$

- *Étape* $i \in [1 : n]$ (cas général). Connaissant $((x_{k+1})_1, \ldots, (x_{k+1})_{i-1}, (x_k)_{i+1}, \ldots, (x_k)_n)$, on calcule $(x_{k+1})_i$ comme solution de l'équation non linéaire (cette solution est supposée exister) :

$$F_i((x_{k+1})_1, \ldots, (x_{k+1})_{i-1}, \boxed{(x_{k+1})_i}, (x_k)_{i+1}, \ldots, (x_k)_n) = 0.$$

La version « par blocs » se définit facilement en considérant des groupes d'équations et d'inconnues, au lieu de considérer, comme ci-dessus, équation et inconnue une par une.

7.3.3 En optimisation

Le principe de la méthode de Gauss-Seidel décrit dans la section précédente s'applique naturellement au problème d'optimisation non linéaire

$$\inf_{x\in X} f(x), \tag{7.14}$$

dans lequel on minimise une fonction $f : \mathbb{R}^n \to \mathbb{R}$ sur un sous-ensemble X de $\mathbb{R}^n$. Nous présentons directement ci-dessous la version « par blocs », qui est la plus utile lorsque le nombre p de blocs est faible (souvent $p = 2$). La méthode de Gauss-Seidel perd en effet de sa pertinence lorsque p est grand, par manque d'efficacité dans ce cas. la version « élément par élément » peut être vue comme un cas particulier de la version par blocs, obtenue en prenant n blocs de cardinal 1.

On suppose donc que l'ensemble des indices $[1:n]$ est *partitionné* en p blocs,

$$[1:n] = I_1 \cup I_2 \cup \cdots \cup I_p, \tag{7.15}$$

et que l'ensemble admissible est un produit cartésien de p ensembles,

$$X = X_1 \times X_2 \times \cdots \times X_p, \tag{7.16}$$

où chaque X_i est un convexe de $\mathbb{R}^{|I_i|}$. La variable $x \in \mathbb{R}^n$ se décomposera comme suit

$$x = (x_{I_1}, x_{I_2}, \ldots, x_{I_p}).$$

Lorsque f est différentiable et que $X = \mathbb{R}^n$, on pourrait obtenir une méthode de Gauss-Seidel en appliquant la méthode de la section 7.3.2 à la condition d'optimalité du premier ordre de ce problème d'optimisation sans contrainte, à savoir

$$\nabla f(x) = 0,$$

qui est un système de n équations non linéaires à n inconnues $x = (x_1, \ldots, x_n)$. Mais on peut préférer, comme ci-dessous, rester dans le domaine de l'optimisation en minimisant f séquentiellement, bloc par bloc. Cette option a l'avantage de pouvoir prendre en compte des contraintes, c'est-à-dire de restreindre les variables à l'ensemble admissible X.

La *méthode de Gauss-Seidel*[1] résout le problème d'optimisation (7.14) de manière itérative, en générant donc une suite $\{x_k\} \subseteq \mathbb{R}^n$. L'algorithme passe d'un itéré au suivant en minimisant f un bloc de variables à la fois, en séquence. On interrompt le calcul de la suite lorsque l'itéré courant, disons x_k, est jugé suffisamment proche d'une solution, par exemple parce que la norme du *gradient projeté* $\|g^{\mathrm{P}}(x_k)\|$ est jugée suffisamment petite (on rappelle que le gradient projeté $g^{\mathrm{P}}(x_k)$ est la projection orthogonale de $\nabla f(x_k)$ sur l'opposé du cône tangent $\mathrm{T}_{x_k} X$ et que celui-ci est nul en une solution de (7.14)).

[1] Cette méthode est appelée *méthode de relaxation* par Glowinski, Lions, Trémolières [252 ; 1976, page 60-68], mais cette appellation est utilisée pour beaucoup trop d'algorithmes pour qu'elle soit suffisamment discriminante.

Algorithme 7.5 (de Gauss-Seidel par bloc en optimisation) Une itération passe de l'itéré courant $x_k \in X$ à l'itéré suivant $x_{k+1} \in X$ en p étapes successives, indicées par $i = 1, \dots, p$:

$$(x_{k+1})_{I_i} \in \operatorname*{arg\,min}_{x_{I_i} \in X_i} f((x_{k+1})_{I_1}, \dots, (x_{k+1})_{I_{i-1}}, x_{I_i}, (x_k)_{I_{i+1}}, \dots, (x_k)_{I_p}). \tag{7.17}$$

La version « élément par élément » se définit facilement en considérant des blocs I_i de cardinal 1 et en minimisant f composante par composante. Ce dernier algorithme porte aussi le nom de *méthode de descente par coordonnée*.

Le résultat suivant montre la convergence de la méthode de Gauss-Seidel lorsque f est de classe C^1, coercive et strictement convexe. Il repose sur la caractérisation de la stricte convexité, donnée à la proposition 3.23.

Proposition 7.6 (convergence de Gauss-Seidel en optimisation) *Si, pour chaque $i \in [1:p]$, X_i est un convexe fermé non vide de $\mathbb{R}^{|I_i|}$ et si f est coercive sur X, strictement convexe sur X et de classe C^1 dans un voisinage de X, alors*
1) *le problème (7.14)-(7.16) a une unique solution $\bar{x}$,*
2) *l'algorithme 13.21 est bien défini et, quel que soit l'itéré initial $x_1 \in X$, il génère une suite $\{x_k\} \subseteq X$ qui converge vers $\bar{x}$.*

Démonstration. 1) L'existence de solution se déduit de la continuité, de la coercivité de f et du fait que l'ensemble admissible est non vide et fermé (proposition 1.4). L'unicité se déduit de la stricte convexité de f et de la convexité de X (proposition 3.5).

2) *Préliminaires.* Pour les mêmes raisons qu'au point 1, tous les *problèmes intermédiaires* (7.17) ont une solution unique, si bien que l'algorithme 13.21 est bien défini. Ces solutions uniques ou *itérés intermédiaires* sont notés comme suit : $x_{k+1,0} := x_k$ et

$$x_{k+1,i} := ((x_{k+1})_{I_1}, \dots, (x_{k+1})_{I_i}, (x_k)_{I_{i+1}}, \dots, (x_k)_{I_p}), \quad \text{pour } i \in [1:p].$$

Donc $x_{k+1,i}$ est la solution du problème intermédiaire (7.17) et $x_{k+1} = x_{k+1,p}$. Par construction

$$f(x_{k+1}) \leqslant f(x_{k+1,p-1}) \leqslant \cdots \leqslant f(x_{k+1,1}) \leqslant f(x_k).$$

Comme f est bornée inférieurement par $f(\bar{x})$, ces inégalités impliquent que

$$\{f(x_k)\} \text{ et les } \{f(x_{k,i})\} \text{ convergent vers la même valeur.} \tag{7.18}$$

Par ailleurs, l'optimalité du problème intermédiaire (7.17) implique que

$$\forall\, x_{I_i} \in X_i : \quad \nabla_{x_i} f(x_{k+1,i})^{\mathsf{T}} (x_{I_i} - (x_{k+1})_{I_i}) \geqslant 0. \tag{7.19}$$

Montrons que $x_{k+1} - x_k \to 0$. La coercivité de f implique la bornitude de ces ensembles de sous-niveau (point (ii) de l'exercice 1.3). Il existe donc un réel $\beta > 0$ tel que

$$\{x : \mathbb{R}^n : f(x) \leqslant f(x_1)\} \subseteq \beta \bar{B},$$

où $\bar{B}$ est la boule unité fermée de $\mathbb{R}^n$, si bien que tous les itérés (intermédiaires ou pas) sont dans $\beta\bar{B}$. Alors, la stricte convexité donne (point (ii) de la proposition 3.23) :

$$\begin{aligned} & f(x_{k+1,i-1}) - f(x_{k+1,i}) \\ & \quad \geqslant \quad f'(x_{k+1,i}) \cdot (x_{k+1,i-1} - x_{k+1,i}) + \tfrac{1}{2} g_\beta(\tfrac{1}{2}\|x_{k+1,i-1} - x_{k+1,i}\|) \\ & \quad = \quad \underbrace{\nabla_{x_{I_i}} f(x_{k+1,i})^{\mathsf{T}}((x_k)_{I_i} - (x_{k+1})_{I_i})}_{\geqslant 0 \text{ par } (7.19)} + \tfrac{1}{2} g_\beta(\tfrac{1}{2}\|x_{k+1,i-1} - x_{k+1,i}\|) \\ & \quad \geqslant \quad \tfrac{1}{2} g_\beta(\tfrac{1}{2}\|x_{k+1,i-1} - x_{k+1,i}\|), \end{aligned}$$

où $\|\cdot\|$ désigne la norme euclidienne. En sommant :

$$f(x_k) - f(x_{k+1}) \geqslant \textstyle\sum_{i=1}^{p} \tfrac{1}{2} g_\beta(\tfrac{1}{2}\|x_{k+1,i-1} - x_{k+1,i}\|).$$

Alors (7.18) et la positivité de g_β impliquent que

$$\forall\, i \in [1:p] : \qquad g_\beta(\tfrac{1}{2}\|x_{k+1,i-1} - x_{k+1,i}\|) \to 0.$$

On utilise alors le fait que $g_\beta(0) = 0$, que $g_\beta(t) > 0$ pour $t > 0$ et la continuité de g_β pour en déduire que $\frac{1}{2}\|x_{k+1,i-1} - x_{k+1,i}\| \to 0$ lorsque $k \to \infty$, pour tout $i \in [1:p]$. Comme p est fini :

$$x_{k+1} - x_k \to 0. \tag{7.20}$$

Montrons que $x_k \to \bar{x}$, ce qui conclura la démonstration. On utilise cette fois le point (iii) de la proposition 3.23 entre x_{k+1} et la solution $\bar{x}$:

$$(f'(x_{k+1}) - f'(\bar{x})) \cdot (x_{k+1} - \bar{x}) \geqslant g_\beta(\|x_{k+1} - \bar{x}\|).$$

Mais $f'(\bar{x}) \cdot (x_{k+1} - \bar{x}) \geqslant 0$ par optimalité de $\bar{x}$ et $x_{k+1} \in X$, si bien qu'en inversant l'inégalité, on a

$$\begin{aligned} g_\beta(\|x_{k+1} - \bar{x}\|) \quad & \leqslant \quad f'(x_{k+1}) \cdot (x_{k+1} - \bar{x}) \\ & = \quad \sum_{i=1}^{p} \nabla_{x_{I_i}} f(x_{k+1})^{\mathsf{T}} (x_{k+1} - \bar{x})_{I_i} \\ & \leqslant \quad \sum_{i=1}^{p} \left(\nabla_{x_{I_i}} f(x_{k+1}) - \nabla_{x_{I_i}} f(x_{k+1,i})\right)^{\mathsf{T}} (x_{k+1} - \bar{x})_{I_i} \\ & \qquad\qquad [\text{par } (7.19) \text{ avec } x_{I_i} = \bar{x}_{I_i}] \\ & \leqslant \quad \sum_{i=1}^{p} \|\nabla_{x_{I_i}} f(x_{k+1}) - \nabla_{x_{I_i}} f(x_{k+1,i})\| \, \|x_{k+1} - \bar{x}\|, \end{aligned}$$

par l'inégalité de Cauchy-Schwarz et $\|(x_{k+1} - \bar{x})_{I_i}\| \leqslant \|x_{k+1} - \bar{x}\|$. Cela permet de conclure. En effet, $\|x_{k+1} - \bar{x}\|$ est borné par 2β. Par ailleurs, $\|x_{k+1} - x_{k+1,i}\| \leqslant \|x_{k+1} - x_k\| \to 0$ par (7.20). Alors la continuité de f', donc de $\nabla_{x_{I_i}} f$, sur le compact $\beta\bar{B}$ se traduit en une continuité uniforme, si bien que, pour tout $i \in [1:p]$, $\nabla_{x_{I_i}} f(x_{k+1}) - \nabla_{x_{I_i}} f(x_{k+1,i}) \to 0$. On en déduit que $g_\beta(\|x_{k+1} - \bar{x}\|) \to 0$ et donc que $x_k \to \bar{x}$ par le même raisonnement que pour obtenir (7.20). □

Remarques 7.7 1) Si l'on applique la proposition 7.6 au cas où $X = \mathbb{R}^n$ et f est la fonction quadratique $x \mapsto \frac{1}{2}x^\mathsf{T}Ax - b^\mathsf{T}x$, on retrouve le résultat affirmant que la méthode de Gauss-Seidel par blocs pour résoudre le système linéaire $Ax = b$ convergence, quels que soient le vecteur b et le point initial, pourvu que A soit définie positive.

2) La méthode de Gauss-Seidel est un algorithme lent (il requiert beaucoup d'itérations), dont la mise en œuvre est coûteuse (chaque itération peut demander beaucoup de temps de calcul, selon les cas). Tel qu'il est présenté, il requiert en effet la minimisation *exacte* de f dans chaque problème intermédiaire et ces p minimisations doivent être réalisées à chaque itération. Son application est donc restreinte au cas où le nombre de blocs est petit.

3) L'algoithme 13.21 ne s'étend pas aisément à des ensembles admissibles plus complexes qu'un produit cartésien d'ensembles convexes. Par exemple si l'on cherche à minimiser composante par composante la fonction linéaire $f : \mathbb{R}^2 \to \mathbb{R} : (x_1, x_2) \mapsto x_1 + x_2$ sur l'ensemble $X := \{x \in \mathbb{R}^2_+ : x_1 x_2 \geqslant 1\}$, qui n'est pas le produit cartésien de deux intervalles, tout point de la frontière de X est bloquant (c'est-à-dire que l'algorithme ne peut y progresser), alors que seul le point $\bar{x} = (1, 1)$ est solution [252] ; voir la figure 7.1.

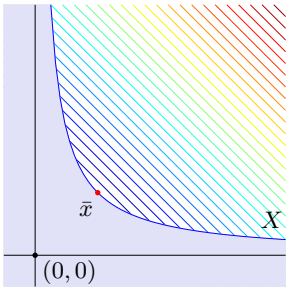


Fig. 7.1. Exemple de problème pour lequel l'algoithme 13.21, de Gauss-Seidel, reste bloqué en tout point de la frontière du domaine admissible (les droites sont les « courbes » de niveau de la fonction objectif dans l'ensemble admissible X).

4) En l'absence de convexité, la méthode de Gauss-Seidel ne converge pas nécessairement, même pour des fonctions de classe C^∞. Powell [488 ; 1973] a en effet construit plusieurs fonctions conduisant à la non-convergence de la méthode de Gauss-Seidel composante par composante, notamment une fonction C^∞ de trois variables pour laquelle les itérés générés ont un cycle limite formé de 6 points en lesquels le gradient n'est pas nul. On peut toutefois avoir convergence si les coordonnées ne sont pas choisies cycliquement [481 ; 1971].

5) D'autres résultats de convergence sont donnés par Luo et Tseng [402 ; 1992]. □

Notes

La convergence de l'algorithme de Gauss-Seidel pour les systèmes linéaires (section 7.3.1) est présentée, par exemple, par Varga [605] et pour les systèmes non linéaires (section 7.3.2) par Ortega et Rockoff [464].

Le résultat de convergence globale de la méthode de Gauss-Seidel en optimisation, la proposition 7.6, est adaptée de [252 ; théorème 1.2, page 66]. Pour d'autres résultats, voir [293] et ses références.

Exercices

7.1. *Terminaison finie de l'algorithme proximal.* Soient $f \in \mathrm{C}\overline{\mathrm{onv}}(\mathbb{E})$ et $\tilde{f}$ sa régularisée de Moreau-Yosida (section 3.7.2). On suppose que $\bar{x} \in (\operatorname{dom} f)^{\circ}$ minimise f et que $0 \in \operatorname{int} \partial f(\bar{x})$ (de manière plus précise : il existe $\varepsilon > 0$ tel que $\bar{B}(0,\varepsilon) \subseteq \partial f(\bar{x})$). On note x_p le point proximal de $x \in \mathbb{E}$. Montrez que

(*i*) $g \in \partial f(x)$ et $\|g\| < \varepsilon \implies x = \bar{x}$,
(*ii*) $\|x - \bar{x}\| \leqslant \varepsilon \implies x_p = \bar{x}$,
(*iii*) $\tilde{f}$ est quadratique sur $\bar{B}(\bar{x},\varepsilon)$.

8 Optimisation quadratique

On s'intéresse dans ce chapitre à la résolution numérique des systèmes d'équations linéaires par des méthodes *itératives finies* ; nous verrons plus loin ce que l'on entend par ces deux qualificatifs. Pour certaines méthodes, on considérera leur extension à la résolution de systèmes d'équations non linéaires. Pour d'autres, cette extension se fera dans d'autres chapitres.

On cherche donc un point $x_* \in \mathbb{R}^n$ solution du système linéaire en x suivant

$$Ax = b, \tag{8.1}$$

où A est une matrice d'ordre n donnée et b est un vecteur donné dans $\mathbb{R}^n$. Nous supposerons toujours que A est inversible. Dans ce cas, la solution unique de (8.1) s'écrit

$$x_* = A^{-1}b,$$

où A^{-1} est la matrice inverse de A. Pour obtenir x_*, on ne calcule jamais A^{-1}. Cela n'est ni nécessaire, ni opportun dès que n dépasse quelques unités (le temps de calcul requis est alors trop important), ni stable numériquement [313 ; 2002, section 14.2]. Rappelons que la résolution d'un système linéaire d'ordre n non structuré par factorisation de la matrice A demande de l'ordre de $O(n^3)$ opérations. L'algorithme le plus simple et le plus utilisé pour ce faire est l'*élimination gaussienne*, équivalente à la *factorisation gaussienne* de la matrice A, avec pivotage partiel. Cet algorithme requiert $2n^3/3 + O(n^2)$ opérations flottantes. Les algorithmes présentés dans cette section ont aussi une complexité opérationnelle en $O(n^3)$. Nous distinguerons deux cas : celui où A est symétrique définie positive et celui où A est non symétrique.

Si la matrice A est symétrique, la solution du système (8.1) est aussi le point stationnaire de la fonction quadratique

$$f(x) = \frac{1}{2}x^\mathsf{T}Ax - b^\mathsf{T}x. \tag{8.2}$$

Si, de plus, A est définie positive, x_* réalise le minimum de f. On peut dans ce cas obtenir des algorithmes de résolution du système (8.1) à partir d'algorithmes de minimisation de f. L'*algorithme du gradient conjugué* est de ce type (section 8.2).

Si A n'est pas symétrique, on peut remplacer la résolution du système (8.1) par la résolution de l'*équation normale* équivalente

$$A^\mathsf{T}Ax = A^\mathsf{T}b. \tag{8.3}$$

Comme la matrice de ce système est symétrique définie positive, on peut songer à utiliser les techniques mentionnées ci-dessus. Cependant, le système (8.3) peut être

beaucoup moins bien conditionné que le système original (8.1), si bien qu'il est souvent préférable d'utiliser des méthodes s'attaquant directement au système (8.1), sans faire appel à des algorithmes de minimisation de fonction. L'idée est alors de générer des approximations de x_* qui sont, dans un sens à préciser, les meilleures approximations sur des sous-espaces affines de dimension croissante. On peut en effet espérer que si l'on a une « bonne » approximation de x_* sur le sous-espace affine x_1+K_p, K_p étant un sous-espace vectoriel de dimension p, il ne sera pas trop difficile d'obtenir une « bonne » approximation sur un sous-espace affine $x_1 + K_{p+1}$, de dimension $p + 1$, contenant $x_1 + K_p$. Dans l'*algorithme du résidu minimal* (section 8.3), les sous-espaces K_p sont obtenus par un procédé particulier : ce sont des *sous-espaces de Krylov* (section 8.1). D'ailleurs, nous verrons que l'algorithme du gradient conjugué génère des itérés qui sont aussi dans un tel sous-espace de Krylov, si bien que ces deux algorithmes font partie de la même famille, celle dite des *méthodes de Krylov*.

Les méthodes de Krylov ont la propriété de trouver la solution du système (8.1) en un nombre fini d'étapes (en artihmétique exacte). Elles s'apparentent sur ce point aux *méthodes directes* de résolution, fondées sur la factorisation de la matrice A : factorisation gaussienne, factorisation QR... En pratique cependant, la présence d'erreurs d'arrondi dans les calculs empêche les méthodes de Krylov de trouver la solution en un nombre fini d'étapes, dès que la dimension du système linéaire dépasse quelques dizaines, si bien qu'on les range souvent dans la famille des méthodes itératives. Ces dernières sont intéressantes pour plusieurs raisons. D'abord, comme elles procèdent par améliorations successives, on peut s'arrêter lorsque la précision est jugée satisfaisante. Ceci peut se produire bien avant que la solution ne soit trouvée. Une solution « acceptable » peut donc être obtenue à un faible coût. Ensuite, elles permettent de tirer parti d'une bonne approximation initiale, ce qui est souvent le cas lorsqu'on doit résoudre une succession d'équations linéaires provenant de la linéarisation d'une même équation non linéaire. Elles ont aussi des inconvénients, par exemple, de ne pouvoir exploiter la « creusité » (caractère creux) éventuelle de A que par des produits matrice-vecteur plus rapides et d'avoir, en pratique, besoin d'un bon préconditionneur pour converger en un nombre raisonnable d'itérations.

Notons pour terminer qu'il existe aussi des méthodes itératives ne trouvant pas la solution en un nombre fini d'itérations : méthodes de Jacobi, de Gauss-Seidel, de relaxation... Les premières ont été introduites au XIX^e siècle pour calculer plus rapidement (à la main, bien sûr) des solutions approchées, alors que la résolution par élimination directe était trop fastidieuse (voir la lettre de Gauss en épigraphe de ce chapitre). Elles sont cependant généralement considérées comme moins efficaces que les méthodes de Krylov [601]. Pour une introduction à ces méthodes, on pourra consulter les livres de Varga [605 ; 1962], de Ciarlet [126 ; 1982] et de Hackbusch [298 ; 1994].

Connaissances supposées. L'introduction de l'algorithme du gradient conjugué se fait en utilisant le procédé d'orthogonalisation de Gram-Schmidt (section B.1.1).

8.1 Sous-espaces de Krylov

On appelle *sous-espace de Krylov* d'ordre p, associé à une matrice A d'ordre n et à un vecteur $r \in \mathbb{R}^n$, le sous-espace vectoriel de $\mathbb{R}^n$ engnedré par les vecteurs r, Ar,

A^2r, ..., $A^{p-1}r$. On le note

$$K_p \equiv K_p(A,r) \equiv \text{vect}\{r, Ar, \ldots, A^{p-1}r\}.$$

La proposition suivante montre que la dimension de K_p vaut p, tant que p reste inférieur ou égal à un certain indice s, à partir duquel sa dimension ne bouge plus (et vaut s). On peut avoir $s < n$. On dit que le sous-espace de Krylov K_s est *saturé* et que $s = s(A,r)$ est l'*indice de saturation* des sous-espaces de Krylov associés à A et à r. La proposition montre aussi que l'indice de saturation est atteint dès que $A^{-1}r$ est « capturé » par les sous-espaces de Krylov.

Proposition 8.1 *Si A est inversible, alors il existe un indice $s \geqslant 1$ tel que*

$$K_1 \subsetneqq \cdots \subsetneqq K_s = K_p, \quad \forall p \geqslant s.$$

L'indice s est caractérisé par le fait que

$$A^{-1}r \in K_s \setminus K_{s-1}.$$

Démonstration. Il suffit de montrer que

$$A^{-1}r \in K_p \iff K_p = K_{p+1}.$$

Le cas où $r = 0$ est trivial. On peut donc supposer que $r \neq 0$.

Si $A^{-1}r \in K_p$, alors il existe des réels α_i, non tous nuls (car $r \neq 0$) tels que

$$A^{-1}r = \sum_{i=0}^{j-1} \alpha_i A^i r, \qquad 1 \leqslant j \leqslant p, \qquad \alpha_{j-1} \neq 0.$$

En appliquant A^{p-j+1}, on en déduit

$$\begin{aligned} A^{p-j}r &= \sum_{i=0}^{j-1} \alpha_i A^{p-j+i+1} r, \\ A^p r &= \frac{1}{\alpha_{j-1}} \left(A^{p-j}r - \sum_{i=0}^{j-2} \alpha_i A^{p-j+i+1} r \right). \end{aligned}$$

Ceci montre que $K_{p+1} = K_p$.

Inversement, si $K_p = K_{p+1}$, on a $A^p r \in K_p$. Il existe donc des réels β_i, non tous nuls, tels que

$$A^p r = \sum_{i=k}^{p-1} \beta_i A^i r, \qquad 0 \leqslant k \leqslant p-1, \qquad \beta_k \neq 0.$$

En appliquant A^{-k-1}, on en déduit

$$A^{-1}r = \frac{1}{\beta_k} \left(- \sum_{i=k+1}^{p-1} \beta_i A^{i-k-1} r + A^{p-k-1} r \right).$$

Ceci montre que $A^{-1}r \in K_{p-k} \subseteq K_p$. □

L'indice s de saturation des sous-espaces de Krylov $K_p(A, r)$ dépend du vecteur r. Par exemple, une conséquence immédiate de la proposition 8.1 est que l'on a $s = 1$ si, et seulement si, r est un vecteur propre de A. La proposition suivante donne une borne supérieure pour $s(A, r)$ ne dépendant que de la matrice A. Nous aurons besoin de la notion suivante (voir [414 ; 1973] par exemple).

On dit que le polynôme

$$p(\xi) = a_0 + a_1\xi + \cdots + a_d\xi^d,$$

de degré d en ξ et à coefficients dans $\mathbb{R}$ *annihile* la matrice A si

$$p(A) \equiv a_0 + a_1A + \cdots + a_dA^d$$

est la matrice nulle. On dit que p est un *polynôme minimal annihilant* A si $p \neq 0$ et s'il n'y a pas de polynôme non nul annihilant A de degré strictement inférieur à celui de p. On dit enfin qu'un polynôme est *unitaire* si le coefficient du terme de degré le plus élevé vaut 1.

On sait qu'il existe un unique polynôme minimal unitaire annihilant A. Il s'écrit

$$\check{p}(\xi) = \prod_{i=1}^{t} (\xi - \lambda_i)^{\beta_i},$$

où $\lambda_1, \ldots, \lambda_t$ sont toutes les valeurs propres distinctes de A et $1 \leqslant \beta_i \leqslant \alpha_i$, α_i étant la multiplicité algébrique de λ_i (sa multiplicité comme racine du polynôme caractéristique de A). La valeur des β_i peut être obtenue en calculant la forme normale de Jordan de A (voir plus loin).

Proposition 8.2 *Si A est inversible, alors pour tout $r \in \mathbb{R}^n$*

$$s(A, r) \leqslant \beta,$$

où β est de degré du polynôme minimal unitaire annihilant A. Cette majoration est optimale : on peut trouver un vecteur r pour lequel on a $s(A, r) = \beta$.

Démonstration. Le polynôme minimal unitaire annihilant A s'écrit

$$\check{p}(\xi) = \prod_{i=1}^{t} (\xi - \lambda_i)^{\beta_i} = a_0 + a_1\xi + \cdots + a_\beta\xi^\beta,$$

où $\beta = \sum_{i=1}^{t} \beta_i$ et $a_\beta = 1$. Le coefficient $a_0 = \prod_{i=1}^{t} (-\lambda_i)^{\beta_i}$ est non nul car, étant inversible, A n'a pas de valeur propre nulle. Comme $\check{p}$ annihile A, on a

$$a_0 + a_1A + \cdots + a_\beta A^\beta = 0.$$

En appliquant A^{-1}, on a quel que soit $r \in \mathbb{R}^n$:

$$A^{-1}r \in \text{vect}\{r, Ar, \ldots, A^{\beta-1}r\} = K_\beta(A, r).$$

D'après la proposition 8.1, cela implique que $s(A, r) \leqslant \beta$.

Enfin, tout polynôme $\tilde{p}$ de degré $\beta - 1$ dont le terme de degré zéro est non nul n'annihile pas A. Il existe donc un vecteur $r \in \mathbb{R}^n$ tel que $\tilde{p}(A)r \neq 0$. Ceci implique que $A^{-1}r \notin K_{\beta-1}$ et par conséquent l'indice de saturation $s(A, r) = \beta$. □

Comme le polynôme caractéristique de A annihile A (théorème de Cayley-Hamilton) et est de degré n, on a nécessairement

$$s \leqslant n,$$

ce que l'on savait déjà. Dans le cas général, le degré

$$\beta = \sum_{i=1}^{t} \beta_i \tag{8.4}$$

du polynôme minimal annihilant A peut être calculé à partir de la forme normale de Jordan de A : il existe une matrice inversible V telle que

$$V^{-1}AV = \text{diag}(J_1^1, \ldots, J_1^{\gamma_1}, J_2^1, \ldots, J_2^{\gamma_2}, \ldots, J_t^1, \ldots, J_t^{\gamma_t}),$$

où chaque J_i^j est un bloc de Jordan de valeur propre λ_i, c'est-à-dire ayant la forme suivante (les éléments en dehors des deux diagonales indiquées sont nuls) :

$$J_i^j = \begin{pmatrix} \lambda_i & 1 & & \\ & \lambda_i & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{pmatrix}.$$

La forme normale de Jordan est unique à une permutation des blocs près. Alors, β_i est l'ordre le plus élevé des blocs de Jordan de valeur propre λ_i :

$$\beta_i = \max_{i \leqslant j \leqslant \gamma_i} \text{ordre}(J_i^j).$$

On en déduit le corollaire suivant :

Corollaire 8.3 *Si A est inversible et non défective, alors pour tout $r \in \mathbb{R}^n$, $s(A, r) \leqslant t$, où t est le nombre de valeurs propres distinctes de A.*

Démonstration. En effet, A est non défective si (par définition) elle est diagonalisable par similitude : il existe une matrice V d'ordre n, inversible telle que

$$V^{-1}AV = \Lambda,$$

où Λ est diagonale. Dans ce cas, les blocs de Jordan sont des matrices d'ordre 1 et $\beta_i = 1$, $\forall i$. On en déduit le résultat en utilisant (8.4) et le théorème. □

Exemples 8.4 La matrice

$$A_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

est diagonale et a deux valeurs propres distinctes. Donc $s(A_1, r) \leqslant 2$, $\forall r \in \mathbb{R}^3$.

La matrice

$$A_2 = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

est défective et sous forme normale de Jordan. A la valeur propre 1 correspond un bloc de Jordan d'ordre 2. Donc $s(A_2, r) \leqslant 3$, $\forall r \in \mathbb{R}^3$. En prenant le vecteur

$$r_2 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$$

on peut voir que $s(A_2, r_2) = 3$. □

Les deux algorithmes itératifs de résolution du système linéaire (8.1) que nous décrivons dans ce chapitre, l'algorithme du gradient conjugué (section 8.2) et l'algorithme GMRES (section 8.3), sont des méthodes de Krylov, ce qui veut dire que l'itété x_k appartient à un certain sous-espace de Krylov $K_k(A, r)$. Dans les deux cas, le vecteur r est le résidu $r_1 = b - Ax_1$ évalué en l'itéré initial x_1. La raison pour laquelle ce choix de r convient est expliqué au début de la section 8.3.1. L'algorithme du gradient conjugué est adapté aux systèmes linéaires dans lesquels la matrice A est symétrique définie positive (ou semi-définie positive), alors que l'algorithme GMRES est plus général (et plus difficile à mettre en œuvre) puisqu'il s'affranchit de toute hypothèse sur A.

8.2 Algorithme du gradient conjugué

Dans cette section, on considère le cas où A est symétrique définie positive. On rappelle que résoudre le système linéaire (8.1) revient alors à minimiser la fonction

$$f(x) = \frac{1}{2} x^\mathsf{T} A x - b^\mathsf{T} x$$

sur $\mathbb{R}^n$.

L'*algorithme du gradient conjugué* est une méthode à direction de descente sur f. Les itérés sont donc générés par

$$x_{k+1} = x_k + \alpha_k d_k,$$

où d_k est une direction de descente de f (elle vérifie $\nabla f(x_k)^\mathsf{T} d_k < 0$) et α_k est un pas positif. On notera

$$E_k = \text{vect}\{d_1, d_2, \ldots, d_k\}$$

($k \geqslant 1$) le sous-espace vectoriel engendré par les k premières directions de descente.

8.2.1 Notion de directions conjuguées

Pour une matrice A symétrique définie positive donnée, la notion de conjugaison équivaut à la notion d'orthogonalité pour le produit scalaire associé à A. De façon plus précise, on a les définitions suivantes.

On dit que u et $v \in \mathbb{R}^n$ sont *conjuguées* si $u^\mathsf{T} A v = 0$. Plus généralement, on dit que $u_1, \dots u_p \in \mathbb{R}^n$ sont conjuguées si $u_i^\mathsf{T} A u_j = 0$, $\forall i \neq j$. Enfin, on dit que u est conjuguée par rapport à un sous-espace vectoriel $E \subseteq \mathbb{R}^n$ si $u^\mathsf{T} A v = 0$, $\forall v \in E$.

Si $v_1, \dots, v_p$ sont des vecteurs linéairement indépendants dans $\mathbb{R}^n$, on pourra leur associer p directions conjuguées $u_1, \dots, u_p$ en utilisant le procédé du Gram-Schmidt (section B.1.1) avec le produit scalaire $\langle u, v \rangle = u^\mathsf{T} A v$, dans lequel on se passe des étapes de normalisation superflues (étapes 1 et 2.3). Il suffit de prendre, pour $k = 1, \dots, p$,

$$u_k = v_k - \sum_{i=1}^{k-1} \beta_{ki} u_i \tag{8.5}$$

avec

$$\beta_{ki} = \frac{v_k^\mathsf{T} A u_i}{u_i^\mathsf{T} A u_i}. \tag{8.6}$$

8.2.2 Algorithme des directions conjuguées

Une méthode de directions conjuguées pour la minimisation d'une fonction quadratique f est une méthode qui à chaque itération prend $x_k \in x_1 + E_k$ de façon à minimiser f sur le sous-espace affine $x_1 + E_k$. Le lien avec la conjugaison des directions est donné par la proposition suivante.

On note

$$g_k = \nabla f(x_k) = A x_k - b$$

le gradient de f en x_k pour le produit scalaire euclidien et

$$y_k = g_{k+1} - g_k.$$

Proposition 8.5 *Si x_k réalise le minimum de f sur $x_1 + E_{k-1}$ ($k \geqslant 2$), $d_k \neq 0$ et $x_{k+1} = x_k + \alpha_k d_k$, alors x_{k+1} réalise le minimum de f sur $x_1 + E_k$ si, et seulement si, l'une des conditions suivantes est réalisée*
(i) $g_k^\mathsf{T} d_k = 0$ *et* $\alpha_k = 0$,
(ii) $g_k^\mathsf{T} d_k \neq 0$, d_k *est conjuguée par rapport à* $d_1, \dots, d_{k-1}$ *et*

$$\alpha_k = -\frac{g_k^\mathsf{T} d_k}{d_k^\mathsf{T} A d_k}. \tag{8.7}$$

Démonstration. Si x_k réalise le minimum de f sur $x_1 + E_{k-1}$, g_k est orthogonal à E_{k-1}, c'est-à-dire :

$$g_k^\mathsf{T} d_i = 0, \quad \text{pour } i = 1, \dots, k-1. \tag{8.8}$$

D'autre part, f étant quadratique, on a

$$g_{k+1} = g_k + A(x_{k+1} - x_k) = g_k + \alpha_k A d_k. \tag{8.9}$$

Alors, comme en (8.8), x_{k+1} réalise le minimum de f sur $x_1 + E_k$ si, et seulement si,

$$g_{k+1}^{\mathsf{T}} d_i = 0, \quad \text{pour } i = 1, \dots, k.$$

D'après (8.9), ceci est équivalent à

$$g_k^{\mathsf{T}} d_i + \alpha_k d_k^{\mathsf{T}} A d_i = 0, \quad \text{pour } i = 1, \dots, k.$$

En prenant $i = k$, on trouve soit que $g_k^{\mathsf{T}} d_k = 0$ et alors $\alpha_k = 0$ ($d_k \neq 0$), soit que $g_k^{\mathsf{T}} d_k \neq 0$ et α_k est donné par (8.7). En prenant $1 \leqslant i \leqslant k-1$ et en tenant compte de (8.8), on trouve dans le cas (ii) que d_k est conjuguée par rapport à $d_1, \dots, d_{k-1}$. Inversement, on voit que les conditions (i) ou (ii) sont suffisantes. □

Le pas α_k donné en (8.7) est optimal dans le sens où il réalise le minimum de l'application

$$\alpha \to f(x_k + \alpha d_k).$$

D'après la proposition 8.5, si l'on veut minimiser f sur les sous-espaces affines successifs $x_1 + E_1$, $x_1 + E_2$..., il suffit de se donner n directions conjuguées $d_1, \dots, d_n$ et de minimiser f le long des droites $\alpha \to x_k + \alpha d_k$. Ces directions conjuguées peuvent s'obtenir à partir de la donnée de n directions linéairement indépendantes $\tilde{d}_1, \dots, \tilde{d}_n$ que l'on utilisera pour construire des directions conjuguées au moyen du procédé de Gram-Schmidt (8.5)–(8.6), où v_i devient $\tilde{d}_i$ et u_i devient d_i. On obtient alors l'algorithme suivant :

Algorithme 8.6 (des directions conjuguées) Une itération passe de l'itéré courant $x_k \in \mathbb{R}^n$ à l'itéré suivant $x_{k+1} \in \mathbb{R}^n$ par les étapes suivantes :

1. *Calcul de la direction* : on choisit une direction arbitraire $\tilde{d}_k \notin \text{vect}\{\tilde{d}_1, \dots, \tilde{d}_{k-1}\}$ et on calcule la direction d_k par la formule

$$d_k = \begin{cases} \tilde{d}_1 & \text{si} \quad k = 1, \\ \tilde{d}_k - \displaystyle\sum_{i=1}^{k-1} \frac{\tilde{d}_k^{\mathsf{T}} A d_i}{d_i^{\mathsf{T}} A d_i} d_i & \text{si} \quad k \geqslant 2. \end{cases} \tag{8.10}$$

2. *Calcul du pas optimal* :

$$\alpha_k = -\frac{g_k^{\mathsf{T}} d_k}{d_k^{\mathsf{T}} A d_k}.$$

3. *Nouveau point* : $x_{k+1} = x_k + \alpha_k d_k$.

Cet algorithme ne demande pas la connaissance explicite de la matrice A. Il suffit de pouvoir calculer le produit de cette matrice par un vecteur arbitraire. Ce produit matrice-vecteur intervient à deux reprises dans l'algorithme ci-dessus : dans le calcul du gradient et dans la détermination du pas optimal. Si cette opération est coûteuse en temps de calcul, on peut n'en faire qu'une à chaque itération, pour calculer Ad_k, et en mettant à jour le gradient par la formule $g_{k+1} = g_k + \alpha_k A d_k$.

La méthode des directions conjuguées peut être améliorée en ce qui concerne l'encombrement mémoire, mais nous allons voir qu'un choix judicieux des directions $\tilde{d}_1$, $\ldots$, $\tilde{d}_n$ conduit à une amélioration beaucoup plus importante encore.

8.2.3 Algorithme du gradient conjugué

La méthode du gradient conjugué est le cas particulier de la méthode précédente ou l'on prend

$$\tilde{d}_k = -g_k. \tag{8.11}$$

Comme on va le voir, ce choix conduit à une simplification considérable de l'algorithme des directions conjuguées.

Proposition 8.7 *Dans la méthode du gradient conjugué, le sous-espace vectoriel E_k engendré par les directions d_1, $\ldots$, d_k est aussi le sous-espace vectoriel engendré par les gradients successifs :*

$$E_k = \mathrm{vect}\{g_1, \ldots, g_k\} \tag{8.12}$$

et ceux-ci sont orthogonaux entre eux :

$$g_k^{\mathsf{T}} g_i = 0, \quad 1 \leqslant i \leqslant k-1. \tag{8.13}$$

Démonstration. En effet, grâce au choix (8.11), les directions d_1, $\ldots$, d_{k-1} sont à présent obtenues en orthogonalisant $-g_1$, $\ldots$, $-g_{k-1}$ (pour le produit scalaire associé à A). La relation (8.12) est alors une conséquence du procédé de Gram-Schmidt.

Ensuite, comme x_k réalise un minimum de f sur $x_1 + E_{k-1}$, g_k est orthogonal à E_{k-1}. Par (8.12), on en déduit (8.13). □

D'après la relation d'orthogonalité des gradients (8.13), on a

$$g_k^{\mathsf{T}} y_i = g_k^{\mathsf{T}} (g_{i+1} - g_i) = 0, \quad \text{pour } 1 \leqslant i \leqslant k-2.$$

Alors, en utilisant (8.9), l'expression du coefficient de d_i dans (8.10) devient

$$\frac{\tilde{d}_k^{\mathsf{T}} y_i}{d_i^{\mathsf{T}} y_i}.$$

Compte tenu de la relation d'orthogonalité des gradients, on voit que seul le dernier terme de la somme dans (8.10) est non nul et que l'opposé de son coefficient s'écrit

$$\beta_k = \frac{\|g_k\|^2}{\|g_{k-1}\|^2}. \tag{8.14}$$

On a utilisé le fait que, d'après (8.13), $-g_k^\mathsf{T} y_{k-1} = \|g_k\|^2$ et $d_{k-1}^\mathsf{T} y_{k-1} = -d_{k-1}^\mathsf{T} g_{k-1} = \|g_{k-1}\|^2$. Ceci conduit à l'algorithme du gradient conjugué pour les fonctions quadratiques.

Algorithme 8.8 (du gradient conjugué) Une itération met à jour l'itéré courant $x_k \in \mathbb{R}^n$, le gradient courant g_k et le carré de la norme de ce dernier $\gamma_k := \|g_k\|_2^2$, par les étapes suivantes.

1. *Test d'arrêt* : si $\gamma_k \simeq 0$, on s'arrête.
2. *Paramètre de conjugaison* : si $k \geqslant 2$, $\beta_k := \gamma_k/\gamma_{k-1}$.
3. *Déplacement en x* :

$$d_k = \begin{cases} -g_1 & \text{si } k = 1 \\ -g_k + \beta_k d_{k-1} & \text{si } k \geqslant 2. \end{cases}$$

4. *Déplacement en g* : $p_k = Ad_k$.
5. *Calcul du pas* : $\alpha_k = \gamma_k/(d_k^\mathsf{T} p_k)$.
6. *Nouveau point* : $x_{k+1} = x_k + \alpha_k d_k$.
7. *Nouveau gradient* : $g_{k+1} = g_k + \alpha_k p_k$ et $\gamma_{k+1} = \|g_{k+1}\|_2^2$.

8.2.4 Propriétés de l'algorithme du gradient conjugué

Terminaison finie

Notons que tant que $g_k \neq 0$, les directions $g_1, \ldots, g_k$ sont linéairement indépendantes. Cela résulte de la condition d'orthogonalité (8.13). Les directions $d_1, \ldots, d_k$ seront donc conjuguées, et le sous-espace vectoriel E_k sera de dimension k. Par conséquent, *l'algorithme du gradient conjugué trouve la solution en au plus n itérations.* On peut être plus précis.

On peut montrer que le sous-espace vectoriel E_k n'est autre que le sous-espace de Krylov

$$K_k(A, g_1) \equiv \text{vect}\{g_1, Ag_1, \ldots, A^{k-1} g_1\},$$

associé à A et g_1 (exercice 8.1). Ces sous-espaces saturent en $k = s$ lorsque

$$A^{-1} g_1 \in K_s.$$

Comme $A^{-1} g_1 = x_1 - x_*$, la saturation se produit lorsque $x_* \in x_1 + K_s$. Comme la méthode du gradient conjugué minimise f sur $x_1 + E_k = x_1 + K_k$, elle s'arrêtera exactement en $k = s(A, g_1)$. Une matrice symétrique définie positive étant non défective, on a, grâce au corollaire 8.3, $s(A, g_1) \leqslant t$, où t est le nombre de valeurs propres distincts de A. Nous avons donc montré la proposition suivante.

Proposition 8.9 *La méthode du gradient conjugué trouve le minimum d'une fonction quadratique strictement convexe en au plus t itérations, où t est le nombre de valeurs propres distinctes de la hessienne de la fonction.*

Vitesse de convergence

Si n est grand (par exemple $n \geqslant 10^3$), il est trop coûteux d'effectuer n itérations de l'algorithme du gradient conjugué pour obtenir la solution. La qualité de l'approximation x_k obtenue peut être estimée par la vitesse de convergence de l'algorithme. Le résultat suivant [Luenberger (1973), pg. 187] donne une estimation de la décroissance de la norme de l'erreur $x_k - x_*$, en termes du conditionnement ℓ_2 de la matrice A :

$$\kappa_2(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}.$$

La norme utilisée est la suivante :

$$\|x\|_A = (x^\mathsf{T} A x)^{1/2}.$$

Proposition 8.10 *Soit κ le conditionnement ℓ_2 de A. La suite $\{x_k\}$ générée par l'algorithme du gradient conjugué vérifie l'estimation*

$$\|x_{k+1} - x_*\|_A \leqslant 2 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k \|x_1 - x_*\|_A.$$

Il s'agit d'une vitesse de convergence r-linéaire, dont le taux est à comparer avec celui de la méthode du gradient à pas optimal (proposition 7.2), qui forme une suite $\{x_k\}$ vérifiant :

$$\|x_{k+1} - x_*\|_A \leqslant \left(\frac{\kappa - 1}{\kappa + 1} \right) \|x_k - x_*\|_A \leqslant \left(\frac{\kappa - 1}{\kappa + 1} \right)^k \|x_1 - x_*\|_A.$$

La convergence est donc plus rapide avec la méthode du gradient conjugué. La figure 8.1 permet de comparer les taux de convergence des méthodes du Gradient et du gradient conjugué en fonction du conditionnement κ.

Si la matrice symétrique A n'est pas définie positive

Si A a des valeurs propres strictement positives et négatives, le dénominateur apparaissant dans le calcul du pas α_k (étape 3.5) n'est plus nécessairement strictement positif ; il peut être arbitrairement proche de zéro ou même nul, ce qui entraîne l'instabilité ou l'échec de l'algorithme. Ce mauvais comportement de l'algorithme

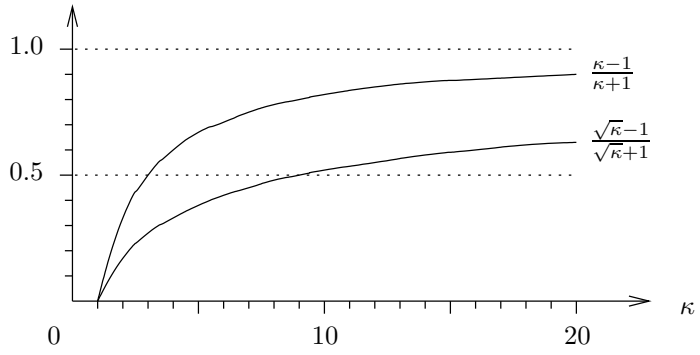


Fig. 8.1. Comparaison des taux de convergence du G et du GC

s'observe en général dès les premières itérations, si bien que l'on ne peut espérer trouver ainsi une solution approchée du système linéaire. Évidemment, si toutes les valeurs propres de A sont strictement négatives, on pourra résoudre par gradient conjugué le système linéaire $(-A)x = -b$, équivalent à (8.1), puisque sa matrice $-A$ est définie positive. En conclusion, *l'algorithme du gradient conjugué ne convient pas pour résoudre des systèmes linéaires dont la matrice est indéfinie, même de manière approchée.*

Le comportement de l'algorithme du GC lorsque $A \succcurlyeq 0$ est plus satisfaisant. Il est décrit dans la proposition suivante. L'algorithme est dit bien défini à l'itération k si $d_k^{\mathsf{T}} A d_k > 0$, si bien que le pas α_k est bien défini à l'étape 3.5 et que l'on peut passer de x_k à x_{k+1}. Enfin, lorsque $b \in \mathcal{R}(A)$, la *solution de norme minimale* de $Ax = b$ est l'unique solution du problème

$$\begin{cases} \min \ \frac{1}{2}\|x\|_2^2 \\ Ax = b, \end{cases}$$

qui est donc la projection de 0 sur le sous-espace affine $\{x \in \mathbb{R}^n : Ax = b\}$. Comme A est symétrique, les conditions d'optimalité de ce problème, qui sont ici nécessaires et suffisantes car le problème est convexe, montrent que c'est l'unique point x vérifiant $Ax = b$ et $x \in \mathcal{R}(A)$.

Proposition 8.11 (GC lorsque $A \succcurlyeq 0$) *On considère l'algorithme du gradient conjugué pour résoudre le système linéaire (8.1) ou minimiser la fonction quadratique f définie en (8.2), avec A symétrique semi-définie positive de rang r.*

1) *Si $b \notin \mathcal{R}(A)$, l'algorithme est bien défini pendant au plus r itérations, disons jusqu'en x_l, où la direction d_l générée est non nulle, dans le noyau de A et telle que $f(x_l + \alpha d_l) = f(x_l) - \alpha\, b^{\mathsf{T}} d_l \to -\infty$ lorsque $\alpha \to +\infty$.*
2) *Si $b \in \mathcal{R}(A)$, l'algorithme est bien défini et converge en au plus r itérations ; de plus, si l'itéré initial est pris dans $\mathcal{R}(A)$ (par exemple $x_1 = 0$), les itérés convergent vers la solution de norme minimale de (8.1).*

DÉMONSTRATION. 1) Si A est semi-définie positive et $b \notin \mathcal{R}(A)$, le système linéaire (8.1) n'a pas de solution. L'algorithme du GC peut itérer tant que le dénominateur $d_k^\mathsf{T} A d_k$ de la fraction donnant le pas α_k à l'étape 3.5 reste strictement positif (avec une instabilité possible si ce dénominateur s'approche de zéro). Alors l'itéré suivant x_{k+1} minimise toujours f sur l'espace de Krylov $K_k(A, g_1)$. Observons que K_k est de dimension k et que $K_k \cap \mathcal{N}(A) = \{0\}$, si bien que cette situation ne peut se poursuivre au-delà de $\dim K_k \leqslant n - \dim \mathcal{N}(A) = \dim \mathcal{R}(A) = r$ itérations : l'algorithme finit par trouver une direction $d_l \in \mathcal{N}(A)$ et doit s'arrêter. Comme $g_l \perp K_{l-1} \ni d_{l-1}$, $g_l^\mathsf{T} d_l = -\|g_l\|_2^2 < 0$ ($g_l \neq 0$ car $b \notin \mathcal{R}(A)$), d_l est une direction de descente (non nulle) de f en x_l. On en déduit que $f(x_l + \alpha d_l) = f(x_l) - \alpha\, b^\mathsf{T} d_l \downarrow -\infty$ lorsque $\alpha \uparrow +\infty$ ($b^\mathsf{T} d_l > 0$ car $0 > g_l^\mathsf{T} d_l = (Ax_l - b)^\mathsf{T} d_l = -b^\mathsf{T} d_l$).

2) Montrons que $x_k \in x_1 + \mathcal{R}(A)$, pour tout $k \geqslant 1$. On obtient ce résultat par récurrence en montrant que toutes les directions de recherche du GC vérifient $d_k \in \mathcal{R}(A)$. C'est clairement vrai pour la première direction qui s'écrit $d_1 = -g_1 = -Ax_1 + b \in \mathcal{R}(A)$, puisque $b \in \mathcal{R}(A)$. Ensuite, pour $k \geqslant 2$, $d_k = -g_k + \beta_k d_{k-1}$ est dans $\mathcal{R}(A)$ par récurrence.

Pour vérifier que le GC est bien défini, il suffit de vérifier que $d_k^\mathsf{T} A d_k$ est strictement positif tant que x_k n'est pas solution de (8.1). Cette propriété résulte du fait que $d_k \in \mathcal{R}(A)$ et que A est définie positive sur $\mathcal{R}(A)$ ($d^\mathsf{T} A d = 0$ et $d \in \mathcal{R}(A)$ impliquent que $d \in \mathcal{N}(A) \cap \mathcal{R}(A) = \{0\}$).

Enfin, comme $\mathcal{R}(A)$ est de dimension r, qu'avant convergence les directions d_k sont linéairement indépendantes et engendrent K_k et que $K_k \subseteq \mathcal{R}(A)$, l'algorithme converge en au plus r itérations.

D'autre part, si x_* est la solution trouvée par l'algorithme, on a de ce qui précède $x_* - x_1 \in \mathcal{R}(A)$. Si $x_1 \in \mathcal{R}(A)$ on en déduit que $x_* \in \mathcal{R}(A)$ et $Ax_* = b$. Dès lors x_* est la solution de norme minimale de (8.1). □

8.2.5 Mise en œuvre de la méthode du gradient conjugué

Encombrement mémoire et comptage des opérations

L'algorithme 8.8 du gradient conjugué est particulièrement économe en place mémoire. C'est l'un de ses principaux attraits. Il ne requiert que le stockage de 4 vecteurs, ceux mémorisant x_k, g_k, d_k et p_k (les trois premiers sont mis à jour à chaque itération en remplaçant le vecteur de l'itération précédente).

Le nombre d'opérations par itération est en $O(10n)$, plus un unique produit matrice-vecteur ($O(2n^2)$ opérations si la matrice est pleine). En arithmétique exacte, nous avons vu que l'algorithme requiert au plus n itérations pour trouver la solution. Il faut donc $O(2n^3)$ opérations pour trouver la solution d'un système linéaire symétrique défini positif plein par l'algorithme du gradient conjugué. On retrouve un nombre d'opérations du même ordre que la factorisation gaussienne.

Le gradient conjugué préconditionné

Dès que le nombre n de variables dépasse quelques dizaines ou centaines (cela dépend du conditionnement de A), l'algorithme du GC peut avoir des difficultés à minimiser la fonction quadratique et, du fait des erreurs d'arrondi, il peut demander

beaucoup plus de n itérations pour avoir une solution acceptable. Pour remédier à cette situation, il faut essayer d'améliorer le conditionnement du problème, c'est-à-dire *préconditionner* le problème ou l'algorithme. Nous présentons ci-après l'approche qui utilise un changement de variables.

Si l'on fait le changement de variables

$$\tilde{x} = Lx, \tag{8.15}$$

au moyen d'une matrice L d'ordre n inversible, la fonction $\tilde{f} : \mathbb{R}^n \to \mathbb{R}$ définie par

$$\tilde{f} = f \circ L^{-1}$$

est telle que $\tilde{f}(\tilde{x}) = f(x)$, si $\tilde{x}$ et x se correspondent par (8.15). Les fonctions f et $\tilde{f}$ atteignent donc la même valeur minimale et ce aux points x_* et $\tilde{x}_* = Lx_*$, respectivement. L'algorithme du GC préconditionné s'obtient en appliquant l'algorithme du GC standard dans l'espace des $\tilde{x}$ et en traduisant l'algorithme résultant dans l'espace des x. L'algorithme est ainsi modifié, amélioré si le changement de variables est bien choisi ; on dit aussi qu'il est sensible à un changement de variables.

Le gradient et la hessienne de $\tilde{f}$ en $\tilde{x} = Lx$ s'écrivent

$$\tilde{g} \equiv \nabla \tilde{f}(\tilde{x}) = L^{-\mathsf{T}} \nabla f(x) \qquad \text{et} \qquad \tilde{A} \equiv \nabla^2 \tilde{f} = L^{-\mathsf{T}} A L^{-1}.$$

Soient $\{\tilde{x}_k\}$ la suite générée par le GC dans l'espace des $\tilde{x}$ et $x_k := L^{-1}\tilde{x}_k$. On note $\tilde{g}_k := \nabla \tilde{f}(\tilde{x}_k)$ et $g_k = \nabla f(x_k)$, si bien que $\tilde{g}_k = L^{-\mathsf{T}} g_k$. La direction de la méthode du gradient conjugué dans l'espace des $\tilde{x}$ s'écrit

$$\tilde{d}_k = -\tilde{g}_k + \tilde{\beta}_k \tilde{d}_{k-1} = -L^{-\mathsf{T}} g_k + \frac{\|L^{-\mathsf{T}} g_k\|^2}{\|L^{-\mathsf{T}} g_{k-1}\|^2} \tilde{d}_{k-1}.$$

Ramenée à l'espace des x, cela donne pour $d_k = L^{-1}\tilde{d}_k$ la formule

$$d_k = \begin{cases} -P g_1 & \text{si} \quad k = 1, \\ -P g_k + \dfrac{g_k^{\mathsf{T}} P g_k}{g_{k-1}^{\mathsf{T}} P g_{k-1}} d_{k-1} & \text{si} \quad k \geqslant 2, \end{cases} \tag{8.16}$$

où

$$P := L^{-1} L^{-\mathsf{T}}.$$

On montre que g_k est encore orthogonal à d_{k-1} (voir l'exercice 8.3), si bien que le pas optimal le long de d_k est donné par la formule

$$\alpha_k = \frac{g_k^{\mathsf{T}} P g_k}{d_k^{\mathsf{T}} A d_k}. \tag{8.17}$$

Si l'on remplace la direction d_k et le pas α_k de la méthode du gradient conjugué par les valeurs données par les formules (8.16) et (8.17), on obtient la *méthode du gradient conjugué préconditionné.*

Algorithme 8.12 (du gradient conjugué préconditionné) Étant donnée une matrice de préconditionnement P, une itération met à jour l'itéré courant $x_k \in \mathbb{R}^n$, le gradient courant g_k et la norme préconditionnée au carré de ce dernier $\gamma_k := g_k^\mathsf{T} P g_k$, par les étapes suivantes.

1. *Test d'arrêt* : si $\gamma_k \simeq 0$, on s'arrête.
2. *Paramètre de conjugaison* : si $k \geqslant 2$, $\beta_k := \gamma_k / \gamma_{k-1}$.
3. *Déplacement en x* :

$$d_k = \begin{cases} -P g_1 & \text{si } k = 1 \\ -P g_k + \beta_k d_{k-1} & \text{si } k \geqslant 2. \end{cases}$$

4. *Déplacement en g* : $p_k = A d_k$.
5. *Calcul du pas* : $\alpha_k = \gamma_k / (d_k^\mathsf{T} p_k)$.
6. *Nouveau point* : $x_{k+1} = x_k + \alpha_k d_k$.
7. *Nouveau gradient* : $g_{k+1} = g_k + \alpha_k p_k$ et $\gamma_{k+1} = g_{k+1}^\mathsf{T} P g_{k+1}$.

Le choix de L, donc de P, est gouverné par le souhait d'avoir $\tilde{A}$ proche de la matrice identité, ce qui a pour effet d'accélérer la convergence de l'algorithme dans l'espace des $\tilde{x}$ (proposition 8.10), donc aussi dans l'espace des x. On voit qu'il est souhaitable de prendre L proche de $A^{1/2}$, ou encore

$$P \simeq A^{-1}.$$

Préconditionneur diagonal

Le *préconditionneur diagonal*, qui consiste à prendre pour P une matrice diagonale, est sans doute le plus simple. On peut par exemple prendre $L = \mathrm{Diag}(A_{ii}^{1/2})$ en (8.15) (on se rappelle que les $A_{ii} > 0$) et donc

$$P = \mathrm{Diag}(A_{11}, \dots, A_{nn})^{-1}. \tag{8.18}$$

Ce dernier préconditionneur diagonal a une propriété de minimalité, dans le sens où il minimise, à un facteur n près, le conditionnement de DAD parmi toutes les matrices diagonales D. En effet, selon van der Sluis [599 ; théorème 4.1], on a

$$\kappa_2(P^{1/2} A P^{1/2}) \leqslant n \left(\min_{\substack{D \text{ diagonale} \\ D \succ 0}} \kappa_2(DAD) \right).$$

S'il a le mérite de la simplicité, sauf cas exceptionnels (celui d'une matrice A diagonale est un cas particulier évident), ce préconditionneur élémentaire n'apporte pas toujours une amélioration notable de la vitesse de convergence de l'algorithme du gradient conjugué. Il ne doit toutefois pas être négligé, si aucun autre préconditionneur ne s'impose.

Si les éléments diagonaux A_{ii}, pour $i \in I$, sont nuls, la formule (8.18) n'est plus bien définie. Cependant, la semi-définie positivité de A implique alors que les colonnes (et les lignes) de A avec indice dans I sont nulles, si bien que les vecteurs de base $\{e^i\}_{i\in I}$ sont dans le noyau de A. Si $b_I = 0$, les itérés de l'algorithme du gradient conjugué sont générés dans $x_1 + \mathrm{vect}\{e^i : i \in I\}^\perp$ et on peut prendre le préconditionneur diagonal P avec $P_{ii} = 0$ si $i \in I$ et $P_{ii} = A_{ii}^{-1}$ sinon. S'il existe un indice $i \in I$ tel que $b_i \neq 0$, la direction $d = \mathrm{sgn}(b_i)e^i$ est une *direction de non bornitude* de la forme quadratique associée $x \mapsto f(x) = \frac{1}{2}x^\mathsf{T}Ax - b^\mathsf{T}x$, dans le sens où, quel que soit $x \in \mathbb{R}^n$, $f(x + td) \to -\infty$ lorsque $t \to \infty$; l'équation $Ax = b$ n'a pas de solution.

Préconditionneur projecteur

Supposons à présent que le préconditionneur symétrique $P \succcurlyeq 0$ soit un projecteur ($P^2 = P$) non trivial ($P \neq I$, donc avec des valeurs propres nulles). L'algorithme du GC préconditionné par P, utilisant les directions (8.16) et les pas (8.17), est encore bien défini et génère une suite dans $x_1 + \mathcal{R}(P)$ (voir l'exercice 8.3). Il porte le nom d'*algorithme du gradient conjugué projeté* pour minimiser le critère quadratique sur le sous-espace affine $x_1 + \mathcal{R}(P)$.

Redémarrage du gradient conjugué

La méthode du gradient conjugué est une méthode dans laquelle les erreurs d'arrondi sont amplifiées au cours des itérations. Peu à peu les relations de conjugaison de d_k avec les premières directions sont perdues parce qu'en présence d'erreurs d'arrondi les relations (8.13) d'orthogonalité des gradients ne sont pas vérifiées exactement.

Notons que les erreurs d'arrondi sont mieux contrôlées si l'on remplace la formule (8.14) de β_k, dite de *Fletcher-Reeves* (1964), par la formule équivalente, dite de *Polak-Ribière* (1969) :

$$\beta_k = \frac{g_k^\mathsf{T} y_{k-1}}{\|g_{k-1}\|^2}.$$

On peut détecter la présence d'erreurs d'arrondi en regardant si

$$|g_k^\mathsf{T} g_{k-1}| \geqslant \nu \|g_k\|^2,$$

où $\nu \simeq 0.2$ (normalement le membre de gauche doit être nul). Ce critère est connu sous le nom de *critère de redémarrage de Powell* (1977). S'il est vérifié pour $k = r$, certains numériciens préconisent de redémarrer l'algorithme du gradient conjugué en x_r dans la direction $-g_r$. Cette pratique n'est plus guère utilisée.

8.2.6 Méthode du gradient conjugué non linéaire

On s'intéresse ici à la minimisation d'une fonction $f : \mathbb{R}^n \to \mathbb{R}$, non nécessairement quadratique :

$$\min_{x\in\mathbb{R}^n} f(x),$$

et on cherche à étendre la méthode du gradient conjugué à ce problème. Il y a plusieurs manières de le faire et peu de critères permettant de dire laquelle est la meilleure. Une

extension possible consiste simplement à reprendre les formules utilisées dans le cas quadratique. On se propose donc d'étudier les méthodes où la direction d_k est définie par la formule de récurrence suivante ($\beta_k \in \mathbb{R}$)

$$d_k = \begin{cases} -g_1 & \text{si} \quad k = 1, \\ -g_k + \beta_k d_{k-1} & \text{si} \quad k \geqslant 2, \end{cases} \tag{8.19}$$

et où $\{x_k\}$ est générée par la formule

$$x_{k+1} = x_k + \alpha_k d_k, \tag{8.20}$$

le pas $\alpha_k \in \mathbb{R}$ étant déterminé par une recherche linéaire.

Ces méthodes sont des extensions de la méthode du gradient conjugué si β_k prend l'une des valeurs

$$\begin{aligned} \beta_k^{\mathrm{FR}} &= \frac{\|g_k\|^2}{\|g_{k-1}\|^2}, \\ \beta_k^{\mathrm{PR}} &= \frac{g_k^{\mathsf{T}} y_{k-1}}{\|g_{k-1}\|^2}, \end{aligned}$$

où $y_{k-1} = g_k - g_{k-1}$. Dans le cas quadratique avec recherche linéaire exacte, on a vu que $\beta_k^{\mathrm{FR}} = \beta_k^{\mathrm{PR}}$. Si f est quelconque, il n'en est plus de même et on parle respectivement de *méthode de Fletcher-Reeves* (1964) ou de *méthode de Polak-Ribière* (1969) selon que l'on utilise β_k^{FR} ou β_k^{PR} à la place de β_k dans (8.19).

Pour que les méthodes ainsi définies soient utilisables, il faut répondre aux deux questions suivantes. Les directions d_k définies par (8.19) sont-elles des directions de descente de f ? Les méthodes ainsi définies sont-elles convergentes ?

En ce qui concerne la première question remarquons que, quel que soit $\beta_k \in \mathbb{R}$, d_k est une direction de descente si l'on fait de la recherche linéaire exacte, c'est-à-dire si le pas α_{k-1} est un point stationnaire de $\alpha \to f(x_{k-1} + \alpha d_{k-1})$. En effet, dans ce cas $g_k^{\mathsf{T}} d_{k-1} = 0$ et on trouve lorsque $g_k \neq 0$:

$$g_k^{\mathsf{T}} d_k = -\|g_k\|^2 < 0.$$

Cependant, il est fortement déconseillé de faire de la recherche linéaire exacte lorsque f n'est pas quadratique : le coût de détermination de α_k est excessif.

Le résultat suivant, que l'on peut trouver dans [245 ; 1992], généralise un résultat de Al-Baali (1985). Il montre que si β_k n'est pas trop grand et si l'on utilise la règle de *recherche linéaire de Wolfe forte* :

$$f(x_{k+1}) \leqslant f(x_k) + \omega_1 \alpha_k g_k^{\mathsf{T}} d_k, \tag{8.21}$$

$$|g_{k+1}^{\mathsf{T}} d_k| \leqslant \omega_2 |g_k^{\mathsf{T}} d_k|, \tag{8.22}$$

avec des constantes positives ω_1 et ω_2 bien choisies, alors d_k est une direction de descente et la méthode converge.

Proposition 8.13 *Supposons que l'ensemble* $\mathcal{L} := \{x \in \mathbb{R}^n : f(x) \leqslant f(x_1)\}$ *soit borné et que* f *soit continûment différentiable dans un voisinage de* $\mathcal{L}$ *avec un gradient lipschitzien. Toute méthode du type* (8.19)–(8.20) *dans laquelle* β_k *vérifie*

$$|\beta_k| \leqslant \beta_k^{FR}, \quad \forall k \geqslant 1,$$

et le pas α_k *est déterminé par la règle de Wolfe forte* (8.21)–(8.22) *avec* $0 < \omega_1 < \omega_2 < 1/2$ *est une méthode de descente* ($g_k^{\mathsf{T}} d_k < 0$, $\forall k \geqslant 1$) *convergente, dans le sens où*

$$\liminf_{k\to\infty} \|g_k\| = 0.$$

D'un point de vue théorique, ce résultat est satisfaisant et suggère que la méthode de Fletcher-Reeves est une bonne méthode. Cependant, en pratique, il est préférable d'utiliser la méthode de Polak-Ribière dont les performances moyennes dépassent de loin celles de la méthode de Fletcher-Reeves. Le résultat suivant est dû à Polak et Ribière (1964).

Proposition 8.14 *Si* f *est fortement convexe, de classe* C^1 *avec un gradient lipschitzien, alors la méthode de Polak-Ribière avec recherche linéaire exacte génère une suite* $\{x_k\}$ *convergeant vers l'unique point* x_* *réalisant le minimum de* f.

Démonstration. Montrons dans un premier temps que

$$\cos\theta_k = \frac{-g_k^{\mathsf{T}} d_k}{\|g_k\|\,\|d_k\|}$$

est uniformément positif. Grâce à la recherche linéaire exacte, on a

$$y_{k-1}^{\mathsf{T}} d_{k-1} = -g_{k-1}^{\mathsf{T}} d_{k-1} = \|g_{k-1}\|^2.$$

La forte convexité de f implique que

$$y_{k-1}^{\mathsf{T}} d_{k-1} = \frac{1}{\alpha_{k-1}} y_{k-1}^{\mathsf{T}}(x_k - x_{k-1}) \geqslant \frac{\gamma}{\alpha_{k-1}} \|x_k - x_{k-1}\|^2,$$

où $\gamma > 0$ est le module de forte convexité de f. On en déduit, en utilisant la constante de lipschitz L de f' :

$$|\beta_k^{\mathrm{PR}}| = \frac{|g_k^{\mathsf{T}} y_{k-1}|}{\|g_{k-1}\|^2} = \frac{|g_k^{\mathsf{T}} y_{k-1}|}{y_{k-1}^{\mathsf{T}} d_{k-1}} \leqslant \frac{\alpha_{k-1} L}{\gamma} \frac{\|g_k\|\|x_k - x_{k-1}\|}{\|x_k - x_{k-1}\|^2} = \frac{L}{\gamma} \frac{\|g_k\|}{\|d_{k-1}\|}.$$

On peut alors borner $\|d_k\|$ par

$$\|d_k\| \leqslant \|g_k\| + |\beta_k^{\mathrm{PR}}|\,\|d_{k-1}\| \leqslant \left(1 + \frac{L}{\gamma}\right) \|g_k\|.$$

Ensuite

$$g_k^T d_k = -\|g_k\|^2 \leqslant -\left(1+\frac{L}{\gamma}\right)^{-1} \|g_k\|\|d_k\|,$$

ou encore $\cos\theta_k \geqslant (1+L/\gamma)^{-1}$.

D'après la proposition 6.10 et la recherche linéaire exacte, la condition de Zoutendijk (6.19) est vérifiée. Mais f et donc $\{f(x_k)\}$ est bornée inférieurement (car f est fortement convexe). On en déduit que $g_k \to 0$ (proposition 6.8). D'autre part, $\{x_k\}$ est bornée (f est fortement convexe) et possède donc des sous-suites convergentes. La limite de celles-ci ne peut être que l'unique minimum x_* de f (car $g_k \to 0$). Donc *toute* la suite $\{x_k\}$ converge vers x_*. □

Si f n'est pas convexe, la méthode de Polak-Ribière peut ne pas converger. Powell (1984) a donné un exemple de fonction pour laquelle l'algorithme génère une suite $\{x_k\}$ dont aucun des points d'adhérence n'est stationnaire. On peut trouver des remèdes simples à ce comportement inattendu [245 ; 1992]. Le plus simple et apparemment le plus efficace est de prendre

$$\beta_k = (\beta_k^{\mathrm{PR}})^+ \equiv \max(\beta_k^{\mathrm{PR}}, 0),$$

ce qui revient à redémarrer l'algorithme chaque fois que $\beta_k^{\mathrm{PR}} < 0$. On peut montrer la convergence globale pour cette valeur de β_k et une recherche linéaire adaptée. Les performances numériques de cette méthode sont très semblables à celles de la méthode de Polak-Ribière.

Comme pour la minimisation de fonctions quadratiques convexes, la méthode du gradient conjugué non linéaire est économe en place mémoire et en temps de calcul propre. Elle s'utilise encore parfois dans les deux situations suivantes : lorsque n est très grand relativement à la place mémoire disponible ou lorsque le coût de calcul de f et de son gradient est très faible (ce qui est plutôt rare dans les problèmes réels) et que l'on ne veut pas perdre de temps avec une méthode plus sophistiquée mais plus gourmande en temps de calcul. Dès que l'espace mémoire disponible est supérieur à, disons $10n$, il est souvent préférable d'utiliser une méthode de quasi-Newton à mémoire limitée, par exemple la méthode l-BFGS (voir la section 11.2.5).

8.3 Algorithme du résidu minimal généralisé (GMRES)

Revenons à la résolution numérique du système linéaire

$$Ax = b, \tag{8.23}$$

lorsque A est inversible mais n'est plus nécessairement symétrique. On note toujours

$$x_* = A^{-1}b$$

la solution unique de (8.23). On appelle *résidu* du système linéaire (8.23) en $x \in \mathbb{R}^n$, le vecteur $r \in \mathbb{R}^n$ défini par

$$r = b - Ax.$$

On note $r_k = b - Ax_k$ le résidu en x_k. Lorsque A est symétrique, le résidu est l'opposé du gradient de $f(x) = \frac{1}{2}x^\mathsf{T}Ax - b^\mathsf{T}x$.

On aimerait, pour les systèmes linéaires non symétriques, disposer d'un algorithme itératif ayant les propriétés attrayantes du gradient conjugué que sont le peu d'encombrement-mémoire (due à une formule de récurrence courte) et une propriété de minimalité sur l'espace de Krylov (qui lui confère de la stabilité). Ceci n'est malheureusement pas possible [611, 196]. Dans l'algorithme GMRES, décrit dans cette section, on ne garde que la propriété de minimalité ; quant à l'encombrement-mémoire, il va croître linéairement avec le nombre d'itérations, si bien que la mise en œuvre de l'algorithme se fait avec redémarrages (section 8.3.6).

L'algorithme GMRES est aujourd'hui l'algorithme itératif standard de résolution des systèmes non symétriques. Il a été proposé par Saad et Schultz [532 ; 1986].

8.3.1 Principe général

Partant d'un point $x_1 \in \mathbb{R}^n$, la *méthode du résidu minimal généralisée* ou *méthode GMRES* (Generalized Minimal RESidual) consiste à générer une suite de points $x_k \in \mathbb{R}^n$ par la formule

$$x_k = x_1 + z_k, \quad k \geqslant 1,$$

où z_k dans le sous-espace de Krylov associé au résidu initial :

$$K_k \equiv K_k(A, r_1) = \text{vect}\{r_1, Ar_1, \ldots, A^{k-1}r_1\}.$$

Cette façon de procéder est motivée par les deux remarques suivantes.

- Si x_1 est une bonne approximation de la solution il y a un sens à approcher x_* par des itérés x_k tels que $x_k - x_1$ appartienne pour k petit à un sous-espace vectoriel de faible dimension (ici un sous-espace de Krylov).
- Si l'on prend z_k dans un sous-espace de Krylov $K_k(A, r)$, pour un certain $r \in \mathbb{R}^n$, et si l'on veut obtenir la solution par cette méthode, il faut que

$$x_* \in x_1 + K_s(A, r), \tag{8.24}$$

où s est l'indice de saturation de ces sous-espaces (voir la section 8.1). Cet indice est caractérisé par le fait que

$$A^{-1}r \in K_s(A, r).$$

On voit qu'en prenant $r = r_1$, cette dernière condition est équivalente à (8.24).

La méthode détermine ensuite x_k dans $x_1 + K_k$, de manière à minimiser le résidu sur $x_1 + K_k$. Il s'agit donc de résoudre à chaque itération

$$\min_{x \in x_1 + K_k} \|b - Ax\|_2, \tag{8.25}$$

où $\|\cdot\|_2$ est la norme ℓ_2 de $\mathbb{R}^n$. Ce problème a une solution unique qui, pour $k = s(A, r_1)$, ne peut être que $x = x_*$ (qui annule le résidu sur $\mathbb{R}^n$). En écrivant $x = x_1 + z$, $z \in K_k$, on a $b - Ax = r_1 - Az$ et le problème (8.25) devient

$$\min_{z \in K_k} \|r_1 - Az\|_2. \tag{8.26}$$

Ceci montre que Az_k est la projection orthogonale de r_1 sur AK_k et donc

$$r_k = r_1 - Az_k \perp Az_k.$$

Le résidu est rendu orthogonal à AK_k.

Remarque 8.15 On peut aussi déterminer x_k dans $x_1 + K_k$ en rendant le résidu r_k orthogonal à K_k. On parle alors de *méthode du résidu orthogonal* ou de *méthode d'Arnoldi*. La méthode du gradient conjugué est de ce type (exercice 8.1). Numériquement les deux méthodes se valent [94]. □

Proposition 8.16 *La méthode GMRES converge en au plus β itérations, où β est le degré du polynôme minimal annihilant A (donc $\beta \leqslant n$). Si A est non défective, β est le nombre de valeurs propres distinctes de A.*

Démonstration. Il suffit de constater que la détermination de x_k par (8.25) donne $x_k = x_*$ dès que $x_* \in x_1 + K_{k-1}$. Ceci aura lieu exactement à la saturation du sous-espace de Krylov K_k, c'est-à-dire pour $k = s(A, r_1)$. Le résultat est alors une conséquence de la proposition 8.2 et de son corollaire. □

Venons-en maintenant au calcul effectif de x_k.

8.3.2 Construction d'une base orthonormale de K_{k+1}

Le problème (8.25) — ou (8.26) — est un problème avec contraintes que l'on peut transformer en problème sans contrainte en se donnant une base de K_k. Tant que $k < s(A, r_1)$, les vecteurs $r_1, Ar_1, \ldots, A^k r_1$ forment une base de K_{k+1}. Cependant, l'utilisation d'une base orthonormale s'avérera utile par la suite.

L'algorithme de Gram-Schmidt permet d'obtenir une base orthonormale de K_{k+1} en « orthonormalisant » les vecteurs $r_1, Ar_1, \ldots, A^k r_1$. Si l'on a déjà une base orthonormale de K_k formée des vecteurs $v_1, \ldots, v_k$, on voit que pour obtenir v_{k+1} il suffit d'orthonormaliser $A^k r_1$ par rapport à $v_1, \ldots, v_k$. On peut aussi orthonormaliser Av_k par rapport à $v_1, \ldots, v_k$ car on verra que, compte tenu de la manière dont les v_i sont calculés, on a

$$A^k r_1 \notin K_k \implies Av_k \in K_{k+1} \setminus K_k, \tag{8.27}$$

Ceci permet d'éviter le stockage de $A^k r_1$ dans un vecteur auxiliaire. La relation (8.27) se montre par récurrence. Mais auparavant, spécifions le calcul des vecteurs de base v_i en adaptant à notre cadre l'algorithme d'orthonormalisation de Gram-Schmidt (section B.1.1). On utilise le produit scalaire euclidien. Si $r_1 = b - Ax_1 = 0$, le point initial x_1 est solution et on s'arrête. Sinon, on prend

$$v_1 = \frac{r_1}{\|r_1\|_2}.$$

Ensuite, pour $k \geqslant 1$, v_{k+1} est calculé comme suit

$$\begin{array}{ll}(i) & h_{i,k} = v_i^{\mathsf{T}} A v_k, \quad \text{pour} \quad i = 1, \ldots, k\,; \\ (ii) & \tilde{v}_{k+1} = A v_k - \sum_{i=1}^{k} h_{i,k} v_i\,; \\ (iii) & h_{k+1,k} = \|\tilde{v}_{k+1}\|_2\,; \\ (iv) & v_{k+1} = \tilde{v}_{k+1}/h_{k+1,k}. \end{array}$$

DÉMONSTRATION. (de (8.27)). L'implication est vraie pour $k = 1$, puisque $Av_1 \parallel Ar_1 \in K_2 \setminus K_1$. Ensuite, si $Av_{k-1} \in K_k \setminus K_{k-1}$ (hypothèse de récurrence), on a

$$Av_{k-1} = \alpha A^{k-1} r_1 + w,$$

avec $\alpha \neq 0$ et $w \in K_{k-1}$. L'étape (ii) de l'algorithme précédent donne

$$\tilde{v}_k = \alpha A^{k-1} r_1 + \tilde{w},$$

avec $\tilde{w} \in K_{k-1}$. Mais $\tilde{v}_k \neq 0$, car $A^{k-1} r_1 \notin K_{k-1}$. D'après l'étape (iv) on trouve alors

$$v_k = \alpha' A^{k-1} r_1 + w',$$

où $\alpha' \neq 0$ et $w' \in K_{k-1}$. Donc $Av_k = \alpha' A^k r_1 + Aw' \in K_{k+1} \setminus K_k$ car $A^k r_1 \in K_{k+1} \setminus K_k$ et $Aw' \in K_k$. □

Remarquons que l'on a

$$Av_k = \sum_{i=1}^{k+1} h_{i,k} v_i.$$

Si l'on note $\bar{H}_k$ la matrice $(k+1) \times k$ dont les éléments h_{ij} sont calculés par l'algorithme précédent et $V_k = (v_1 \ \ldots \ v_k)$ la matrice $n \times k$ dont les colonnes sont formées des vecteurs de la base orthonormale de K_k, on a la relation

$$AV_k = V_{k+1} \bar{H}_k. \tag{8.28}$$

Comme

$$(\bar{H}_k)_{ij} = 0, \quad \text{pour} \quad i \geqslant j + 2,$$

les k premières lignes de $\bar{H}_k$ forment une matrice H_k de Hessenberg supérieure :

$$\bar{H}_k = \begin{pmatrix} H_k \\ 0 \cdots 0 \ h_{k+1,k} \end{pmatrix}.$$

8.3.3 Calcul de l'itéré x_k

Il s'agit de résoudre le problème (8.26). Comme $z \in K_k$ peut s'exprimer dans la base $v_1, \ldots, v_k$ par $z = V_k y$, $y \in \mathbb{R}^k$, (8.26) peut se récrire comme un problème de minimisation de

$$J(y) = \|r_1 - AV_k y\|_2,$$

pour $y \in \mathbb{R}^k$.

En utilisant (8.28) et le fait que $r_1 = \beta V_{k+1} e_1$, $e_1 = (1, 0, \ldots, 0)^{\mathsf{T}} \in \mathbb{R}^{k+1}$ et $\beta = \|r_1\|$, on a

$$J(y) = \|V_{k+1}(\beta e_1 - \bar{H}_k y)\|_2.$$

Compte tenu de l'orthonormalité des colonnes de V_{k+1} on obtient finalement

$$J(y) = \|\beta e_1 - \bar{H}_k y\|_2, \tag{8.29}$$

fonction qu'il s'agit de minimiser. On verra à la section 8.3.5 une méthode de résolution de ce problème de moindres-carrés, basée sur la factorisation QR de $\bar{H}_k$.

8.3.4 L'algorithme GMRES

Après la phase d'initialisation (étapes 1 à 3), l'algorithme GMRES se déroule en deux temps. Dans un premier temps (étape 4), on calcule une base de $K_s(A, r_1)$, l'indice s se détermine en détectant la saturation des espaces de Krylov. Dans un second temps (étape 5), on calcule la solution x_* du système linéaire en résolvant le problème (8.29).

Algorithme 8.17 (GMRES)

1. Initialisation : choix de $x_1 \in \mathbb{R}^n$ et calcul du résidu $r_1 = b - Ax_1$;
2. Si $r_1 = 0$ on s'arrête ;
3. $v_1 = r_1/\|r_1\|$;
4. Pour $k = 1, 2, \ldots$ faire :
   4.1. $h_{i,k} = v_i^{\mathsf{T}} A v_k$, pour $i = 1, \ldots, k$;
   4.2. $\tilde{v}_{k+1} = Av_k - \sum_{i=1}^{k} h_{i,k} v_i$;
   4.3. $h_{k+1,k} = \|\tilde{v}_{k+1}\|_2$;
   4.4. Si $h_{k+1,k} = 0$ aller en 5 ;
   4.5. $v_{k+1} = \tilde{v}_{k+1}/h_{k+1,k}$;
5. Calcul de la solution :
   5.1. $y_k = \arg\min_{y \in \mathbb{R}^k} J(y)$, où $J(y)$ donné par (8.29) ;
   5.2. $x_k = x_1 + V_k y_k$.

À l'étape 4.4, lorsque $h_{k+1,k} = 0$, on a $Av_k \in K_k$. D'après l'implication (8.27), cela entraîne que $K_{k+1} = K_k$: l'espace de Krylov K_k est saturé. Par conséquent, x_k calculé à l'étape 5 sera la solution x_*.

8.3.5 L'algorithme GMRES/QR

L'étape 5 de l'algorithme GMRES ne spécifie pas comment on résout le problème de minimisation de la fonction J donnée en (8.29). Dans cette section, nous montrons comment procéder en utilisant une factorisation QR de $\bar{H}_k$:

$$\bar{H}_k = Q_k \bar{R}_k, \quad \bar{R}_k = \begin{pmatrix} R_k \\ 0 \end{pmatrix},$$

où Q_k est une matrice orthogonale d'ordre $k+1$ et $\bar{R}_k$ est une matrice $(k+1) \times k$ dont les k premières lignes forment une matrice R_k triangulaire supérieure et la $(k+1)$-ième ligne est nulle.

Avec cette factorisation, J s'écrit

$$J(y) = \|Q_k^{\mathsf{T}}(\beta e_1 - \bar{H}_k y)\|_2 = \|\bar{q}_k - \bar{R}_k y\|_2,$$

où

$$\bar{q}_k = \beta Q_k^{\mathsf{T}} e_1. \tag{8.30}$$

On voit que le vecteur $y_k \in \mathbb{R}^k$ minimisant $J(y)$ sur $\mathbb{R}^k$ est solution du système triangulaire supérieur :

$$R_k y = q_k,$$

où $q_k \in \mathbb{R}^k$ est formé des k premiers éléments de $\bar{q}_k$. Remarquons aussi que la norme du résidu en x_k s'obtient comme la composante $k+1$ de $\bar{q}_k$:

$$\|r_k\| = J(y_k) = (\bar{q}_k)_{k+1}. \tag{8.31}$$

On a

$$\bar{q}_k = \begin{pmatrix} q_k \\ \|r_k\| \end{pmatrix}.$$

Voyons à présent comment calculer $\bar{R}_{k+1}$ et $\bar{q}_{k+1}$ à partir de $\bar{R}_k$ et $\bar{q}_k$ en n'utilisant qu'une rotation de Givens. Supposons que l'on dispose de la factorisation QR de $\bar{H}_k$:

$$Q_k^{\mathsf{T}} \bar{H}_k = \bar{R}_k = \begin{pmatrix} R_k \\ 0 \end{pmatrix},$$

et que la matrice orthogonale Q_k^{T} d'ordre $k+1$ soit le produit de p rotations de Givens d'ordre $k+1$ (en fait, on aura $p = k$)

$$Q_k^{\mathsf{T}} = G_p \cdots G_1.$$

La matrice $\bar{H}_{k+1}$ s'obtient à partir de $\bar{H}_k$ en lui adjoignant une ligne et une colonne supplémentaire

$$\bar{H}_{k+1} = \begin{pmatrix} & & & \times \\ & \bar{H}_k & & \vdots \\ & & & \times \\ 0 & \cdots & 0 & h_{k+2,k+1} \end{pmatrix},$$

où $\times$ désigne des éléments éventuellement non nuls. Les matrices d'ordre $k+2$

$$\tilde{G}_i = \begin{pmatrix} & & & 0 \\ & G_i & & \vdots \\ & & & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix}, \quad 1 \leqslant i \leqslant p$$

sont encore des rotations de Givens. L'application de $\tilde{Q}_{k+1}^{\mathsf{T}} = \tilde{G}_p \cdots \tilde{G}_1$ à $\bar{H}_{k+1}$ laissera sa dernière ligne inchangée et l'on aura

$$\tilde{Q}_{k+1}^{\mathsf{T}} \bar{H}_{k+1} = \begin{pmatrix} & & & \times \\ & \bar{R}_k & & \vdots \\ & & & \times \\ 0 & \cdots & 0 & h_{k+2,k+1} \end{pmatrix}.$$

On voit que pour annuler la dernière ligne de la matrice dans le membre de droite, il suffira d'utiliser une rotation de Givens supplémentaire $\tilde{G}_{p+1}$.

Quant au vecteur $\bar{q}_{k+1}$ défini en (8.30), il s'obtient à partir de $\bar{q}_k$, comme suit ($e_1 \in \mathbb{R}^{k+1}$)

$$\bar{q}_{k+1} = \tilde{G}_{p+1} \tilde{G}_p \cdots \tilde{G}_1 \begin{pmatrix} \beta e_1 \\ 0 \end{pmatrix} = \tilde{G}_{p+1} \begin{pmatrix} \bar{q}_k \\ 0 \end{pmatrix}.$$

En pratique, on modifie l'algorithme GMRES pour contrôler son arrêt au moyen d'un seuil de tolérance $\varepsilon > 0$: on s'arrête si $\|r_k\| \leqslant \varepsilon$. Ce contrôle peut facilement être pris en compte dans le test d'arrêt de l'étape 2. Quant à l'arrêt au cours de l'étape 4.4, il peut se faire sans le calcul explicite de x_k, en contrôlant le résidu par $(\bar{q}_k)_{k+1}$; voir (8.31).

8.3.6 Algorithme GMRES avec redémarrage

Pour la résolution de grands systèmes linéaires, l'algorithme GMRES a l'inconvénient de devoir mémoriser un nombre de vecteurs croissant avec le nombre d'itérations. Il faut en effet mémoriser V_k et les k rotations de Givens donnant $Q_k^\intercal$ et R_k. Pour remédier à cet inconvénient, on peut redémarrer la méthode toutes les m itérations.

Cependant, une certaine prudence s'impose dans l'utilisation d'un tel algorithme, car il n'est pas nécessairement convergent. On peut facilement construire un exemple dans le cas où $m = 1$. Il suffit en effet que sur $x_1 + K_1$, le résidu soit minimal en x_1. Alors $x_2 = x_1$ et on ne quitte pas ce point si l'on redémarre la méthode à chaque itération. C'est le cas de l'exemple suivant

$$A = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad x_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

On trouve

$$r_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Alors $x_2 = x_1 + \bar{\alpha} r_1$, où $\bar{\alpha} = 0$ est la solution de

$$\min_\alpha \|b - \alpha A r_1\|_2 = \min_\alpha \left\| \begin{pmatrix} 1-\alpha \\ 1+\alpha \end{pmatrix} \right\|_2.$$

Notes

Rappelons qu'il existe des algorithmes de résolution de systèmes linéaires par factorisation de la matrice, demandant moins de $O(n^3)$ opérations flottantes. Un des premiers à descendre sous cette complexité est dû à Strassen [573], avec une borne en $O(n^{2.807})$, mais il est numériquement instable [313 ; 2002, section 26.3.2]. Le schéma de Coppersmith et Winograd [139] est en $O(n^{2.495})$.

La méthode du GC a été proposée par Hestenes et Stiefel [312 ; 1952] pour la résolution de systèmes linéaires symétriques et reprise par Fletcher et Reeves [216 ; 1964] pour l'optimisation de fonctions quadratiques strictement convexes. L'effet des erreurs d'arrondi sur la vitesse de convergence est analysé par Greenbaum [279 ; 1989] et Notay [462 ; 1993] et sur la précision de la solution par Greenbaum et Strakoš [283 ; 1992], Greenbaum [280 ; 1994] et Greenbaum [282 ; 1997]. Van der Vorst [600 ; 1990] analyse l'effet des erreurs d'arrondi sur le gradient conjugué préconditionné par factorisation de Cholesky incomplète. La dérivée des itérés par rapport au second membre b est exhibée dans [276].

Andrei [14 ; 2020] propose une synthèse sur les algorithmes de gradient conjugué pour la minimisation sans contrainte de fonctions non linéaires.

La méthode GMRES a été principalement développée par Saad et Schultz [532 ; 1986]. La stabilité numérique de l'algorithme est analysée dans [180 ; 1995].

Pour en savoir plus sur les méthodes itératives de résolution de systèmes linéaires, on pourra consulter les livres ou synthèses de Greenbaum [281 ; 1997], Gutknecht [297 ; 1997], van der Vorst [602 ; 2003] et Saad [531 ; 2003]. Saad et van der Vorst [533, 601 ; 2000] donnent un apreçu historique intéressant du développement des méthodes itératives de résolution de systèmes linéaires au cours du XX-ième siècle.

Exercices

8.1. *Le GC comme méthode de Krylov.* Soient A une matrice d'ordre n, symétrique, définie positive et $b \in \mathbb{R}^n$. Soient d_1, d_2, ..., les directions générées par la méthode du GC pour minimiser $f(x) = \frac{1}{2}x^\mathsf{T}Ax - b^\mathsf{T}x$ et $E_p := \text{vect}\{d_1, d_2, \ldots, d_p\}$. On note $K_p := \text{vect}\{g_1, Ag_1, \ldots, A^{p-1}g_1\}$ le sous-espace de Krylov d'ordre p, associé à A et au gradient initial $g_1 = Ax_1 - b$. Montrez que $E_p = K_p$.

8.2. *Propriétés de monotonie dans l'algorithme du gradient conjugué.* Soient $\{x_k\}_{k=1}^p$ les itérés générés par le GC (donc $x_p = x_*$). On suppose que $p \geqslant 2$.

(*i*) Montrez que $\{\|x_k - x_1\|_2\}_{k=1}^p$ est strictement croissante et que $\{\|x_k - x_*\|_2\}_{k=1}^p$ est strictement décroissante.

(*ii*) Montrez que l'angle entre $x_2 - x_1$ et $x_k - x_1$ est strictement croissant.

8.3. *Le GC préconditionné.* On considère l'algorithme du gradient conjugué préconditionné de la section 8.2.5. Soit P la matrice symétrique définie positive utilisée pour préconditionner l'algorithme. Montrez que les propriétés suivantes ont lieu.

(*i*) C'est un algorithme à directions conjuguées dans lequel $\tilde{d}_k = -Pg_k$.

(*ii*) On a $\text{vect}\{d_1, \ldots, d_k\} = P\,\text{vect}\{g_1, \ldots, g_k\} = K_k(PA, Pg_1)$ (l'espace de Krylov associé à la matrice PA et au vecteur Pg_1).

(*iii*) On a $g_k^\mathsf{T}Pg_i = 0$ et $g_k^\mathsf{T}d_i = 0$, pour $1 \leqslant i \leqslant k-1$.

(*iv*) Montrez que si $P^{1/2}AP^{1/2} = \alpha I + E$, où $\alpha > 0$ et E est une matrice de rang m, alors le GC préconditionné converge en au plus m itérations.

8.4. *Algorithmes du GC réduit et du GC projeté.* On cherche à minimiser par GC une fonction quadratique strictement convexe $x \in \mathbb{R}^n \mapsto f(x) := \frac{1}{2}x^\mathsf{T}Ax - b^\mathsf{T}x$ sur un sous-espace affine $\mathcal{A}$ de dimension p de $\mathbb{R}^n$. Dans ce but, on se donne $x_1 \in \mathcal{A}$ et une matrice Z de type $n \times p$ orthogonale (c.-à-d., $Z^\mathsf{T}Z = I$) telle que tout point de $x \in \mathcal{A}$ puisse s'écrire sous la forme $x = x_1 + Zu$, avec $u \in \mathbb{R}^p$. On note $P = ZZ^\mathsf{T}$ le projecteur orthogonal sur $\mathcal{A}$. On considère deux formes de l'algorithme du GC pour résoudre ce problème.

(A1) L'*algorithme du GC réduit* consiste à minimiser par GC la *fonction réduite*

$$u \in \mathbb{R}^p \mapsto f^r(u) := \frac{1}{2}u^\mathsf{T}Z^\mathsf{T}AZu + (Ax_1 - b)^\mathsf{T}Zu,$$

obtenue en remplaçant x par $x_1 + Zu$ dans $f(x)$.

(A2) L'*algorithme du GC projeté* consiste à minimiser f dans $\mathbb{R}^n$ par l'algorithme du gradient conjugué préconditionné de la section 8.2.5 qui utilise le préconditionneur singulier P.

Montrez que les deux algorithmes sont identiques (dans un sens à préciser) si le premier est démarré en $u_1 = 0$ et le second en x_1.

8.5. *Algorithme du GC projeté préconditionné.* On se place dans le cadre défini à l'exercice 8.4, dans lequel on cherche à minimiser une fonction quadratique strictement convexe $x \in \mathbb{R}^n \mapsto f(x) := \frac{1}{2}x^\mathsf{T}Ax - b^\mathsf{T}x$ sur un sous-espace affine $\mathcal{A}$ de dimension p de $\mathbb{R}^n$. On note Z une matrice de type $n \times p$ orthogonale (c.-à-d., $Z^\mathsf{T}Z = I$) telle que tout point de $x \in \mathcal{A}$ puisse s'écrire sous la forme $x = x_1 + Zu$, avec $u \in \mathbb{R}^p$. Soit L une approximation de $A^{1/2}$ que l'on cherche à utiliser pour préconditionner le GC projeté. Cet algorithme s'obtient en faisant un changement de variable $\tilde{x} = Lx$, en appliquant le GC projeté dans l'espace des $\tilde{x}$, puis en revenant dans l'espace des x. Montrez que l'algorithme résultant est l'algorithme du gradient conjugué préconditionné de la section 8.2.5 qui utilise le préconditionneur singulier $P = Z(Z^\mathsf{T}L^\mathsf{T}LZ)^{-1}Z^\mathsf{T}$.

8.6. *Application de GMRES et comparaison avec le GC.* Utilisez l'algorithme GMRES pour trouver la solution x_* de $Ax = b$, en partant de $x_1 = 0$, lorsque

$$A = \begin{pmatrix} -1 & 2 & 0 \\ -2 & 3 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{et} \quad b = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}.$$

Comparez le nombre d'itérations de cette méthode avec le nombre d'étapes que peut demander la méthode du GC pour calculer x_* comme solution de l'équation normale $A^\mathsf{T}Ax = A^\mathsf{T}b$.

9 Méthodes à régions de confiance

Connaissances supposées. CNS d'optimalité pour les problèmes avec contraintes d'inégalité (pour la section 9.3); méthodes à directions de descente (pour les comparaisons); algorithme du gradient conjugué (pour la résolution numérique du sous-problème quadratique à la section 9.4).

9.1 L'algorithme de référence

On suppose dans cette section que l'on cherche à résoudre numériquement le problème d'optimisation sans contrainte

$$\min_{x\in\mathbb{R}^n} f(x). \tag{9.1}$$

La fonction f est supposée (Fréchet-)différentiable.

La technique des régions de confiance peut être vue comme une méthode de *globalisation de la convergence* d'algorithmes pour la résolution du problème (9.1), ce qui veut dire qu'elle permet de forcer la convergence des itérés lorsque le démarrage des algorithmes se fait en un point éloigné d'une solution (il ne s'agit aucunement ici de trouver des minima globaux de f). De ce point de vue, cette technique concurrence les méthodes à directions de descente. Un peu plus difficile à mettre en œuvre que ces dernières, elles ont divers avantages, notamment en ce qui concerne les résultats de convergence, qui sont plus forts sous des hypothèses plus faibles. Elles ont aussi la réputation d'être plus robustes, c'est-à-dire de pouvoir résoudre des problèmes plus difficiles, moins bien conditionnés. Elles offrent enfin un moyen naturel de faire face à la non-convexité éventuelle de f.

9.1.1 Présentation de l'algorithme

Les méthodes à régions de confiance sont fondées sur le principe suivant, qui est très général et peut donc s'appliquer à d'autres contextes. Au point courant $x_k \in \mathbb{R}^n$, on suppose donné un modèle de la *variation* de f pour un incrément s de x_k. En optimisation différentiable, il est raisonnable de considérer un modèle quadratique de la forme

$$\psi_k(s) = g_k^\mathsf{T} s + \frac{1}{2} s^\mathsf{T} M_k s,$$

où $g_k = \nabla f(x_k)$ est le gradient de f en x_k et M_k est une matrice symétrique qui peut être $\nabla^2 f(x_k)$, la hessienne de f en x_k, ou une approximation de celle-ci. Gradient

et hessienne sont supposés être calculés pour le produit scalaire euclidien $\langle u, v\rangle = \sum_i u_i v_i$. On n'impose pas à la matrice M_k d'être définie positive (en particulier, elle peut être nulle). Cette souplesse contribue à la robustesse de cette approche.

En optimisation non différentiable, on ne dispose pas des gradients de f. On pourra alors construire des modèles différents, comme celui obtenu à partir des plans tangents à l'épigraphe de f aux points $(x_k, f(x_k))$. On est alors très proche de l'algorithme connu sous le nom de *méthode de faisceaux* (voir [317 ; Tome 2]).

Si $M_k = \nabla^2 f(x_k)$, on a le développement limité à l'ordre 2 :

$$f(x_k + s) \simeq f(x_k) + \psi_k(s).$$

Dans ce cas, le modèle quadratique est adéquat pour s voisin de 0. Dans les méthodes à régions de confiance, on considère que ψ_k est un modèle de la variation de f qui est acceptable dans un voisinage de la forme :

$$\{x \in \mathbb{R}^n : \ \|x\| \leqslant \Delta_k\},$$

où $\Delta_k > 0$ et $\|\cdot\|$ est la norme euclidienne. Ce domaine est appelé la *région de confiance* du modèle ψ_k et Δ_k est appelé le *rayon de confiance*.

Pour trouver l'accroissement s_k à donner à x_k, on minimise le modèle quadratique ψ_k sur la région de confiance. On doit donc résoudre le *sous-problème quadratique*

$$(RC_k) \quad \left\{ \begin{array}{l} \min \ \psi_k(s) \\ \|s\| \leqslant \Delta_k. \end{array} \right.$$

Soit s_k une solution (en général approchée) de ce sous-problème quadratique. Si $f(x_k + s_k)$ est suffisamment plus petit que $f(x_k)$, on accepte le pas s_k et on passe à l'itération suivante avec $x_{k+1} = x_k + s_k$. Dans le cas contraire, on résout à nouveau (RC_k) avec un rayon de confiance plus petit, car on considère que le modèle courant n'est pas fiable sur la boule $\bar{B}(0, \Delta_k)$. Nous verrons que, comme f et ψ_k coïncident au premier ordre, l'algorithme finit par trouver un rayon de confiance $\Delta_k > 0$ pour lequel la solution de (RC_k) est acceptée.

On ajuste ensuite le modèle à l'itéré suivant x_{k+1} : calcul de g_{k+1}, de M_{k+1} et de Δ_{k+1}. L'ajustement du rayon de confiance est basé sur la comparaison entre la décroissance réelle de f en passant de x_k à x_{k+1} et la décroissance prédite par le modèle, c'est-à-dire $-\psi_k(s_k)$. On peut alors passer à l'itération suivante.

Décrivons de manière précise une itération de la méthode à régions de confiance, celle passant de l'itéré x_k à x_{k+1}. Nous commenterons ses différentes étapes par la suite. On se donne des constantes indépendantes de l'indice k des itérations, qui sont :

- des seuils de succès : $0 < \omega_1 < \omega_2 < 1$,
- des facteurs de mise à jour de Δ_k : $0 < \tau_1 \leqslant \tau_2 < 1 < \tau_3$.

Des valeurs typiques sont :

$$\omega_1 = 10^{-2}, \quad \omega_2 = 0.8, \quad \tau_1 = \tau_2 = 0.5 \quad \text{et} \quad \tau_3 \in [2, 4].$$

Étant donnés x_k, g_k, M_k et Δ_k, on s'y prend alors de la manière suivante.

Schéma 9.1 (des régions de confiance — une itération)

1. *Test de convergence* : si $\|\nabla f(x_k)\| \simeq 0$ on s'arrête.

2. *Calcul du déplacement* s_k :
   2.1. On calcule une solution *approchée* s_k de (RC_k) telle que
   $$\psi_k(s_k) < 0.$$
   Si ce n'est pas possible, on s'arrête.
   2.2. On calcule le ratio
   $$\rho_k = \frac{f(x_k) - f(x_k + s_k)}{-\psi_k(s_k)}.$$
   2.3. Si $\rho_k \leqslant \omega_1$, on retourne alors à l'étape 2.1 après avoir réduit Δ_k en le prenant dans $[\tau_1\|s_k\|, \tau_2\|s_k\|]$.
3. *Itéré suivant* : $x_{k+1} = x_k + s_k$.
4. *Mise à jour du rayon de confiance* :
$$\Delta_{k+1} \in \begin{cases} [\tau_2\Delta_k, \Delta_k] & \text{si } \rho_k \leqslant \omega_2 \\ [\Delta_k, \tau_3\Delta_k] & \text{sinon.} \end{cases}$$
5. *Mise à jour du modèle* : calcul de g_{k+1} et M_{k+1}.

Levons une ambiguïté de notation : on note dorénavant (s_k, Δ_k) la solution approchée et le rayon de confiance obtenus à la sortie de l'étape 2. On ne tient donc pas compte des valeurs intermédiaires prises par ce couple lorsque la condition $\rho_k > \omega_1$ n'est pas vérifiée du premier coup. Certains auteurs considèrent les essais de déplacements comme de vraies itérations, motivés sans doute par le fait que le calcul le plus coûteux de l'algorithme est la résolution des problèmes quadratiques. Cependant, dans les grands problèmes industriels, le coût d'une simulation (calcul de g_k et M_k à l'étape 5) prédomine souvent. Nous considérerons donc ici qu'il y a une itération à chaque renouvellement de modèle. D'autre part, comme nous le verrons ci-dessous, cela permet de faciliter la comparaison des régions de confiance et de la recherche linéaire.

Si à l'étape 2.1, il n'est pas possible de réaliser $\psi(s_k) < 0$, c'est que $g_k = 0$ et M_k est semi-définie positive. Ceci devrait vouloir dire qu'on ne peut pas faire décroître f non plus. Donc si l'on s'impose un modèle ψ_k convexe (M_k semi-définie positive), on ne pourra trouver qu'un point stationnaire de f. Si f est convexe, c'est aussi un minimum. Si f n'est pas convexe, il est préférable de ne pas maintenir artificiellement M_k semi-définie positive mais d'y incorporer des informations sur la hessienne de f, afin d'éviter les points stationnaires qui ne sont pas des minima.

Cette observation met en évidence une différence importante avec les méthodes à directions de descente. Dans ces dernières, la direction de descente s'obtient également en minimisant un modèle quadratique, mais sur $\mathbb{R}^n$ tout entier. Dans ce cas, il est essentiel d'avoir M_k (semi-)définie positive, sinon le sous-problème quadratique n'a pas de solution. De plus, en prenant M_k définie positive, on est sûr que la direction est de descente pour f. Les conditions assurant le bon fonctionnement des régions de confiance sont donc moins restrictives. C'est exactement sur ce point, grâce au fait qu'elles ne demandent pas d'avoir un modèle fortement convexe, que les régions de confiance gagnent en robustesse. Cela se traduira par un affaiblissement des hypothèses

garantissant la convergence de cette approche. Le prix à payer est la résolution d'un problème quadratique non trivial.

Le ratio ρ_k calculé à l'étape 2.1 est appelé *concordance*. Celle-ci apprécie l'adéquation entre la fonction f et son modèle ψ_k courant. Il s'agit en effet du rapport entre la décroissance réelle de f $(f(x_k) - f(x_k + s_k))$ et la décroissance prédite par le modèle $(-\psi_k(s_k) = \psi_k(0) - \psi_k(s_k) > 0)$. Si ce rapport n'est pas assez grand ($\rho_k \leqslant \omega_1$, ρ_k sera négatif si le pas s_k ne fait pas décroître f), l'algorithme considère que le modèle ne permet pas de prédire la variation réelle de f et ne convient donc pas. En diminuant Δ_k on accroît l'importance de la partie linéaire de ψ_k qui coïncide avec l'approximation du premier ordre de f : $\nabla\psi_k(0) = \nabla f(x_k)$. On montre alors (Proposition 9.11) que si $g_k \neq 0$ et si la résolution de (RC_k) est suffisamment précise on finira par avoir $\rho_k > \omega_1$ (pour Δ_k petit) et par passer à l'étape 3. On montre le même résultat (Proposition 9.14) lorsque $g_k = 0$, à condition que l'on prenne $M_k = \nabla^2 f(x_k)$ et que $\nabla^2 f(x_k)$ ait une valeur propre < 0. Lorsqu'il n'y a pas de bouclage à l'étape 2 de l'algorithme, on dit que *l'itération k est un succès*. Sinon, on dit que *l'itération k est un échec*.

Lorsqu'on passe à l'étape 3, on a $\rho_k \geqslant \omega_1$, ce qui s'écrit encore

$$f(x_k + s_k) \leqslant f(x_k) + \omega_1\, \psi_k(s_k). \tag{9.2}$$

Cette condition rappelle la condition d'Armijo pour les méthodes à directions de descente et recherche linéaire, mais le terme contrôlant la décroissance de f, ici $\omega_1\psi_k(s_k) \leqslant 0$, n'a pas le caractère linéaire qu'il a dans (6.9). Comme dans la règle d'Armijo (technique de rebroussement), on force la réalisation de cette condition en faisant décroître Δ_k. La structure du schéma 9.1 met clairement en évidence l'analogie entre les deux approches. L'étape 2 qui détermine Δ_k dans le schéma 9.1 correspond à l'étape de calcul du pas α_k en recherche linéaire. La différence essentielle est que la valeur de Δ_k a une incidence sur la direction $s_k \in \mathbb{R}^n$, alors que α_k n'affecte pas la direction $d_k \in \mathbb{R}^n$ suivant laquelle on se déplace en recherche linéaire.

La règle de mise à jour du rayon de confiance à l'étape 4 est simple. On sait alors que la concordance $\rho_k > \omega_1 > 0$. Si la concordance est moyenne $(\omega_1 < \rho_k \leqslant \omega_2)$ on diminue (éventuellement) Δ_k; si elle est bonne $(\rho_k > \omega_2)$ on augmente (éventuellement) Δ_k. Remarquons que, contrairement aux méthodes à directions de descente, il n'y a aucun mécanisme permettant d'accroître Δ_k au cours de la même itération. L'accroissement éventuel de Δ_k se fait ici d'une itération à l'autre. Ceci est acceptable car il est raisonnable de faire l'hypothèse que la région de confiance du modèle varie peu d'une itération à l'autre (ce qui n'est pas le cas pour le pas dans les méthodes à directions de descente lorsque le modèle n'est pas newtonien).

On peut prendre d'autres normes que la norme euclidienne pour définir la région de confiance, mais cela conduit à un problème (RC_k) d'une autre nature. Par exemple, on pourrait prendre la norme ℓ_∞, qui donne à (RC_k) n contraintes de borne. Bien que ces contraintes soient linéaires, le problème est plus difficile à résoudre exactement du fait de son aspect combinatoire (quelles sont les contraintes actives en la solution ?). L'intérêt de la norme euclidienne vient de ce que (RC_k) n'a qu'une seule contrainte (il n'y a donc qu'un seul multiplicateur à trouver) et que cette contrainte est différentiable (en passant au carré, elle est donc quadratique). On verra une méthode efficace pour résoudre (RC_k) de manière approchée demandant en moyenne moins de deux itérations pour les problèmes de taille moyenne et des critères d'arrêt raisonnables.

9.1.2 Préconditionnement des méthodes à régions de confiance

Dans les problèmes d'optimisation, il est classique de préconditionner une méthode numérique, soit en faisant un changement de variables et en utilisant l'algorithme dans l'espace transformé, soit en utilisant un produit scalaire différent du produit scalaire euclidien. Sous certaines conditions les deux approches donnent le même préconditionnement. Il en est de même ici.

Changement de variables

Considérons d'abord le changement de variables $x \mapsto \tilde{x}$ sur $\mathbb{R}^n$, défini par

$$\tilde{x} = C_k x, \tag{9.3}$$

où C_k est une matrice régulière pouvant dépendre de l'itéré courant x_k. On introduit dans l'espace des $\tilde{x}$ la fonction $\tilde{f}_k := f \circ C_k^{-1}$, pour laquelle on a donc

$$\tilde{f}_k(\tilde{x}) = f(C_k^{-1}\tilde{x}), \quad \text{pour tout } \tilde{x} \in \mathbb{R}^n.$$

Donc, si x et $\tilde{x}$ sont liés par (9.3), on a $\tilde{f}_k(\tilde{x}) = f(x)$, pour tout $x \in \mathbb{R}^n$. On voit aisément que sous la correspondance (9.3), on a

$$\nabla_{\tilde{x}} \tilde{f}_k(\tilde{x}) = C_k^{-\mathsf{T}} \nabla_x f(x) \quad \text{et} \quad \nabla^2_{\tilde{x}\tilde{x}} \tilde{f}_k(\tilde{x}) = C_k^{-\mathsf{T}} \nabla^2_{xx} f(x) C_k^{-1}. \tag{9.4}$$

Pour que le problème dans l'espace transformé soit bien conditionné, il est judicieux de choisir C_k de manière à avoir $\nabla^2_{\tilde{x}\tilde{x}} \tilde{f}_k(\tilde{x}_*) \simeq I$. Ceci induit la condition sur C_k suivante

$$\boxed{C_k^{\mathsf{T}} C_k \simeq \nabla^2_{xx} f(x_k).} \tag{9.5}$$

Dans l'espace transformé, le problème (RC_k) s'écrit

$$\begin{cases} \min \ \tilde{g}_k^{\mathsf{T}} \tilde{s} + \frac{1}{2} \tilde{s}^{\mathsf{T}} \tilde{M}_k \tilde{s} \\ \|\tilde{s}\| \leqslant \Delta_k, \end{cases}$$

où $\tilde{g}_k = \nabla_{\tilde{x}} \tilde{f}_k(\tilde{x}_k)$ et $\tilde{M}_k \simeq \nabla^2_{\tilde{x}} \tilde{f}_k(\tilde{x}_k)$. Ramené à l'espace de départ en utilisant (9.3) et (9.4), ce problème devient

$$\boxed{\begin{cases} \min \ g_k^{\mathsf{T}} s + \frac{1}{2} s^{\mathsf{T}} M_k s \\ \|C_k s\| \leqslant \Delta_k, \end{cases}} \tag{9.6}$$

où $g_k = \nabla_x f(x_k)$ et $M_k = C_k^{\mathsf{T}} \tilde{M}_k C_k \simeq \nabla^2_x f(x_k)$.

En conclusion, le préconditionnement de (RC_k) s'obtient en modifiant la région de confiance sphérique $\bar{B}_2(0, \Delta_k)$ qui devient l'ellipsoïde $\{s : \|C_k s\|_2 \leqslant \Delta_k\}$ dont la forme est donnée par une matrice régulière C_k vérifiant au mieux (9.5).

Changement de produit scalaire

Soit à présent un produit scalaire $\langle \cdot, \cdot \rangle_{(k)}$ sur $\mathbb{R}^n$, qui peut dépendre de l'itéré courant, et $\| \cdot \|_{(k)}$ la norme associée. Alors, il est « naturel » de prendre comme sous-problème quadratique à l'itération k :

$$\begin{cases} \min \langle \bar{g}_k, s \rangle + \frac{1}{2} \langle \bar{M}_k s, s \rangle_{(k)} \\ \|s\|_{(k)} \leqslant \Delta_k, \end{cases}$$

où $\bar{g}_k$ est le gradient de f en x_k et $\bar{M}_k$ est une approximation de la hessienne de f en x_k *pour le produit scalaire* $\langle \cdot, \cdot \rangle_{(k)}$. Ce produit scalaire peut s'exprimer en fonction du produit scalaire euclidien par

$$\boxed{\langle u, v \rangle_{(k)} = u^{\mathsf{T}} P_k v,}$$

où P_k est une matrice symétrique définie positive. Alors, on voit aisément que

$$\bar{g}_k = P_k^{-1} g_k$$

et il est normal d'avoir (il s'agit d'approximations)

$$\bar{M}_k = P_k^{-1} M_k.$$

Dès lors le problème précédent devient

$$\begin{cases} \min \ g_k^{\mathsf{T}} s + \frac{1}{2} s^{\mathsf{T}} M_k s \\ (s^{\mathsf{T}} P_k s)^{1/2} \leqslant \Delta_k. \end{cases}$$

On retrouve une matrice de préconditionnement P_k qui n'affecte que la forme de la région de confiance. Si

$$\boxed{P_k = C_k^{\mathsf{T}} C_k,} \tag{9.7}$$

on retrouve le problème préconditionné (9.6). On réalise donc le même préconditionnement avec ce nouveau produit scalaire qu'avec le changement de variables (9.3) si P_k et C_k sont liés par (9.7).

9.2 Exemples de modèles quadratiques

9.2.1 Modèle linéaire

On prend $M_k = 0$, ce qui fournit un modèle linéaire en s. À chaque itération, il faut résoudre

$$\begin{cases} \min \ g_k^{\mathsf{T}} s \\ \|s\| \leqslant \Delta_k. \end{cases}$$

En écrivant les conditions d'optimalité, on voit que la solution, disons s_k, doit vérifier

$$\lambda_k s_k = -g_k,$$

où $\lambda_k \geqslant 0$ est le multiplicateur associé à la contrainte $\frac{1}{2}\|s\|^2 \leqslant \frac{1}{2}\Delta_k^2$. Si $g_k \neq 0$, le multiplicateur est non nul et la contrainte doit donc être active : s_k est obtenu en faisant un pas dans la direction de la plus profonde descente $-g_k$, de manière à venir toucher le bord de la région de confiance. Si $g_k = 0$, λ_k doit être nul et s_k peut être arbitraire dans la boule $\bar{B}(0, \Delta_k)$.

Avec ce modèle, on est très proche de l'algorithme du gradient (section 6.2.1). Le pas le long de $-g_k$ est déterminé en adaptant la région de confiance (schéma 9.1), d'une manière très semblable à ce que ferait une recherche linéaire.

9.2.2 Modèle newtonien

Au point $x_k \in \mathbb{R}^n$, on prend $M_k = \nabla^2 f(x_k)$. Il s'agit donc de résoudre à l'itération k :

$$\begin{cases} \min\ g_k^\mathsf{T} s + \dfrac{1}{2} s^\mathsf{T} \nabla^2 f(x_k) s \\ \|s\| \leqslant \Delta_k. \end{cases}$$

On observera que ce problème a toujours une solution (critère continu et ensemble admissible compact), même si la hessienne n'est pas définie positive, mais il n'est en général plus possible d'en donner une solution analytiquement. Comme nous le verrons à la section 9.3, il y a en effet parfois un problème de valeur propre à résoudre. Des méthodes numériques de résolution seront proposées et étudiées à la section 9.4.

9.2.3 Modèle gauss-newtonien

On se place ici dans le cadre déjà rencontré à la section 6.2.5, où la fonction à minimiser est la norme ℓ_2 d'une fonction $r : \mathbb{R}^n \to \mathbb{R}^m$, appelée résidu :

$$\min_{x \in \mathbb{R}^n} \frac{1}{2}\|r(x)\|_2^2.$$

De manière à éviter le calcul des dérivées secondes des résidus r_i, ce que le modèle newtonien exigerait, on se contente de la partie de la hessienne qui ne fait intervenir que la jacobienne $J_k = r'(x_k)$ des résidus, à savoir $J_k^\mathsf{T} J_k$. Il s'agit donc de résoudre à l'itération k :

$$\begin{cases} \min\ r_k^\mathsf{T} J_k s + \dfrac{1}{2}\|J_k s\|^2 \\ \|s\| \leqslant \Delta_k. \end{cases}$$

L'approche par régions de confiance fournit alors une *relecture* intéressante de l'algorithme de Levenberg-Marquardt, qui sera approfondie à la section 19.3.3.

9.3 Analyse du sous-problème quadratique

Le sous-problème quadratique à résoudre à chaque itération est de la forme suivante (on laisse tomber l'indice k)

$$(RC) \quad \begin{cases} \min\ \psi(s) = g^\mathsf{T} s + \dfrac{1}{2} s^\mathsf{T} M s \\ \|s\| \leqslant \Delta, \end{cases} \tag{9.8}$$

où $g \in \mathbb{R}^n$, $M \in \mathbb{R}^{n\times n}$ est symétrique (sans être nécessairement définie positive) et $\Delta > 0$. On note

$$\lambda_1 \leqslant \cdots \leqslant \lambda_n,$$

les valeurs propres de M (elles sont réelles) et $E_1 = \mathcal{N}(M - \lambda_1 I)$ l'espace propre correspondant à la plus petite valeur propre λ_1.

9.3.1 Conditions d'optimalité

Le résultat suivant est fondamental pour l'étude des méthodes à régions de confiance. Il est aussi remarquable car il donne des conditions nécessaires *et suffisantes* d'optimalité pour un problème *non* convexe (on ne suppose pas que M est semi-définie positive), ce qui est assez rare. Cela vient de la structure particulière du problème : un critère quadratique (non nécessairement convexe) et une seule contrainte quadratique convexe ($\|s\|^2 \leqslant \Delta^2$). On note

$$\ell : (s,\lambda) \in \mathbb{R}^n \times \mathbb{R} \mapsto \ell(s,\lambda) = g^{\mathsf{T}} s + \tfrac{1}{2}\, s^{\mathsf{T}} M s + \tfrac{1}{2}\,\lambda(\|s\|^2 - \Delta^2),$$

le lagrangien du problème (RC), dans lequel la contrainte à été rendue différentiable en remplaçant $\|s\| \leqslant \Delta$ par la contrainte équivalente $\|s\|^2 \leqslant \Delta^2$. Selon le théorème 4.32, les conditions de KKT en un couple stationnaire $(\hat{s}, \hat{\lambda})$ s'écrivent (on utilise la notation compacte (4.32))

$$(M + \hat{\lambda} I)\hat{s} = -g \qquad \text{et} \qquad 0 \leqslant \hat{\lambda} \perp (\Delta - \|\hat{s}\|) \geqslant 0. \tag{9.9}$$

On rappelle que, lorsque A est une matrice symétrique, la notation $A \succcurlyeq 0$ signifie que A est semi-définie positive.

Théorème 9.2 (caractérisation d'une solution de (RC)) *On considère le problème de région de confiance (9.8) avec $\Delta > 0$. Soit $\hat{s} \in \mathbb{R}^n$. Les propositions suivantes sont équivalentes :*

(i) $\hat{s}$ est solution de (RC),
(ii) il existe $\hat{\lambda} \in \mathbb{R}$ tel que $(\hat{s}, \hat{\lambda})$ vérifie (9.9) et $(M + \hat{\lambda} I) \succcurlyeq 0$,
(iii) il existe $\hat{\lambda} \in \mathbb{R}$ tel que $(\hat{s}, \hat{\lambda})$ vérifie (9.9) et $\hat{s}$ minimise $\ell(\cdot, \hat{\lambda})$ sur $\mathbb{R}^n$.

Démonstration. [$(i) \Rightarrow (ii)$] Soit $\hat{s}$ une solution de (RC). Si la contrainte $\frac{1}{2}(\|\cdot\|^2 - \Delta^2)$ est active en $\hat{s}$, alors $\hat{s} \neq 0$ et la contrainte y est qualifiée car $\nabla c(\hat{s}) = \hat{s} \neq 0$. Par le théorème 4.32, il existe un multiplicateur $\hat{\lambda} \in \mathbb{R}$ tel que l'on ait (9.9). Il reste à montrer que $(M + \hat{\lambda} I) \succcurlyeq 0$.

D'après la condition du second ordre,

$$h^{\mathsf{T}}(M + \hat{\lambda} I)h > 0,$$

pour tout h non nul dans le cône critique $C(\hat{s})$. Si $\|\hat{s}\| < \Delta$, la contrainte n'est pas active, $\hat{\lambda} = 0$ et le cône critique $C(\hat{s}) = \mathbb{R}^n$: on a donc $(M + \hat{\lambda} I) = M \succcurlyeq 0$. Si $\|\hat{s}\| = \Delta$, alors $C(\hat{s})$ contient les $h \in \mathbb{R}^n$ vérifiant $\hat{s}^{\mathsf{T}} h = 0$. Donc

$$\forall h \text{ tel que } \hat{s}^{\mathsf{T}} h = 0 : \quad h^{\mathsf{T}}(M + \hat{\lambda} I)h \geqslant 0.$$

Il reste à montrer cette inégalité pour les autres directions h. Le point-clé de la démonstration est de remarquer que pour tout $s \in \mathbb{R}^n$ tel que $\|s\| = \|\hat{s}\|$, on a

$$\begin{aligned} 0 &\leqslant \psi(s) - \psi(\hat{s}) \quad [\text{car } \hat{s} \text{ est optimal}] \\ &= g^T s + \frac{1}{2} s^T M s - g^T \hat{s} - \frac{1}{2} \hat{s}^T M \hat{s} \\ &= -\hat{s}^T (M + \hat{\lambda} I) s + \frac{1}{2} s^T M s + \hat{s}^T (M + \hat{\lambda} I) \hat{s} - \frac{1}{2} \hat{s}^T M \hat{s} \qquad [(9.9)_1] \\ &= -\hat{s}^T (M + \hat{\lambda} I) s + \frac{1}{2} s^T (M + \hat{\lambda} I) s + \frac{1}{2} \hat{s}^T (M + \hat{\lambda} I) \hat{s} \qquad [\|\hat{s}\| = \|s\|]. \end{aligned}$$

Donc

$$\forall s \text{ tel que } \|s\| = \|\hat{s}\| : \quad 0 \leqslant \frac{1}{2} (s - \hat{s})^T (M + \hat{\lambda} I)(s - \hat{s}). \tag{9.10}$$

Soit $h \in \mathbb{R}^n$ tel que $\hat{s}^T h \neq 0$. Alors l'équation en α,

$$\|\hat{s} + \alpha h\|^2 = \|\hat{s}\|^2,$$

a deux racines réelles. En notant α la racine non nulle et en prenant $s = \hat{s} + \alpha h$, on a $\|s\| = \|\hat{s}\|$ et (9.10) montre que

$$\forall h \text{ tel que } \hat{s}^T h \neq 0 : \quad h^T (M + \hat{\lambda} I) h \geqslant 0.$$

$[(ii) \Rightarrow (iii)]$ Comme $\ell(\cdot, \hat{\lambda})$ est quadratique de hessienne $M + \hat{\lambda} I \succcurlyeq 0$, cette fonction est convexe. Par $(9.9)_1$, son gradient s'annule en $\hat{s}$, si bien que $\hat{s}$ minimise $\ell(\cdot, \hat{\lambda})$.

$[(iii) \Rightarrow (i)]$ Par $(9.9)_2$ et la proposition 4.73, $\hat{\lambda}$ maximise $\ell(\hat{x}, \cdot)$. Par ailleurs, il est supposé dans (iii) que $\hat{s}$ minimise $\ell(\cdot, \hat{\lambda})$. Dès lors, $(\hat{s}, \hat{\lambda})$ est un point-selle du lagrangien. Par la proposition 4.74, $\hat{s}$ est alors solution (globale) de (RC). □

Un résultat voisin du théorème 9.2 pour les problèmes homogènes est proposé à l'exercice 14.11. La dualisation lagrangienne du problème de région de confiance (RC) est considérée à l'exercice 14.9.

Il sera utile de détailler les conditions caractérisant une solution $\hat{s}$ du problème de région de confiance (9.8) au moyen d'un multiplicateur associé $\hat{\lambda}$ (point (ii) du théorème 9.2) :

$$(M + \hat{\lambda} I)\hat{s} = -g, \tag{9.11a}$$
$$\|\hat{s}\| \leqslant \Delta, \tag{9.11b}$$
$$\hat{\lambda} \geqslant 0, \tag{9.11c}$$
$$\hat{\lambda}(\Delta - \|\hat{s}\|) = 0, \tag{9.11d}$$
$$(M + \hat{\lambda} I) \text{ est semi-définie positive.} \tag{9.11e}$$

On notera que, quelle que soit la matrice M, toute solution $\hat{s}$ du sous-problème quadratique (RC) est une direction de descente (non stricte) de f en x, puisque $g^T \hat{s} \leqslant 0$ grâce à (9.11a) et (9.11e).

9.3.2 Le chemin de recherche

D'après l'équation d'optimalité (9.11e), lorsque $\lambda > \hat{\lambda}$, $(M+\lambda I)$ est définie positive et on peut définir de manière unique un point $s_\lambda \in \mathbb{R}^n$ comme solution du problème linéaire

$$(M + \lambda I)s_\lambda = -g. \tag{9.12}$$

Si λ est également pris positif, le couple (s_λ, λ) vérifie les conditions d'optimalité (9.11a), (9.11c) et (9.11e). On peut donc songer à satisfaire les deux autres conditions en se déplaçant le long de la courbe définie par l'application

$$\lambda \in]-\lambda_1, +\infty[\mapsto s_\lambda \in \mathbb{R}^n.$$

C'est ce que fait l'algorithme de résolution de Moré-Sorensen décrit en section 9.4.2.

Dans cette section 9.3.2, nous allons étudier l'application $\lambda \mapsto s_\lambda$, ce qui apportera une meilleure compréhension du sous-problème quadratique (RC).

Proposition 9.3 *(i) L'application $\lambda \in]-\lambda_1, +\infty[\mapsto s_\lambda \in \mathbb{R}^n$ est de classe C^∞.*

(ii) $s_\lambda \to 0$ lorsque $\lambda \to +\infty$.

(iii) La tangente en l'origine à la courbe définie par l'application $\lambda \mapsto s_\lambda$ est colinéaire à $-g$.

(iv) Si $v_1, \ldots, v_n$ sont n vecteurs propres orthonormaux de M, associés aux valeurs propres $\lambda_1, \ldots, \lambda_n$, on a

$$s_\lambda = -\sum_{i=1}^{n} \frac{v_i^T g}{\lambda_i + \lambda} v_i. \tag{9.13}$$

Démonstration. L'application $\lambda \mapsto s_\lambda$ apparaît comme fonction implicite de l'équation $\Psi(s, \lambda) = 0$, où

$$\Psi(s, \lambda) = (M + \lambda I)s + g.$$

Comme Ψ est C^∞ et que $\Psi'_s(s, \lambda) = M + \lambda I$ est inversible pour $\lambda \in]-\lambda_1, +\infty[$, la fonction implicite est C^∞ sur cet intervalle.

En multipliant scalairement l'équation (9.12) par s_λ, on obtient $(\lambda_1 + \lambda)\|s_\lambda\|^2 \leqslant -g^T s_\lambda$ et par l'inégalité de Cauchy-Schwarz

$$(\lambda_1 + \lambda)\|s_\lambda\| \leqslant \|g\|.$$

On en déduit que $s_\lambda \to 0$ pour $\lambda \to \infty$.

Si pour $\lambda > \max(0, -\lambda_1)$ on fait le changement de variable $\mu = 1/\lambda$ et si l'on note $r_\mu = s_\lambda$, on a

$$(\mu M + I)r_\mu = -\mu g.$$

Cette équation définit une fonction implicite $\mu \mapsto r_\mu$, de classe C^∞ dans un voisinage de 0, dont l'image pour $\mu > 0$ se confond avec l'image de l'application $\lambda \mapsto s_\lambda$. On a $r_0 = 0$ et en dérivant l'équation ci-dessus on obtient $r'_0 = -g$.

Soit $V = (v_1, \ldots, v_n)$ la matrice orthogonale formée des vecteurs propres considérés de M et $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. On a $M + \lambda I = V(\Lambda + \lambda I)V^\mathsf{T}$ et

$$(M + \lambda I)^{-1} = V(\Lambda + \lambda I)^{-1}V^\mathsf{T}.$$

Alors (9.12) conduit à (9.13). □

La figure 9.1 représente le chemin image de l'application $\lambda \mapsto s_\lambda$.

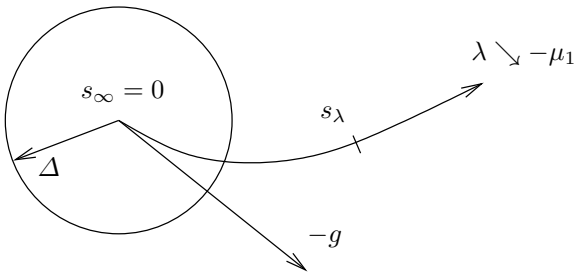


Fig. 9.1. Le chemin $\{s_\lambda : \lambda > -\lambda_1\}$.

Proposition 9.4 *On a*

$$\|s_\lambda\|^2 = \sum_{i=1}^{n} \frac{(v_i^\mathsf{T} g)^2}{(\lambda_i + \lambda)^2}. \tag{9.14}$$

Si $g \neq 0$, les applications $\lambda \mapsto \|s_\lambda\|$ et $\lambda \mapsto \|s_\lambda\|^2$ sont strictement décroissantes et strictement convexes sur $]-\lambda_1, +\infty[$.

Démonstration. La relation (9.14) se déduit aisément de (9.13). Notons

$$\varphi(\lambda) = \|s_\lambda\|^2 \quad \text{et} \quad \nu(\lambda) = \|s_\lambda\|,$$

et supposons que $g \neq 0$. Donc $v_j^\mathsf{T} g \neq 0$ pour au moins un indice j. Alors, d'après (9.14),

$$\varphi'(\lambda) = -2\sum_{i=1}^{n} \frac{(v_i^\mathsf{T} g)^2}{(\lambda_i + \lambda)^3} < 0.$$

Donc φ est strictement décroissante. Comme $\nu'(\lambda) = \varphi'(\lambda)/(2\nu(\lambda))$, il en est de même de ν.

On a aussi

$$\varphi''(\lambda) = 6\sum_{i=1}^{n} \frac{(v_i^\mathsf{T} g)^2}{(\lambda_i + \lambda)^4} > 0,$$

d'où la stricte convexité de φ. D'autre part, comme $\varphi'' = 2(\nu')^2 + 2\nu\nu'' = (\varphi')^2/(2\nu^2) + 2\nu\nu''$, on a

$$\begin{aligned}\nu''(\lambda) &= \frac{1}{4\nu(\lambda)^3}[2\varphi(\lambda)\varphi''(\lambda) - \varphi'(\lambda)^2] \\ &= \frac{1}{4\nu(\lambda)^3}\left[12\sum_{i=1}^{n}\frac{(v_i^{\mathsf{T}}g)^2}{(\lambda_i+\lambda)^2}\sum_{i=1}^{n}\frac{(v_i^{\mathsf{T}}g)^2}{(\lambda_i+\lambda)^4} - 4\left(\sum_{i=1}^{n}\frac{(v_i^{\mathsf{T}}g)^2}{(\lambda_i+\lambda)^3}\right)^2\right].\end{aligned}$$

En utilisant l'inégalité de Cauchy-Schwarz, on en déduit

$$\nu''(\lambda) \geqslant \frac{2}{\nu(\lambda)^3}\left[\sum_{i=1}^{n}\frac{(v_i^{\mathsf{T}}g)^2}{(\lambda_i+\lambda)^2}\sum_{i=1}^{n}\frac{(v_i^{\mathsf{T}}g)^2}{(\lambda_i+\lambda)^4}\right] = \frac{\varphi(\lambda)\varphi''(\lambda)}{3\nu(\lambda)^3} > 0.$$

Ceci implique la stricte convexité de ν. □

La relation (9.14) montre que

$$\lim_{\lambda \searrow -\lambda_1} \|s_\lambda\| = +\infty \quad \Longleftrightarrow \quad P_{E_1} g \neq 0.$$

On a noté $P_{E_1}g$, la projection orthogonale de g sur l'espace propre E_1 correspondant à la valeur propre λ_1. Les allures possibles pour la courbe $\lambda \mapsto \|s_\lambda\|$ correspondant aux cas $P_{E_1}g \neq 0$ et $P_{E_1}g = 0$ sont données à la figure 9.2.

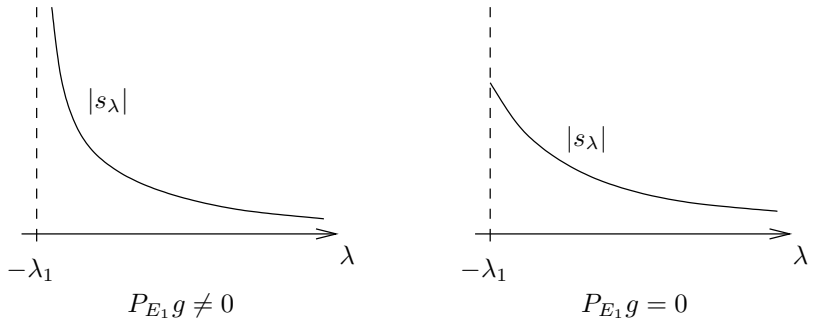


Fig. 9.2. Allures des applications $\lambda \mapsto \|s_\lambda\|$.

La proposition suivante est utile pour analyser le cas où $P_{E_1}g = 0$.

Proposition 9.5 *Si $P_{E_1}g = 0$, on a les propriétés suivantes :*
(i) $s_\lambda \perp E_1$, pour tout $\lambda > -\lambda_1$,
(ii) l'équation en s

$$(M - \lambda_1 I)s = -g, \tag{9.15}$$

a des solutions et $\|s\| \geqslant \|s_\lambda\|$, pour toute solution s de (9.15) et pour tout $\lambda > -\lambda_1$,

(*iii*) $\lim_{\lambda \searrow -\lambda_1} s_\lambda = s^+_{-\lambda_1}$, *où*

$$s^+_{-\lambda_1} = -(M - \lambda_1 I)^+ g,$$

$(\cdot)^+$ *désignant l'inverse généralisé.*

DÉMONSTRATION. 1) Soit $v \in E_1$. Alors $v^T g = 0$ et (9.12) donne

$$0 = v^T (M + \lambda I) s_\lambda = (\lambda_1 + \lambda) v^T s_\lambda.$$

Donc $v^T s_\lambda = 0$, puisque $\lambda > -\lambda_1$.

2) Comme $g \in E_1^\perp = \mathcal{R}(M - \lambda_1 I)$, l'équation (9.15) a des solutions. Soit $\tilde{s}$ une de ces solutions. Montrons que

$$\|s_\lambda\| \leqslant \|\tilde{s}\|, \quad \forall \lambda > -\lambda_1. \tag{9.16}$$

D'après (9.12) et la définie positivité de $(M + \lambda I)$ pour $\lambda > -\lambda_1$, s_λ est solution du problème convexe

$$\min_s \ \psi(s) + \frac{\lambda}{2}\|s\|^2.$$

De même, $\tilde{s}$ est solution de ce problème avec $\lambda = -\lambda_1$. On en déduit

$$\psi(s_\lambda) + \frac{\lambda}{2}\|s_\lambda\|^2 \leqslant \psi(\tilde{s}) + \frac{\lambda}{2}\|\tilde{s}\|^2$$

$$\psi(\tilde{s}) - \frac{\lambda_1}{2}\|\tilde{s}\|^2 \leqslant \psi(s_\lambda) - \frac{\lambda_1}{2}\|s_\lambda\|^2$$

En sommant ces deux inégalités, on trouve (9.16).

3) D'après (*ii*), la suite $\{s_\lambda\}_{\lambda \searrow -\lambda_1}$ est bornée. Ses points d'adhérence $\bar{s}$ vérifient (9.15) avec $s = \bar{s}$ et $\|\bar{s}\| \leqslant \|\tilde{s}\|$ pour toute solution $\tilde{s}$ de (9.15). Ceci implique qu'il n'y a qu'un seul point d'adhérence qui est la solution de norme minimale de (9.15), ou encore $s^+_{-\lambda_1}$. □

Proposition 9.6 *Si (s, λ) est solution des équations d'optimalité de (RC), alors λ est déterminé de façon unique. Il en est de même de s si $\lambda > -\lambda_1$.*

DÉMONSTRATION. La seconde partie de l'énoncé se déduit de la première et de l'équation d'optimalité (9.11a) puisqu'alors $(M + \lambda I)$ est inversible.

Montrons à présent l'unicité de λ. Supposons que l'on ait deux solutions (s', λ') et (s'', λ'') avec $\lambda' < \lambda''$. Comme $\lambda' \geqslant -\lambda_1$, on a $\lambda'' > -\lambda_1$. Donc $M + \lambda'' I$ est régulière et $s'' = s_{\lambda''}$. Analysons successivement les cas où $g = 0$ et $g \neq 0$.

Si $g = 0$, alors $s'' = 0$ et, par (9.11d), $\lambda'' = 0$. Or ceci contredit le fait que $\lambda'' > \lambda'$ puisque $\lambda' \geqslant 0$.

Si $g \neq 0$, l'application $\lambda \mapsto \|s_\lambda\|$ est strictement décroissante sur $]-\lambda_1, \infty[$ (Proposition 9.4). Donc

$$\Delta \geqslant \|s'\| \geqslant \|s_{\frac{\lambda'+\lambda''}{2}}\| > \|s_{\lambda''}\|,$$

la seconde inégalité venant de la proposition 9.5-(ii) lorsque $\lambda' \neq -\lambda_1$ et $P_{E_1} g = 0$. Grâce à (9.11d), on en déduit encore que $\lambda'' = 0$, ce qui conduit à la même contradiction que dans le cas où $g = 0$. □

9.3.3 Discussion des solutions

Les résultats de la section 9.3.2 précédente permettent de discuter les solutions de (RC) en fonction de Δ, λ_1 et $P_{E_1} g$. Nous distinguons les cas où M est ou n'est pas définie positive.

Cas A : M est définie positive ($\lambda_1 > 0$). Alors quel que soit $\lambda \geqslant 0$, $(M + \lambda I)$ est définie positive et, vu l'unicité de λ, (RC) a une solution unique. Celle-ci est située sur la courbe définie par l'application $\lambda \mapsto s_\lambda$, qui est bien définie pour $\lambda \in [0, \infty[$. Il y a deux sous-cas.

A-1. $\|s_0\| \leqslant \Delta$. Alors $(s_0, 0)$ vérifie les conditions d'optimalité. s_0 est le point minimisant ψ sur $\mathbb{R}^n$. Cette situation est illustrée à la figure 9.3.

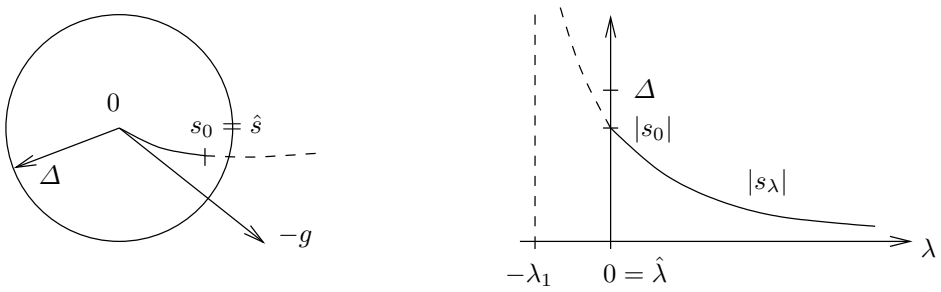


Fig. 9.3. Le cas A-1.

A-2. $\|s_0\| > \Delta$. La solution $\hat{s}$ est donnée par l'intersection de la sphère de rayon Δ et de la courbe définie par $\lambda \mapsto s_\lambda$. Cette situation est illustrée à la figure 9.4.

Cas B : M n'est pas définie positive ($\lambda_1 \leqslant 0$). La recherche de $\hat{\lambda}$ se fait sur $[-\lambda_1, \infty[$ mais s_λ n'est bien défini que pour $\lambda \in]-\lambda_1, \infty[$. On distingue trois sous-cas.

B-1. $P_{E_1} g \neq 0$. Alors $\|s_\lambda\| \to +\infty$ quand $\lambda \searrow -\lambda_1$. On peut donc trouver $\hat{\lambda} \in]-\lambda_1, \infty[$ tel que $\|s_{\hat{\lambda}}\| = \Delta$. Cette situation est illustrée à la figure 9.5.

B-2. $P_{E_1} g = 0$ et $\|s^+_{-\lambda_1}\| > \Delta$. Cette fois $\{s_\lambda\}_{\lambda > -\lambda_1}$ est bornée mais on peut encore trouver $\hat{\lambda} \in]-\lambda_1, \infty[$ tel que $\|s_{\hat{\lambda}}\| = \Delta$. Cette situation est illustrée à la figure 9.6.

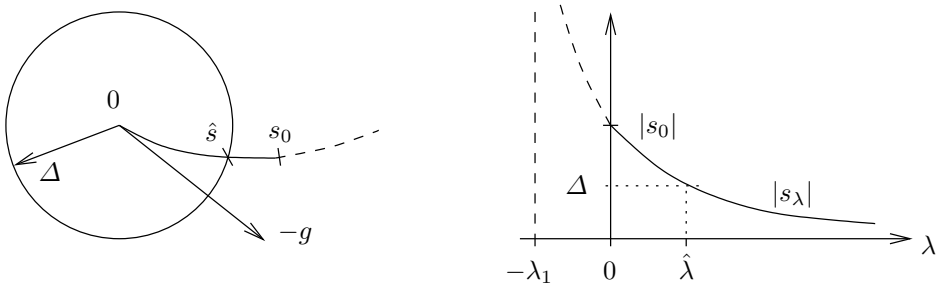


Fig. 9.4. Le cas A-2.

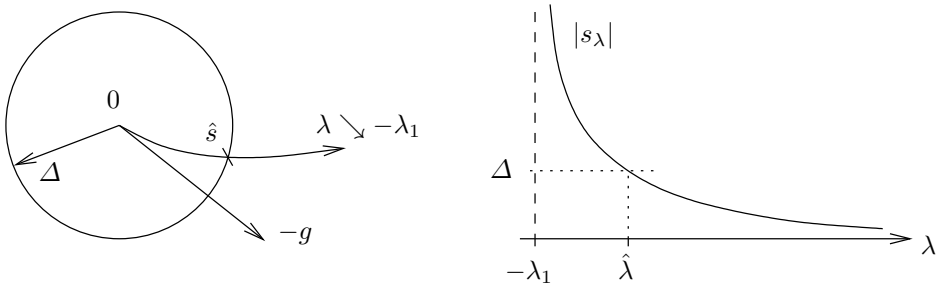


Fig. 9.5. Le cas B-1.

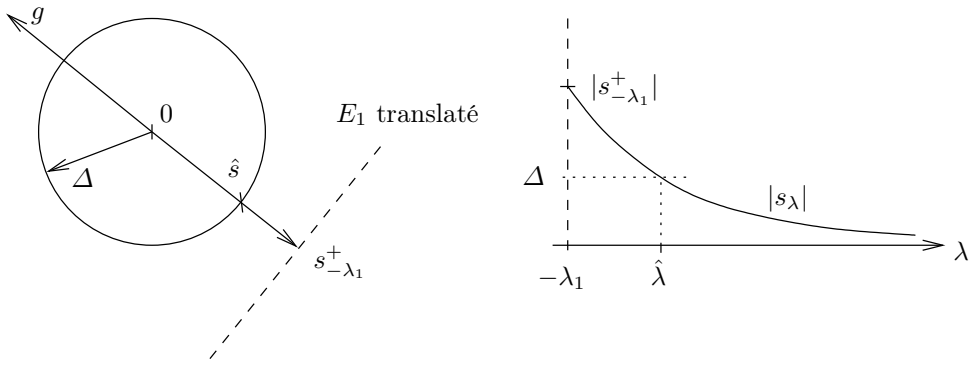


Fig. 9.6. Le cas B-2.

B-3. $P_{E_1} g = 0$ et $\|s^+_{-\lambda_1}\| \leqslant \Delta$. C'est le « cas difficile ». Il y a un problème de valeur propre à résoudre. Il faut prendre $\hat{\lambda} = -\lambda_1$ et trouver $\hat{s}$ tel que $\|\hat{s}\| \leqslant \Delta$, $(M - \lambda_1 I)\hat{s} = -g$ et $\lambda_1(\Delta - \|\hat{s}\|) = 0$. Donc $\hat{s}$ est de la forme

$$\hat{s} = s^+_{-\lambda_1} + t v_1, \tag{9.17}$$

où $s^+_{-\lambda_1}$ est défini dans la proposition 9.5, $t \in \mathbb{R}$ et v_1 est un vecteur propre de M associé à la valeur propre λ_1. Cette situation est illustrée à la figure 9.7.

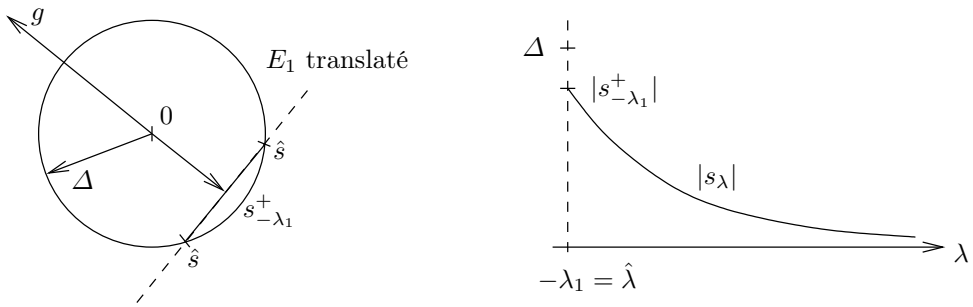


Fig. 9.7. Le cas B-3.

Seul le cas A-1 a une solution qui peut s'obtenir exactement. C'est le minimum de ψ sans contrainte : $s_0 = -M^{-1}g$. Dans les cas A-2, B-1 et B-2, il faut mettre en œuvre une méthode itérative pour trouver une racine (en λ) de l'équation

$$\|s_\lambda\| = \Delta.$$

Le cas B-3 est appelé le *cas difficile* : il a un problème de valeur propre à résoudre puisqu'il faut trouver λ_1. Si en plus $\lambda_1 < 0$, il faut trouver $\hat{s}$ tel que $\|\hat{s}\| = \Delta$. Cela peut se faire en ajustant le pas t dans (9.17) de manière à avoir $\|s^+_{-\lambda_1} + t v_1\| = \Delta$, ce qui est possible car $\|s^+_{-\lambda_1}\| \leqslant \Delta$.

La discussion qui précède permet de mettre en évidence un *mécanisme de repli* des méthodes à régions de confiance qui n'est pas sans attrait. Lorsque le modèle ψ_k s'écarte du modèle quadratique de la variation de f, l'algorithme décroît Δ_k, ce qui conduit à prendre un point s_λ avec une valeur plus grande de λ, là où la courbe $\lambda \mapsto s_\lambda$ est plus proche de la tangente issue de l'origine $-g$. Par conséquent, en cas de difficulté, on recolle à la méthode de la plus forte pente. Ce mécanisme sécurisant contribue à la robustesse des méthodes à régions de confiance et a servi de motivation à leur introduction (voir Marquardt [407 ; 1963]).

9.4 Résolution numérique du sous-problème quadratique

Il est clair d'après la section 9.3 qu'en général on ne peut pas résoudre le sous-problème quadratique (RC_k) exactement. Le seul cas simple se présente lorsque M_k

est définie positive et Δ_k est suffisamment grand pour que le minimum global de ψ_k, ou *point de Newton*

$$s_k^{\mathrm{N}} = -M_k^{-1} g_k,$$

soit dans la région de confiance ($\|s_k^{\mathrm{N}}\| \leqslant \Delta_k$). Autrement, il y a un problème non linéaire à résoudre et parfois un problème de valeur propre. Nous allons décrire dans cette section quelques méthodes numériques de résolution de (RC_k).

La convergence des méthodes à régions de confiance dépend aussi de l'algorithme utilisé pour résoudre le sous-problème quadratique, en particulier lorsque ceux-ci en calculent une solution très grossière, ce qui était passé sous silence dans le schéma algorithmique 9.1. En fait, il faut très peu en faire si l'on n'est intéressé que par le convergence des gradients vers zéro, c'est-à-dire si la stationnarité seule des points d'adhérence de $\{x_k\}$ est déjà satisfaisante. Nous discutons à la section 9.4.1 le rôle que joue le *point de Cauchy* dans cette convergence « au premier ordre » et nous y décrivons quelques algorithmes qui s'y réfèrent.

Les algorithmes à régions de confiance ont une propriété remarquable que n'ont pas les algorithmes à directions de descente[1] : ils peuvent forcer les itérés à s'échapper d'un point stationnaire qui n'est pas un minimum local, pourvu que le sous-problème quadratique soit résolu avec une précision suffisante. C'est l'intérêt principal des méthodes de résolution présentées à la section 9.4.2. On montrera à la section 9.5.2 qu'avec une résolution fine des sous-problèmes quadratiques utilisant $M_k = \nabla^2 f(x_k)$, les algorithmes à régions de confiance convergent vers des minima locaux (pas seulement des points stationnaires).

9.4.1 Méthodes fondées sur le point de Cauchy

Le point de Cauchy

Définition 9.7 On appelle *point de Cauchy* du sous-problème quadratique (RC_k), le point noté s_k^{C}, solution de :

$$\begin{cases} \min\ \psi_k(s) \\ \|s\| \leqslant \Delta_k, \\ s = -\alpha g_k, \quad \alpha \in \mathbb{R}. \end{cases}$$

□

C'est donc le point minimisant ψ_k dans la région de confiance, le long de la droite de plus forte pente de ψ_k (ou de f, puisque l'on a supposé que $\nabla\psi_k(0) = \nabla f(x_k)$).

Le point de Cauchy peut être calculé à un coût très faible, comme le montre la formule ci-dessous.

Proposition 9.8 *Le point de Cauchy s_k^{C} est unique et donné par*

[1] À moins que l'on ne soit à même de gérer un jour avec soin les directions à courbure négative du critère, sujet que nous n'avons pas abordé et qui est très peu stabilisé, faute d'idées algorithmiques qui s'imposent.

$$s_k^{\mathrm{C}} = \begin{cases} 0 & \text{si } g_k = 0 \\ -\dfrac{\Delta_k}{\|g_k\|} g_k & \text{si } g_k \neq 0 \text{ et } g_k^{\mathsf{T}} M_k g_k \leqslant 0 \\ -\min\left(\dfrac{\Delta_k}{\|g_k\|}, \dfrac{\|g_k\|^2}{g_k^{\mathsf{T}} M_k g_k}\right) g_k & \text{sinon.} \end{cases} \tag{9.18}$$

DÉMONSTRATION. En effet, le cas où $g_k = 0$ est clair. Supposons à présent que $g_k \neq 0$. Si $g_k^{\mathsf{T}} M_k g_k \leqslant 0$, alors $\alpha \mapsto \psi_k(-\alpha g_k)$ est concave et $\nabla\psi_k(0)^{\mathsf{T}}(-g_k) = -\|g_k\|^2 < 0$, si bien qu'il faut donc prendre α le plus grand possible tout en gardant $\|-\alpha g_k\| \leqslant \Delta_k$. Ceci donne $\alpha = \Delta_k/\|g_k\|$. Il reste à examiner le cas où $g_k^{\mathsf{T}} M_k g_k > 0$. Le résultat se déduit du fait que, dans ce cas, le minimum de $\alpha \mapsto \psi_k(-\alpha g_k)$ est obtenu en $\alpha = \|g_k\|^2 / g_k^{\mathsf{T}} M_k g_k$. □

Le point de Cauchy joue un rôle important dans la convergence des méthodes à régions de confiance. Pour que celles-ci convergent (au sens où la suite des gradients tend vers 0), il suffit en effet de résoudre le sous-problème quadratique de manière très approximative, en trouvant un point qui minimise ψ_k autant que le point de Cauchy. On pourrait évidemment prendre le point de Cauchy lui-même comme « solution » approchée très économique. En pratique, ce point n'est cependant pas très bon en ce qui concerne la *vitesse* de convergence et on a souvent intérêt à exploiter davantage le modèle quadratique que ne le fait le point de Cauchy. C'est ce que tentent de faire les algorithmes de résolution approchée présentés ci-après.

Nous allons traduire cette condition assurant la convergence que nous venons d'énoncer de manière formelle et nous verrons plus tard que les algorithmes de résolution du sous-problème quadratique vérifient effectivement cette condition. Celle-ci accepte d'être affaiblie en utilisant des constantes strictement positives β_1 et β_2. On dit que (RC_k) est résolu en s_k avec une *condition de décroissance suffisante* si

$$\begin{cases} \psi_k(s_k) \leqslant \beta_1 \psi_k(s_k^{\mathrm{C}}) \\ \|s_k\| \leqslant \beta_2 \Delta_k. \end{cases} \tag{9.19}$$

Comme annoncé, il faut que s_k fasse décroître ψ_k autant que le fait point de Cauchy s_k^{C}, à un facteur constant β_1 près sur ψ_k (en général $0 < \beta_1 \leqslant 1$) et avec une certaine tolérance sur l'appartenance de s_k à la région de confiance (en général $\beta_2 \geqslant 1$).

Tout algorithme de résolution de (RC_k) qui est capable de réaliser la condition (9.19) à chaque itération assurera la convergence de l'algorithme (voir la section 9.5.1). On a évidemment tout intérêt à choisir une méthode de résolution s'adaptant bien au problème à résoudre. En ce qui concerne les algorithmes généralistes, nous en décrirons deux parmi les plus courants.

Algorithme dogleg de Powell

Cet algorithme demande que M_k *soit définie positive*, ce que nous supposerons donc dans cette section, de telle sorte que le minimum de ψ_k sans contrainte puisse

être calculé. Ce minimum est le point de Newton, déjà rencontré

$$s_k^{\mathrm{N}} = -M_k^{-1} g_k.$$

Dans l'*algorithme dogleg*[2] *de Powell*, on cherche une solution approchée du sous-problème quadratique, en explorant que le chemin linéaire par morceaux qui rejoint l'origine au point de Cauchy s_k^{C}, puis au point de Newton s_k^{N} (voir figure 9.8). Si on

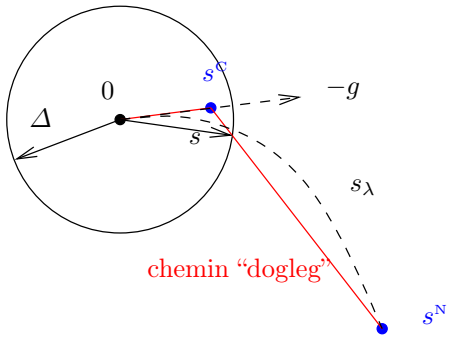


Fig. 9.8. L'algorithme « dogleg » de Powell (on a translaté x en 0).

note

$$\alpha_k^{\mathrm{C}} := \frac{\|s_k^{\mathrm{C}}\|}{\|s_k^{\mathrm{C}}\| + \|s_k^{\mathrm{N}} - s_k^{\mathrm{C}}\|},$$

ce chemin continu se décrit de manière précise en fonction d'une abscisse $\alpha \in [0,1]$ proportionnelle à son abscisse curviligne (ce qui n'est pas essentiel) de la manière suivante :

$$\alpha \mapsto s_k(\alpha) := \begin{cases} \dfrac{\alpha}{\alpha_k^{\mathrm{C}}}\, s_k^{\mathrm{C}} & \text{si } 0 \leqslant \alpha \leqslant \alpha_k^{\mathrm{C}} \\ \dfrac{1-\alpha}{1-\alpha_k^{\mathrm{C}}}\, s_k^{\mathrm{C}} + \dfrac{\alpha - \alpha_k^{\mathrm{C}}}{1-\alpha_k^{\mathrm{C}}}\, s_k^{\mathrm{N}} & \text{si } \alpha_k^{\mathrm{C}} < \alpha \leqslant 1. \end{cases}$$

On se contente alors de trouver *la* solution de

$$\begin{cases} \min_\alpha \ \psi_k(s_k(\alpha)) \\ \|s_k(\alpha)\| \leqslant \Delta_k. \end{cases}$$

On minimise donc ψ_k sur la partie du chemin qui est dans la région de confiance. Montrons que lorsque $g_k \neq 0$, ce problème a bien une unique solution.

– D'abord, $\psi_k(s(\alpha))$ *décroît strictement avec* α.

Cette propriété est clairement vraie pour la partie du chemin comprise entre 0 et s_k^{C} puisque $-g_k$ est une direction de descente de ψ_k et que s_k^{C} minimise ψ_k

[2] L'appellation « dogleg » (littéralement « patte de chien ») est sensé faire référence à la forme du chemin le long duquel on recherche la solution approchée dans cet algorithme.

le long de $\alpha \to -\alpha g_k$ dans la région de confiance. Elle est également vraie pour la partie du chemin comprise entre s_k^{C} et s_k^{N} (supposé différent de s_k^{C}, sinon il n'y a pas lieu de poursuivre le raisonnement), puisque $\psi_k(s_k^{\mathrm{C}}) > \psi_k(s_k^{\mathrm{N}})$ (ψ_k a s_k^{N} comme unique minimum et $s_k^{\mathrm{C}} \neq s_k^{\mathrm{N}}$) et que ψ_k est convexe (on utilise alors l'inégalité de convexité : $\psi_k'(s_k^{\mathrm{C}}; s_k^{\mathrm{N}} - s_k^{\mathrm{C}}) \leqslant \psi_k(s_k^{\mathrm{N}}) - \psi_k(s_k^{\mathrm{C}}) < 0$).

– Il reste à montrer que $\|s_k(\alpha)\|$ *croît strictement avec* α, puisqu'alors le problème ci-dessus revient à minimiser une fonction strictement décroissante $\psi_k(s_k(\cdot))$ sur un intervalle, problème dont l'existence et l'unicité de la solution sont alors claires.

  La croissance de $\|s_k(\alpha)\|$ est une conséquence évidente de l'inégalité $(s_k^{\mathrm{N}} - s_k^{\mathrm{C}})^{\mathsf{T}} s_k^{\mathrm{C}} \geqslant 0$. Comme celle-ci s'écrit aussi $(s_k^{\mathrm{N}})^{\mathsf{T}} s_k^{\mathrm{C}} \geqslant \|s_k^{\mathrm{C}}\|^2$, ou encore

$$g_k^{\mathsf{T}} M_k^{-1} g_k \geqslant \alpha_k^{\mathrm{C}} \|g_k\|^2 = \frac{\|g_k\|^4}{g_k^{\mathsf{T}} M_k g_k},$$

  elle a lieu comme une conséquence de l'inégalité de Cauchy-Schwarz.

Notons que la première propriété sur la décroissance de $\psi_k(s(\alpha))$ implique que le point sélectionné par dogleg vérifie (9.19) et par là la convergence globale de l'algorithme à régions de confiance utilisant cette méthode de résolution de (RC_k).

Algorithme du gradient conjugué tronqué de Steihaug

On peut voir cette approche comme une amélioration de l'algorithme dogleg, dans le sens où l'on n'a plus besoin de supposer que M_k est définie positive et où le chemin de recherche, aussi linéaire par morceaux, est vraisemblablement une meilleure approximation du chemin de recherche $\lambda \mapsto s_\lambda$ décrit à la section 9.3.2.

L'idée est simple. On utilise l'algorithme du gradient conjugué (GC) pour minimiser $\psi_k(s)$, en partant de $s_{k,0} = 0$. On note $s_{k,i}$ les itérés du GC, i désignant l'indice des itérations, et $v_{k,i}$ la direction conjuguée menant de $s_{k,i}$ à $s_{k,i+1}$. Cependant, comme M_k n'est pas nécessairement définie positive et que l'on veut rester dans la région de confiance, on arrête les itérations dès que l'un des événements suivants se produit :

(i) l'itéré courant $s_{k,i}$ est dans la région de confiance, mais pas le suivant $s_{k,i+1}$; on prend alors comme solution approchée, l'intersection de la droite $t \mapsto s_{k,i} + t v_{k,i}$ avec la frontière de la région de confiance; on a donc

$$s_k := s_{k,i} + t v_{k,i},$$

où $t \geqslant 0$ est tel que $\|s_{k,i} + t v_{k,i}\|^2 = \Delta^2$ (une racine d'un polynôme du second degré à calculer);

(ii) l'algorithme rencontre une direction conjuguée $v_{k,i} \neq 0$, qui est *à courbure négative* pour M_k, c'est-à-dire

$$v_{k,i}^{\mathsf{T}} M_k v_{k,i} \leqslant 0;$$

dans ce cas, les itérations du GC ne peuvent être poursuivies (si $v_{k,i}^{\mathsf{T}} M_k v_{k,i} = 0$ on ne peut pas définir le pas le long de $v_{k,i}$, permettant de passer à l'itéré suivant $s_{k,i+1}$) et on prend à nouveau comme solution approchée, l'intersection de $t \mapsto s_{k,i} + t v_{k,i}$ avec la frontière de la région de confiance.

Il faut à présent montrer que cet algorithme a bien un sens, qu'il est non ambigu, et qu'il permet d'assurer la convergence de la méthode à régions de confiance à laquelle il contribue. Deux propriétés de l'algorithme du GC jouent un rôle-clé dans le bien-fondé de cet algorithme.

– *Tant que l'algorithme du GC ne rencontre pas de direction à courbure négative,* $\{\psi_k(s_{k,i})\}_i$ *décroît strictement.* C'est une propriété fondatrice du GC (voir par exemple la proposition 8.5). On en déduit que, dans la situation (i) ci-dessus, la solution approchée calculée s_k vérifie (9.19), ce qui assure la convergence globale. Dans la situation (ii), bien que $v_{k,i}$ soit à courbure négative, on a encore $\nabla\psi_k(s_{k,i})^T v_{k,i-1} = 0$ du fait de l'itération $i-1$, si bien que

$$\nabla\psi_k(s_{k,i})^T v_{k,i} = \nabla\psi_k(s_{k,i})^T\Big(-\nabla\psi_k(s_{k,i}) + \beta_{k,i} v_{k,i-1}\Big) = -\|\nabla\psi_k(s_{k,i})\|^2 < 0.$$

Dès lors, on fait décroître ψ_k en allant chercher s_k sur la frontière de la région de confiance le long de $v_{k,i}$. On a donc la même conclusion que précédemment.

– *Tant que l'algorithme du GC ne rencontre pas de direction à courbure négative,* $\{\|s_{k,i}\|\}_i$ *croît strictement.* Cette propriété a été proposée en exercice au chapitre 8 (l'exercice 8.2–(i)). Le fait de démarrer le GC en 0 est essentiel pour garantir cette propriété. Dès lors, au cours du GC, les itérés $s_{k,i}$ s'éloignent de 0 de façon monotone, ce qui implique que le chemin dessiné par le GC ne peut croiser la frontière de la région de confiance qu'une seule fois. Autrement dit, il n'y a aucune chance de trouver un point faisant décroître ψ_k davantage tout en restant dans la région de confiance, dès qu'un itéré généré par le GC a franchi la frontière de la région de confiance. Ceci motive l'arrêt des itérations du GC dans la situation (i) décrite ci-dessus.

9.4.2 Méthodes de résolution fines

Nous avons dit en introduction de la section 9.4 que les algorithmes à régions de confiance ont la propriété de pouvoir forcer les itérés à s'échapper d'un point stationnaire qui n'est pas un minimum. On peut comprendre l'origine de cette propriété en examinant le système d'optimalité (9.11). Supposons que M soit la hessienne de f et que l'itéré courant x soit un point stationnaire ($g = g(x) = 0$) sans être un minimum (M n'est pas semi-définie positive; elle a au moins une valeur propre strictement négative). Alors, d'après (9.11e), $\hat{\lambda}$ ne peut être nul et (9.11d) implique alors que $\hat{s}$ ne peut être la solution nulle de (9.11a), mais doit être un vecteur propre associé à la plus petite valeur propre de M. Dès lors $\hat{s} \neq 0$ et on s'échappe du point stationnaire x. Cette propriété est remarquable et le raisonnement qui précède montre que, pour l'obtenir, M doit être assez proche de $\nabla^2 f(x)$ et le sous-problème quadratique doit être résolu avec suffisamment de précision (on a utilisé les conditions d'optimalité (9.11) qui sont nécessaires et suffisantes).

Soient β_1 et β_2 deux constantes vérifiant $0 < \beta_1 \leqslant 1$ et $0 < \beta_2$. On dit que (RC_k) est résolu en s_k avec une *condition de décroissance forte* si l'on a

$$\begin{cases} \psi_k(s_k) \leqslant \beta_1 \widehat{\psi}_k \\ \|s_k\| \leqslant \beta_2 \Delta_k, \end{cases} \tag{9.20}$$

où $\widehat{\psi}_k = \min\{\psi_k(s) : \|s\| \leqslant \Delta_k\}$. La condition (9.20) exprime qu'il faut résoudre (RC_k) exactement à un facteur constant β_1 près sur la valeur de ψ_k, avec une certaine tolérance sur l'appartenance de s_k à la région de confiance (en général $\beta_2 \geqslant 1$). La condition (9.20) est clairement plus forte que la condition de décroissance suffisante (9.19).

Algorithme de Moré-Sorensen

L'algorithme de Moré-Sorensen est adapté aux problèmes dont la taille n'est pas trop grande, car il s'autorise des factorisations de Cholesky de matrices d'ordre n. Il permet de résoudre le sous-problème quadratique avec toute la précision souhaitée, même dans le cas difficile. L'algorithme est piloté par la recherche du multiplicateur optimal $\hat{\lambda}$ associé à la région de confiance.

Soit λ_1 la plus petite valeur propre de M (on laisse tomber l'indice k de l'itération courante). La dernière condition d'optimalité de (9.11) montre que $\hat{\lambda} \geqslant -\lambda_1$. On peut donc espérer trouver le multiplicateur optimal en s'en approchant avec des valeurs de $\lambda > -\lambda_1$. Dans ce cas, la matrice $(M + \lambda I)$ est définie positive et on peut résoudre le système linéaire correspondant à la première condition de (9.11), à savoir

$$(M + \lambda I)s_\lambda = -g.$$

Dans l'algorithme de Moré-Sorensen, on recherche $\hat{\lambda}$ en l'approchant par des $\lambda > -\lambda_1$, que l'on détermine de manière à ce que l'on ait

$$\|s_\lambda\| = \Delta. \tag{9.21}$$

Par cette approche, on peut prendre en compte un grand nombre de cas, ceux pour lesquels le chemin $(\lambda \searrow -\lambda_1) \mapsto s_\lambda$ rencontre la frontière de la région de confiance. Si l'on s'en réfère à la section 9.3.3, sont donc exclus de cette approche : le cas A-1 (M est définie positive et le point de Newton est intérieur à la région de confiance; ce cas ne pose pas de difficultés et sera traité d'emblée par l'algorithme) et le cas B-3 (c'est le cas difficile qui demande la résolution d'un problème de valeur propre).

D'après (9.14), $\|s_\lambda\| \sim (\lambda_1 + \lambda)^{-1}$, pour λ proche de $-\lambda_1$, si bien que dans cette région l'équation (9.21) a l'inconvénient d'être fortement non linéaire. Dans l'algorithme de Moré-Sorensen, (9.21) est résolue par des itérations de Newton sur l'équation équivalente

$$\frac{1}{\|s_\lambda\|} = \frac{1}{\Delta}, \tag{9.22}$$

qui a l'avantage de se comporter linéairement en λ dans le voisinage de $-\lambda_1$. Si on définit l'application φ par $\varphi(\lambda) := \|s_\lambda\|^{-1} - \Delta^{-1}$, l'itération de Newton met évidemment à jour le multiplicateur par la formule

$$\lambda^+ = \lambda - \frac{\varphi(\lambda)}{\varphi'(\lambda)}.$$

On vérifie aisément que $\varphi'(\lambda) = -s_\lambda^{\mathsf{T}} \dot{s}_\lambda / \|s_\lambda\|^3$, où la dérivée $\dot{s}_\lambda$ de $\lambda \mapsto s_\lambda$ est donnée par

$$(M + \lambda I)\dot{s}_\lambda = -s_\lambda.$$

Il y a donc deux systèmes linéaires utilisant la même matrice $(M + \lambda I)$ à résoudre à chaque essai de multiplicateur, ce qui justifie la factorisation de cette matrice.

L'algorithme de Moré-Sorensen ci-dessous est dit être dans sa version simplifiée, car le cas difficile n'y est pas traité.

Algorithme 9.9 (de Moré-Sorensen — version simplifiée)

0. On se donne un tolérance $\varepsilon > 0$.
1. Si M est définie positive (factorisable par Cholesky) et $s^{\mathrm{N}} := -M^{-1}g$ vérifie $\|s^{\mathrm{N}}\| \leqslant \Delta$, alors la solution primale-duale est $(\hat{s}, \hat{\lambda}) = (s^{\mathrm{N}}, 0)$. On arrête.
2. Choisir $\lambda_0 > 0$. On pose $\underline{\lambda} = 0$ et $\overline{\lambda} = +\infty$.
3. Pour $i = 0, 1, 2, \ldots$

   3.1. *Factorisation de Cholesky*:

   $$M + \lambda_i I = R_i^{\mathsf{T}} R_i.$$

   Si cette factorisation ne peut pas se faire, poser $\underline{\lambda} = \lambda_i$, choisir $\lambda_{i+1} \in]\underline{\lambda}, \overline{\lambda}[$ et passer à l'itération suivante.

   3.2 Résoudre en s_i et q_i :

   $$R_i^{\mathsf{T}} R_i s_i = -g \quad \text{et} \quad R_i^{\mathsf{T}} q_i = s_i.$$

   3.3 *Test d'arrêt* : Si $|\,\|s_i\| - \Delta\,| \leqslant \varepsilon$, on prend comme solution primale-duale (s_i, λ_i). On s'arrête.

   3.4 Si $\|s_i\| > \Delta$, alors $\underline{\lambda} := \lambda_i$, sinon $\overline{\lambda} := \lambda_i$.

   3.5 *Nouveau multiplicateur* :

   $$\lambda_{i+1} = \lambda_i + \left(\frac{\|s_i\|}{\|q_i\|}\right)^2 \left(\frac{\|s_i\| - \Delta}{\Delta}\right).$$

   Si $\lambda_{i+1} \notin]\underline{\lambda}, \overline{\lambda}[$, choisir une nouvelle valeur de λ_{i+1} dans cet intervalle.

Aux étapes 3.1 et 3.5 de cet algorithme, où il faut choisir un nouveau multiplicateur dans l'intervalle $]\underline{\lambda}, \overline{\lambda}[$, on veillera à ne pas générer des points d'accumulation artificiellement. Il suffira, par exemple, de prendre $(\underline{\lambda}+\overline{\lambda})/2$ lorsque $\overline{\lambda}$ est fini et un élément de $[\underline{\lambda} + 1, \overline{\lambda}[$ dans le cas contraire.

Il a été vérifié expérimentalement que l'algorithme 9.9 converge en très peu d'itérations : 2 ou 3 selon les essais des auteurs, pour autant que l'on ne soit pas trop exigeant ($\varepsilon > 0$ pas trop petit) et que l'on ne soit pas proche du cas difficile.

Dans le cas difficile (qui se rencontre rarement), l'algorithme 9.9 génère une suite de $\lambda_i \searrow -\lambda_1 \geqslant 0$, sans jamais réussir à satisfaire $\|s_i\| \simeq \Delta$. Dans l'algorithme de Moré-Sorensen complet, chaque itération se prépare à cette éventualité en générant des vecteurs destinés à approcher un vecteur propre de M associé à la valeur propre λ_1. L'algorithme complet est détaillé dans l'article [440].

9.5 Résultats de convergence

Dans cette section sont donnés des résultats de convergence des méthodes à régions de confiance. On considère dans un premier temps (section 9.5.1), le cas où le modèle est du premier ordre. On veut dire par là que le résultat de convergence ne s'intéresse pas à la matrice M_k (la seule chose demandée est que la suite $\{M_k\}$ soit bornée), mais plutôt au point de Cauchy. C'est la condition de descente suffisante (9.19) qui joue le rôle-clé. Ensuite (section 9.5.2), on considère les modèles du second ordre, dans lesquels $M_k = \nabla^2 f(x_k)$ et la condition de décroissante forte (9.20) est assurée par l'algorithme de résolution du sous-problème quadratique.

9.5.1 Convergence avec modèle du premier ordre

On suppose dans cette section que le sous-problème quadratique (RC_k) est résolu de façon approchée, de manière toutefois à satisfaire la condition de décroissance suffisante (9.19), que nous rappelons ci-dessous

$$\begin{cases} \psi_k(s_k) \leqslant \beta_1 \psi_k(s_k^{\mathrm{C}}) \\ \|s_k\| \leqslant \beta_2 \Delta_k. \end{cases} \tag{9.23}$$

Les constantes β_1 et β_2 sont supposées être strictement positives. On note

$$\mathcal{N}_1 = \{x \in \mathbb{R}^n : \ f(x) \leqslant f(x_1)\},$$

l'ensemble de sous-niveau $f(x_1)$ de f, fixé par le premier itéré.

La proposition suivante montre que la condition de décroissance suffisante (9.23) implique l'inégalité (9.24), connue sous le nom de *condition de Powell*. Celle-ci joue, pour les méthodes à régions de confiance, le même rôle que la condition de Zoutendijk pour les méthodes à directions de descente : elle résume la contribution de la méthode de détermination du pas à la convergence de l'algorithme.

Proposition 9.10 *Le point de Cauchy s_k^{C} vérifie la condition de Powell suivante avec $s_k = s_k^{\mathrm{C}}$ et $C = 1/2$:*

$$\psi_k(s_k) \leqslant -C\|g_k\| \min\left(\Delta_k, \frac{\|g_k\|}{\|M_k\|}\right). \tag{9.24}$$

On en déduit immédiatement qu'un point s_k vérifiant la condition de décroissance suffisante (9.23) vérifie également la condition (9.24), avec $C = \beta_1/2$.

Démonstration. Le cas où $g_k = 0$ est clair car alors $s_k^{\mathrm{C}} = 0$ et $\psi_k(s_k^{\mathrm{C}}) = 0$. Supposons à présent que $g_k \neq 0$. Si $g_k^{\mathsf{T}} M_k g_k \leqslant 0$, $s_k^{\mathrm{C}} = -(\Delta_k/\|g_k\|)g_k$ d'après la formule (9.18). Donc

$$\psi_k(s_k^{\mathrm{C}}) = -\Delta_k\|g_k\| + \frac{\Delta_k^2}{2\|g_k\|^2} g_k^{\mathsf{T}} M_k g_k \leqslant -\Delta_k\|g_k\|$$

et (9.24) est vérifiée avec $C = 1$.

Il reste à examiner le cas où $g_k \neq 0$ et $g_k^T M_k g_k > 0$. Si $\Delta_k \leqslant \|g_k\|^3 / g_k^T M_k g_k$, $s_k^C = -(\Delta_k / \|g_k\|) g_k$ et on a

$$\psi_k(s_k^C) = -\Delta_k \|g_k\| + \frac{\Delta_k^2}{2\|g_k\|^2} g_k^T M_k g_k \leqslant -\frac{\Delta_k}{2}\|g_k\|.$$

Donc (9.24) est vérifiée avec $C = 1/2$. Sinon

$$s_k^C = -\frac{\|g_k\|^2}{g_k^T M_k g_k} g_k$$

et

$$\psi_k(s_k^C) = -\frac{\|g_k\|^4}{2(g_k^T M_k g_k)} \leqslant -\frac{\|g_k\|^2}{2\|M_k\|}.$$

Donc (9.24) est vérifiée avec $C = 1/2$. □

La proposition suivante montre que si $g_k \neq 0$, si le rayon de confiance Δ_k est assez petit, et si la solution approchée s_k de (RC_k) vérifie la condition de décroissance suffisante (9.23) alors la décroissante de f, $f(x_k) - f(x_k + s_k)$, est suffisante pour avoir (9.2), c'est-à-dire $\rho_k \geqslant \omega_1$. Dans ces conditions, il n'y a pas de bouclage à l'étape 2 de l'algorithme.

Proposition 9.11 *Supposons que f soit dérivable en x_k et que $g_k \neq 0$. Il existe $\Delta_k^0 > 0$ tel que si $0 < \Delta_k \leqslant \Delta_k^0$ et si la solution approchée de (RC_k) vérifie la condition de décroissance suffisante (9.23), alors le ratio $\rho_k > \omega_1$.*

Démonstration. Raisonnons par l'absurde. Si ce n'est pas le cas, il existe une suite de rayons de confiance $\Delta_{k,i} \to 0$ pour $i \to \infty$ et des solutions approchées $s_{k,i}$ vérifiant

$$f(x_k + s_{k,i}) \geqslant f(x_k) + \omega_1 \psi_k(s_{k,i}), \tag{9.25}$$

$$\psi_k(s_{k,i}) \leqslant \beta_1 \psi_k(s_{k,i}^C), \tag{9.26}$$

$$\|s_{k,i}\| \leqslant \beta_2 \Delta_{k,i}, \tag{9.27}$$

où $s_{k,i}^C$ est le point de Cauchy correspondant au rayon de confiance $\Delta_{k,i}$. Comme f est différentiable en x_k, on a en développant $f(x_k + s_{k,i})$ autour de x_k :

$$\begin{aligned} f(x_k + s_{k,i}) &= f(x_k) + g_k^T s_{k,i} + o(\|s_{k,i}\|) \\ &= f(x_k) + \psi_k(s_{k,i}) + o(\|s_{k,i}\|). \end{aligned}$$

Alors (9.25) et $0 < \omega_1 < 1$ donnent

$$\psi_k(s_{k,i}) = o(\|s_{k,i}\|).$$

D'autre part, comme $g_k \neq 0$, on peut supposer que $\Delta_{k,i} \leqslant \|g_k\| / \|M_k\|$. Alors (9.26), la proposition 9.10 et (9.27) montrent que

$$|\psi_k(s_{k,i})| = -\psi_k(s_{k,i}) \geqslant \frac{\beta_1}{2}\|g_k\|\Delta_{k,i} \geqslant \frac{\beta_1}{2\beta_2}\|g_k\|\|s_{k,i}\|.$$

Ceci est en contradiction avec l'estimation précédente ($g_k \neq 0$) et cette contradiction prouve la proposition. □

Comme on le verra à la section 9.5, la condition de Powell (9.24) et donc la condition de décroissance suffisante (9.23) permettent d'avoir la convergence de l'algorithme. Ce choix est très semblable à la méthode de la plus forte pente avec recherche linéaire, qui converge également pour des recherches linéaires raisonnables. Seule la manière de déterminer le pas le long de $-g_k$ change. Comme la méthode de la plus forte pente n'est pas une méthode convergeant rapidement, on comprend que si l'on veut une bonne vitesse de convergence, il faut résoudre (RC_k) de manière plus précise. Nous verrons qu'une résolution plus fine de (RC_k) muni d'un modèle du second ordre permet de converger vers des points vérifiant les conditions nécessaires d'optimalité du second ordre, qui ont de fortes chances d'être des minima locaux.

Le théorème 9.12 donne des conditions pour qu'une sous-suite des gradients tende vers 0. Dans la situation (i) de ce théorème, l'algorithme boucle en un point où le gradient est nul. Il ne peut pas faire mieux avec un modèle du premier ordre.

Le théorème 9.13 montre qu'avec un peu plus de régularité sur f (l'uniforme continuité de ∇f), on a convergence de toute la suite des gradients.

Théorème 9.12 (convergence globale avec résolution approchée des sous-problèmes quadratiques) *Supposons que f soit bornée inférieurement et de classe C^1 dans un voisinage de $\mathcal{N}_1$. Si dans la méthode à régions de confiance, $\{M_k\}$ est bornée et la solution approchée s_k du sous-problème quadratique (RC_k) vérifie les conditions de décroissance suffisante (9.23), alors l'une des situations suivantes a lieu :*
(i) il y a échec d'une itération k_0, en un point x_{k_0} où $\nabla f(x_{k_0}) = 0$;
(ii) $\liminf \|\nabla f(x_k)\| = 0$.

Démonstration. On raisonne par l'absurde, en supposant que la situation (i) n'a pas lieu (donc une suite $\{x_k\}$ est générée) et qu'il existe un indice k_1 et une constante $\gamma > 0$ tels que

$$\|g_k\| \geqslant \gamma, \quad \forall k \geqslant k_1. \tag{9.28}$$

Montrons que

$$\Delta_k \to 0 \quad \text{et} \quad \sum_{k \geqslant 1} \|s_k\| < +\infty. \tag{9.29}$$

D'après la condition de Powell (9.24), la supposition de départ (9.28) et la bornitude de $\{M_k\}$, on a

$$\psi_k(s_k) \leqslant -C\|g_k\| \min\left(\Delta_k, \frac{\|g_k\|}{\|M_k\|}\right) \leqslant -C \min\left(\Delta_k, 1\right), \tag{9.30}$$

où C est une constante absorbante strictement positive indépendante de k. En utilisant (9.2), on obtient alors

$$f(x_{k+1}) \leqslant f(x_k) - C \min\left(\Delta_k, 1\right),$$

$$C \min(\Delta_k, 1) \leqslant f(x_k) - f(x_{k+1}).$$

Comme f est bornée inférieurement, la suite $\{f(x_k) - f(x_{k+1})\}$ est sommable et on déduit de l'inégalité ci-dessus que $\sum_{k\geqslant 1} \Delta_k < +\infty$. Comme $\|s_k\| \leqslant \beta_1 \Delta_k$, les assertions de (9.29) s'en déduisent.

Montrons que le ratio $\rho_k \to 1$. D'après (9.29), $\{x_k\}$ est une suite de Cauchy, donc elle converge et on a

$$\frac{|f(x_k + s_k) - f(x_k) - g_k^T s_k|}{\|s_k\|} \to 0. \tag{9.31}$$

D'autre part, comme $\{M_k\}$ est bornée, on a

$$|\psi_k(s_k) - g_k^T s_k| \leqslant C\|s_k\|^2. \tag{9.32}$$

Enfin, on peut écrire successivement

$$\rho_k = \frac{f(x_{k+1}) - f(x_k)}{\psi_k(s_k)} = \frac{f(x_{k+1}) - f(x_k) - g_k^T s_k}{\psi_k(s_k)} + \frac{g_k^T s_k - \psi_k(s_k)}{\psi_k(s_k)} + 1,$$

$$|\rho_k - 1| \leqslant \frac{|f(x_{k+1}) - f(x_k) - g_k^T s_k|}{|\psi_k(s_k)|} + \frac{|\psi_k(s_k) - g_k^T s_k|}{|\psi_k(s_k)|}.$$

D'après (9.30), $\Delta_k \to 0$ et $\|s_k\| \leqslant \beta_2 \Delta_k$, on a $|\psi_k(s_k)| \geqslant C\|s_k\|$. Alors, la dernière inégalité, (9.31) et (9.32) montrent que $\rho_k \to 1$.

Dès lors $\rho_k > \omega_2$ pour k assez grand. D'après la règle de mise à jour de Δ_k (étape 4 de l'algorithme), cela implique $\Delta_k \geqslant \Delta > 0$, qui est en contradiction avec la première assertion de (9.29). Cette contradiction prouve le théorème. □

La conclusion du théorème 9.12 reste vraie avec une hypothèse beaucoup plus faible sur la suite $\{M_k\}$. Il suffit, en effet, de supposer que $\{\|M_k\|\}$ ne croît pas trop vite dans le sens où

$$\sum_{k\geqslant 1} \frac{1}{1 + \max\limits_{1\leqslant i\leqslant k} \|M_i\|} = +\infty.$$

Par exemple, cette hypothèse est vérifiée si la croissance est au plus linéaire : $\|M_k\| \leqslant Ck$, pour une constante $C > 0$. On trouvera une démonstration de ce résultat chez Powell (1984). Il est important car il permet d'étudier la convergence des méthodes à régions de confiance lorsque la matrice M_k est mise à jour par une méthode de quasi-Newton.

Théorème 9.13 *Supposons, en plus des hypothèses et conditions du théorème 9.12, que ∇f soit uniformément continue sur $\mathcal{N}_1$. Alors soit la situation (i) du théorème 9.12 a lieu, soit $g_k \to 0$.*

Démonstration. On raisonne par l'absurde en supposant que la situation (i) du théorème 9.12 n'a pas lieu (donc une suite $\{x_k\}$ est générée) et qu'il existe une constante $\gamma > 0$ et une sous-suite $\{k_j\}_{j\geqslant 1} \subseteq \mathbb{N}$ telles que

$$\|g_{k_j}\| \geqslant \gamma, \quad \forall j \geqslant 1. \tag{9.33}$$

Comme $\liminf \|g_k\| = 0$, en extrayant une sous-suite au besoin, on peut supposer que pour tout j on peut trouver ℓ_j tel que

$$\begin{cases} k_j < \ell_j < k_{j+1}, \\ \|g_k\| \geqslant \gamma/2, \quad \text{pour } k = k_j, \ldots, \ell_{j-1}, \\ \|g_{\ell_j}\| < \gamma/2. \end{cases} \tag{9.34}$$

En utilisant la condition de Powell (9.24) et la bornitude de $\{M_k\}$, on a comme dans la démonstration du théorème 9.12 :

$$f(x_{k+1}) - f(x_k) \leqslant -C\|g_k\| \min(\Delta_k, \|g_k\|).$$

Donc pour $k = k_j, \ldots, \ell_{j-1}$ et $j \geqslant 1$, on a

$$f(x_{k+1}) - f(x_k) \leqslant -C \min(\Delta_k, 1).$$

Comme $f(x_{k+1}) - f(x_k) \to 0$, $\min(\Delta_k, 1) = \Delta_k$ pour k assez grand. En utilisant le fait que $\|s_k\| \leqslant \beta_2 \Delta_k$, on a finalement

$$\|s_k\| \leqslant C(f(x_k) - f(x_{k+1})), \quad k = k_j, \ldots, \ell_{j-1},$$

$$\|x_{\ell_j} - x_{k_j}\| \leqslant \sum_{k=k_j}^{\ell_j - 1} \|s_k\| \leqslant C(f(x_{k_j}) - f(x_{\ell_j})).$$

Donc $\|x_{\ell_j} - x_{k_j}\| \to 0$, quand $j \to \infty$. Alors par uniforme continuité de ∇f, on a :

$$\|g_{\ell_j} - g_{k_j}\| \to 0, \quad \text{quand } j \to \infty.$$

Ceci est en contradiction avec le fait que, par (9.33) et (9.34)

$$\|g_{\ell_j} - g_{k_j}\| \geqslant \|g_{k_j}\| - \|g_{\ell_j}\| \geqslant \gamma/2.$$

Cette contradiction prouve le théorème. □

En ce qui concerne les conditions de convergence, on notera la différence suivante entre les méthodes à directions de descente et les méthodes à régions de confiance. Pour obtenir la convergence des premières, on est toujours amené à montrer que les suites $\{M_k\}$ et $\{M_k^{-1}\}$ sont bornées ou n'explosent pas trop vite. Pour les méthodes à régions de confiance, aucune hypothèse sur $\{M_k^{-1}\}$ n'est nécessaire, seul le comportement de $\{M_k\}$ importe. En particulier, M_k peut être singulière ou nulle. D'ailleurs, dans ce dernier cas, on est ramené à la méthode de la plus forte pente qui a de bonnes propriétés de convergence globale (bien qu'une mauvaise *vitesse* de convergence). Cette propriété contribue à la robustesse des méthodes à régions de confiance.

9.5.2 Convergence avec modèle du second ordre

Dans cette section, nous étudions la convergence et les propriétés des points d'adhérence des suites générées par les méthodes à régions de confiance lorsque le

sous-problème quadratique (RC_k) est résolu de façon fine, de manière à satisfaire la condition de décroissance forte (9.20), et lorsque le modèle ψ_k est du second ordre, c'est-à-dire lorsqu'on prend

$$M_k = \nabla^2 f(x_k).$$

On suppose donc que la fonction f est deux fois continûment dérivable et qu'on en calcule la hessienne. Si ceci est coûteux en temps de calcul, cela peut en valoir la peine. On montre en effet (voir Théorème 9.16 pour les hypothèses précises) que la suite $\{x_k\}$ a au moins un point d'adhérence où la hessienne est semi-définie positive et qu'en tout point d'adhérence isolé, la hessienne est semi-définie positive. On montre aussi (Théorème 9.17) que si $\{x_k\}$ a un point d'adhérence où la hessienne est définie positive alors toute la suite converge vers ce point.

Ces bons résultats sont à comparer avec ceux des méthodes à directions de descente qui, sans utilisation de directions à courbure négative, ne peuvent trouver que des points stationnaires.

La condition de décroissance forte (9.20) est clairement plus forte que la condition de décroissance suffisante (9.19). En particulier, elle implique également la condition de Powell (9.24) et le résultat de la proposition 9.11. Le résultat suivant est plus précis. Il montre qu'il n'y a pas de bouclage dans l'algorithme à régions de confiance avec modèle du second ordre ($M_k = \nabla^2 f(x_k)$) si x_k ne vérifie pas les conditions nécessaires d'optimalité du second ordre du problème (9.1). Si x_k vérifie ces conditions, l'itération k est interrompue, car on ne peut pas réaliser $\psi_k(s_k) < 0$.

Proposition 9.14 *Supposons que f soit de classe C^1 dans un voisinage de x_k et deux fois dérivable en x_k et que soit $g_k \neq 0$, soit $\lambda_1(\nabla^2 f(x_k)) < 0$. Supposons également que dans dans (RC_k), on prenne $M_k = \nabla^2 f(x_k)$. Dans ce cas, il existe $\Delta_k^0 > 0$ tel que si $0 < \Delta_k \leqslant \Delta_k^0$ et si la solution approchée de (RC_k) vérifie la condition de décroissance forte (9.20), alors le ratio $\rho_k > \omega_1$.*

Démonstration. Le cas où $g_k \neq 0$ est réglé par la proposition 9.11, puisque la condition de décroissance forte implique la condition de décroissance suffisante (9.19). Supposons à présent que $g_k = 0$ et donc, d'après les hypothèses, $\lambda_1(\nabla^2 f(x_k))$ est strictement négatif.

Comme dans la démonstration de la proposition 9.11, on raisonne par l'absurde, en supposant qu'il existe une suite de rayons de confiance $\Delta_{k,i} \to 0$ pour $i \to \infty$ et des solutions approchées $s_{k,i}$ vérifiant

$$f(x_k + s_{k,i}) \geqslant f(x_k) + \omega_1 \psi_k(s_{k,i}), \tag{9.35}$$

$$\begin{aligned} \psi_k(s_{k,i}) &\leqslant \beta_1 \widehat{\psi}_{k,i}, \\ \|s_{k,i}\| &\leqslant \beta_2 \Delta_{k,i}, \end{aligned} \tag{9.36}$$

où $\widehat{\psi}_{k,i} = \min\{\psi_k(s) : \|s\| \leqslant \Delta_{k,i}\}$. Comme $g_k = 0$, on a en développant $f(x_k + s_{k,i})$ autour de x_k :

$$\begin{aligned} f(x_k + s_{k,i}) &= f(x_k) + \frac{1}{2} s_{k,i}^T M_k s_{k,i} + o(\|s_{k,i}\|^2) \\ &= f(x_k) + \psi_k(s_{k,i}) + o(\|s_{k,i}\|^2). \end{aligned}$$

Alors (9.35), $\psi_k(s_{k,i}) < 0$ et $\omega_1 < 1$ donnent

$$\psi_k(s_{k,i}) = o(\|s_{k,i}\|^2). \tag{9.37}$$

Notons $(\hat{s}_{k,i}, \hat{\lambda}_{k,i})$ un couple optimal de $(RC_{k,i})$. En utilisant le fait que $g_k = 0$ et les conditions d'optimalité (9.11a), (9.11d) et (9.11e) de $(RC_{k,i})$, on trouve

$$\widehat{\psi}_{k,i} = \frac{1}{2}\hat{s}_{k,i}^{\mathsf{T}} M_k \hat{s}_{k,i} = -\frac{1}{2}\hat{\lambda}_{k,i}\|\hat{s}_{k,i}\|^2 = -\frac{1}{2}\hat{\lambda}_{k,i}\Delta_{k,i}^2 \leqslant \frac{1}{2}\lambda_1(M_k)\Delta_{k,i}^2,$$

et en utilisant (9.36)

$$|\psi_k(s_{k,i})| \geqslant \frac{\beta_1}{2}\|\lambda_1(M_k)\|\Delta_{k,i}^2 \geqslant C\|s_{k,i}\|^2,$$

où $C = \frac{\beta_1}{2\beta_2^2}\|\lambda_1(M_k)\| > 0$ est une constante indépendante de i. Cette inégalité contredit (9.37), ce qui termine la démonstration. □

Le lemme technique suivant concerne des suites générales de $\mathbb{R}^n$.

Lemme 9.15 *Soit $\bar{x}$ un point d'adhérence isolé d'une suite $\{x_k\}$ de $\mathbb{R}^n$. Si $\{x_k\}$ ne converge pas vers $\bar{x}$, il existe $\varepsilon > 0$ et une sous-suite $\{x_{k_i}\}_{i\geqslant 1}$ qui converge vers $\bar{x}$ et vérifie*

$$\|x_{k_i+1} - x_{k_i}\| \geqslant \varepsilon, \quad \forall i \geqslant 1.$$

Démonstration. Si $\bar{x}$ est isolé, il existe $\varepsilon > 0$ tel que $B(\bar{x}, 3\varepsilon)$ ne contienne pas d'autre point d'adhérence que $\bar{x}$. Comme $\bar{x}$ est un point d'adhérence de $\{x_k\}$, on peut trouver une sous-suite $\{x_{k_i}\}_{i\geqslant 1}$ contenue dans $B(\bar{x}, 2\varepsilon)$ telle que $x_{k_i+1} \notin B(\bar{x}, 2\varepsilon)$ (sinon toute la suite convergerait vers $\bar{x}$). Comme $x_{k_i} \to \bar{x}$ (car $B(\bar{x}, 3\varepsilon)$ n'a pas d'autre point d'adhérence que $\bar{x}$), le lemme est démontré en ne conservant que les itérés $x_{k_i} \in B(\bar{x}, \varepsilon)$. □

Théorème 9.16 (convergence globale avec résolution fine des sous-problèmes quadratiques) *Supposons que f soit bornée inférieurement, de classe C^2 dans un voisinage de $\mathcal{N}_1$ et que sa hessienne $\nabla^2 f$ soit bornée sur $\mathcal{N}_1$. On suppose que l'on prend $M_k = \nabla^2 f(x_k)$ dans la méthode à régions de confiance et que la solution approchée s_k de (RC_k) vérifie la condition de décroissance forte (9.20). Alors, soit il y a échec d'une itération k_0 en un point x_{k_0} vérifiant les conditions nécessaires d'optimalité du second ordre ($\nabla f(x_{k_0}) = 0$ et $\nabla^2 f(x_{k_0})$ est semi-définie positive), soit une suite $\{x_k\}$ est générée et*

(i) $g_k \to 0$;

(*ii*) *si* $\{x_k\}$ *est bornée, alors* $\limsup \lambda_1(\nabla^2 f(x_k)) \geqslant 0$, *ce qui revient à dire que* $\{x_k\}$ *a au moins un point d'adhérence* $\bar{x}$ *tel que* $\nabla^2 f(\bar{x})$ *soit semi-définie positive;*

(*iii*) *si* $\bar{x}$ *est un point d'adhérence isolé de* $\{x_k\}$, *alors* $\nabla^2 f(\bar{x})$ *est semi-définie positive.*

Démonstration. Si l'itération k_0 échoue, ce ne peut être qu'en un point vérifiant les conditions nécessaires d'optimalité du second ordre (Proposition 9.14). Dans le cas contraire, une suite $\{x_k\}$ est générée, ce que nous supposons à présent.

Comme $\nabla^2 f$ est bornée sur $\mathcal{N}_1$, ∇f est uniformément continue et donc le théorème 9.13 implique que $g_k \to 0$. Ceci montre (i).

Pour (ii), on raisonne par l'absurde en supposant qu'il existe $\varepsilon > 0$ et un indice k_1 tels que

$$\lambda_k^1 := \lambda_1(\nabla^2 f(x_k)) \leqslant -\varepsilon, \quad \forall k \geqslant k_1.$$

Soit v_k un vecteur propre correspondant à λ_k^1, tel que $g_k^T v_k \leqslant 0$ et $\|v_k\| = \Delta_k$. Alors

$$\begin{aligned} \psi_k(s_k) &\leqslant \beta_1 \widehat{\psi}_k \\ &\leqslant \beta_1 \psi_k(v_k) \\ &= \beta_1 \left(g_k^T v_k + \frac{1}{2} \lambda_k^1 \|v_k\|^2 \right) \\ &\leqslant -\frac{\varepsilon \beta_1}{2} \Delta_k^2. \end{aligned} \tag{9.38}$$

D'après (9.2), $\psi_k(s_k) \to 0$. Donc, l'inégalité ci-dessus montre que $\Delta_k \to 0$.

On obtient une contradiction (grâce à l'étape 4 de l'algorithme) en montrant que $\rho_k \to 1$. On a en effet avec $M_k = \nabla^2 f(x_k)$

$$|f(x_{k+1}) - f(x_k) - \psi_k(s_k)| \leqslant \frac{\|s_k\|^2}{2} \max_{0 \leqslant t \leqslant 1} \|\nabla^2 f(x_k + t s_k) - \nabla^2 f(x_k)\| = o(\|s_k\|^2),$$

car f est de classe C^2, $\{x_k\}$ est bornée et $s_k \to 0$. Alors, avec (9.38), on a :

$$|\rho_k - 1| = \frac{|f(x_{k+1}) - f(x_k) - \psi_k(s_k)|}{|\psi_k(s_k)|} \to 0.$$

La contradiction obtenue ($\Delta_k \to 0$ et Δ_k uniformément positif) prouve le point (ii).

Soit $\bar{x}$ un point d'adhérence isolé de $\{x_k\}$. Si $x_k \to \bar{x}$, alors (iii) se déduit de (ii). Supposons à présent que $x_k \not\to \bar{x}$. Alors $\bar{x}$ étant isolé, le lemme 9.15 montre qu'il existe $\varepsilon > 0$ et une sous-suite $x_{k_i} \to \bar{x}$ telle que $\|s_{k_i}\| \geqslant \varepsilon$, $\forall i$. Donc

$$\beta_2 \Delta_{k_i} \geqslant \|s_{k_i}\| \geqslant \varepsilon > 0. \tag{9.39}$$

Soit $\hat{\lambda}_k$ le multiplicateur associé à la contrainte de (RC_k) et $\hat{s}_k$ une solution de (RC_k). D'après les conditions d'optimalité de $(\hat{s}_k, \hat{\lambda}_k)$ (théorème 9.2), on a :

$$\begin{aligned}
\psi_k(s_k) &\leqslant \beta_1 \widehat{\psi}_k \\
&= \beta_1 \left(g_k^T \hat{s}_k + \frac{1}{2} \hat{s}_k^T \nabla^2 f(x_k) \hat{s}_k \right) \\
&= \beta_1 \left(-\hat{\lambda}_k \|\hat{s}_k\|^2 - \frac{1}{2} \hat{s}_k^T \nabla^2 f(x_k) \hat{s}_k \right), \qquad \text{[par (9.11a)]} \\
&\leqslant -\frac{\beta_1}{2} \hat{\lambda}_k \|\hat{s}_k\|^2, \qquad \text{[par (9.11e)]} \\
&= -\frac{\beta_1}{2} \hat{\lambda}_k \Delta_k^2, \qquad \text{[par (9.11d)]}.
\end{aligned}$$

En utilisant cette inégalité, (9.39) et (9.11e) :

$$\psi_{k_i}(s_{k_i}) \leqslant \frac{\varepsilon^2 \beta_1}{2\beta_2^2} \lambda_{k_i}^1.$$

Comme $\psi_{k_i}(s_{k_i}) \to 0$, en passant à la limite dans cette inégalité, on voit que $\nabla^2 f(\bar{x})$ est semi-définie positive. □

Dans le théorème suivant, on suppose seulement que le problème (RC_k) est résolu avec condition de décroissance suffisante (9.19).

Théorème 9.17 *On suppose que f vérifie les mêmes propriétés que dans le théorème 9.16, que $M_k = \nabla^2 f(x_k)$ et que la solution approchée s_k de (RC_k) vérifie la condition de décroissance suffisante (9.19). Alors, soit il y a échec d'une itération k_0 en un point x_{k_0} tel que $\nabla f(x_{k_0}) = 0$, soit une suite $\{x_k\}$ est générée et*

(i) $g_k \to 0$,

(ii) si $\bar{x}$ est un point d'adhérence de $\{x_k\}$ tel que $\nabla^2 f(\bar{x})$ soit définie positive, alors $x_k \to \bar{x}$ et Δ_k est uniformément > 0.

Démonstration. Si une itération k_0 échoue, alors la proposition 9.11 montre que $\nabla f(x_{k_0}) = 0$. Supposons à présent que toutes les itérations réussissent. Alors, une suite $\{x_k\}$ est générée et, d'après le théorème 9.13, $g_k \to 0$.

On montre que $x_k \to \bar{x}$ en utilisant le lemme 9.15. Comme $\nabla^2 f(\bar{x})$ est définie positive, $\bar{x}$ est un point d'adhérence isolé, car pour x voisin de $\bar{x}$

$$\nabla f(x)^T (x - \bar{x}) = \int_0^1 f''(\bar{x} + t(x - \bar{x})) \cdot (x - \bar{x})^2 \, \mathrm{d}t \geqslant C \|x - \bar{x}\|^2.$$

Donc x voisin de $\bar{x}$ et $\nabla f(x) = 0$ impliquent que $x = \bar{x}$. D'autre part, pour x_k voisin de $\bar{x}$, on a en utilisant $\psi_k(s_k) < 0$ et l'inégalité de Cauchy-Schwarz :

$$C\|s_k\|^2 \leqslant \frac{1}{2} s_k^T \nabla^2 f(x_k) s_k < -g_k^T s_k \leqslant \|g_k\| \, \|s_k\|. \tag{9.40}$$

Comme $g_k \to 0$, on voit que pour toute sous-suite $\{x_{k_i}\}$ convergeant vers $\bar{x}$, on a $s_{k_i} \to 0$. D'après le lemme 9.15, cela implique que $x_k \to \bar{x}$.

Il reste à montrer que Δ_k est uniformément > 0. De (9.40) et de la convergence de $\{x_k\}$ vers $\bar{x}$, on déduit que $\|g_k\| \geqslant C\|s_k\|$ pour k grand. Comme la condition de

décroissance suffisante implique la condition de Powell (Proposition 9.10), on a en utilisant $\|g_k\| \geqslant C\|s_k\|$ et $\Delta_k \geqslant \|s_k\|/\beta_2$:

$$\psi_k(s_k) \leqslant -C\|s_k\|^2, \quad \text{pour } k \text{ grand.}$$

On peut en déduire que $\rho_k \to 1$, puisque

$$f(x_{k+1}) - f(x_k) - \psi_k(s_k) = o(\|s_k\|^2)$$

et donc

$$|\rho_k - 1| = \frac{|f(x_{k+1}) - f(x_k) - \psi_k(s_k)|}{|\psi_k(s_k)|} \to 0.$$

Alors l'étape 4 de l'algorithme montre que $\Delta_k \geqslant \Delta > 0$. □

9.6 Comparaison avec la recherche linéaire

On trouvera dans le tableau 9.1 une comparaison des différentes caractéristiques des algorithmes à directions de descente, c'est-à-dire ceux utilisant la recherche linéaire (chapitre 6), et des algorithmes à régions de confiance étudiés dans ce chapitre.

Recherche linéaire	**Région de confiance**
On se donne une *direction de descente* d_k de f en x_k	On se donne un *modèle* ψ_k de f en x_k
On adapte le *pas* $\alpha_k > 0$ le long de d_k pour faire décroître f	On adapte le *rayon de confiance* $\Delta_k > 0$ pour faire décroître f
Le déplacement $s_k = \alpha_k d_k$ est aligné sur d_k (recherche *linéaire*)	Le déplacement s_k change d'orientation avec Δ_k (recherche *curviligne*)
Facile à mettre en œuvre	*Difficile* à mettre en œuvre
Résultats de convergence *faibles*	Résultats de convergence *renforcés*

Tableau 9.1. Comparaison de la recherche linéaire et des régions de confiance

Notes

Le concept de région de confiance semble remonter à Powell [487 ; 1970], qui l'introduit pour la résolution de systèmes non linéaires par *moindres-carrés*, après avoir constaté l'échec de la recherche linéaire. Cette approche a des points communs avec l'algorithme de Levenberg-Marquardt [389, 407], mais le pilotage de l'algorithme est fait ici par le réglage du rayon de confiance Δ_k à chaque itération, alors que c'est le multiplicateur associé à la région de confiance qui joue le rôle principal dans l'algorithme de Levenberg-Marquardt. Cette quantité duale est en général plus difficile à déterminer.

Ce chapitre doit beaucoup au compte-rendu éclairant de Moré [438 ; 1983].

Nous avons supposé que $g_k = \nabla f(x_k)$ dans le modèle quadratique. On peut relâcher cette hypothèse et ne supposer que la convergence $(g_k - \nabla f(x_k)) \to 0$ lorsque $\{x_k\}$ converge [438]. Ceci permet de prendre en compte le cas où le gradient est obtenu par différences divisées.

Les conditions nécessaires et suffisantes d'optimalité pour le sous-problème quadratique des régions de confiance ont été données par Gay [234] et Sorensen [563]. Le théorème 9.12 est dû à Powell [489] bien que l'algorithme qu'il analyse diffère sensiblement de celui que nous avons étudié. Le théorème 9.13 est dû à Thomas [582]

La complexité de la résolution du sous-problème (9.8) des méthodes à régions de confiance est considérée dans [632, 570, 511, 224 ; 1992-1997].

D'autres développements peuvent être trouvés chez Byrd, Schnabel et Schultz [549], Martínez [410]. Notons enfin que l'utilisation des régions de confiance pour les problèmes avec contraintes est un sujet toujours en développement et les meilleures approches ne font pas encore l'objet d'un consensus.

Exercices

9.1. *Généralisation de la condition de Powell* (9.18). Soient $g \in \mathbb{R}^n$, M une matrice symétrique non nulle, C une matrice non nulle, $\Delta > 0$ et $d \in \mathbb{R}^n$ vérifiant $g^T d \leqslant 0$ et $Cd \neq 0$. Montrez que la solution $\hat{s}$ du sous-problème quadratique

$$(RC) \quad \left\{ \begin{array}{l} \min\ \psi(s) = g^T s + \frac{1}{2} s^T B s \\ \|Cs\| \leqslant \Delta \end{array} \right.$$

vérifie

$$\psi(\hat{s}) \leqslant -\frac{|g^T d|}{2\|d\|} \min\left(\Delta \frac{\|d\|}{\|Cd\|}, \frac{|g^T d|}{\|M\|\,\|d\|} \right).$$

10 Algorithmes de Newton

Soit l'équation $x^3 - 2x - 5 = 0$ dont on cherche une racine. Prenez un nombre comme 2, qui diffère de moins de 10 % de la vraie valeur d'une racine. Écrivez $x = 2 + d_1$ et remplacez x par $2 + d_1$ dans l'équation. Vous aurez $d_1^3 + 6d_1^2 + 10d_1 - 1 = 0$, dont il faut trouver la racine pour l'ajouter à 2. Négligez $d_1^3 + 6d_1^2$ à cause de sa petitesse ; il restera $10d_1 - 1 = 0$ ou $d_1 = 0.1$, ce qui est très près de la vraie valeur de d_1. C'est pourquoi, j'écris $d_1 = 0.1 + d_2$ et substituant comme auparavant j'ai $d_2^3 + 6.3d_2^2 + 11.23d_2 + 0.061 = 0$. Négligeant les deux premiers termes, il reste $11.23d_2 + 0.061 = 0$ ou $d_2 = -0.0054$ à peu près. [...] *Et je continue ainsi les opérations aussi longtemps qu'il convient.*

I. Newton (1736). Methodus fluxionum et serierum infinitorum. (Voir les notes en fin de chapitre.)

The central idea in this essay is that narcissism is an advantageous trait for succeeding in science. Scientists with a high ego are better able to convince others of the importance of their research. [...] Narcissists emerge as charismatic leaders but the cost of their attitude is invisible, paid by others.

B. Lemaitre [386].

On donne aujourd'hui le nom de *méthode de Newton* à toute approche algorithmique procédant par *linéarisation* des fonctions définissant le *système* dont on cherche une solution. C'est faire un grand honneur, peut-être excessif, à Isaac Newton, cet important contributeur de la science. Le terme *système* est pris ici dans un sens très large puisqu'il peut s'agir d'équations, d'inéquations, d'inclusions, d'équations différentielles ou aux dérivées partielles, d'inéquations variationnelles, *etc.* De même, le terme *linéarisation* doit être pris dans un sens étendu, car on utilise aussi la méthode de Newton pour résoudre des systèmes définis par des fonctions non différentiables dans le sens classique. On peut donc mesurer le chemin parcouru depuis l'algorithme proposé au XVII$^{\text{e}}$ siècle par Newton pour déterminer une racine d'un polynôme réel d'une variable réelle, décrit dans les quelques lignes de l'épigraphe de ce chapitre, alors que la notion de dérivée n'existait pas encore. Il aurait d'ailleurs été préférable d'utiliser le nom de Simpson pour décrire ces méthodes (voir les notes de fin de cha-

pitre), mais l'usage actuel en a décidé autrement. Nous aurions aussi pu utiliser la locution *méthodes de linéarisation*, mais nous n'avons pas franchi le pas et avons suivi la tradition.

Il y a de nombreuses monographies consacrées à l'algorithme de Newton ou à un aspect de cette approche par linéarisation (voir les notes en fin de chapitre), si bien que notre présentation ne pourra être que partielle, se concentrant sur des sujets qui nous paraissent essentiels ou en rapport direct avec l'esprit de cet ouvrage. Notre description commencera par le cas simple et instructif dans lequel on cherche à résoudre un système d'équations non linéaires, à en trouver un zéro (section 10.1.1). Ce cas est important en optimisation pour au moins deux raisons. D'abord il se présente lorsqu'on cherche à minimiser la fonction nulle sous des contraintes d'égalité. Par ailleurs, l'algorithme de Newton en optimisation sans contrainte est un cas particulier du précédent, si bien que certaines de ses propriétés, non attractives pour l'optimisation, trouveront leur origine dans le fait que cette approche est d'abord destinée à la résolution d'équations non linéaires. Nous verrons ensuite comment adapter l'algorithme à la minimisation de fonctions sans contrainte (section 10.1.2) ; le cas des problèmes avec contraintes sera examiné en détail au chapitre 15.

La propriété la plus attrayante de l'algorithme de Newton, qui en fait une référence, est sa convergence quadratique locale (théorèmes 10.2 et 10.3). Il a malheureusement aussi beaucoup de défauts ; nous les détaillerons. Comme remède à ces imperfections nous examinerons en détail les méthodes inexactes (section 10.2) et la globalisation de la convergence (section 10.3).

Connaissances supposées. Conditions d'optimalité pour les problèmes sans contrainte (section 4.2) ; algorithme du gradient conjugué (chapitre 8, utile pour l'algorithme de Newton tronqué à la section 10.3.1).

10.1 Méthodes locales

10.1.1 Systèmes d'équations

On s'intéresse ici à la recherche d'un *zéro* d'un système d'équations non linéaires, c'est-à-dire d'un point $x \in \mathbb{R}^n$ qui vérifie

$$F(x) = 0, \tag{10.1}$$

où $F : \mathbb{R}^n \to \mathbb{R}^n$ est une fonction différentiable. Il faut donc que $F_i(x) = 0$ pour tout $i = 1, \ldots, n$. Le système (10.1) étant formé de n équations aux n inconnues $x = (x_1, \ldots, x_n)$, il a quelques chances d'être bien posé.

L'algorithme de Newton génère une suite $\{x_k\}$ par une idée très simple, qui est illustrée à la figure 10.1 dans le cas où $n = 1$. On commence par linéariser l'équation en l'itéré courant x_k, ce qui donne la fonction $x \mapsto F(x_k) + F'(x_k) \cdot (x - x_k)$ dont le graphe est représenté par la ligne en tirets à la figure 10.1. Puis on cherche un zéro de cette fonction linéaire, s'il existe. C'est un opération simple puisqu'il suffit de résoudre un système linéaire. Ce zéro est l'itéré suivant x_{k+1}, qui est donc défini par

$$F(x_k) + F'(x_k) \cdot (x_{k+1} - x_k) = 0.$$

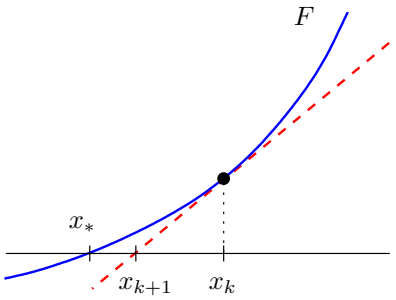


Fig. 10.1. Une itération de Newton

Cette équation peut certainement être résolue si $F'(x_k)$ est inversible. On renforcera l'analogie avec les méthodes à directions de descente du chapitre 6 en écrivant

$$x_{k+1} = x_k + d_k, \tag{10.2}$$

où d_k est la solution de l'*équation de Newton*, qui est le système linéaire suivant

$$F'(x_k)d_k = -F(x_k). \tag{10.3}$$

On peut maintenant décrire l'algorithme de Newton, que l'on qualifie de *local* car, comme on le verra, sa convergence n'est garantie que si le premier itéré est proche d'un zéro régulier de F.

Algorithme 10.1 (Newton local pour système non linéaire) On suppose qu'au début de l'itération k, on dispose d'un itéré $x_k \in \mathbb{R}^n$.

1. *Test d'arrêt.* Si $F(x_k) \simeq 0$, arrêt de l'algorithme.
2. *Direction.* Calculer d_k comme solution de (10.3).
3. *Nouvel itéré.* $x_{k+1} := x_k + d_k$.

Le coût de l'itération repose essentiellement sur l'évaluation de la jacobienne $F'(x_k)$ et sur la résolution du système linéaire (10.3) à l'étape 2. Cet algorithme ramène donc la résolution du système non linéaire (10.1) à une *suite* de systèmes linéaires, plus simples à résoudre.

L'intérêt principal de l'algorithme Newton est de générer des suites q-quadratiquement convergentes, c'est ce que nous allons montrer dans le théorème 10.2 ci-dessous. Les conditions assurant un tel comportement sont à peine plus fortes que celles requises pour que la méthode soit bien définie : il faut que F ait une dérivée lipschitzienne (alors que seule la dérivabilité de F est nécessaire à la définition de l'algorithme) et que F' soit inversible en la solution x_* recherchée (alors $F'(x_k)$ sera inversible pour un itéré x_k proche de x_*).

Le théorème suivant analyse la convergence d'une méthode un peu plus générale que l'algorithme 10.1, dans laquelle, à l'étape 2, la direction d_k est solution du système linéaire

$$M_k d_k = -F(x_k), \tag{10.4}$$

où M_k est une matrice inversible, pouvant être différente de $F'(x_k)$. Les méthodes de quasi-Newton entrent dans ce cadre (voir le chapitre 11).

Théorème 10.2 (convergence locale de l'algorithme de Newton) *On suppose que F a un zéro x_*, que F est de classe C^1 dans un voisinage Ω de x_* et que $F'(x_*)$ est inversible.*

1) *Alors, il existe $\varepsilon_x > 0$ et $\varepsilon_M > 0$ tels que si*

$$\|x_1 - x_*\| \leqslant \varepsilon_x \qquad \text{et} \qquad \|M_k - F'(x_k)\| \leqslant \varepsilon_M, \quad \forall k \geqslant 1,$$

l'algorithme de Newton avec d_k solution de (10.4), plutôt que de (10.3), est bien défini et génère une suite $\{x_k\}$ convergeant q-linéairement vers x_*.

2) *Si de plus*

$$\big(M_k - F'(x_*)\big)(x_k - x_*) = o(\|x_k - x_*\|),$$

alors la convergence est q-superlinéaire.

3) *Si de plus F' est lipschitzienne sur Ω et*

$$\big(M_k - F'(x_*)\big)(x_k - x_*) = O(\|x_k - x_*\|^2),$$

alors la convergence est q-quadratique.

Démonstration. On note $\beta := \|F'(x_*)^{-1}\|$ et on choisit $\varepsilon_M > 0$ tel que $\beta\varepsilon_M < 1$ et

$$r := \frac{3\beta\varepsilon_M}{1 - \beta\varepsilon_M} < 1.$$

On détermine ensuite $\varepsilon_x > 0$ tel que $\bar{B}(x_*, \varepsilon_x) \subseteq \Omega$ et tel que $\|x - x_*\| \leqslant \varepsilon_x$ implique que $\|F'(x) - F'(x_*)\| \leqslant \varepsilon_M$ (possible par la continuité de F').

Si une matrice M vérifie $\|M - F'(x_*)\| \leqslant \varepsilon_M$, alors $\|F'(x_*)^{-1}(M - F'(x_*))\| \leqslant \beta\varepsilon_M < 1$ et, par le lemme A.2 de perturbation de Banach, la matrice M est inversible et vérifie $\|M^{-1}\| \leqslant \beta/(1 - \beta\varepsilon_M)$. En appliquant cela à $M = M_k$ ou $M = F'(x)$, on trouve que, pour tout $k \geqslant 1$ et tout $x \in \bar{B}(x_*, \varepsilon_x)$, M_k et $F'(x)$ sont inversibles et

$$\|M_k^{-1}\| \quad \text{et} \quad \|F'(x)^{-1}\| \leqslant \frac{\beta}{1 - \beta\varepsilon_M}.$$

Dans ce cas, la formule (10.4) définit bien la direction d_k.

En utilisant $F(x_*) = 0$ et le fait que F est de classe C^1 sur $\bar{B}(x_*, \varepsilon_x)$ (ce qui autorise un développement de Taylor avec reste intégral), on a si $x_k \in \bar{B}(x_*, \varepsilon_x)$

$$\begin{aligned} x_{k+1} - x_* &= x_k - x_* + d_k \\ &= M_k^{-1}\big(M_k(x_k - x_*) - F(x_k)\big) \\ &= M_k^{-1}\big(M_k - F'(x_k)\big)(x_k - x_*) \\ &\quad + M_k^{-1}\int_0^1 \Big(F'(x_k) - F'(x_* + t(x_k - x_*))\Big)(x_k - x_*)\,\mathrm{d}t. \end{aligned}$$

En utilisant le fait que la norme d'une intégrale est plus petite que l'intégrale de la norme de l'intégrant, on en déduit que $\|x_{k+1} - x_*\| \leqslant r\|x_k - x_*\|$. Dès lors, par récurrence, toute la suite $\{x_k\} \subseteq \bar{B}(x_*, \varepsilon_x)$ si $x_1 \in \bar{B}(x_*, \varepsilon_x)$ (car $r \leqslant 1$). De plus $x_k \to x_*$ (car $r < 1$). Ceci démontre le point 1 du théorème.

Sous la condition additionnelle du point 2, l'estimation de l'erreur $x_{k+1} - x_*$ ci-dessus montre que $x_{k+1} - x_* = o(\|x_k - x_*\|)$, c'est-à-dire la convergence superlinéaire de $\{x_k\}$. Sous les conditions additionnelles du point 3, on trouve à partir de l'estimation de l'erreur $x_{k+1} - x_*$ ci-dessus que, pour une constante C, $\|x_{k+1} - x_*\| \leqslant C\|x_k - x_*\|^2$; on obtient la convergence quadratique de $\{x_k\}$. □

Le résultat de convergence ci-dessus s'applique directement à l'algorithme de Newton, c'est-à-dire lorsque $M_k = F'(x_k)$ pour tout $k \geqslant 1$. En particulier, on voit que sous les conditions de régularité de F spécifiées dans les trois parties du théorème, l'algorithme est bien défini et génère une suite convergeant quadratiquement, dès que le premier itéré x_1 est pris assez proche de x_*.

Le théorème de Kantorovitch offre une autre manière de montrer la convergence de l'algorithme de Newton. Il a la particularité intéressante de ne pas supposer l'existence d'un zéro de F, mais de l'affirmer. Ce résultat offre donc aussi un moyen de démontrer l'existence d'un zéro d'une équation non linéaire. Il est d'ailleurs apparenté à des théorèmes d'existence de points fixes (voir les notes en fin de chapitre). Le résultat s'exprime simplement : si x_1 est presqu'un zéro ($F(x_1) \simeq 0$) et si F' est inversible en x_1 et ne change pas trop vite, alors il doit y avoir un zéro dans un voisinage de x_1 ; de plus l'algorithme de Newton démarrant en x_1 converge vers ce zéro.

Théorème 10.3 (Kantorovitch) *Supposons que F soit différentiable sur un ouvert convexe $\Omega \subseteq \mathbb{R}^n$. On suppose également qu'en $x_1 \in \Omega$, $F'(x_1)$ est inversible, que $F'(x_1)^{-1}F'(\cdot)$ est lipschitzienne de module $L > 0$ sur Ω et que, pour $\delta := \|F'(x_1)^{-1}F(x_1)\|$ et $r := (1 - \sqrt{1-2\delta L})/L$, on a*

$$2\delta L \leqslant 1 \qquad \text{et} \qquad \bar{B}(x_1, r) \subseteq \Omega.$$

Alors,

1) *F a un zéro $x_* \in \bar{B}(x_1, r)$,*
2) *F n'a pas d'autre zéro que x_* dans $(\bar{B}(x_1, r) \cup B(x_1, r_+)) \cap \Omega$, où $r_+ := (1 + \sqrt{1-2\delta L})/L$,*
3) *l'algorithme de Newton démarrant en x_1 est bien défini et génère une suite $\{x_k\} \subseteq \bar{B}(x_1, r)$ convergeant vers x_*.*

Démonstration. □

Concluons cette section par une propriété de l'algorithme de Newton importante pour les applications : l'algorithme est invariant par changement de variables. De manière plus précise, supposons que l'on fasse le changement de variables

$$\tilde{x} = Ax,$$

où A est une matrice d'ordre n inversible. Soit $\tilde{F} = F \circ A^{-1}$ l'expression de F dans l'espace des $\tilde{x}$; donc $\tilde{F}(\tilde{x}) = F(x)$ si $\tilde{x}$ et x sont reliés par la relation ci-dessus. On a le résultat suivant.

Proposition 10.4 (invariance par changement de variables) *Dans les conditions décrites ci-dessus, si $\{x_k\}$ [resp. $\{\tilde{x}_k\}$] est la suite des itérés générés par l'algorithme de Newton pour résoudre le système non linéaire $F(x) = 0$ [resp. $\tilde{F}(\tilde{x}) = 0$] à partir d'un premier itéré x_1 [resp. $\tilde{x}_1 = Ax_1$], alors $\tilde{x}_k = Ax_k$ pour tout $k \geqslant 1$.*

Démonstration. Soit d_k la direction de Newton sur F en x_k et $\tilde{d}_k$ la direction de Newton sur $\tilde{F}$ en $\tilde{x}_k$. Si $\tilde{x}_k = Ax_k$, on a

$$\tilde{d}_k = -\tilde{F}'(\tilde{x}_k)^{-1}\tilde{F}(\tilde{x}_k) = -AF'(x)^{-1}F(x) = Ad_k,$$

car $\tilde{F}'(\tilde{x}_k) = F'(x_k)A^{-1}$ et $\tilde{F}(\tilde{x}_k) = F(x_k)$. On en déduit que

$$\tilde{x}_{k+1} = \tilde{x}_k + \tilde{d}_k = A(x_k + d_k) = Ax_{k+1}.$$

Le résultat s'en ensuit alors par récurrence. □

On obtient un résultat d'invariance analogue si, au lieu de pré-composer la fonction F par une application linéaire inversible, on la post-compose : $\tilde{F} = A \circ F$. Ces résultats nous montrent qu'il ne sert à rien de préconditionner l'algorithme de Newton par pré- ou post-composition avec une application linéaire inversible, puisque les itérés générés n'en seraient pas affectés. Un préconditionnement peut toutefois avoir une incidence en arithmétique flottante et dans les algorithmes de Newton tronqués, dans lesquels le système linéaire n'est résolu que partiellement (section **??**).

10.1.2 Optimisation

On considère le problème d'optimisation non linéaire sans contrainte suivant :

$$\begin{cases} \min f(x) \\ x \in \mathbb{R}^n, \end{cases} \tag{10.5}$$

dans lequel f est supposée régulière. Son équation d'optimalité s'écrit :

$$\nabla f(x) = 0,$$

où $\nabla f(x)$ est le gradient de f en x pour un produit scalaire arbitraire donné (on le note $\langle \cdot, \cdot \rangle$). Il s'agit d'un système de n équations non linéaires à n inconnues, que l'on peut résoudre par l'algorithme de Newton de la section 10.1.1, avec $F \equiv \nabla f$. Dans ce cas, l'équation de Newton (10.3) s'obtient en linéarisant en x_k l'équation d'optimalité ci-dessus, qui s'écrit aussi $f'(x) \cdot h = 0$, pour tout $h \in \mathbb{R}^n$. Cela donne $f'(x_k) \cdot h + f''(x_k) \cdot (d_k, h) = 0$, pour tout $h \in \mathbb{R}^n$; ou encore

$$\nabla^2 f(x_k) d_k = -\nabla f(x_k). \tag{10.6}$$

Dans (10.6), $\nabla^2 f(x_k)$ est donc la hessienne de f en x_k pour le produit scalaire ayant servi à calculer le gradient $\nabla f(x_k)$. On adapte ainsi aisément l'algorithme de la section 10.1.1 au cas de l'optimisation.

Algorithme 10.5 (Newton local en optimisation) On suppose qu'au début de l'itération k, on dispose d'un itéré $x_k \in \mathbb{R}^n$.

1. *Test d'arrêt.* Si $\nabla f(x_k) \simeq 0$, arrêt de l'algorithme.
2. *Direction.* Calculer d_k comme solution de (10.6).
3. *Nouvel itéré.* $x_{k+1} := x_k + d_k$.

Une autre approche conduisant au même résultat est la suivante. Étant donné l'itéré x_k, on cherche à trouver x_{k+1} en minimisant l'approximation quadratique de f. Ceci conduit au *problème quadratique osculateur* en x_k, qui est le problème en d suivant

$$\min_{d\in\mathbb{R}^n} \; f(x_k) + \nabla f(x_k)^\mathsf{T} d + \frac{1}{2} d^\mathsf{T} \nabla^2 f(x_k) d. \tag{10.7}$$

S'il a un *point stationnaire*, disons d_k, on prend alors $x_{k+1} = x_k + d_k$. Il est aisé de montrer qu'il s'agit du même algorithme de Newton : l'équation d'optimalité de (10.7) n'est autre que (10.6).

Il est important d'observer que l'algorithme de Newton construit des suites convergeant vers des points stationnaires, sans faire de distinction entre les minima ou les maxima, par exemple. Ceci est dû au fait qu'il est conçu pour trouver des zéros de $\nabla f(x) = 0$. Par conséquent, sans modification adéquate, si le premier itéré est proche d'un point stationnaire « régulier », la suite générée convergera vers ce point stationnaire. On comprend que, si l'on cherche à minimiser f, converger vers un maximum local n'est pas une propriété satisfaisante. Il sera donc nécessaire de modifier l'algorithme de Newton de manière à le contraindre à éviter les points stationnaires qui ne sont pas des minima. Ce n'est pas une tâche facile : si $\nabla f(x_1) = 0$, la direction de Newton est nulle et donc l'itéré suivant x_2 est identique au premier ! Cette question est toujours un objet d'études. Nous en reparlerons aux sections 10.3.1 et 10.3.2.

10.1.3 Défauts et remèdes

Les inconvénients de la méthode de Newton pour résoudre des systèmes d'équations non linéaires [resp. des problèmes d'optimisation] sont bien connus :

1. il faut calculer les dérivées premières de F [resp. les dérivées secondes de f], ce qui peut être coûteux en temps de calcul (n^2 éléments à évaluer), en effort humain (l'expression analytique de ces dérivées n'est pas toujours simple à obtenir) et en espace mémoire ;
2. l'algorithme n'est pas globalement convergent (si le premier itéré est éloigné d'une solution, le comportement des itérés suivants est souvent erratique) ;
3. l'algorithme n'est pas nécessairement défini aux points x où $F'(x)$ [resp. $\nabla^2 f(x)$] est singulière ;
4. pour les problèmes d'optimisation, si f n'est pas fortement convexe, l'algorithme ne génère pas nécessairement des directions de descente de f ;

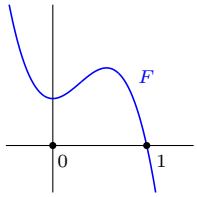
5. un système linéaire d'ordre n doit être résolu à chaque itération.

Voilà bien des défauts pour un algorithme aux propriétés locales tant souhaitées (il converge localement quadratiquement). Dans les sections suivantes, nous étudions des modifications de la méthode de Newton qui remédient partiellement à ces points faibles, tout en essayant de conserver son intérêt majeur qui est de générer des suites convergeant rapidement.

Les remèdes foisonnent et il n'est pas aisé de les exposer de manière concise. On peut en effet s'intéresser à la résolution de systèmes non linéaires ou à l'optimisation ; la globalisation de la convergence peut se faire par recherche linéaire, par région de confiance ou des méthodes de suivi de chemin ; le système de Newton peut être résolu exactement ou de manière approchée, par des méthodes directes ou itératives ; les algorithmes peuvent faire l'effort de ne pas utiliser la transposée de la jacobienne $F'(x)$ ou pas. Voilà donc beaucoup de possibilités à décrire et elles peuvent toutes être intéressantes en fonction du problème à résoudre. Nous serons brefs sur certaines approches si elles peuvent se déduire de méthodes déjà décrites ailleurs. Par ailleurs, nous ne considérerons pas les cas singuliers, où la jacobienne n'est pas inversible en la solution, où la fonction est non lisse, où la solution recherchée n'est pas isolée, *etc*, qui sont tous très importants pour pouvoir aborder des problèmes plus généraux que les deux considérés ici (annuler une fonction non linéaire et l'optimisation), comme les problèmes d'inéquations variationnelles ou de complémentarité, l'optimisation sous contrainte, *etc*.

Mais soyons clair : nonobstant cette abondance, il n'y a pas d'algorithme newtonien qui garantisse la convergence vers un zéro d'une fonction non linéaire F arbitraire quel que soit l'itéré initial. Cependant, les techniques numériques que nous allons présenter dans les sections suivantes améliorent grandement les qualités (efficacité et robustesse) des algorithmes en pratique, si bien qu'on ne peut les négliger. La difficulté se rencontre déjà en dimension 1, pour la fonction suivante

$$F(x) = \tfrac{1}{2} + 3x^2 - \tfrac{7}{2}x^3, \tag{10.8}$$



laquelle a un unique zéro en $x = 1$. Si l'on prend comme itéré initial $x_0 = 0$, les algorithmes échouent lamentablement. Il faut noter que la jacobienne de F y est nulle et donc que la direction de Newton n'y est pas définie. Par ailleurs, la fonction $x \mapsto |F(x)|$ a un minimum local en zéro, ce qui rend ce point attrayant aux yeux de beaucoup d'algorithmes. En réalité, il n'y a pas aujourd'hui de remède universel à cette difficulté fondamentale, qui trouve son origine dans le fait que l'algorithme de Newton est une méthode locale (en chaque itéré, elle n'utilise que les valeurs de F et de sa dérivée) alors que la détermination d'un zéro de F est de nature globale (dans l'exemple ci-dessus, en n'examinant F que dans le voisinage de 0, il est très difficile de savoir s'il faut s'éloigner de 0 en partant vers la gauche ou vers la droite — ce n'est pas impossible de faire le bon choix lorsque F est analytique et que l'on dispose des dérivées de tous ordres de F en zéro, mais en pratique il n'est possible d'utiliser qu'une quantité finie d'information).

On notera enfin que la situation est beaucoup plus favorable si les composantes de F sont des *polynômes*. La nature globale des zéros de F ne pose alors pas de difficulté aux techniques algébriques (via l'utilisation de *base de Gröbner* par exemple [147]) ou numériques (en utilisant des méthodes d'*optimisation globale* [375]) pourvu que le nombre de variables ou le degré des polynômes reste faible.

10.2 Méthodes inexactes ▲

10.2.1 Systèmes d'équations

Dans les problèmes de grande taille, il peut être coûteux de résoudre les systèmes linéaires de Newton (10.3) avec précision. Souvent même, une résolution exacte n'est pas possible, si bien qu'il faut définir un seuil de tolérance. Par ailleurs, on conviendra également qu'un calcul précis, qui fait entièrement confiance à la linéarisation de F, n'est probablement pas utile lorsque l'itéré courant x_k est éloigné d'un zéro de F, parce qu'en de tels points la direction de Newton d_k est généralement grande et qu'alors $F(x_k + d_k)$ est souvent éloigné de la valeur nulle prédite par le modèle linéarisé de F. Si le nombre de variables est important, les systèmes linéaires sont en général résolus par des méthodes itératives, dont l'arrêt est contrôlé par un test ; il est alors naturel d'avoir un test permissif, autorisant un important résidu $F(x_k) + F'(x_k)d_k$, lorsque $F(x_k)$ est grand et un test plus contraignant lorsque $F(x_k)$ est petit. Ces différentes considérations conduisent à la notion suivante.

On parle de *méthode de Newton inexacte* lorsque l'algorithme cherche à calculer des directions d_k, dites *de Newton inexactes*, vérifiant la condition suivante :

$$\|F(x_k) + F'(x_k)d_k\| \leqslant \eta_k \|F(x_k)\|, \tag{10.9}$$

où $\|\cdot\|$ est une norme arbitraire et $\eta_k \in [0, 1[$ est appelé le *facteur d'inexactitude*. Il est naturel de prendre $\eta_k < 1$ de manière à ne pas accepter une direction nulle. Par ailleurs, la direction de Newton, quand elle existe, annule le membre de gauche, si bien que (10.9) peut être vu comme une condition acceptant davantage de directions que celle de Newton. Comme annoncé, la condition (10.9) contrôle la précision avec laquelle il faut résoudre le système linéaire de Newton (10.3) au moyen de la grandeur $\|F(x_k)\|$, qui mesure la précision avec laquelle l'itéré courant résout le système non linéaire (10.1).

La condition (10.9) n'est pas nécessairement réalisable. La proposition suivante montre que l'on peut trouver une direction de Newton inexacte pour une norme arbitraire, essentiellement lorsque la direction de Newton elle-même existe, c'est-à-dire lorsque $F(x_k) \in \mathcal{R}(F'(x_k))$. Cependant, si la direction de Newton n'existe pas, on pourra parfois trouver une direction de Newton inexacte pour une norme particulière et un facteur d'inexactitude assez grand. Par exemple, si $F : \mathbb{R}^2 \to \mathbb{R}^2$ est la fonction linéaire définie par

$$F(x) = \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} x,$$

on a $F(0) \notin \mathcal{R}(F'(0))$, alors que $\|F(0) + F'(0)d\|_2 \leqslant \eta \|F(0)\|_2$ est réalisable pourvu que $\eta \in [\sqrt{2}/2, 1[$.

Proposition 10.6 (existence d'une direction de Newton inexacte) *Supposons que F soit différentiable en un itéré x_k tel que $F(x_k) \neq 0$. Alors, les propriétés suivantes sont équivalentes :*

(i) $F(x_k) \in \mathcal{R}(F'(x_k))$,
(ii) *pour toute norme $\|\cdot\|$ et tout $\eta_k \in [0,1[$, il existe un d_k vérifiant (10.9),*
(iii) *pour toute norme $\|\cdot\|$ associée à un produit scalaire, il existe un $\eta_k \in [0,1[$ et un d_k vérifiant (10.9).*

DÉMONSTRATION. $[(i) \Rightarrow (ii)]$ Si $F(x_k) \in \mathcal{R}(F'(x_k))$, on peut trouver un d_k tel que $F(x_k) + F'(x_k)d_k = 0$. Cette direction de Newton d_k vérifie évidemment (10.9) quels que soient la norme et le $\eta_k \in [0,1[$.

$[(ii) \Rightarrow (iii)]$ Évident.

$[(iii) \Rightarrow (i)]$ Si $F(x_k) \notin \mathcal{R}(F'(x_k))$, on peut construire une base de $\mathbb{R}^n$ en complétant une base de $\mathcal{R}(F'(x_k))$ à laquelle on joint le vecteur $F(x_k)$. On prend sur $\mathbb{R}^n$ le produit scalaire $\langle\cdot,\cdot\rangle$ associé à cette base, qui est le produit scalaire euclidien des coordonnées dans cette base, et la norme associée, que l'on note $\|\cdot\|$. Alors $F(x_k)$ est orthogonal à $\mathcal{R}(F'(x_k))$, ce qui s'écrit

$$F'(x_k)^* F(x_k) = 0.$$

On en déduit que $d = 0$ minimise la fonction convexe différentiable $d \mapsto \|F(x_k) + F'(x_k)d\|$, c'est-à-dire que $\|F(x_k) + F'(x_k)d\| \geqslant \|F(x_k)\|$ pour tout $d \in \mathbb{R}^n$. Dès lors, quel que soit $\eta_k \in [0,1[$, (10.9) n'est pas réalisable pour la norme $\|\cdot\|$. □

10.3 Globalisation de la convergence

Grâce au théorème 10.2, on sait que la convergence de l'algorithme de Newton local 10.1 est garantie si l'itéré initial est « suffisamment » proche d'une solution (un zéro de F ou un minimum de f). Si le premier itéré est « éloigné » d'une solution, l'algorithme pourra générer une suite au comportement erratique, qui pourra accidentellement se retrouver dans le voisinage d'une solution et converger vers celle-ci, mais qui le plus souvent divergera (voir [43 ; 2012] pour un cas de cyclage, avec une fonction non différentiable, que l'on pourrait facilement lisser). En général, il est difficile de dire si un itéré initial est dans le voisinage d'une solution qui garantit la convergence de l'algorithme de Newton. Il est donc important de disposer de techniques permettant d'éviter le comportement désordonné indésirable probable de ses suites générées.

On entend par *globalisation de la convergence* de l'algorithme de Newton toute technique permettant d'améliorer la convergence des itérés vers une solution du problème, même si l'itéré initial est éloigné d'une solution. Cette notion n'a donc pas de lien avec la recherche d'un minimum global d'une fonction.

10.3.1 Recherche linéaire

Newton inexact ▲

On appelle *fonction de mérite*, toute fonction réelle qui atteint un minimum (si possible global) en une solution du problème que l'on cherche à résoudre. Si le problème considéré est celui de la minimisation sans contrainte (10.5), la fonction de mérite idéale est la fonction coût elle-même. Dans le cas où l'on recherche un zéro de l'équation (10.1), une fonction de mérite naturelle est la *fonction de moindres-carrés* $\varphi : \mathbb{R}^n \to \mathbb{R}$, définie par

$$\varphi(x) = \frac{1}{2}\|F(x)\|_2^2. \tag{10.10}$$

Le facteur $\frac{1}{2}$ n'est utile que pour simplifier l'expression de la dérivée ; la norme ℓ_2 et le carré assurent quant à eux la différentiabilité de φ.

La fonction φ atteint une valeur optimale nulle en toute solution de (10.1). Elle peut toutefois avoir des minima locaux qui ne sont pas solutions de (10.1). Ceux-ci vérifient la condition d'optimalité du premier ordre

$$F'(x)^\mathsf{T} F(x) = 0.$$

Ces points stationnaires seront donc des solutions de (10.1) si $F'(x)$ y est inversible, ce qui est loin d'être toujours le cas. Ce raisonnement simple met en évidence le rôle critique que jouera, dans cette approche, le lieu des points où la jacobienne $F'(x)$ est singulière :

$$\mathcal{S} := \{x \in \mathbb{R}^n : F'(x) \text{ est singulière}\}.$$

La situation est cependant plus compliquée : même si $\mathcal{S}$ est vide, certains algorithmes utilisant φ comme fonction de mérite pourront rencontrer des difficultés lorsque la fonction $x \mapsto F'(x)^{-1}$ n'est pas bornée. Autrement dit, une matrice $F'(x)$ « singulière à l'infini » peut aussi être une source de difficultés.

Les techniques de globalisation de la convergence utilisent souvent de telles fonctions de mérite, car on sait comment forcer la convergence d'itérés vers des minima locaux de fonctions, par recherche linéaire (chapitre 6) ou par régions de confiance (chapitre 9), alors que l'on ne connaît pas de méthode systématique permettant de trouver un zéro d'une fonction. On peut dire qu'en adoptant une telle approche, ces techniques renoncent à trouver un zéro de F et se contentent d'un point stationnaire ou d'un minimum local de φ. Dans ce cadre, cette recherche de zéro revient à trouver un minimum global de la fonction φ ci-dessus, tâche considérée aujourd'hui comme très difficile, parfois impossible, et de toutes façons très coûteuse en toute généralité.

Nous nous intéressons donc, dans cette section, à la globalisation de la convergence de l'algorithme de Newton pour résoudre le système (10.1), au moyen de la fonction de mérite φ définie ci-dessus. Cette approche permet d'ailleurs de résoudre de manière approchée l'équation de Newton (10.3). On se satisfait en effet d'une direction d_k qui vérifie

$$\|F(x_k) + F'(x_k)d_k\|_2 \leqslant \eta_k \|F(x_k)\|_2, \tag{10.11}$$

où $0 \leqslant \eta_k \leqslant \eta < 1$ (η est une constante). En général, on prend η_k proche de 1 lorsque x_k est loin d'une solution, de manière à ne pas passer trop de temps dans la résolution d'un système linéaire qui n'est sans doute pas un bon modèle de F dans ce cas, et l'on

prend η_k proche de zéro lorsque x_k se rapproche d'une solution, de manière à bénéficier de la convergence rapide de l'algorithme de Newton dans le voisinage d'une solution. Rappelons que le fait que l'on puisse trouver une direction d_k vérifiant (10.11) cache une hypothèse sur $F'(x_k)$; voir la proposition 10.6.

A priori, on ne voit pas pourquoi la direction de Newton (inexacte) serait une direction de descente de φ, laquelle est définie de manière naturelle, mais sans lien évident avec l'algorithme de Newton. Le fait qu'il en soit ainsi est le premier miracle du couple Newton-φ (voir la proposition 10.15 pour le second).

Proposition 10.7 (descente) *Si $F(x_k) \neq 0$, toute direction d_k vérifiant (10.11) est une direction de descente (non nulle) de φ en x_k car on a*

$$\nabla\varphi(x_k)^{\mathsf{T}} d_k = F'(x_k)^{\mathsf{T}} F(x_k) d_k \leqslant -2(1-\eta_k)\varphi(x_k) < 0. \tag{10.12}$$

Démonstration. On a en effet $\nabla\varphi(x_k) = F'(x_k)^{\mathsf{T}} F(x_k)$. Puis en utilisant l'inégalité de Cauchy-Schwarz :

$$\nabla\varphi(x_k)^{\mathsf{T}} d_k = F(x_k)^{\mathsf{T}} \left(F(x_k) + F'(x_k)d_k\right) - \|F(x_k)\|_2^2 \leqslant -(1-\eta_k)\|F(x_k)\|_2^2 < 0.$$

Forcément, comme toute direction de descente, d_k ne peut être nulle. □

Comme on cherche à annuler F et qu'un zéro de F est un minimum global de φ, la propriété remarquable précédente légitime le fait de trouver l'itéré suivant x_k en faisant de la recherche linéaire le long de d_k. C'est ce que fait l'algorithme ci-dessous, qui porte le nom d'*algorithme de Newton inexact*, malgré la connotation péjorative de cette appellation. Cette approche, que l'on retrouvera pour l'algorithme de Newton en optimisation, semble providentielle. Nous verrons cependant qu'elle a ses propres limites.

Algorithme 10.8 (Newton inexact) On suppose qu'au début de l'itération k, on dispose d'un itéré $x_k \in \mathbb{R}^n$.

1. *Test d'arrêt.* Si $F(x_k) \simeq 0$, arrêt de l'algorithme.
2. *Direction.* Calculer d_k vérifiant (10.11). Si ce n'est pas possible l'algorithme échoue.
3. *Recherche linéaire.* Déterminer un pas $\alpha_k > 0$ « suffisamment grand » le long de d_k de manière à faire décroître φ « suffisamment ».
4. *Nouvel itéré.* $x_{k+1} := x_k + \alpha_k d_k$.

La description de la recherche linéaire utilisée à l'étape 3 est vague et sera précisée dans les résultats de convergence ci-dessous. Il est courant cependant d'utiliser la règle d'Armijo (section 6.3.3) : pour ω et $\beta \in]0,1[$, le pas α_k est pris égal à $\beta^{i_k}\alpha_k^1$ où α_k^1 plus grand qu'une constante strictement positive et i_k est le plus petit entier positif tel que

$$\varphi(x_k + \alpha_k d_k) \leqslant \varphi(x_k) - 2\omega\alpha_k(1-\eta_k)\varphi(x_k). \tag{10.13}$$

En utilisant la proposition 10.7, on voit facilement qu'un tel pas existe.

Le premier énoncé de convergence globale que nous donnons ci-après montre que les points d'adhérence *réguliers* (dans un sens précisé dans l'énoncé) de la suite $\{x_k\}$ générée par l'algorithme 10.8 sont des zéros de F. Il faut se garder de penser que la question de la convergence globale de l'algorithme de Newton est réglée avec ce résultat, car il se peut très bien que de tels points stationnaires réguliers n'existent pas et que la suite générée converge vers un point non régulier qui n'est pas un zéro de F. Néanmoins, un tel résultat est une première indication sur la bonne conception de l'algorithme et nous l'énonçons et le démontrons pour cette raison.

Proposition 10.9 (Newton inexact et points d'adhérence) *On considère l'algorithme de Newton inexact 10.8 pour résoudre le système $F(x) = 0$ où $F : \mathbb{R}^n \to \mathbb{R}^n$ est différentiable avec $F'(x_k)$ inversible en tout itéré x_k généré par l'algorithme. On suppose que l'algorithme utilise la règle de recherche linéaire d'Armijo décrite autour de (10.13). Alors, s'il existe un point d'adhérence $\bar{x}$ de $\{x_k\}$ tel que $F'(\bar{x})$ est inversible et si F' y est continue, il s'ensuit que $\varphi(x_k) \to 0$ et $F(\bar{x}) = 0$.*

Démonstration. Si $\bar{x}$ est un point d'adhérence de $\{x_k\}$, il existe une sous-suite d'indices $\mathcal{K}$ tel que $x_k \to \bar{x}$ lorsque $k \to \infty$ dans $\mathcal{K}$.

Montrons que $\{d_k\}_{k\in\mathcal{K}}$ *est bornée* (c'est une conséquence de la régularité de $\bar{x}$). On raisonne par l'absurde, en supposant que $\{d_k\}$ n'est pas bornée. Alors, en extrayant une sous-suite au besoin, on peut supposer que $\|d_k\| \to \infty$ et $d_k/\|d_k\| \to d \neq 0$ lorsque $k \to \infty$ dans $\mathcal{K}$. En divisant les deux membres de l'inégalité (10.11) par $\|d_k\|$ et en passant à la limite lorsque $k \to \infty$, on trouve que $F'(\bar{x})d = 0$, ce qui contredit l'inversibilité supposée de $F'(\bar{x})$ puisque $d \neq 0$.

Par la règle d'Armijo, la suite $\{\varphi(x_k)\}$ est décroissante. Comme elle est aussi borné inférieurement (par zéro), elle converge. Alors, la règle d'Armijo et $\eta_k \leqslant \eta < 1$ impliquent que

$$\alpha_k\varphi(x_k) \to 0.$$

On poursuit en examinant deux cas complémentaires.

1) Supposons d'abord que $\alpha_k \not\to 0$. Alors, $\varphi(x_k) \to 0$ pour une sous-suite d'indices tendant vers l'infini. Par la décroissance de la suite $\{\varphi(x_k)\}$, on voit que toute la suite $\{\varphi(x_k)\} \to 0$. Dès lors, tout point d'adhérence $\bar{x}$ de $\{x_k\}$ vérifie $\varphi(\bar{x}) = 0$ ou $F(\bar{x}) = 0$.

2) Considérons à présent le cas plus difficile où $\alpha_k \to 0$. On peut supposer que $\alpha_k < 1$, ce qui veut dire que le pas $\hat{\alpha}_k := \alpha_k/\beta$ n'est pas accepté par la règle d'Armijo ou encore, qu'au point $\hat{x}_k := x_k + \hat{\alpha}_k d_k$, on a

$$\varphi(\hat{x}_k) > \varphi(x_k) - 2\omega\hat{\alpha}_k(1-\eta_k)\varphi(x_k). \tag{10.14}$$

Notons que $\hat{\alpha}_k \to 0$ et que, par la bornitude de $\{d_k\}_{k\in\mathcal{K}}$, $\hat{x}_k \to \bar{x}$ pour $k \to \infty$ dans $\mathcal{K}$. On peut estimer l'écart $\varphi(\hat{x}_k) - \varphi(x_k)$ comme suit. Par le théorème des accroissements finis (corollaire C.13), on a

$$\|F(\hat{x}_k) - F(x_k) - F'(x_k)(\hat{x}_k - x_k)\| \leqslant \left(\sup_{z \in]x_k, \hat{x}_k[} \|F'(z) - F'(x_k)\| \right) \|\hat{x}_k - x_k\|.$$

Par la continuité supposée de F' en $\bar{x}$, le facteur entre parenthèses du membre de droite tend vers zéro quand $k \to \infty$ dans $\mathcal{K}$. En utilisant $\hat{x}_k - x_k = \hat{\alpha}_k d_k$, on obtient $F(\hat{x}_k) = F(x_k) + \hat{\alpha}_k F'(x_k) d_k + o(\hat{\alpha}_k)$ et donc

$$\begin{aligned} \varphi(\hat{x}_k) &= \varphi(x_k) + \hat{\alpha}_k F(x_k)^{\mathsf{T}} F'(x_k) d_k + o(\hat{\alpha}_k) \\ &\leqslant \varphi(x_k) - 2\hat{\alpha}_k (1 - \eta_k) \varphi(x_k) + o(\hat{\alpha}_k) \qquad [(10.12)]. \end{aligned}$$

Alors (10.14) conduit à

$$0 \leqslant 2(1 - \omega)\hat{\alpha}_k (1 - \eta_k) \varphi(x_k) \leqslant o(\hat{\alpha}_k).$$

En divisant chaque membre de ces inégalités par $\hat{\alpha}_k > 0$, en utilisant $\omega < 1$, en extrayant une sous-suite convergente de $\{\eta_k\}_{k \in \mathcal{K}} \subseteq]0, \eta]$ et en passant à la limite lorsque $k \to \infty$ dans $\mathcal{K}$, on obtient que $\varphi(\bar{x}) = 0$ ou $F(\bar{x}) = 0$. □

La recherche linéaire est considérée comme raisonnable dans le résultat de convergence qui suit, si elle permet d'obtenir la condition de Zoutendijk (6.19). La règle d'Armijo (algorithme 6.3) avec le pas initial $\alpha_k^1 = 1$ est souvent utilisée. Selon le chapitre 6, il faut souvent que φ soit $\mathcal{C}^{1,1}$ pour que la condition de Zoutendijk soit vérifiée par les recherches linéaires qui y sont étudiées. Dès lors, exiger la condition de Zoutendijk cache une hypothèse de régularité sur F.

Proposition 10.10 (convergence globale de Newton inexact) *Considérons l'algorithme de Newton inexact 10.8, avec une recherche linéaire vérifiant la condition de Zoutendijk (6.19), et supposons qu'il génère une suite $\{x_k\}$ telle que le conditionnement $\kappa_2(F'(x_k))$ soit borné. Alors*
1) $\nabla\varphi(x_k) \to 0$,
2) *si, de plus, la suite $\{F'(x_k)^{-1}\}$ est bornée, alors $F(x_k) \to 0$.*

Démonstration. Comme $\varphi(x_k)$ est bornée inférieurement, la proposition 6.8 montre que (6.21) a lieu. Le point 1 sera démontré si l'on prouve que le cosinus de l'angle θ_k entre $\nabla\varphi(x_k) = F'(x_k)^{\mathsf{T}} F(x_k)$ et $-d_k$ est uniformément positif. On a par (10.11)

$$\|d_k\|_2 \leqslant \|F'(x_k)^{-1}\|_2 \, \|F'(x_k) d_k\|_2 \leqslant (1 + \eta_k) \|F'(x_k)^{-1}\|_2 \, \|F(x_k)\|_2.$$

Dès lors, si C est une borne sur $\kappa_2(F'(x_k))$, on a

$$\cos\theta_k = \frac{-\nabla\varphi(x_k)^{\mathsf{T}} d_k}{\|\varphi(x_k)\|_2 \, \|d_k\|_2} \geqslant \frac{1 - \eta_k}{1 + \eta_k} \frac{1}{\|F'(x_k)\|_2 \, \|F'(x_k)^{-1}\|_2} \geqslant \frac{1 - \eta}{2C}.$$

Pour le point 2, on déduit de $\nabla\varphi(x_k) = F'(x_k)^{\mathsf{T}} F(x_k) \to 0$ et du caractère borné de $\{F'(x_k)^{-1}\}$ que

$$\|F(x_k)\|_2 \leqslant \|F'(x_k)^{-\mathsf{T}}\|_2 \, \|F'(x_k)^{\mathsf{T}} F(x_k)\|_2 \to 0. \qquad \square$$

Proposition 10.11 (convergence locale de Newton inexact) *On suppose que F a un zéro x_*, que F est de classe C^1 dans un voisinage Ω de x_* et que $F'(x_*)$ est inversible. On considère l'algorithme de Newton inexact 10.8, avec pas unité. Alors ...*

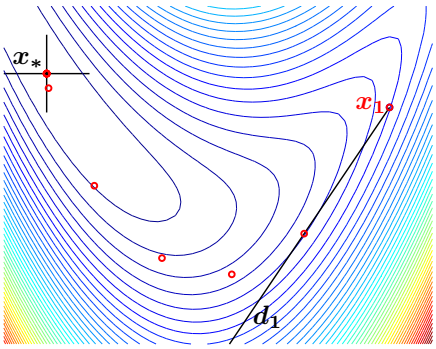
DÉMONSTRATION. □

D'après la proposition 10.10, tout se passe de manière très satisfaisante si $F'(x_k)$ forme une suite bornée d'inverses bornés (il ne suffit pas que $F'(x_k)$ soit inversible, comme le montrera l'exemple 10.16). Cette hypothèse est vérifiée dans beaucoup de problèmes, mais il n'est pas difficile de la violer et de piéger l'algorithme 10.8. C'est ce que l'on s'attache à mettre en évidence dans les exemples ci-dessous, qui sont des variations autour de l'exemple très favorable suivant.

Exemple 10.12 On considère la fonction $F : \mathbb{R}^2 \to \mathbb{R}^2$ définie par

$$F(x) = \begin{pmatrix} x_1 \\ -(x_1-2)^2 + x_2 + 4 \end{pmatrix}$$

$$F'(x) = \begin{pmatrix} 1 & 0 \\ -2(x_1-2) & 1 \end{pmatrix}.$$



La fonction F a un unique zéro en $x_* = 0$. □

La proposition 10.10 a mis en évidence le rôle joué par $F'(x)$. Si l'algorithme 10.8 est utilisé dans l'exemple ci-dessus, on a $\|F(x_k)\|_2 \leqslant \|F(x_1)\|_2$, ce qui implique que $\{x_k\}$ est bornée et donc aussi $\{F'(x_k)\}$ et $\{F'(x_k)^{-1}\}$. D'après la proposition 10.10, $F(x_k) \to 0$ et même $x_k \to x_*$ si l'on reprend le raisonnement de la démonstration du théorème **??**. C'est bien ce que l'on observe dans le tracé à droite ci-dessus, dans lequel on a indiqué les itérés générés par l'algorithme 10.8 avec recherche linéaire inexacte, les courbes de niveaux de φ et la direction de Newton d_1 en x_1.

Exemple 10.13 Dans cet exemple, on modifie le terme x_2 de $F_2(x)$ de l'exemple précédent de manière à introduire une singularité dans F' :

$$F(x) = \begin{pmatrix} x_1 \\ -(x_1-2)^2 + (x_2-1)^2 + 3 \end{pmatrix}, \quad F'(x) = \begin{pmatrix} 1 & 0 \\ -2(x_1-2) & 2(x_2-1) \end{pmatrix}. \tag{10.15}$$

La fonction F a toujours un zéro en $x_* = 0$ (et un autre en $(0, 2)$), mais $F'(x)$ est singulière sur la droite $\mathcal{S} := \{x \in \mathbb{R}^2 : x_2 = 1\}$. □

Évidemment, l'algorithme de Newton n'est pas défini en un $x \in \mathcal{S}$, mais la situation est bien plus délicate que cela. Comme le montre le tracé de gauche à la figure 10.3.1, ce lieu de singularités $\mathcal{S}$ est attrayant pour les itérés, qui peuvent s'y précipiter en

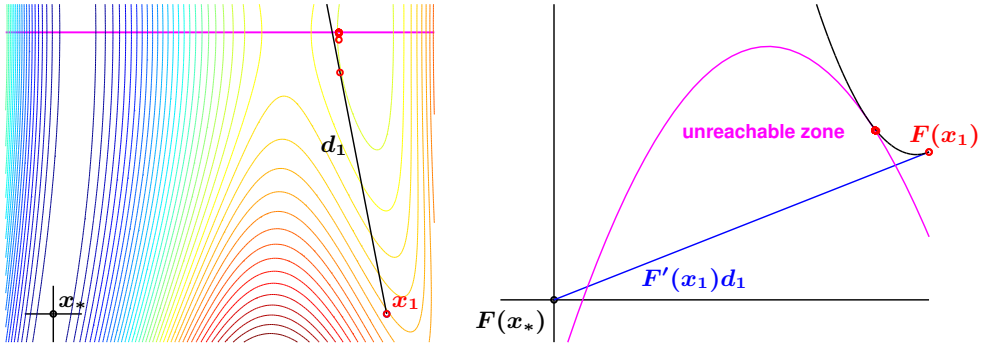


Fig. 10.2. Itérés de Newton avec recherche linéaire pour le système (10.15)

quelques itérations. Les courbes de niveau de φ apportent un premier éclairage sur ce comportement indésirable : les itérés sont piégés dans un bassin de φ, qui ne contient pas x_* et qu'ils ne peuvent quitter car $\varphi(x_k)$ décroît à chaque itération. Plus étrange, $\nabla\varphi(\bar{x}) \neq 0$ au point $\bar{x}$ vers lequel les itérés convergent. Un autre éclairage est apporté par l'observation de l'image $F(x_k)$ des itérés dans l'espace image (tracé de droite). Dans cet exemple, F n'est pas surjective.

Exemple 10.14 Dans ce dernier exemple, on modifie le terme x_2 de $F_2(x)$ de l'exemple 10.12 de manière à rendre $F'(x)$ inversible en tout $x \in \mathbb{R}^2$, mais d'inverse non borné :

$$F(x) = \begin{pmatrix} x_1 \\ -(x_1-2)^2 + e^{x_2} + 3 \end{pmatrix}, \quad F'(x) = \begin{pmatrix} 1 & 0 \\ -2(x_1-2) & e^{x_2} \end{pmatrix}. \tag{10.16}$$

La fonction F a un unique zéro en $x_* = 0$ et $F'(x)$ présente une « singularité à l'infini » (c.-à-d., $\|F'(x)^{-1}\|_2$ explose pour $x_2 \to -\infty$). □

Nous examinons ci-après la question de l'*admissibilité asymptotique du pas unité* par la recherche linéaire dans le voisinage d'un zéro de F. La question que l'on se pose est la suivante. Dans quelles conditions peut-on garantir que le pas $\alpha = 1$ est admis par l'inégalité de décroissance suffisante

$$\varphi(x + \alpha d) \leqslant \varphi(x) + \omega\,\alpha\,\varphi'(x)d, \tag{10.17}$$

si l'itéré courant x est proche d'un zéro de F ? Dans cette inégalité, φ est la fonction de moindres-carrés définie par (10.10). Cette propriété est importante, car elle permet localement à l'algorithme avec recherche linéaire d'avoir la convergence quadratique locale de l'algorithme de Newton. Comme φ n'a pas de lien évident avec l'algorithme de Newton, on ne voit pas trop pourquoi il en serait ainsi. C'est le second miracle du couple Newton-φ (voir la proposition 10.7 pour le premier). Le résultat suivant donne des conditions pour que cette propriété d'admissibilité du pas unité ait lieu.

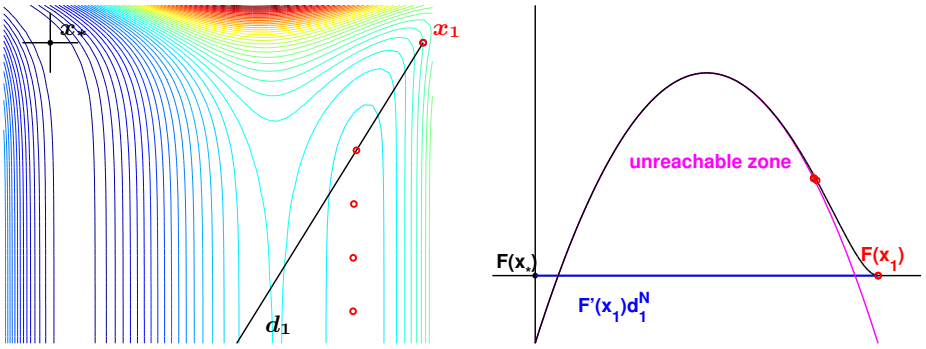


Fig. 10.3. Itérés de Newton avec recherche linéaire pour le système (10.16)

Proposition 10.15 (admissibilité asymptotique du pas unité – Newton) *On suppose que F est de classe C^1 dans le voisinage d'un zéro x_* de F et que $F'(x_*)$ est inversible. Si $\omega \in]0, \frac{1}{2}[$, si x est proche de x_* et si $d := -F'(x)^{-1}F(x)$ est la direction de Newton en x, alors l'inégalité (10.17) est vérifiée avec $\alpha = 1$.*

Démonstration. Si x est suffisamment proche de x_*, $F'(x)$ est inversible (parce que $F'(x_*)$ est inversible par hypothèse, que F' est continue dans un voisinage de x_* par hypothèse et que l'ensemble des opérateurs inversibles est un ouvert, une conséquence du lemme A.2). Il s'ensuit que la direction de Newton d est bien définie en des points proches de x_*. On supposera ci-dessous que les x voisins de x_* considérés sont différents de x_*, car autrement $d = 0$ et (10.17) est trivialement vérifiée.

De l'inversibilité de $F'(x_*)$ et de la proximité de x et x_*, on déduit l'existence d'une constante positive C telle que pour x voisin de x_*, on a

$$\|d\| \leqslant C\|F(x)\|. \tag{10.18}$$

En effet $F'(x_*)d = F'(x)d + o(\|d\|) = -F(x) + o(\|d\|)$ et $2C^{-1}\|d\| \leqslant \|F'(x_*)d\|$. On a noté ici et on notera ci-dessous $o(\|d\|^\sigma)$ une quantité qui dépend de x et telle que pour tout $\varepsilon > 0$, il existe un voisinage V de x_* tel que $o(\|d\|^\sigma)/\|d\|^\sigma < \varepsilon$ lorsque $x \in V \setminus \{x_*\}$ (on observera que $d \neq 0$ lorsque x est voisin et différent de x_*).

On cherche à présent à montrer la négativité de $\varphi(x+d) - \varphi(x) - \omega\,\varphi'(x)d$. En utilisant la dérivabilité de F, on a

$$\|F(x+d) - F(x) - F'(x)d\| \leqslant \left(\sup_{z \in]x, x+d[} \|F'(z) - F'(x)\| \right) \|d\|.$$

Par (10.18), on voit que $d \to 0$ lorsque $x \to \bar{x}$. On déduit déduit alors de l'inégalité précédente que

$$F(x+d) = F(x) + F'(x)d + o(\|d\|) \tag{10.19}$$

où $o(\|d\|)$ désigne un terme tel que $o(\|d\|)/\|d\| \to 0$ lorsque $x \to \bar{x}$ avec $x \neq \bar{x}$ (ce n'est donc pas le « petit o » de la différentiabilité, puisque x est également modifié

dans cette estimation). Pour l'algorithme de Newton, $F(x) + F'(x)d = 0$. On obtient alors les estimatioins suivantes

$$F(x+d) = o(\|d\|), \qquad [(10.19)]$$
$$\varphi(x+d) = \frac{1}{2}\|F(x+d)\|_2^2 = o(\|d\|^2).$$

Comme $\varphi(x) = \frac{1}{2}\|F(x)\|_2^2$ et $\varphi'(x)d = F(x)^\mathsf{T} F'(x)d = -\|F(x)\|_2^2$, on obtient finalement

$$\varphi(x+d) - \varphi(x) - \omega\,\varphi'(x)d = -(\tfrac{1}{2} - \omega)\|F(x)\|_2^2 + o(\|F(x)\|_2^2),$$

où on a aussi utilisé (10.18) poue transformer le $o(\|d\|^2)$ en $o(\|F(x)\|_2^2)$. Comme $\omega < \frac{1}{2}$, le membre de droite est négatif lorsque x est suffisamment proche de x_*. Ceci montre que l'inégalité (10.17) est vérifiée avec $\alpha = 1$ dans ce cas. □

Newton modifié ▲

On modifie la hessienne, pour en construire une approximation auto-adjointe définie positive. Cela peut se faire, soit en en calculant le spectre de $\nabla^2 f(x)$ (opération coûteuse), soit en modifiant sa factorisation de Cholesky.

Les résultats numériques ont montré que ces techniques ne sont pas très robustes (convergence souvent lente, voire inexistante en pratique) et qu'il est préférable d'utiliser l'approche par régions de confiance (section 10.3.2), qui elles aussi modifient la hessienne, mais avec une interprétation géométrique dans l'espace primal claire.

Newton tronqué

Les algorithmes étudiés dans cette section apportent un remède aux inconvénients de l'algorithme de Newton original sur les deux points suivants : (1) les problèmes de consistance et de convergence de la recherche linéaire et (2) le coût de résolution du système linéaire requis à chaque itération de l'algorithme de Newton. L'idée est de résoudre de manière partielle ce système linéaire (d'où le mot *tronqué*), ce qui permettra du même coup d'obtenir une direction de descente de qualité. On suppose toutefois que des dérivées premières de F (ou secondes de f en optimisation) sont évaluées, mais il n'est pas nécessaire de calculer toute la jacobienne $F'(x)$ (toute la hessienne $\nabla^2 f(x)$ en optimisation).

On peut décrire l'*algorithme de Newton tronqué* brièvement, comme suit. C'est une méthode à directions de descente, dans laquelle les directions sont déterminées en résolvant de manière approchée l'*équation de Newton*, qui est l'équation linéaire en $d_k \in \mathbb{R}^n$ suivante

$$H_k d_k = -g_k. \qquad (10.20)$$

On y a noté $H_k := \nabla^2 f(x_k)$ la hessienne de f en x_k et $g_k := \nabla f(x_k)$ son gradient en x_k. L'algorithme fait ensuite de la recherche linéaire le long de d_k pour déterminer un pas $\alpha_k > 0$. Ceci conduit au nouveau point $x_{k+1} := x_k + \alpha_k d_k$.

Ce que l'on vient de décrire est une *itération externe* de l'algorithme. La résolution approchée de l'équation de Newton se fait en général par un processus itératif (le plus souvent il s'agit d'itérations de gradient conjugué) que l'on arrête avant d'avoir trouvé

la solution et dont les itérations sont dites *internes*. Il y a dans ce cas deux processus itératifs imbriqués. On dit que l'algorithme de Newton est tronqué, pour exprimer le fait que le processus interne est interrompu avant convergence. Certains auteurs utilisent le terme « inexact » pour exprimer que (10.20) n'est pas résolue exactement à chaque itération (voir []), mais ce terme peut laisser penser que la méthode n'est pas très précise, ce qui n'est pas le cas.

Cette approche est justifiée par les considérations suivantes. La résolution précise de l'équation de Newton (10.20) peut prendre beaucoup de temps de calcul (pensez au cas où $n = 10^3 \dots 10^6$ et au fait qu'un système linéaire général se résout en $O(n^3)$ opérations), si bien qu'il est tentant d'en calculer une solution approchée à un coût inférieur. D'autre part, si la direction de Newton est bonne près d'une solution, il n'en est pas de même si l'itéré en est éloigné. Il est donc raisonnable de penser que l'on va être plus efficace et réduire le temps de calcul total en résolvant (10.20) grossièrement lorsqu'on est loin de la solution et avec plus de précision lorsqu'on s'en rapproche. En pratique, c'est la stratégie qu'il faut suivre, mais l'on voit que le choix du nombre d'itérations internes à exécuter par itération externe est délicat. C'est le talon d'Achille de la méthode : il faut que l'algorithme « sente » la proximité d'une solution pour bien doser l'effort à faire à chaque itération externe. Ceci demande souvent un réglage qui peut dépendre du problème.

On utilise souvent le gradient conjugué (GC) pour résoudre le système (10.20) de manière approchée et c'est avec ce processus itératif interne que nous présenterons l'algorithme. Par là on cherche à annuler ou à faire décroître le *résidu* (c'est le gradient de la fonction quadratique $\varphi_k(d) = \frac{1}{2} d^\mathsf{T} H_k d + g_k^\mathsf{T} d$)

$$r_k := H_k d_k + g_k.$$

Ceci revient aussi à minimiser partiellement le problème quadratique osculateur (10.7). De plus, l'algorithme du GC est démarré avec $d_k^0 = 0$. Dans ce cas, le résidu initial est $r_k^0 = g_k$ et la première direction de recherche est $-g_k$. Si l'algorithme s'arrête après la première itération, d_k sera approché par une direction parallèle à $-g_k$, si bien que l'algorithme de Newton tronqué se ramène à la méthode de la plus forte pente. D'autre part, plus on fait d'itérations internes, plus l'algorithme de Newton tronqué se rapproche de l'algorithme de Newton. Il s'agit donc d'une méthode intermédiaire entre ces deux extrêmes. Si on suit la stratégie mentionnée ci-dessus, la méthode est proche de l'algorithme du gradient dans les premières itérations et obtient la convergence rapide de l'algorithme de Newton proche de la solution. Cet algorithme converge si f est régulière et si $\{\nabla^2 f(x_k)\}$ reste bornée. Il n'est pas nécessaire que $\{\nabla^2 f(x_k)^{-1}\}$ soit bornée.

Voyons cela de manière plus précise. Soit $\{x_k\}$ la suite générée. L'algorithme a besoin que l'on spécifie deux constantes (indépendantes de l'itération k), $\gamma \in]0, 1[$ et $\omega_1 \in]0, \frac{1}{2}[$, qui sont utilisées dans la recherche linéaire. Ensuite, l'algorithme doit détecter quand est-ce qu'une direction interne générée par le gradient conjugué correspond à une courbure positive de f trop proche de zéro. Ceci se fait au moyen d'une valeur-seuil $\nu = \nu_k > 0$ qui pourra éventuellement être modifiée au cours des itérations externes. On dit alors qu'une direction v est *à courbure quasi-négative* pour f en x si

$$v^\mathsf{T} \nabla^2 f(x) v < \nu \|v\|_2^2. \tag{10.21}$$

Nous pouvons maintenant décrire une itération de l'algorithme, celle qui démarre en $x_k \in \mathbb{R}^n$.

Algorithme 10.16 (Newton tronqué en optimisation) On suppose qu'au début de l'itération k, on dispose d'un itéré $x_k \in \mathbb{R}^n$.

1. *Test d'arrêt.* Si $\nabla f(x_k) \simeq 0$, arrêt de l'algorithme.
2. *Direction.* On calcule d_k par i_k itérations (internes) de gradient conjugué qui démarrent en $d_k^0 := 0$. Pour $j \geqslant 0$:

   2.1. Calcul de la direction conjuguée interne ($r_k^j := H_k d_k^j + g_k$) :

   $$v_k^j := \begin{cases} -r_k^0 & (= -g_k) & \text{si } j = 0 \\ -r_k^j + \beta_k^j v_k^{j-1} & (\beta_k^j := \|r_k^j\|_2^2 / \|r_k^{j-1}\|_2^2) & \text{si } j \geqslant 1. \end{cases}$$

   2.2. Test d'arrêt : on interrompt les itérations internes (et on va au point 3) quand on veut, mais certainement quand $r_k^j = 0$ ou quand la direction interne v_k^i est à courbure « quasi-négative », c'est-à-dire si elle vérifie (10.21) avec $v = v_k^j$ et $\nu = \nu_k$. Dans ce cas, on prend

   $$d_k := \begin{cases} -g_k & \text{si } j = 0 \\ d_k^j & \text{si } j \geqslant 1 \end{cases}$$

   et on passe à l'étape 3.

   2.3. Nouvel itéré interne

   $$d_k^{j+1} := d_k^j + t_k^j v_k^j,$$

   où le pas $t_k^j > 0$ est calculé par la formule habituelle

   $$t_k^j := -\frac{(r_k^j)^\mathsf{T} v_k^j}{(v_k^j)^\mathsf{T} H_k v_k^j}.$$

3. *Calcul du pas.* On calcule un pas $\alpha_k > 0$ par la règle d'Armijo : α_k est le premier nombre (et le plus grand) dans $\{1, \gamma, \gamma^2, \gamma^3, \ldots\}$ tel que l'on ait
   $$f(x_k + \alpha_k d_k) \leqslant f(x_k) + \omega_1 \alpha_k g_k^\mathsf{T} d_k.$$
4. *Nouvel itéré.* $x_{k+1} := x_k + \alpha_k d_k$.

Voici quelques remarques sur l'algorithme.

– Il n'utilise de la hessienne $H_k = \nabla^2 f(x_k)$ que ses produits $H_k v$ par une direction conjuguée v. Il n'est donc pas nécessaire de calculer la hessienne complètement. Une routine qui calcule ces produits suffira. Rappelons que $H_k v$ est la dérivée du gradient en x_k et dans la direction v.

– Le contrôle du nombre d'itérations internes par itération externe est une tâche délicate. Nous avons donné les tests d'arrêt minimal. Comme on peut s'arrêter quand on veut (pour avoir convergence, d'après la proposition 10.17), on peut ajouter d'autres conditions d'arrêt librement. C'est dans ce sens que l'algorithme décrit ci-dessus est dit être dans sa version minimale.

– À la première étape, l'algorithme du GC ne voit pas la non définie positivité éventuelle de H_k, puisque $(v_k^j)^T H_k v_k^j \geqslant \nu_k \|v_k^j\|_2^2$ pour toute direction interne v_k^j acceptée. Il est donc bien défini.

  Si la première direction direction interne du GC, qui n'est autre que $-g_k$, est à courbure quasi-négative, l'algorithme ne la rejette pas (comme c'est le cas dans les itérations internes suivantes), mais la prend : $d_k = -g_k$. Donc, même si $\nabla^2 f(x_k) = 0$, cette étape de l'algorithme est bien définie et fournit $d_k = -g_k$ comme direction de recherche.

– On n'a pas précisé comment choisir le seuil ν_k au cours des itérations externes. C'est clairement un point délicat. Le résultat de convergence ci-dessous autorise plusieurs règles. On peut maintenir ν_k supérieur à un seuil constant $\nu > 0$, mais c'est assez restrictif, car il est difficile de savoir quelle est la bonne valeur de ν. La proposition analyse aussi le cas où ν_k n'est décru que lorsque le pas unité est accepté par la recherche linéaire. On a alors un résultat plus faible ($\liminf \|g_k\| = 0$), mais si l'on maintient ν_k supérieur à un seuil proportionnel à $\|g_k\|^p$ (p étant une constante positive), on retrouve un résultat de convergence satisfaisant ($g_k \to 0$). Cette dernière règle permet à ν_k de décroître dans le voisinage d'une solution, ce qui permet de ne pas empêcher la convergence quadratique de l'algorithme.

  Au lieu de contrôler par ν_k la petitesse des quotients de Rayleigh $v^T H_k v / \|v\|_2^2$ de H_k, on peut aussi contrôler celle de

$$\cos\theta_k := \frac{-g_k^T d_k}{\|g_k\| \, \|d_k\|}$$

  qui, contrairement aux quotients de Rayleigh, a l'élégance de décroître de façon monotone au cours des itérations internes (voir exercice 8.2).

L'algorithme de Newton tronqué permet d'avoir un résultat de convergence relativement fort ($g_k \to 0$), sous la seule condition que les hessiennes $\nabla^2 f(x_k)$ forment une suite bornée. On n'a pas besoin que l'inverse des hessiennes (qui n'existent peut être pas !) forment une suite bornée.

Proposition 10.17 (convergence de Newton tronqué) *Supposons que f soit deux fois dérivable. On considère l'algorithme de Newton tronqué décrit ci-dessus.*

(i) Si $x_k \in \mathbb{R}^n$ n'est pas un point stationnaire de f, la direction d_k est de descente pour f en x_k et l'algorithme est bien défini en x_k.

(*ii*) *Supposons que la suite* $\{f(x_k)\}$ *soit bornée inférieurement, que la suite* $\{\nabla^2 f(x_k)\}$ *soit bornée et qu'aucun itéré* x_k *généré ne soit un point stationnaire de* f.
   (*a*) *Si* ν_k *est maintenu plus grand qu'une constante* $\nu > 0$, *alors* $\nabla f(x_k) \to 0$.
   (*b*) *Si* ν_k *n'est décru que si le pas unité est accepté par la recherche linéaire à l'étape* k, *alors* $\liminf_{k\to\infty} \|\nabla f(x_k)\| = 0$.
   (*c*) *Si* ν_k *n'est décru que si le pas unité est accepté par la recherche linéaire à l'étape* k *et si* ν_k *vérifie*

$$\nu_k \geqslant \nu \|\nabla f(x_k)\|^p, \tag{10.22}$$

où $\nu > 0$ *et* $p \geqslant 0$ *sont des constantes, alors* $\nabla f(x_k) \to 0$. *Le même résultat a lieu si l'on prend* $\cos\theta_k$ *au lieu de* ν_k *dans* (10.22).

Démonstration. Commençons par donner une formule de la direction d_k. Pour cela, on constate que si $i_k \geqslant 1$, on a pour $0 \leqslant j \leqslant i_k$

$$(v_k^j)^\mathsf{T} H_k d_k^j = (v_k^j)^\mathsf{T} H_k \left(\sum_{l=0}^{j-1} t_k^l v_k^l \right) = 0,$$

parce que les directions internes v_k^l d'indices l différents sont conjuguées. Dès lors $(v_k^j)^\mathsf{T} r_k^j = (v_k^j)^\mathsf{T}(H_k d_k^j + g_k) = (v_k^j)^\mathsf{T} g_k$. On en déduit que

$$d_k = \sum_{j=0}^{i_k-1} t_k^j v_k^j = -\sum_{j=0}^{i_k-1} \frac{v_k^j (v_k^j)^\mathsf{T} g_k}{(v_k^j)^\mathsf{T} H_k v_k^j} = -J_k g_k,$$

où J_k est la matrice semi-définie positive de rang i_k donnée par la formule

$$J_k := \sum_{j=0}^{i_k-1} \frac{v_k^j (v_k^j)^\mathsf{T}}{(v_k^j)^\mathsf{T} H_k v_k^j}.$$

Si $i_k = 0$, on a aussi $d_k = -J_k g_k$, avec cette fois $J_k = I$.

On voit alors facilement que, si x_k n'est pas stationnaire, d_k est une direction de descente de f en x_k. En effet, si $i_k = 0$, $g_k^\mathsf{T} d_k = -\|g_k\|_2^2$. Si $i_k \geqslant 1$, en utilisant le fait que $(v_k^j)^\mathsf{T} H_k v_k^j > 0$ et que $v_k^0 = -g_k$, on a

$$g_k^\mathsf{T} d_k = -\sum_{j=0}^{i_k-1} \frac{(g_k^\mathsf{T} v_k^j)^2}{(v_k^j)^\mathsf{T} H_k v_k^j} \leqslant -\frac{(g_k^\mathsf{T} v_k^0)^2}{(v_k^0)^\mathsf{T} H_k v_k^0} = -\frac{\|g_k\|_2^4}{g_k^\mathsf{T} H_k g_k} \leqslant -\frac{\|g_k\|_2^2}{\|H_k\|_2}.$$

En rassemblant les deux cas:

$$g_k^\mathsf{T} d_k \leqslant -\min\left(1, \frac{1}{\|H_k\|_2}\right) \|g_k\|_2^2. \tag{10.23}$$

Donc $g_k^\mathsf{T} d_k < 0$ si $g_k \neq 0$.

Supposons à présent que, pour tout $k \geqslant 1$, $\nu_k \geqslant \nu$, où $\nu > 0$ est une constante. Montrons qu'il existe une constante $C > 0$ telle que

$$f(x_{k+1}) \leqslant f(x_k) - C\|g_k\|^2. \tag{10.24}$$

La convergence de $g_k \to 0$ s'en déduit du fait que $\{f(x_k)\}$ est décroissante et bornée inférieurement. D'après la proposition 6.11, il existe une constante $C_1 > 0$ telle que $\forall k \geqslant 1$, on ait soit

$$f(x_{k+1}) \leqslant f(x_k) - C_1|g_k^{\mathsf{T}} d_k|, \tag{10.25}$$

soit

$$f(x_{k+1}) \leqslant f(x_k) - C_1\|g_k\|^2 \cos^2\theta_k, \tag{10.26}$$

où $\cos\theta_k = -(g_k^{\mathsf{T}} d_k)/(\|g_k\|_2\|d_k\|_2)$. Si la première inégalité (10.25) a lieu, on a par l'estimation (10.23) de $g_k^{\mathsf{T}} d_k$:

$$f(x_{k+1}) \leqslant f(x_k) - C_1 \min\left(1, \frac{1}{\|H_k\|_2}\right)\|g_k\|_2^2. \tag{10.27}$$

On en déduit (10.24) du fait que $\{H_k\}$ est supposée bornée. Supposons à présent que la seconde inégalité (10.26) ait lieu. Notons d'abord que $\|uu^{\mathsf{T}}\|_2 = \|u\|_2^2$ et que $(v_k^j)^{\mathsf{T}} H_k v_k^j \geqslant \nu_k \|v_k^j\|_2^2$ pour tout $j = 0, \ldots, i_k$. Dès lors $\|J_k\|_2 \leqslant \max(1, n\nu_k^{-1})$ et par (10.23)

$$\cos\theta_k = \frac{-g_k^{\mathsf{T}} d_k}{\|g_k\|_2\|d_k\|_2} \geqslant \min\left(1, \frac{1}{\|H_k\|_2}\right)\frac{\|g_k\|_2}{\|d_k\|_2} \geqslant \min\left(1, \frac{1}{\|H_k\|_2}\right)\min\left(1, \frac{\nu_k}{n}\right).$$

La suite $\{H_k\}$ étant supposée bornée, le cosinus de θ_k est uniformément positif et on obtient également (10.24).

Si ν_k n'est décru que lorsque le pas unité est accepté par la recherche linéaire, deux cas peuvent se présenter. Soit $\liminf \nu_k > 0$ et on est ramené au point $(ii\text{-}a)$, selon lequel $g_k \to 0$. Soit il existe une sous-suite d'itérés pour lesquels le pas unité est accepté. Pour les indices k correspondants, on a (10.25) avec $C_1 = \omega_1$, donc (10.27), et du fait que $\{H_k\}$ est bornée, cela implique que $g_k \to 0$ pour les indices k considérés.

Considérons pour terminer le cas où ν_k n'est décru que lorsque le pas unité est accepté par la recherche linéaire et où (10.22) a lieu (éventuellement avec $\cos\theta_k$ au lieu de ν_k). On sait déjà que $\liminf \|g_k\| = 0$. Si toute la suite $\{g_k\}$ ne converge pas vers zéro, on peut trouver une constante $\gamma > 0$ et une suite d'indices $\{l_k\}_{k\geqslant 0}$ strictement croissante telle que pour tout $k \geqslant 0$:

$$\|g_{l_{2k}}\| \geqslant \gamma \quad \text{et} \quad \|g_{l_{2k+1}}\| \leqslant \gamma/2.$$

Pour $l_{2k} \leqslant l < l_{2k+1}$, en utilisant la borne inférieure sur $\cos\theta_k$ ci-dessus et (10.22), on obtient

$$f(x_{l+1}) \leqslant f(x_l) - \omega_1\|g_l\|\|s_l\|\cos\theta_l \leqslant f(x_l) - C\|s_l\|,$$

où $s_l = x_{l+1} - x_l$ et $C > 0$ est une constante indépendante de k et de l. On en déduit

$$\|x_{l_{2k+1}} - x_{l_{2k}}\| \leqslant \sum_{l=l_{2k}}^{l_{2k+1}-1}\|s_l\| \leqslant \frac{1}{C}\Big(f(x_{l_{2k}}) - f(x_{l_{2k+1}})\Big).$$

Dès lors $\|x_{l_{2k+1}} - x_{l_{2k}}\| \to 0$ et, par l'uniforme continuité de ∇f, $\|g_{l_{2k+1}} - g_{l_{2k}}\| \to 0$, ce qui contredit le fait que $\|g_{l_{2k+1}} - g_{l_{2k}}\| \geqslant \gamma/2$. □

10.3.2 Régions de confiance ▲

Cette globalisation de la convergence offre plus de robustesse (résultats de convergence meilleurs, moins de problème à l'utilisation), mais elle n'est pas toujours utilisable pour résoudre des systèmes non linéaires de très grande taille.

Systèmes non linéaires

Présenter l'algortithme classique qui minimise $\|F(\cdot)\|$ ou $\frac{1}{2}\|F(\cdot)\|_2^2$ avec le pas de Cauchy $-\alpha^c F'(x)^\top F(x)$. Discuter des méthodes avec résolution directe ou itérative du système de Newton. Pour les résolutions itératives, discuter des méthodes qui permettent d'avoir la croissance du pas de Newton approché au cours des itérations internes.

Cet algorithme classique a l'inconvénient de requérir le calcul du produit de la *transposée* de la jacobienne par un vecteur pour estimer le pas de Cauchy. Ceci peut être un inconvénient majeur pour les grands problèmes dans lesquels les produits jacobienne-vecteur sont estimés par différences finies. Le seul algorithme n'utilisant pas la transposée de la jacobienne semble être celui de Brown et Saad [95 ; 1990], mais sa convergence n'est pas démontrée.

Optimisation

Tout un chapitre est consacré à cette méthode importante (le chapitre 9). Mentionnons seulement ici son principe.

10.3.3 Autres méthodes

Nous évoquons succintement dans cette section d'autres approches de globalisation de la convergence, sans en étudier lleur convergence.

Réduction du pas de temps dans la résolution d'équations différentielles

Supposons que l'on cherche à résoudre l'équation différentielle

$$\frac{\mathrm{d}x}{\mathrm{d}t} + \phi(x) = 0, \quad x(0) = x_0, \tag{10.28}$$

où l'état initial $x_0 \in \mathbb{R}^n$ est donné et $\phi : \mathbb{R}^n \to \mathbb{R}^n$ est une fonction non linéaire.

Un schéma de discrétisation en temps de (10.28) est *implicite* si l'état (approché) x_{i+1} au temps $t_{i+1} > 0$ est solution d'une équation faisant intervenir $\phi(x_{i+1})$. Ainsi, dans le *schéma d'Euler implcite*, x_{i+1} est solution de l'équation non linéaire en x suivante :

$$\frac{x - x_i}{\delta_i} + \phi(x) = 0, \tag{10.29}$$

où $\delta_i := t_{i+1} - t_i > 0$ est un *pas de temps* que l'on se donne.

On cherche parfois une solution de l'équation non linéaire (10.29) par des itérations de Newton. Si δ_i est petit, l'*état précédent* x_i ou l'*état prédit*

$$x_i - \phi(x_i)\,\delta_i \quad \left[= x_i + \frac{\mathrm{d}x}{\mathrm{d}t}(t_i)\,\delta_i\right]$$

sont en général de bons points de départ pour ces itérations. Si la solution de l'équation différentielle (10.28) dépend continûment du temps, ces points de départ seront d'autant meilleurs que δ_i est petit (on calcule alors une approximation de $x(t_i+\delta_i)$, qui dépend du choix de δ_i). Un moyen d'obtenir la convergence des itérés de Newton est de prendre un pas de temps δ_i suffisamment petit et de le réduire si la convergence ne se produit pas en quelques itérations (10 par exemple).

Méthode du régime pseudo-transitoire

Bien que rappelant la section précédente par certains aspects, l'approche décrite ici est bien différente. L'idée est de chercher à calculer un zéro de F comme un *état stationnaire* (c.-à-d., ne dépendant pas du temps) de l'équation différentielle

$$\frac{\mathrm{d}x}{\mathrm{d}t} + F(x) = 0, \quad x(0) = x_0. \tag{10.30}$$

On constate en effet qu'il y a une bijection entre les états stationnaires de (10.30) et les zéros de F. L'équation (10.30) rappelle l'équation différentielle (10.28), mais elle est introduite ici de manière artificielle. De plus, on n'est pas intéressé ici par l'évolution de l'état $x(t)$, la solution de (10.30), au cours du temps fictif t, mais seulement par l'état asymptotique, lorsque $t \uparrow \infty$.

Observons que l'approche du régime pseudo-stationnaire n'est pas symétrique dans le sens suivant. Si l'équation non linéaire $F(x) = 0$ ne change pas si on remplace F par $-F$ (on garde les mêmes zéros), l'équation différentielle (10.30) est sensible au fait de remplacer F par son opposé. Par exemple, si F est donnée par (10.8) et $x_0 = 0$, on a $x(t) < 0$ pour tout $t > 0$ et la trajectoire s'écarte de l'unique état stationnaire $x_* = 1$ lorsque t augmente ; par contre si on change le signe de F, la trajectoire se dirige vers $x_* = 1$ lorsque $t \uparrow \infty$. Techniquement et pratiquement, la convergence de l'approche du régime transitoire ne pourra être garantie que si l'on peut faire l'hypothèse que la trajectoire issue de x_0 converge vers un zéro de F lorsque $t \uparrow \infty$.

Voici la méthode. Dans un premier temps, on discrétise l'équation différentielle (10.30) par un *schéma d'Euler implcite* : en l'itéré x_k ($k \geqslant 0$), on s'intéresse à la solution de l'équation non linéaire

$$\frac{x - x_k}{\delta_k} + F(x) = 0, \tag{10.31}$$

où $\delta_k > 0$ est un *pas de temps*. Le plus souvent, l'itéré suivant est obtenu en faisant une unique itération de Newton pour résoudre cette équation, ce qui conduit à prendre x_{k+1} qui vérifie

$$F(x_k) + \left[\delta_k^{-1} I + F'(x_k)\right](x_{k+1} - x_k) = 0.$$

Si $\delta_k^{-1} I + F'(x_k)$ est inversible, on obtient

$$x_{k+1} = x_k - \left[\delta_k^{-1} I + F'(x_k)\right]^{-1} F(x_k).$$

On retrouve l'algorithme de Newton lorsque $\delta_k = \infty$. Il est coutumier de choisir les pas de temps par des variantes de la règle suivante [446, 608, 352]

$$\delta_k = \frac{\|F(x_{k-1})\|}{\|F(x_k)\|}\, \delta_{k-1},$$

qui fait croître δ_k autant que $\|F(x_k)\|$ décroît. On peut aussi plafonner le pas de temps δ_k par $\delta_{\max} > 0$ si la valeur donnée par la formule précédente dépasse le seuil fixé $\delta_{\max}$ [337] ou le prendre infini dans les mêmes circonstances [195].

Il faut noter que dans l'approche du régime transitoire la suite $\{\|F(x_k)\|\}$ n'est pas nécessairement décroissante, ce qui permet parfois d'éviter les minima locaux de $\|F(\cdot)\|$, une propriété que n'ont pas la recherche linéaire et les régions de confiance.

Des conditions de convergence de cette technique sont données dans [351].

Méthodes de continuation

Les *méthodes de continuation* peuvent constituer une approche intéressante lorsque l'équation non linéaire à résoudre $F(x) = 0$ contient un paramètre $p \in \mathbb{R}$ qui peut atténuer la difficulté du problème lorsqu'on change sa valeur. De façon plus précise, l'équation originale correspond à la valeur $p = p_1$ du paramètre :

$$\forall x \in \mathbb{R}^n : \quad F(x) = \Phi(x, p_1),$$

où $\Phi : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$, et le système

$$\Phi(x, p) = 0 \tag{10.32}$$

est « facile » à résoudre lorsque $p = p_0$ (que l'on va supposer $< p_1$). On note x_0 une solution (approchée) de (10.32) avec $p = p_0$. Un exemple typique en mécanique des fluides est celui des équations de Navier-Stokes, dans lesquelles le nombre de Reynolds peut jouer le rôle du paramètre p ci-dessus.

La version la plus simple des méthodes de continuation suppose qu'il existe une fonction implicite $p \in \mathbb{R} \mapsto x(p)$ telle que $\Phi(x(p), p) = 0$ pour tout $p \in [p_0, p_1]$ et elle cherche à suivre approximativement le chemin $x([p_0, p_1])$ en faisant croître progressivement p de p_0 à p_1. Connaissant une solution approchée x_i de (10.32) avec $p = p_i$ (indice i fractionnaire dans $[0, 1]$), un bon point de départ pour calculer une solution de (10.32) avec $p = p_{i+1}$ est obtenu par la *phase de prédiction* suivante :

$$x^0_{i+1} = x_i + x'(p_i)(p_{i+1} - p_i).$$

On obtient la dérivée $x'(p)$ de la fonction implicite $p \mapsto x(p)$ en différentiant l'identité $\Phi(x(p), p) \equiv 0$ par rapport à p, ce qui donne

$$x'(p_i) = -\left(\frac{\partial \Phi}{\partial x}(x_i, p_i)\right)^{-1} \frac{\partial \Phi}{\partial p}(x_i, p_i).$$

On peut alors calculer le point de prédiction x^0_{i+1} à partir duquel quelques itérations de Newton sur le système $\Phi(\cdot, p_{i+1}) = 0$ permettent de trouver x_{i+1}.

Le problème est plus compliqué si, le long du chemin suivi, on rencontre des points de bifurcation, de rebroussement, *etc.* Comme points d'entrée sur ce sujet à peine ébauché, citons [367, 616, 10, 11, 129].

Notes

Isaac Newton (1642-1727) était intéressé par le calcul de zéro de polynôme et sa méthode, exposée dans l'épigraphe de ce chapitre, était sensiblement différente de

l'algorithme de Newton tel que nous le connaissons aujourd'hui, celui présenté à la section 10.1.1. Même si les itérés générés sont identiques dans les deux approches, celle de Newton ne s'étend pas aisément aux fonctions non polynomiales. Le texte donné en épigraphe est tiré de *Methodus fluxionum et serierum infinitorum*, qui fut écrit en latin entre 1664 et 1671, édité en anglais en 1736 ; l'algorithme fut également exposé dans *De analysi per aequationes numero terminorum infinitas*, ouvrage composé en 1669 mais seulement publié en 1711. [112]

On associe souvent le nom de *Joseph Raphson* (peut-être 1648-1712) à celui de Newton pour nommer l'algorithme 10.1. Raphson s'intéressait aussi au calcul de zéro de polynôme. Sa contribution, qui date de 1690 et 1697 [506], a été d'écrire l'algorithme sous la forme $x_{k+1} = \varphi(x_k)$, où la fonction rationnelle φ est construite à partir du polynôme considéré, mais sans faire intervenir sa dérivée. [362, 639]

Les apports de *Thomas Simpson* (1710-1761) à l'algorithme 10.1 ont trop souvent été oubliés. Ils furent pourtant essentiels. On peut en citer trois. Le premier est d'avoir fait intervenir la dérivée de la fonction (qu'il appelle *fluxion*, comme Newton) dans le calcul du nouvel itéré, permettant ainsi d'appliquer l'algorithme à des fonctions non polynomiales, ce qu'il fit [553 ; 1740, p. 83-84]. Sa seconde contribution est d'avoir montré comment on pouvait utiliser l'algorithme pour résoudre un système de 2 équations à 2 inconnues, en résolvant un système linéaire dont la matrice est la jacobienne de la fonction en l'itéré courant [553 ; 1740, p. 82]. Enfin, il donne sans doute le premier exemple de maximisation d'une fonction de plusieurs variables sans contrainte, par recherche d'un zéro de son gradient [552 ; 1737]. [639]

Ypma [639] attribue l'absence de reconnaissance aux autres contributeurs à l'algorithme de Newton au livre influent de Fourier [220 ; 1831], lequel l'appelait la *méthode newtonienne*, sans faire référence à Raphson ou Simpson.

En 1939, Kantorovitch [341] a présenté un résultat préliminaire de convergence de l'algorithme de Newton, qu'il améliora substantiellement en 1948/49 [342, 343]. La version du théorème proposée (théorème 10.3, [169, 345]) est parfois qualifiée d'*invariante par transformation linéaire* (« affine invariant »), parce qu'elle est invariante lorsqu'on pré-compose F avec une application linéaire bijective (F devient $F \circ A$, avec A linéaire inversible), comme l'est l'algorithme de Newton (proposition 10.4). Les hypothèses du théorème de Kantorovitch sont plus fortes que celles d'autres théorèmes d'existence de zéro, tels que ceux de Miranda, de Moore, de Borsuk [5, 4] et d'autres théorèmes de point fixe, mais elles donnent aussi plus d'informations, à savoir la convergence des itérés de Newton et donc un moyen numérique de calculer le point fixe. Pour une revue de l'évolution de l'analyse de la convergence de l'algorithme de Newton, on pourra consulter [631].

Le premier exemple de système non linéaire $F(x) = 0$ pour lequel l'algorithme de Newton *avec recherche linéaire* génère des points convergeant vers un point singulier de F' qui n'est ni un zéro de F ni un point stationnaire de $\|F(\cdot)\|_2^2$ est dû à Powell [487 ; 1970]. Cet exemple a motivé l'introduction des méthodes à régions de confiance pour globaliser l'algorithme de Newton pour ces problèmes, approche qui ne présente pas le même inconvénient. En optimisation aussi, l'algorithme de Newton avec recherche linéaire (celle de Wolfe par exemple) présente le même type de défaut : il peut générer des itérés x_k convergeant vers un point où le gradient n'est pas nul, alors que la hessienne est définie positive en tout itéré [412 ; 2008].

Le comportement de l'algorithme de Newton pour résoudre un système d'équations non linéaires dont la jacobienne est singulière en la solution a souvent été exploré, notamment en dimension un [638]. La revue de Griewank [285 ; 1985] considère le cas multidimensionnel et présente quelques modifications de la méthode de Newton pour faire face aux problèmes de convergence et de stabilité numérique que cette singularité entraîne ; mentionnons une technique de *sur-relaxation*, dans laquelle $x_{k+1} = x_k + \alpha_k d_k$, où d_k est la direction de Newton et le pas α_k est pris dans l'intervalle $[1, 2[$. Une autre possibilité, explorée par Schnabel et ses collaborateurs [81 ; 1998], sont les méthodes dites *tensorielles*, dans lesquelles on ajoute à l'approximation linéaire de F, quelques termes d'ordre 2 (des tenseurs), qui sont approchés par des techniques quasi-newtoniennes.

Les *méthodes de Newton inexactes* ont été beaucoup étudiées, car elles sont très utilisées pour résoudre les grands systèmes non linéaires issus de la discrétisation d'équations aux dérivées partielles. L'article fondateur est [162 ; 1982] et on trouvera de nombreux articles de synthèse et de monographies sur cette question (par exemple [356]). Pour la proposition 10.9, nous avons repris les arguments de [184], eux-mêmes inspirés de [472, 301], qui considèrent la situation plus complexe d'équation non lisse. L'*algorithme de Newton tronqué* décrit à la section 10.3.1, qui s'inscrit dans la veine des méthodes inexactes, est dû à Dembo et Steihaug [163 ; 1983].

L'effet de l'arithmétique flottante sur l'algorithme de Newton a été étudié par divers auteurs ; citons Dennis et Walker [167 ; 1984] et Tisseur [587 ; 2001].

L'extension de l'algorithme de Newton à la résolution du système d'équations non linéaires $F(x) = 0$ dans lequel F est *non différentiable* s'est faite suivant plusieurs directions. Le cas des fonctions C^1 par morceaux est analysé par Kojima et Shindo [360 ; 1986] qui montrent que la convergence quadratique locale est préservée par l'algorithme qui utlise une quelconque des jacobiennes des fonctions actives au point courant, pourvu que soient vérifiées des hypothèses naturelles (au vu du théorème 10.2) incluant l'inversibilité des jacobiennes des fonctions actives en la solution ; nous ne connaissons pas de résultat de convergence globale pour cet algorithme. On a ensuite étudié le cas fréquemment rencontré des *fonctions B-différentiables*, qui sont celles qui vérifient l'estimation (C.9) des fonctions Fréchet-différentiables, mais avec une application $h \mapsto Lh \equiv F'(x)h$ qui n'est plus que positivement homogène de degré 1 (on perd la linéarité). Des résultats de convergence locale et globale par recherche linéaire peuvent être obtnus [471 ; 1990], mais l'équation de Newton à résoudre à chaque itération, $F(x) + F'(x)d = 0$, est cette fois non linéaire, ce qui complique l'algorithme. Le cas où F est *semi-lisse* a commencé a être exploré par Qi et Sun [503 ; 1993], qui ont proposé un algorithme ne requérant que la résolution d'un système linéaire à chaque itération, ce qui est attractif, mais dont la globalisation de la convergence est plus difficile à mettre au point. On pourra lire sur ce thème la synthèse très complète de Facchinei et Pang [198], qui appliquent les algorithmes présentés à la résolution des *problèmes de complémentarité* ou d'*inéquations variationnelles*.

Une extension de l'algorithme de Newton à la recherche de zéro de fonction non lisse, avec zéro non isolé et en présence de contrainte est proposée dans [197 ; 2014].

Autres monographies sur l'algorithme de Newton : Kelley [348, 349, 350 ; 1995-2003], dont la dernière référence contient de nombreux conseils sur la mise en œuvre et le contrôle de l'algorithme ; Higham [313 ; 2002, § 2.5] donne une analyse d'erreur ; Deuflhard [168 ; 2004] décrit l'utilisation des algorithmes de Newton dans la résolution

de problèmes gouvernés par des équations différentielles ; Dedieu [160 ; 2006] présente la théorie en dimension infinie (avec des résultats de Smale) ; Argyros [18 ; 2008] ; Ulbrich [598 ; 2011] fait une synthèse en dimension infinie sur la méthode de Newton semi-lisse, Izmailov and Soldov [330 ; 2014] traitent des problèmes d'optimisation et d'inéquations variationnelles.

Exercices

10.1. Montrez qu'en un point non stationnaire, lorsqu'elle est bien définie, la direction de Newton pour minimiser f est une direction de descente de $x \mapsto \varphi(x) := \|\nabla f(x)\|$, où $\|\cdot\|$ est une norme quelconque.

Remarque. On sait qu'au contraire la direction de Newton n'est pas nécessairement une direction de descente de f en un point éloigné d'une solution forte. On pourrait donc penser qu'il est préférable de globaliser la méthode de Newton en cherchant à minimiser φ. Il n'en est rien. Cela vient du fait que φ est moins bien conditionnée que f (pensez au cas quadratique), si bien que loin d'une solution le pas accepté par φ peut être très petit, au point d'empêcher tout progrès significatif vers la solution.

10.2. On considère l'algorithme de Newton pour résoudre l'équation non linéaire $F(x) = 0$, où $F : \mathbb{R}^n \to \mathbb{R}^n$ est $\mathcal{C}^{1,1}$ dans le voisinage de x_*, une solution telle que $F'(x_*)$ soit inversible. Pour mesurer le progrès vers la solution, on utilise la fonction $x \mapsto \varphi(x) := \|F(x)\|$, où $\|\cdot\|$ est une norme quelconque. Montrez que le pas unité le long de la direction de Newton $d = -F'(x)^{-1}F(x)$ est accepté localement par l'inégalité d'Armijo : si x est voisin de x_* et $\omega \in]0, 1[$, on a $\varphi(x + d) \leqslant \varphi(x) + \omega\varphi'(x; d)$.

Montrez qu'il en est de même si $\varphi(x) = \|F(x)\|^p$, avec $p \geqslant 1$ (norme arbitraire).

10.3. *Lignes de flux de Newton* [306]. On considère le problème de la minimisation d'une fonction $f : \mathbb{R}^n \to \mathbb{R}$. Les *lignes de flux de Newton* sont les courbes $t \mapsto x(t)$, solutions de l'équation différentielle

$$\dot{x} = -H(x)^{-1}g(x), \qquad x(0) = x_0,$$

où $\dot{x}$ désigne la dérivée de $x(\cdot)$ par rapport à t, $H(x) := \nabla^2 f(x)$ est supposé inversible aux points visités $x(t)$, $g(x) := \nabla f(x)$ et x_0 est une condition initiale arbitraire. Ces courbes ont des propriétés remarquables dont certaines sont aisées à vérifier.

1) Le gradient le long d'une ligne de flux vérifie $g(x(t)) = e^{-t}g(x_0)$ et donc, si $g(x_0) \neq 0$, le gradient normalisé $g(x)/\|g(x)\|$ y est constant (la norme $\|\cdot\|$ est arbitraire).

Soient $x_0 \in \mathbb{R}$ et $\mathcal{N}_0 := \{x \in \mathbb{R}^n : f(x) \leqslant f(x_0)\}$. On suppose à présent que $H(x) \succ 0$ pour tout $x \in \mathcal{N}_0$ et qu'il existe un point $x_* \in \mathcal{N}_0$ tel que $\nabla f(x_*) = 0$.

2) Montrez que $\lim_{t\to\infty} x(t) = x_*$.
3) Montrez que $\lim_{t\to\infty} e^t\dot{x}(t) = -H(x^*)^{-1}g(x_0)$.
4) Montrez que l'application $x \in \mathcal{N}_0 \mapsto g(x)$ est injective.

On se donne à présent une autre fonction $\tilde{f} : \mathbb{R}^n \to \mathbb{R}$ telle que $\nabla^2\tilde{f}(x) \succ 0$ pour $x \in \mathcal{N}_0$ et qui ne diffère de f que sur un ouvert $\Omega := \{x \in \mathbb{R}^n : f(x) \neq \tilde{f}(x)\}$ tel que $x_0 \notin \overline{\Omega}$. On note $\tilde{x} : [0, +\infty[\mapsto \mathbb{R}^n$ la ligne de flux de Newton associée à $\tilde{f}$, issue de $x_0 := \tilde{x}(0)$.

5) Montrez que $\tilde{x}(t) = x(t)$ si $x(t) \notin \Omega$ ou si $\tilde{x}(t) \notin \Omega$.

Conclusion. On a donc le résultat étonnant suivant : le flux de Newton associé à une perturbation modérée $\tilde{f}$ de f (telle que $\nabla^2\tilde{f}(x) \succ 0$) n'est différent de celui associé

à f qu'aux points x tels que $f(x) \neq \tilde{f}(x)$. Dès lors, après la zone perturbée, le flux de Newton recolle au flux de la fonction non perturbée.

11 Algorithmes de quasi-Newton

The key ideas that led me to the development of variable-metric algorithms were 1) to update a metric in the space of gradients during the search for an optimum, rather than waiting until the search was over, and 2) to accelerate convergence by using each updated metric to choose the next search direction.

W.C. Davidon, dans la nouvelle introduction de son article fondateur des méthodes de quasi-Newton, écrit en 1959, mais qui ne fut publié qu'en 1991, lors de la parution du premier numéro de la revue *SIAM Journal on Optimization* [155, 156].

I was able to minimize some nonquadratic functions of 100 variables by Davidon's method in 1962. I mentioned this fact at a meeting that was held at Imperial College then, but this remark was impromptu, it was made in a discussion that included the view from a senior person that 10-variable problems were usually too difficult and I was nervous.

M.J.D. Powell [495 ; 2003].

Dans ce chapitre, on s'intéresse à la résolution du problème d'optimisation sans contrainte

$$\begin{cases} \min\ f(x) \\ x \in \mathbb{R}^n \end{cases}$$

par des algorithmes à directions de descente particuliers. On note $\{x_k\}_{k\geqslant 1}$ la suite des itérés et $g_k = \nabla f(x_k)$ le gradient de f en x_k. Les itérés sont donc générés par la récurrence

$$x_{k+1} = x_k + \alpha_k d_k, \quad k \geqslant 1,$$

où le pas $\alpha_k > 0$ est déterminé par recherche linéaire (section 6.3) et d_k est une direction de descente. Dans les méthodes de quasi-Newton, d_k est de la forme

$$d_k = -M_k^{-1} g_k. \tag{11.1}$$

Si $M_k = \nabla^2 f(x_k)$, on retrouve la méthode de Newton, si bien que le nom donné aux méthodes décrites ci-dessous apparaît naturel. Cependant, tout algorithme ayant une direction de descente de la forme (11.1) ne se retrouve pas pour autant dans cette classe de méthodes. Le qualificatif *quasi-Newton* fait en effet référence à un ensemble

de techniques mises au point à partir des années 1960 permettant de générer les matrices M_k à partir des gradients g_k.

Dans beaucoup de problèmes, on cherche à éviter le calcul des dérivées secondes, pour les raisons suivantes :

- l'évaluation de $\nabla^2 f(x_k)$ ou le produit de cete hessienne par un vecteur comme dans l'algorithme de Newton tronqué (section 10.3.1) peut demander trop de temps de calcul,
- on ne dispose pas toujours des dérivées secondes et leur calcul peut demander un investissement humain trop important, alors que l'on voudrait obtenir un résultat rapidement par des algorithmes peut-être plus lents que l'algorithme de Newton mais ne demandant que le calcul des dérivées premières,
- la fonction peut ne pas être deux fois dérivable,
- on ne dispose pas de place mémoire pour stocker les $O(n^2)$ éléments d'une matrice (c'est la moins bonne des raisons, voir section 10.3.1).

Ce sont des situations dans lesquelles les algorithmes de quasi-Newton sont très utiles. Dans ceux-ci, la matrice M_k dans (11.1) n'est pas égale à $\nabla^2 f(x_k)$, mais est générée par des formules qui cherchent à ce que cette matrice soit proche de la hessienne de f. On parle de *formules de mise à jour*. Celles-ci ne font intervenir que les dérivées premières de f. C'est en utilisant la variation du gradient de f d'une itération à l'autre que ces formules permettent d'engranger de l'information sur la hessienne.

Dans ce chapitre, on se propose d'introduire et d'étudier ces formules de mise à jour et les algorithmes de quasi-Newton qui les utilisent. Voici quelques propriétés de ces algorithmes.

- Localement (proche d'une solution), les algorithmes de quasi-Newton convergent moins rapidement que l'algorithme de Newton. Dans les implémentations correctes, les itérés convergent toutefois q-superlinéairement.
- Chaque itération demande moins de calcul au simulateur que dans l'algorithme de Newton : il ne faut pas évaluer les dérivées secondes.
- Dans leur version standard, ces algorithmes peuvent être utilisés pour un nombre de variables qui n'est pas trop grand, disons $n \leqslant 500$ pour fixer les idées. Cette borne sur n vient d'une part du coup de l'itération (de l'ordre de $O(n^2)$ opérations dans l'optimiseur) et du fait que les algorithmes deviennent plus lents lorsque n augmente. Notons qu'il existe des adaptations de ces algorithmes quasi-newtoniens pouvant être utilisées pour résoudre des problèmes de très grande taille, avec plusieurs millions de variables, par exemple. Ces méthodes sont dites *à mémoire limitée* car elles ne demandent pas que l'on garde M_k en mémoire. L'algorithme ℓ-BFGS de la section 11.2.5 est de ce type. Il est très utilisé.

Hypothèse générale à ce chapitre. Nous développerons la théorie lorsque le produit scalaire que l'on se donne sur $\mathbb{R}^n$ est le produit scalaire euclidien : $(u, v) \mapsto u^{\mathsf{T}} v$. Dans ce cas, $\nabla f(x)$ est le vecteur des dérivées partielles de f. On peut se placer dans un cadre plus général en ne faisant aucune hypothèse sur le produit scalaire $(u, v) \mapsto \langle u, v \rangle$ utilisé. Il faut alors utiliser le produit tensoriel associé $(u, v) \mapsto u \otimes v$, où $(u \otimes v)$ est la matrice d'ordre n définie par

$$(u \otimes v)d := \langle v, d \rangle \, u, \quad \text{pour tout } d \in \mathbb{R}^n.$$

Pour obtenir des formules de mise à jour tenant compte du produit scalaire utilisé (et donc du préconditionnement que cela implique), il suffit le plus souvent de remplacer dans les formules obtenues une matrice de la forme uv^T par la matrice $u \otimes v$. Voir par exemple [244] et ses références pour plus détails.

11.1 Système d'équations

11.1.1 Formules de mise à jour

11.1.2 Convergence linéaire locale

Lemme 11.1 (détérioration bornée)

Démonstration. □

11.2 Optimisation

11.2.1 Formules de mise à jour

Principes

Supposons que M_k soit connue et cherchons quelle valeur donner à M_{k+1} pour que cette matrice approche $\nabla^2 f(x_{k+1})$. On cherche à ce que M_{k+1} soit proche de la hessienne de f pour que l'algorithme hérite des bonnes propriétés de convergence locale de l'algorithme de Newton.

Soient

$$s_k := x_{k+1} - x_k \quad \text{et} \quad y_k := g_{k+1} - g_k,$$

où $g_k = \nabla f(x_k)$ comme d'habitude. Le développement de Taylor avec reste intégral s'écrit

$$y_k = \left(\int_0^1 \nabla^2 f(x_k + t s_k) \, \mathrm{d}t \right) s_k.$$

Si on veut que la nouvelle matrice M_{k+1} approche la hessienne de f, il semble raisonnable de lui imposer de satisfaire l'équation vérifiée par la hessienne *moyenne* $\int_0^1 \nabla^2 f(x_k + t s_k) \, \mathrm{d}t$, à savoir

$$y_k = M_{k+1} s_k. \tag{11.2}$$

Cette relation porte le nom d'*équation de quasi-Newton*. D'autre part, comme la hessienne de f est symétrique, il est normal d'imposer également cette propriété à M_{k+1} :

$$M_{k+1} = M_{k+1}^\mathsf{T}. \tag{11.3}$$

Rien n'assure a priori que cette idée simple d'imposer à M_{k+1} de vérifier des propriétés facilement exprimable de la hessienne moyenne donnera une matrice avec de bonnes propriétés, mais l'expérience a montré que l'inspiration est excellente.

Si $n > 1$, les conditions (11.2) et (11.3) ne suffisent pas pour déterminer M_{k+1} : il y a $n(n+1)/2$ inconnues et n équations seulement. Ce manque de conditions laisse une liberté dont ne se sont pas privé les spécialistes en optimisation numérique. Les années 1960-80 ont vu ainsi naître de nombreuses formules de mise à jour. Les plus importantes en optimisation sont la formule SR1 et la formule de BFGS.

Avant d'introduire ces formules, notons que lorsque $n = 1$, les conditions (11.2) et (11.3) déterminent M_{k+1}, qui est le scalaire y_k/s_k (on suppose $x_{k+1} \neq x_k$). L'algorithme de quasi-Newton avec pas unité associé donne (en décalant les indices d'une unité et en supposant $g_{k-1} \neq g_k$ et $g_k \neq 0$) :

$$x_{k+1} = x_k - \frac{s_{k-1}}{y_{k-1}} g_k \qquad \text{ou} \qquad \frac{x_k - x_{k+1}}{g_k - 0} = \frac{x_{k-1} - x_k}{g_{k-1} - g_k}.$$

On reconnaît l'*algorithme de la sécante*. Les méthodes de quasi-Newton portent d'ailleurs parfois ce nom. La figure 11.1 rappelle ce qu'est l'algorithme de la sécante en

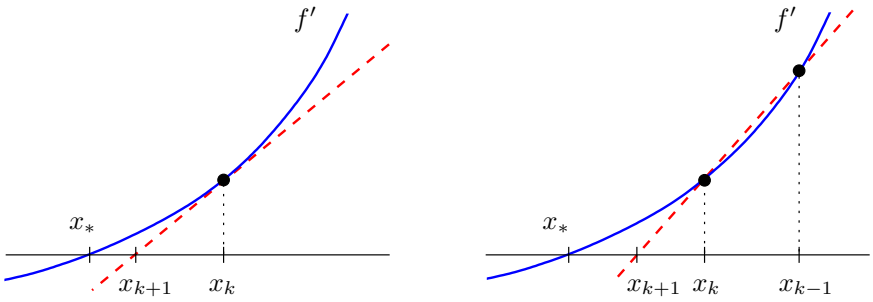


Fig. 11.1. Comparaison des algorithmes de Newton (à gauche) et de quasi-Newton (à droite) lorsque $n = 1$

dimension 1 et le compare à l'algorithme de Newton lorsqu'on cherche à minimiser une fonction $x \in \mathbb{R} \mapsto f(x) \in \mathbb{R}$ en annulant sa dérivée. Dans l'algorithme de Newton (à gauche), le nouvel itéré x_{k+1} est obtenu en calculant l'intersection avec l'axe des abscisses de la droite tangente à la courbe $x \mapsto f'(x)$ en $(x_k, f'(x_k))$ (c'est le zéro de la fonction f' linéarisée en x_k), tandis que dans l'algorithme de la sécante (à droite), on prend l'intersection du même axe avec la droite passant par $(x_{k-1}, f'(x_{k-1}))$ et $(x_k, f'(x_k))$. Dans ce dernier algorithme, on ne doit pas linéariser f' (c'est-à-dire calculer la dérivée seconde de f), mais on doit utiliser la valeur de f' aux deux points x_{k-1} et x_k.

Formule SR1

Une première idée est de rechercher parmi les matrices vérifiant (11.2) et (11.3), une matrice M_{k+1} suffisamment proche de M_k en imposant que la différence entre les deux matrices soit de rang 1. On prend une correction de faible rang de manière à assurer une certaine stabilité à la suite $\{M_k\}$. Il semble en effet raisonnable de souhaiter que ces matrices n'oscillent pas trop au cours des itérations.

On cherche donc une mise à jour de la forme

$$M_{k+1} = M_k + uv^\mathsf{T},$$

où u et v sont deux vecteurs de $\mathbb{R}^n$ à déterminer. Si on veut que M_{k+1} vérifie l'équation de quasi-Newton (11.2), il faut que u et v satisfassent

$$y_k = M_k s_k + (v^\mathsf{T} s_k)u.$$

Si $v^\mathsf{T} s_k = 0$, soit M_k vérifie déjà l'équation de quasi-Newton, soit une correction de rang 1 de M_k ne permet pas d'obtenir une matrice vérifiant cette équation. Si $v^\mathsf{T} s_k \neq 0$, la relation ci-dessus permet de déterminer u, ce qui conduit à

$$M_{k+1} = M_k + \frac{(y_k - M_k s_k)v^\mathsf{T}}{v^\mathsf{T} s_k}.$$

Pour que M_{k+1} soit symétrique (on suppose que M_k l'est), il faut donc prendre $v = y_k - M_k s_k$, ce qui donne

$$\boxed{M_{k+1} = M_k + \frac{(y_k - M_k s_k)(y_k - M_k s_k)^\mathsf{T}}{(y_k - M_k s_k)^\mathsf{T} s_k}.} \tag{11.4}$$

Cette formule porte le nom de *formule SR1* (Symétrique de Rang 1). Elle n'est bien définie que si $(y_k - M_k s_k)^\mathsf{T} s_k \neq 0$. En pratique, cela nécessite l'utilisation de garde-fous corrigeant le dénominateur dans (11.4) s'il est trop petit. La formule SR1 est souvent utilisée lorsqu'il n'est pas possible ou qu'il n'est pas nécessaire que M_{k+1} soit définie positive. Si cette propriété est souhaitable, on recourt à la formule de BFGS décrite ci-après, laquelle a de meilleures propriétés que la formule SR1.

Formule de BFGS

Une autre manière d'imposer à M_{k+1} d'être proche de M_k est de minimiser l'« écart » entre M_{k+1} et M_k, toujours en requérant que M_{k+1} soit symétrique et vérifie l'équation de quasi-Newton (11.2). On est donc conduit à considérer le problème en la variable matricielle $M \in \mathbb{R}^{n\times n}$ suivant :

$$\begin{cases} \min \ \text{« écart »}(M, M_k) \\ y_k = M s_k \\ M = M^\mathsf{T}. \end{cases} \tag{11.5}$$

On dit alors que la matrice est obtenue par une *approche variationnelle*. La fonction « écart » utilisée dans ce problème est spécifiée ci-dessous.

Il est souvent intéressant d'imposer également la définie positivité des matrices M_k. En effet, dans ce cas, d_k donnée par (11.1) est une direction de descente de f en x_k. Cette exigence n'est pas dépourvue de fondement puisque M_k doit approcher $\nabla^2 f(x_*)$ qui est semi-définie positive en la solution et définie positive en des solutions fortes (celles vérifiant les conditions d'optimalité du second ordre).

Pour obtenir des matrices ayant cette propriété de définie positivité, il ne suffit pas d'ajouter cette contrainte au problème (11.5). En effet, le cône $\mathcal{S}^n_{++}$ des matrices symétriques définies positives est un ouvert dans l'espace vectoriel des matrices $\mathbb{R}^{n\times n}$, si bien qu'avec cette contrainte additionnelle, (11.5) peut ne pas avoir de solution.

Imposer la semi-définie positivité (on a alors un fermé) n'est pas satisfaisant non plus, car M_{k+1} ne serait pas nécessairement inversible et d_k ne serait pas bien défini par (11.1). On va donc chercher à ce que ce soit le critère dans (11.5) qui impose la définie positivité de la matrice solution.

Dans ce but, on commence par introduire une fonction $\psi : \mathcal{S}^n \to \mathbb{R} \cup \{+\infty\}$, dont le domaine est $\mathcal{S}^n_{++}$ et qui forme une « barrière » au bord du cône $\mathcal{S}^n_{++}$ (elle tend vers l'infini lorsque son argument se rapproche du bord de $\mathcal{S}^n_{++}$) ainsi qu'à l'infini :

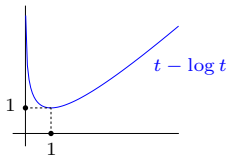
$$\psi(M) = \operatorname{tr} M + \operatorname{ld} M, \tag{11.6}$$

où la fonction log-déterminant $\operatorname{ld} : \mathcal{S}^n \to \mathbb{R} \cup \{+\infty\}$ est définie en $M \in \mathcal{S}^n$ par

$$\operatorname{ld}(M) = \begin{cases} -\log\det M & \text{si } M \in \mathcal{S}^n_{++} \\ +\infty & \text{sinon} \end{cases}$$

Les propriétés annoncées de ψ peuvent se voir sur son expression suivante. Si on note $\{\lambda_i\}$ les valeurs propres de M, on a $\operatorname{tr} M = \sum_{i=1}^n \lambda_i$, $\det M = \prod_{i=1}^n \lambda_i$ et donc

$$\psi(M) = \sum_{i=1}^n (\lambda_i - \log \lambda_i), \quad \text{si } M \in \mathcal{S}^n_{++}. \tag{11.7}$$



Étant donné l'allure de la fonction $t \in \mathbb{R}_{++} \mapsto t - \log t$, $\psi(M)$ tend vers l'infini si l'une des valeurs propres de M tend vers zéro ou vers l'infini.

La formule (11.7) montre aussi que l'unique minimiseur de ψ est la matrice identité ($\lambda_i = 1$ pour tout i). Afin de minimiser l'écart entre M et M_k, on va chercher à ce que $M_k^{-1/2} M M_k^{-1/2}$ soit proche de I. Ceci peut être obtenu en minimisant $\psi(M_k^{-1/2} M M_k^{-1/2})$. On est donc conduit à résoudre le problème suivant sur l'espace vectoriel $\mathcal{S}^n$ des matrices symétriques :

$$\begin{cases} \min\ \psi(M_k^{-1/2} M M_k^{-1/2}) \\ y_k = M s_k \\ M \in \mathcal{S}^n_{++} \qquad \text{(contrainte implicite).} \end{cases} \tag{11.8}$$

Si $s_k = 0$, de deux choses l'une : soit $y_k \neq 0$ et le problème ci-dessus n'a pas de solution (son ensemble admissible est vide), soit $y_k = 0$ et sa solution est M_k (ψ prend alors sa valeur minimale n sur $\mathcal{S}^n_{++}$). Le cas non trivial où $s_k \neq 0$ est examiné dans la proposition suivante.

Proposition 11.2 *Supposons que M_k soit symétrique définie positive et que $s_k \neq 0$. Alors, le problème (11.8) a une solution si, et seulement si, $y_k^{\mathsf{T}} s_k > 0$. Sous cette condition la solution M_{k+1} de (11.8) est unique et est donnée par l'une des formules suivantes :*

$$M_{k+1} = M_k + \frac{y_k y_k^\mathsf{T}}{y_k^\mathsf{T} s_k} - \frac{M_k s_k s_k^\mathsf{T} M_k}{s_k^\mathsf{T} M_k s_k}, \tag{11.9}$$

$$W_{k+1} = \left(I - \frac{s_k y_k^\mathsf{T}}{y_k^\mathsf{T} s_k}\right) W_k \left(I - \frac{y_k s_k^\mathsf{T}}{y_k^\mathsf{T} s_k}\right) + \frac{s_k s_k^\mathsf{T}}{y_k^\mathsf{T} s_k}, \tag{11.10}$$

où on a noté $W_k := M_k^{-1}$ *et* $W_{k+1} := M_{k+1}^{-1}$.

Démonstration. D'abord, il est clair que si (11.8) a une solution M_{k+1}, qui est une matrice symétrique définie positive, on a $y_k^\mathsf{T} s_k = s_k^\mathsf{T} M_{k+1} s_k > 0$. On suppose dorénavant que $y_k^\mathsf{T} s_k > 0$.

Montrons d'abord que le problème (11.8) a au plus une solution. L'ensemble admissible de (11.8) étant clairement convexe, il suffit de vérifier que ψ est strictement convexe, ce que l'on fait en calculant ses dérivées secondes. Rappelons que, pour une matrice inversible A, $\det'(A) \cdot H = (\det A)\,\mathrm{tr}(A^{-1}H)$. Alors, pour tout $M \in \mathcal{S}^n_{++}$ et tout $H \in \mathcal{S}^n$ non nul, on a avec $W := M^{-1}$:

$$\psi'(M) \cdot H = \mathrm{tr}\, H - \mathrm{tr}(WH),$$

$$\psi''(M) \cdot H^2 = \mathrm{tr}(W^{1/2} H W^{1/2})(W^{1/2} H W^{1/2}) > 0.$$

On achève la démonstration en calculant explicitement la solution $M = M_{k+1}$ de (11.8) au moyen des conditions d'optimalité du premier ordre. On travaille dans l'espace vectoriel $\mathcal{S}^n$ des matrices symétriques d'ordre n, muni du produit scalaire $\langle A, B\rangle = \mathrm{tr}\, AB$. On note $\lambda \in \mathbb{R}^n$ le multiplicateur de Lagrange associé à la contrainte affine de (11.8), si bien que le lagrangien s'écrit

$$\ell(M, \lambda) = \psi(W_k^{1/2} M W_k^{1/2}) + \lambda^\mathsf{T}(y_k - M s_k).$$

On cherche $M \in \mathcal{S}^n_{++}$ telle que pour tout $H \in \mathcal{S}^n$ la dérivée directionnelle de ℓ en M dans la direction H soit nulle, c'est-à-dire

$$\mathrm{tr}(W_k^{1/2} H W_k^{1/2}) - \mathrm{tr}(M_k^{1/2} W M_k^{1/2})(W_k^{1/2} H W_k^{1/2}) - \lambda^\mathsf{T} H s_k = 0.$$

En utilisant la relation $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, on obtient pour tout $H \in \mathcal{S}^n$:

$$0 = \mathrm{tr}(W_k H) - \mathrm{tr}(WH) - \mathrm{tr}(s_k \lambda^\mathsf{T} H) = \mathrm{tr}\left(\left[W_k - W - \frac{s_k \lambda^\mathsf{T} + \lambda s_k^\mathsf{T}}{2}\right] H\right).$$

Comme le facteur de H ci-dessus est dans $\mathcal{S}^n$, on a

$$W = W_k - \frac{s_k \lambda^\mathsf{T} + \lambda s_k^\mathsf{T}}{2}.$$

Pour calculer λ, on utilise l'équation de quasi-Newton, qui donne

$$s_k = W_k y_k - \frac{y_k^\mathsf{T}\lambda}{2} s_k - \frac{y_k^\mathsf{T} s_k}{2}\lambda.$$

En prenant le produit scalaire avec y_k et en utilisant le fait que $y_k^\mathsf{T} s_k \neq 0$, on trouve $y_k^\mathsf{T}\lambda = y_k^\mathsf{T}(W_k y_k - s_k)/(y_k^\mathsf{T} s_k)$ qui, injecté dans la relation ci-dessus, fournit la valeur du multiplicateur

$$\lambda = \frac{-2}{y_k^\mathsf{T} s_k}(s_k - W_k y_k) + \frac{y_k^\mathsf{T}(s_k - W_k y_k)}{(y_k^\mathsf{T} s_k)^2} s_k.$$

En utilisant celle-ci dans la formule de W ci-dessus, on trouve l'expression de W_{k+1} suivante

$$W_{k+1} = W_k + \frac{(s_k - W_k y_k)s_k^\mathsf{T} + s_k(s_k - W_k y_k)^\mathsf{T}}{y_k^\mathsf{T} s_k} - \frac{y_k^\mathsf{T}(s_k - W_k y_k)}{(y_k^\mathsf{T} s_k)^2} s_k s_k^\mathsf{T}. \quad (11.11)$$

Par un calcul laborieux, mais mécanique, on vérifie que cette formule est équivalente à (11.10) et que M_{k+1} donné par (11.9) est bien l'inverse de W_{k+1} donné par (11.10) (il suffit de vérifier que $M_{k+1}W_{k+1} = I$).

Il reste à montrer que la matrice obtenue est bien dans $\mathcal{S}^n_{++}$ (le problème (11.8) a des points stationnaires en dehors de cet ensemble), si $y_k^\mathsf{T} s_k > 0$. Pour un vecteur $v \in \mathbb{R}^n$ arbitraire, la formule (11.9) donne

$$v^\mathsf{T} M_{k+1} v = v^\mathsf{T} M_k v - \frac{(s_k^\mathsf{T} M_k v)^2}{s_k^\mathsf{T} M_k s_k} + \frac{(y_k^\mathsf{T} v)^2}{y_k^\mathsf{T} s_k}.$$

Le dernier terme est positif. On voit qu'il en est de même de la somme des deux premiers termes, car d'après l'inégalité de Cauchy-Schwarz

$$(s_k^\mathsf{T} M_k v)^2 \leqslant \|M_k^{1/2} s_k\|_2^2 \, \|M_k^{1/2} v\|_2^2 = (s_k^\mathsf{T} M_k s_k)(v^\mathsf{T} M_k v).$$

Donc $v^\mathsf{T} M_{k+1} v \geqslant 0$ et ce produit ne peut être nul que si $v \perp y_k$ et $v \parallel s_k$, ce qui n'est vrai que si $v = 0$, car $y_k^\mathsf{T} s_k > 0$. □

Les formules (11.9) et (11.10) portent le nom de *formules de BFGS* directe et inverse. L'appellation BFGS vient des initiales des auteurs (Broyden, Fletcher, Goldfarb et Shanno) qui ont proposé cette formule dans des articles parus en 1970 [96, 209, 257, 545]. On notera les formules (11.9) et (11.10) de la manière suivante :

$$M_{k+1} = \mathrm{BFGS}(M_k, y_k, s_k) \quad \text{et} \quad W_{k+1} = \overline{\mathrm{BFGS}}(W_k, y_k, s_k).$$

Notons que la première de ces formules peut aussi s'écrire sous la forme de produits de matrices [92, 258] :

$$M_{k+1} = \left(I + v_k \bar{s}_k^\mathsf{T}\right) M_k \left(I + \bar{s}_k v_k^\mathsf{T}\right),$$

où $v_k := -M\bar{s}_k + \bar{y}_k$, $\bar{s}_k := s_k/(s_k^\mathsf{T} M_k s_k)^{1/2}$ et $\bar{y}_k := y_k/(y_k^\mathsf{T} s_k)^{1/2}$. Observons enfin que les corrections $M_{k+1} - M_k$ et $W_{k+1} - W_k$ apportées par la formule de BFGS sont des matrices de rang 2.

La proposition 11.2 a pour corollaire immédiat que si $s \neq 0$

$$\exists\, M \in \mathcal{S}^n_{++} : y = Ms \quad \Longleftrightarrow \quad y^\mathsf{T} s > 0. \tag{11.12}$$

Cette proposition a une extension naturelle au cas où l'on impose plus d'une équation de quasi-Newton, qui doivent toutefois être compatibles : $M \in \mathcal{S}^n_{++}$ doit vérifier $Y = MS$, où Y, $S \in \mathbb{R}^{n\times p}$ et S est *injective* (voir l'exercice 11.1). L'équivalence (11.12) devient alors

$$\exists\, M \in \mathcal{S}^n_{++} : Y = MS \quad \Longleftrightarrow \quad Y^\mathsf{T} S \in \mathcal{S}^n_{++}.$$

On utilise rarement une formule de mise à jour qui génère une matrice M vérifiant le système $Y = MS$ rassemblant plus d'une équation de quasi-Newton, car on ne sait pas comment satisfaire la condition $Y^\mathsf{T} S \succ 0$ qui assure la définie positivité de M. La situation est différente si l'on n'a qu'une seule équation de quasi-Newton à satisfaire, car nous verrons qu'il est aisé de vérifier la condition $y^\mathsf{T} s > 0$ apparaissant dans (11.12) (voir la section 11.2.2).

Propriétés algébriques

Voici une propriété de la formule de BFGS qui justifie l'utilisation de cette formule pour construire des préconditionneurs symétriques définis positifs de systèmes linéaires. Elle nous apprend que, si au cours de la minimisation d'une fonction quadratique strictement convexe sur $\mathbb{R}^n$ par l'algorithme du gradient conjugué, on met à jour une matrice par le formule de BFGS (resp. par la formule de BFGS inverse), en utilisant les couples (Au_i, u_i) formés à partir des directions conjuguées u_i, la matrice obtenue après n itérations est la hessienne de la fonction minimisée (resp. son inverse).

Proposition 11.3 *Soit A une matrice symétrique et $u_1, \ldots, u_p$ des directions non nulles, conjuguées par rapport à cette matrice (c'est-à-dire : $u_i^\mathsf{T} A u_j = 0$, pour $i \neq j$ et $u_i^\mathsf{T} A u_i > 0$). On se donne une matrice M_1 symétrique définie positive et on définit $M_{i+1} = \mathrm{BFGS}(M_i, Au_i, u_i)$, pour $i = 1, \ldots, p$. Alors,*
(i) $M_{k+1} u_i = Au_i$, pour $k = 1, \ldots, p$ et $i = 1, \ldots, k$,
(ii) si $p = n$, on a $M_{n+1} = A$.

Démonstration. La première propriété se démontre par récurrence. Elle est vérifiée pour $k = 1$: $M_2 u_1 = Au_1$ (c'est l'équation de quasi-Newton vérifiée par M_2). Supposons qu'elle le soit pour un $k = 1, \ldots, l-1$, avec $2 \leqslant l \leqslant p$, et démontrons la pour $k = l$. Soit $1 \leqslant i < l$. Par conjugaison et récurrence, $u_l^\mathsf{T} A u_i = 0$ et $u_l^\mathsf{T} M_l u_i = u_l^\mathsf{T} A u_i = 0$. Dès lors, la formule de BFGS (11.9) donne $M_{l+1} u_i = M_l u_i = Au_i$. Si $i = l$, on a $M_{l+1} u_i = Au_i$ (c'est l'équation de quasi-Newton vérifiée par M_{l+1}).

Si $p = n$, M_{n+1} prend la même valeur que A sur les n vecteurs linéairement indépendants $u_1, \ldots, u_n$. Donc $M_{n+1} = A$. □

Terminons cette section par deux formules qui nous seront utiles pour étudier la convergence de l'algorithme de BFGS, celles de la trace et du déterminant.

Proposition 11.4 *Si M_{k+1} et M_k sont reliés par le formule de BFGS (11.9) avec M_k définie positive et $y_k^\mathsf{T} s_k > 0$, on a*

$$\operatorname{tr} M_{k+1} = \operatorname{tr} M_k + \frac{\|y_k\|_2^2}{y_k^\mathsf{T} s_k} - \frac{\|M_k s_k\|_2^2}{s_k^\mathsf{T} M_k s_k} \quad \textit{et} \quad \det M_{k+1} = \det M_k \left(\frac{y_k^\mathsf{T} s_k}{s_k^\mathsf{T} M_k s_k} \right).$$

Démonstration. La formule de la trace se déduit directement de (11.9). Pour celle du déterminant, on écrit

$$\det M_{k+1} = \det(M_k) \det \left(I + \frac{M_k^{-1} y_k y_k^\mathsf{T}}{y_k^\mathsf{T} s_k} - \frac{s_k s_k^\mathsf{T} M_k}{s_k^\mathsf{T} M_k s_k} \right).$$

Pour évaluer le dernier déterminant, on utilise le point (*iii*) de l'exercice B.16. □

Propriétés asymptotiques ▲

Supposons que l'on ait deux suites $\{y_k\}_{k\geqslant 1}$ et $\{s_k\}_{k\geqslant 1}$ de vecteurs de $\mathbb{R}^n$, vérifiant $y_k^\mathsf{T} s_k > 0$ pour tout $k \geqslant 1$, et une matrice d'ordre n symétrique définie positive M_1. On génère alors la suite $\{M_k\}_{k\geqslant 1}$ par la formule de BFGS : $M_{k+1} = \mathrm{BFGS}(M_k, y_k, s_k)$ pour tout $k \geqslant 1$. On s'intéresse ici aux propriétés que peuvent induire les vecteurs y_k et s_k sur la suite $\{M_k\}$. Ces propriétés seront utiles pour étudier le comportement asymptotique des suites générées par l'algorithme de BFGS.

On sait déjà, d'après la proposition 11.2, que les matrices M_k sont toutes symétriques définies positives.

Formule de mise à jour pour matrices creuses ▲

11.2.2 L'algorithme de BFGS

D'après la proposition 11.2, on aura un algorithme à directions de descente si dans (11.1), M_k est générée par la formule de BFGS à partir d'une matrice initiale M_1 définie positive et si à chaque itération on réalise l'inégalité $y_k^\mathsf{T} s_k > 0$. La formule de BFGS génère alors des matrices M_k définies positives. Il est tout à fait remarquable que l'inégalité $y_k^\mathsf{T} s_k > 0$ soit satisfaite lorsque la recherche linéaire détermine le pas $\alpha_k > 0$ par la règle de Wolfe (section 6.3.4 ; rappelons que l'on a choisit le produit scalaire euclidien $\langle u, v\rangle := u^\mathsf{T} v$). En effet, si l'on retranche $g_k^\mathsf{T} d_k$ dans les deux membres de l'inégalité (6.11b), on obtient

$$y_k^\mathsf{T} s_k \geqslant (\omega_2 - 1)\, g_k^\mathsf{T} s_k > 0,$$

car $\omega_2 < 1$ et $g_k^\mathsf{T} s_k = \alpha_k g_k^\mathsf{T} d_k < 0$ (d_k est une direction de descente). Pour minimiser une fonction non linéaire (non quadratique) au moyen de la formule de BFGS, on utilise donc *toujours* la recherche linéaire de Wolfe. Ceci conduit à l'algorithme parfaitement bien défini suivant.

Algorithme 11.5 (de BFGS)

0. On se donne deux constantes ω_1 et ω_2 pour la recherche linéaire de Wolfe : $0 < \omega_1 < \frac{1}{2}$ et $\omega_1 < \omega_2 < 1$.
   Choix d'un itéré initial $x_1 \in \mathbb{R}^n$ et d'une matrice initiale W_1 définie positive (approximation de l'*inverse* de la hessienne $\nabla^2 f(x_1)$).
   Initialisation : $k := 1$.
1. *Test d'arrêt* : si $\nabla f(x_k) = 0$, arrêt de l'algorithme.
2. *Calcul de la direction de descente* : $d_k = -W_k g_k$.
3. *Recherche linéaire de Wolfe* : trouver un pas $\alpha_k > 0$ tel que l'on ait (6.11a) et (6.11b).
4. $x_{k+1} := x_k + \alpha_k d_k$.
5. Mettre à jour la matrice W_k par la formule (11.10) dans laquelle on prend $s_k = x_{k+1} - x_k$ et $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$.
6. Accroître k de 1 et aller en 1.

En pratique, on arrête l'algorithme à l'étape 1 lorsque $\nabla f(x_k)$ est suffisamment petit. On notera $M_k = W_k^{-1}$ l'inverse des matrices générées par l'algorithme ci-dessus, qui approchent donc la hessienne $\nabla^2 f$. On trouvera à la section 11.2.4 plus de détails sur la mise en œuvre de l'algorithme de BFGS. Analysons d'abord ses principales propriétés.

11.2.3 Propriétés de l'algorithme de BFGS

Convergence globale

> *My favorite results are his proof of the global convergence of the BFGS method for convex functions, and the introduction of the least change approach for derivative-free optimization.*
>
> J. Moré (2015), dans un hommage à M.J.D. Powell [439].

Un des plus beaux résultats de convergence en optimisation sans contrainte est le théorème 11.6 ci-dessous. Il donne des conditions assurant la convergence globale de l'algorithme de BFGS lorsque la fonction à minimiser est convexe. On notera qu'il n'est pas nécessaire de supposer la *forte* convexité du critère. La beauté du résultat tient en particulier au fait qu'il est rare de pouvoir démontrer la convergence d'un algorithme quasi-newtonien, avec si peu d'hypothèses sur les objets générés par celui-ci. On suppose seulement que l'algorithme génère une suite $\{x_k\}$ (ce qui signifie qu'il ne s'arrête pas à l'étape 1 parce que le gradient y est nul ; auquel cas, il n'y a rien à démontrer) et que la suite $f(x_k)$ est bornée inférieurement (il existe une constante C telle que pour tout indice k, $f(x_k) \geqslant C$). On ne fait aucune hypothèse sur la suite des matrices $\{M_k\}$.

Théorème 11.6 (convergence globale de BFGS) *Supposons que $f : \mathbb{R}^n \to \mathbb{R}$ soit convexe et $\mathcal{C}^{1,1}$ dans un voisinage convexe de $\{x \in \mathbb{R}^n : f(x) \leqslant f(x_1)\}$, où $x_1 \in \mathbb{R}^n$. On considère l'algorithme de BFGS démarrant en x_1 avec une matrice M_1 symétrique définie positive et on suppose que celui-ci génère une suite $\{x_k\}$ telle que $\{f(x_k)\}$ soit bornée inférieurement. Alors,* $\liminf_{k\to\infty} \|g_k\| = 0$.

Démonstration. On note $C_1, C_2, \ldots$ des constantes strictement positives. L'idée de la démonstration est de contrôler le comportement des matrices $\{M_k\}$ en obtenant une majoration de leur trace et une minoration de leur déterminant, ce qui permettra d'avoir une borne inférieure sur le pas α_k. On note

$$q_k := \frac{s_k^\mathsf{T} M_k s_k}{\|s_k\|_2^2} \quad \text{et} \quad \cos\theta_k := \frac{-g_k^\mathsf{T} d_k}{\|g_k\|_2 \|d_k\|_2} = \frac{s_k^\mathsf{T} M_k s_k}{\|M_k s_k\|_2 \|s_k\|_2}$$

le quotient de Rayleigh de M_k dans la direction s_k et le cosinus de l'angle entre d_k et l'opposé du gradient.

Première étape. On utilise la formule de la trace pour montrer que pour tout $k \geqslant 1$:

$$\operatorname{tr} M_{k+1} \leqslant C_1 k \quad \text{et} \quad \sum_{i=1}^{k} \frac{\|M_i s_i\|_2^2}{s_i^\mathsf{T} M_i s_i} = \sum_{i=1}^{k} \frac{q_i}{\cos^2\theta_i} \leqslant C_1 k. \tag{11.13}$$

En effet, comme $f \in \mathcal{C}^{1,1}$, on a $\|y_k\|_2^2/(y_k^\mathsf{T} s_k) \leqslant C_2$ (proposition 3.69) et donc la formule de la trace (proposition **??**) donne

$$\operatorname{tr} M_{k+1} \leqslant \operatorname{tr} M_k + C_2 - \frac{\|M_k s_k\|_2^2}{s_k^\mathsf{T} M_k s_k} \leqslant \operatorname{tr} M_1 + C_2 k - \sum_{i=1}^{k} \frac{\|M_i s_i\|_2^2}{s_i^\mathsf{T} M_i s_i}.$$

On en déduit la première estimation de (11.13) avec $C_1 := \operatorname{tr} M_1 + C_2$. D'autre part, comme M_{k+1} est définie positive, on a $\operatorname{tr} M_{k+1} \geqslant 0$ et l'inégalité précédente donne la seconde estimation de (11.13).

Deuxième étape. En utilisant la majoration de la trace (11.13) et une minoration du déterminant, on obtient une minoration de la plus petite valeur propre de M_{k+1}, donc une minoration du pas. On montre en fait que cela conduit à la borne inférieure suivante sur le pas moyen : pour tout $k \geqslant 1$,

$$\left(\prod_{i=1}^{k} \alpha_i\right)^{\frac{1}{k}} \geqslant C_3. \tag{11.14}$$

En effet, par la formule du déterminant (proposition **??**) et la règle de Wolfe (6.11b) :

$$\begin{aligned}\det M_{k+1} &= \frac{y_k^{\mathsf{T}} s_k}{s_k^{\mathsf{T}} M_k s_k} \det M_k \\ &\geqslant \frac{(1-\omega_2)}{\alpha_k} \det M_k \\ &\geqslant \frac{(1-\omega_2)^k}{\prod_{i=1}^k \alpha_i} \det M_1 \\ &\geqslant \frac{C_4^k}{\prod_{i=1}^k \alpha_i},\end{aligned}$$

avec $C_4 := (1-\omega_2)\min(1, \det M_1)$. D'autre part, en utilisant l'inégalité géométrico-arithmétique (3.76), (11.13) et le fait que $k^n \leqslant (e^n)^k$:

$$\det M_{k+1} \leqslant \left(\frac{1}{n} \operatorname{tr} M_{k+1}\right)^n \leqslant \left(\frac{C_1}{n} k\right)^n \leqslant C_5^k,$$

avec $C_5 := \max(1, C_1/n)^n e^n$. Les deux estimations ci-dessus conduisent à (11.14) avec $C_2 := C_4/C_5$.

Troisième étape. En exploitant la deuxième estimation de (11.13), on obtient une majoration du pas moyen pondéré. On se défera de la pondération dans le raisonnement par l'absurde de la quatrième étape. Pour tout $k \geqslant 1$:

$$\left(\prod_{i=1}^k \alpha_i \|g_i\|_2^2\right)^{\frac{1}{k}} \leqslant \frac{C_6}{k}. \tag{11.15}$$

En effet,

$$\sum_{i=1}^k \frac{\|M_i s_i\|_2^2}{s_i^{\mathsf{T}} M_i s_i} = \sum_{i=1}^k \frac{\alpha_i \|g_i\|_2^2}{-g_i^{\mathsf{T}} s_i} \leqslant C_1 k. \tag{11.16}$$

On utilise ensuite deux fois l'inégalité géométrico-arithmétique (3.76) et la règle de Wolfe (6.11a)

$$\begin{aligned}\prod_{i=1}^k \alpha_i \|g_i\|_2^2 &\leqslant C_1^k \prod_{i=1}^k (-g_i^{\mathsf{T}} s_i) \qquad [(11.16) \text{ et } (3.76)] \\ &\leqslant \left(\frac{C_1}{k} \sum_{i=1}^k (-g_i^{\mathsf{T}} s_i)\right)^k \qquad [(3.76)] \\ &\leqslant \left(\frac{C_1}{\omega_1 k} \sum_{i=1}^k (f(x_i) - f(x_{i+1}))\right)^k \qquad [(6.11a)] \\ &= \left(\frac{C_1}{\omega_1 k} (f(x_1) - f(x_{k+1}))\right)^k.\end{aligned}$$

On en déduit (11.15), avec $C_6 := C_1(f(x_1) - f_{\min})/\omega_1$, où $f_{\min} \in \mathbb{R}$ est un minorant de $\{f(x_k)\}$.

Quatrième étape. On conclut par un raisonnement par l'absurde. Supposons que le résultat ne soit pas vrai. Alors $\|g_k\| \geqslant C_7 > 0$ et (11.15) donne

$$\left(\prod_{i=1}^{k}\alpha_i\right)^{\frac{1}{k}} \leqslant \frac{C_6}{C_7^2 k},$$

qui est en contradiction avec (11.14). □

Observons que si l'on avait supposé f fortement convexe, on n'aurait pu utiliser de la recherche linéaire de Wolfe que ce qui est contenu dans la condition de Zoutendijk (6.21). En effet la formule du déterminant aurait donné pour une constante générique strictement positive C :

$$\det M_{k+1} \geqslant C\frac{\|s_k\|^2}{s_k^{\mathrm{T}} M_k s_k}\det M_k \geqslant \frac{C^k}{\prod_{i=1}^{k} q_i},$$

qui avec la formule de la trace aurait conduit à

$$\left(\prod_{i=1}^{k} q_i\right)^{\frac{1}{k}} \geqslant C.$$

Alors la seconde estimation de (11.13) et l'inégalité géométrico-arithmétique (3.76) auraient donné

$$\left(\prod_{i=1}^{k}\frac{q_i}{\cos^2\theta_i}\right)^{\frac{1}{k}} \leqslant \frac{1}{k}\sum_{i=1}^{k}\frac{q_i}{\cos^2\theta_i} \leqslant C.$$

Donc

$$C \leqslant \left(\prod_{i=1}^{k} q_i\right)^{\frac{1}{k}} \leqslant C\left(\prod_{i=1}^{k}\cos^2\theta_i\right)^{\frac{1}{k}} \leqslant \frac{C}{k}\sum_{i=1}^{k}\cos^2\theta_i.$$

On en aurait déduit que $\sum_{k\geqslant 0}\cos^2\theta_i = \infty$ et, avec (6.21), que $\liminf \|g_k\| = 0$. Lorsque f est seulement convexe, nous avons vu que la règle de Wolfe intervenait à plusieurs endroits dans la démonstration.

Corollaire 11.7 *Sous les hypothèses du théorème 11.6, si la fonction f est fortement convexe, alors la suite $\{x_k\}$ converge vers l'unique minimum de f.*

Démonstration. Si f est fortement convexe, il existe un unique point x_* minimisant f sur $\mathbb{R}^n$.

Montrons qu'il existe une sous-suite de $\{x_k\}$ qui converge vers x_*. La suite $\{x_k\}$ est bornée (elle est dans $\{x : f(x) \leqslant f(x_1)\}$ qui est borné). Dès lors, d'après le théorème, on peut trouver une sous-suite convergente $\{x_k\}_{k\in\mathcal{K}}$, telle que $g_k \to 0$. On en déduit que $x_k \to x_*$ lorsque $k \to \infty$ dans $\mathcal{K}$ (ça ne peut pas être un autre point du fait de la continuité du gradient et de l'unicité du point annulant le gradient).

On a alors que $f(x_k) \to f(x_*)$. En effet, la suite $\{f(x_k)\}$ est décroissante et bornée inférieurement. Donc elle converge et ça ne peut être que vers $f(x_*)$, car une sous-suite converge vers cette valeur.

Il reste à conclure. De toute sous-suite $\{x_k\}_{k\in\mathcal{K}'}$, on peut extraire une sous-suite convergente (elle est bornée), qui ne peut converger que vers x_* (car $f(x_k)$ converge vers $f(x_*)$). Alors, toute la suite $\{x_k\}$ converge vers x_*. □

Convergence locale

Terminaison quadratique

> *Indeed for many years the only intrusion of theory on my research was the point of view that, if an algorithm does not perform well when the objective function is quadratic, then it is unlikely that it will be efficient in general use. This rule of thumb and trying to make good use of available information were the main considerations that helped me to develop several successful algorithms for unconstrained calculations.*
>
> M.J.D. Powell [495 ; 1991].

L'algorithme de BFGS peut être interprété comme une extension de l'algorithme du gradient conjugué (GC) préconditionné. Si on l'utilise pour minimiser une fonction quadratique strictement convexe, donc de la forme

$$x \in \mathbb{R}^n \mapsto f(x) = \frac{1}{2} x^\mathsf{T} A x - b^\mathsf{T} x,$$

où la matrice A est d'ordre n symétrique définie positive et $b \in \mathbb{R}^n$, *en faisant de la recherche linéaire exacte*, alors l'algorithme génère les mêmes itérés que l'algorithme du GC préconditionné par la matrice $W_1 = M_1^{-1}$. Cette propriété est connue sous le nom de *terminaison quadratique*.

Proposition 11.8 (terminaison quadratique) *Si $f : \mathbb{R}^n \to \mathbb{R}$ est quadratique strictement convexe. Alors l'algorithme de BFGS avec recherche linéaire exacte est bien défini et génère les mêmes itérés que l'algorithme du gradient conjugué préconditionné par la matrice $W_1 = M_1^{-1}$. En particulier, il converge en au plus n itérations.*

Démonstration. On montre par récurrence que l'on a pour tout $k \geqslant 1$ tel que $g_k \neq 0$:

$$W_i g_k = W_1 g_k, \quad i = 1, \ldots, k-1, \tag{11.17}$$

$$d_k = d_k^{\mathrm{GC}} := \begin{cases} -W_1 g_1 & \text{si } k = 1 \\ -W_1 g_k + \frac{g_k^\mathsf{T} W_1 g_k}{g_{k-1}^\mathsf{T} W_1 g_{k-1}} d_{k-1} & \text{si } k \geqslant 2, \end{cases} \tag{11.18}$$

où d_k^{GC} est la direction du GC préconditionné par la matrice W_1.

Si $g_1 \neq 0$ et $k = 1$, les deux identités sont clairement vérifiées puisque la première direction de l'algorithme de BFGS est $d_1 = -W_1 g_1$.

Supposons à présent que les identités (11.17) et (11.18) aient lieu jusqu'à un indice $k \geqslant 1$ et montrons qu'elles ont également lieu pour l'indice $k+1$ si $g_{k+1} \neq 0$. Sous l'hypothèse de récurrence, les itérés $x_1, \ldots, x_{k+1}$ de l'algorithme de BFGS sont identiques à ceux générés par le GC préconditionné par W_1. Dès lors $s_i^\mathsf{T} g_{k+1} = 0$ pour $i = 1, \ldots, k$ (exercice 8.3) et, par la formule (11.10),

$$W_{i+1} g_{k+1} = \left(I - \frac{s_i y_i^\mathsf{T}}{y_i^\mathsf{T} s_i} \right) W_i g_{k+1}, \quad \text{pour } i = 1, \ldots, k.$$

On a aussi $g_i^T W_1 g_{k+1} = 0$ pour $i = 1, \ldots, k$ (exercice 8.3), et donc $y_i^T W_1 g_{k+1} = 0$ pour $i = 1, \ldots, k-1$. L'identité ci-dessus avec $i = 1, \ldots, k-1$ permet alors d'obtenir (11.17) par récurrence sur i. L'identité ci-dessus avec $i = k$ donne

$$d_{k+1} = -W_{k+1} g_{k+1} = -W_1 g_{k+1} + \frac{y_k^T W_1 g_{k+1}}{y_k^T d_k} d_k.$$

On en déduit (11.18) en notant que $y_k^T W_1 g_{k+1} = g_{k+1}^T W_1 g_{k+1}$ et $y_k^T d_k = -g_k^T d_k = g_k^T W_1 g_k$ (on utilise l'hypothèse de récurrence et $g_k^T d_{k-1} = 0$). □

11.2.4 Mise en œuvre de l'algorithme de BFGS

Formules de mise à jour directe ou inverse

A priori, il parait plus intéressant de mettre à jour W_k, l'inverse de M_k par (11.10), que M_k par (11.9), de manière à ne pas devoir résoudre le système linéaire $M_k d_k = -g_k$ pour déterminer la direction de descente d_k à l'étape 2 de l'algorithme. Dans cette approche, la direction $d_k = -W_k g_k$ est obtenue en $O(n^2)$ opérations, alors que la résolution d'un système linéaire en demande $O(n^3)$ en général. On notera aussi que la mise à jour de W_k par (11.10) peut se faire en $O(n^2)$, par la suite d'opérations suivantes :

$$\begin{aligned} \widetilde{W}_k &:= \left(I - \frac{s_k y_k^T}{y_k^T s_k}\right) W_k = W_k - \bar{s}_k (W_k \bar{y}_k)^T \\ W_{k+1} &:= \widetilde{W}_k \left(I - \frac{y_k s_k^T}{y_k^T s_k}\right) + \frac{s_k s_k^T}{y_k^T s_k} = \widetilde{W}_k - (\widetilde{W}_k \bar{y}_k)\bar{s}_k^T + \bar{s}_k \bar{s}_k^T, \end{aligned}$$

dans lesquelles on a noté $\bar{s}_k := s_k/(y_k^T s_k)^{1/2}$ et $\bar{y}_k := y_k/(y_k^T s_k)^{1/2}$. Évidemment, $\widetilde{W}_k$ et W_{k+1} peuvent être stockées dans W_k. Dans les problèmes avec contraintes, on doit mettre à jour la matrice directe M_k et cela se fait parfois par la formule suivante

$$M_{k+1} = M_k + u_k v_k^T + v_k u_k^T + v_k v_k^T, \tag{11.19}$$

qui est équivalente à (11.9) si on prend $\bar{s}_k := s_k/(s_k^T M_k s_k)^{1/2}$, $\bar{y}_k := y_k/(y_k^T s_k)^{1/2}$, $u_k := M_k \bar{s}_k$ et $v_k := \bar{y}_k - M_k \bar{s}_k$.

En pratique il est préférable d'utiliser (11.10) plutôt que (11.11). Cette dernière formule est en effet instable, du fait des erreurs d'arrondi qui y sont moins bien contrôlées et qui peuvent rendre W_{k+1} indéfinie. Cette affirmation relève de notre propre expérience avec ces formules, mais nous ne connaissons pas d'étude précise sur le sujet. On observe toutefois que par le second terme de (11.11), on *retranche* une matrice semi-définie positive, alors que dans (11.10) on fait la *somme* de deux matrices semi-définies positives. Dans la procédure de calcul proposée ci-dessus, dans laquelle on évite de devoir former la matrice $(I - \bar{s}_k \bar{y}_k^T)$ et de faire un produit de deux matrices (ce qui requiert $O(n^3)$ opérations), on ne bénéficie pas entièrement de cette bonne propriété, car on retrouve des différences de matrices pouvant entraîner de l'instabilité. De ce point de vue, on pourra aussi préférer la formule (11.19) pour la mise à jour de la matrice directe.

Certains auteurs [248, 258] préfèrent mettre à jour les facteurs de Cholesky L_k de $M_k = L_k L_k^\mathsf{T}$ de manière à contrôler la définie positivité de cette matrice qui peut être perdue du fait des erreurs d'arrondi (pour les problèmes difficiles seulement). Cette mise à jour des facteurs de Cholesky de M_k peut se faire en $O(n^2)$ opérations. Le calcul de la direction de descente d_k demande aussi $O(n^2)$ opérations (il y a deux systèmes linéaires triangulaires à résoudre), alors qu'il en faudrait $O(n^3)$ si M_k n'était pas connue par ses facteurs triangulaires.

D'autres auteurs enfin [259, 494, 550] utilisent la troisième approche qui consiste à mettre à jour des facteurs Z_k de $W_k = Z_k Z_k^\mathsf{T}$ (qui ne sont pas de Cholesky). Cela permet d'avoir un meilleur contrôle de la définie positivité de W_k. Cette approche jette un nouvel éclairage sur les liens avec l'algorithme du gradient conjugué (voir à ce sujet les propositions 11.3 et 11.8).

Choix de la matrice initiale

Le choix de la matrice symétrique définie positive initiale M_1 est crucial pour un bon fonctionnement de l'algorithme de BFGS. On conçoit en effet que prendre $M_1 = I$ peut avoir des effets désastreux, parce que la matrice identité peut être beaucoup trop grande ou trop petite, selon les cas, et que, $M_{k+1} - M_k$ étant une matrice de rang 2, la suite $\{M_k\}$ ne va pas évoluer rapidement (on tient d'ailleurs beaucoup à cette propriété de stabilité de la formule de BFGS).

En toute logique, il faudrait prendre comme matrice initiale M_1 une matrice symétrique définie positive approchant convenablement $\nabla^2 f(x_1)$. On dispose parfois d'une telle matrice. Par exemple, cela peut être une approximation définie positive de la diagonale de cette hessienne (on rappelle qu'en toute généralité, celle-ci est aussi coûteuse à évaluer que la hessienne complète, voir l'exercice **??**) ou une partie définie positive de $\nabla^2 f(x_1)$ (comme $J_1^\mathsf{T} J_1$ dans les problèmes de moindres-carrés lorsque la jacobienne $J_1 = r'(x_1)$ du résidu au point initial est calculable à peu de frais et est injective, voir la section 19.3).

Parfois, il est intéressant de prendre pour M_1, la matrice obtenue en fin de résolution par l'algorithme de BFGS d'un problème voisin. Dans ce dernier cas, on parle de *démarrage à chaud*.

Le plus souvent, cependant, on ne dispose d'aucune matrice « naturelle » M_1 et il faut *démarrer à froid*. On prend alors $M_1 = I$ ou $W_1 = I$. Ensuite $M_2 = \mathrm{BFGS}(\sigma_1 I, y_1, s_1)$ ou $W_2 = \overline{\mathrm{BFGS}}(\bar\sigma_1 I, y_1, s_1)$, avec

$$\sigma_1 := \frac{y_1^\mathsf{T} s_1}{\|s_1\|_2^2} \quad \text{et} \quad \bar\sigma_1 := \frac{y_1^\mathsf{T} s_1}{\|y_1\|_2^2}. \tag{11.20}$$

Pour $k \geqslant 2$, on prend comme d'habitude $M_{k+1} = \mathrm{BFGS}(M_k, y_k, s_k)$ ou $W_{k+1} = \overline{\mathrm{BFGS}}(W_k, y_k, s_k)$. Les facteurs scalaires σ_1 et $\bar\sigma_1$, connus sous le nom de *facteurs d'Oren-Luenberger*, sont déterminés par le souci d'avoir $\sigma_1 I$ et $\bar\sigma_1 I$ les plus proches possibles, dans des sens très variés, du sous-espace affine $\{M \in \mathcal{S}^n : y_1 = M s_1\}$ des matrices symétriques vérifiant l'équation de quasi-Newton (voir l'exercice 11.3). On notera que σ_1 et $\bar\sigma_1$ sont tous les deux strictement positifs (on suppose $y_1^\mathsf{T} s_1 > 0$). On notera également que, par l'inégalité de Cauchy-Schwarz, $\sigma_1 \leqslant (\bar\sigma_1)^{-1}$ sans que l'on ait nécessairement égalité, si bien que générer les matrices directes ou inverses avec les initialisations ci-dessus ne donnent pas le même algorithme. On pourrait d'ailleurs

aussi prendre $\sigma_1 = \|y_1\|_2^2/(y_1^T s_1)$ et $\bar{\sigma}_1 = \|s_1\|_2^2/(y_1^T s_1)$. On ne connaît pas de critère faisant l'unanimité permettant de faire le bon choix des facteurs σ_1 ou $\bar{\sigma}_1$; on utilise souvent (11.20).

Mise à l'échelle répétée

Nous avons dit que la suite $\{M_k\}$ était très stable. Hélas parfois trop ! Si l'initialisation n'est pas satisfaisante et que les itérés progressent rapidement, on aimerait que les matrices évoluent plus vite pour s'adapter à la hessienne courante. Il peut donc être intéressant de *catalyser* la mise à jour en multipliant M_k ou W_k par un facteur scalaire *avant* leur mise à jour, mais *après* leur utilisation. On prend donc $M_{k+1} = \text{BFGS}(\sigma_k M_k, y_k, s_k)$ ou $W_{k+1} = \overline{\text{BFGS}}(\bar{\sigma}_k W_k, y_k, s_k)$, avec des scalaires $\sigma_k > 0$ et $\bar{\sigma}_k > 0$ valant

$$\sigma_k := \frac{y_k^T s_k}{s_k^T M_k s_k} \quad \text{et} \quad \bar{\sigma}_k := \frac{y_k^T s_k}{y_k^T W_k y_k}. \tag{11.21}$$

Ceux-ci sont obtenus par un raisonnement analogue à celui conduisant à (11.20)

Ce qui peut ne pas fonctionner

L'algorithme de BFGS avec recherche linéaire en optimisation demande que le gradient soit calculé avec une bonne précision. Si ce n'est pas le cas, en particulier lorsque le gradient est estimé par différences finies, l'algorithme pourra s'interrompre prématurément parce que la direction générée ne sera pas une direction de descente. En effet, celle-ci s'écrit $d_k = -W_k g_k$, où g_k est le gradient *calculé* et W_k est une matrice définie positive dont on ne contrôle pas le conditionnement $\kappa_2(W_k)$. Comme, selon (6.4), l'inverse de ce dernier minore l'angle θ_k entre d_k et $-g_k$, on pourra avoir un angle θ_k arbitrairement proche de $\pi/2$. Si g_k n'est pas suffisamment précis, il se pourra que la direction d_k calculée par l'algorithme de BFGS ne soit pas de descente en x_k, entraînant l'échec de la recherche linéaire. Si la précision obtenue alors n'est pas suffisante, il n'y a qu'un seul remède : calculer le gradient avec plus de précision (éviter la discrétisation d'un gradient en dimension infinie, faire de la différentiation automatique ou manuelle, chasser les erreurs d'arrondi).

Mentionnons aussi, mais sans insister, que des exemples de non-convergence de l'algorithme de BFGS avec recherche linéaire ont été découverts [411 ; 2004], soit environ 35 ans après son introduction. Il y a peu de chance toutefois que cette non-convergence s'observe en pratique.

11.2.5 L'algorithme ℓ-BFGS

Si le problème est de grande taille, disons avec un nombre de variables n supérieur à 500 (ce seuil dépend en partie des performances des ordinateurs), les méthodes de quasi-Newton telles qu'on les a présentées ne sont plus adaptées, pour deux raisons au moins. D'abord parce qu'il peut être impossible de mettre en mémoire une matrice d'ordre n. Ensuite, même si le stockage d'une telle matrice est possible, sa mise à jour peut prendre beaucoup de temps (celle-ci requiert en effet de l'ordre de n^2 opérations) et son adaptation à la hessienne courante peut être lente (dans les formules de

mise à jour vues, la différence entre deux matrices successives est de faible rang, si bien qu'il faut normalement un nombre de mises à jour proportionnel à n pour que l'approximation courante devienne correcte et que l'algorithme devienne efficace).

Une belle découverte, très utile en pratique, a été de constater que l'on peut *adapter* beaucoup de méthodes de quasi-Newton à des problèmes de très grande taille en appliquant les principes suivants :

1. on ne forme l'approximation de la hessienne qu'en utilisant un nombre limité de couples (y_i, s_i), disons m, avec $m \ll n$ (typiquement $m \simeq 5 \dots 10$) ;
2. on ne mémorise que ces couples, pas la matrice elle-même, et les mises à jour se font à partir d'une matrice initiale peu encombrante en mémoire (typiquement une matrice diagonale) ; la matrice approchant la hessienne (ou son inverse) est donc définie de manière implicite, sans représentation en mémoire ;
3. on calcule le produit de cette matrice (ou de son inverse) par un vecteur (comme cela est requis par la formule (11.1)) par un algorithme qui doit être efficace et peut se passer d'une représentation explicite de la matrice W_k.

Une technique de mise à jour reposant sur ces principes est dite *à mémoire limitée*. La sélection des bons couples (y_i, s_i) est un problème délicat qui n'a pas trouvé de réponse satisfaisante, si bien que l'on sélectionne en général les m plus récents, ceux sensés donner la meilleure information sur la hessienne courante. De même, le choix de la matrice initiale est encore aujourd'hui un sujet de recherche ; certains auteurs [399] choisissent la matrice identité multipliée par un des scalaires de (11.21) ; d'autres [244] utilisent des matrices diagonales, elles-mêmes mises à jour au cours des itérations. Enfin, nous verrons qu'une conséquence de la différentiation automatique en mode inverse (section 5.5.3) est que le produit de la matrice implicite par un vecteur pourra se faire de manièrre efficace si la forme quadratique associée à cette matrice s'évalue rapidement.

Soyons plus concret et décrivons la mise à jour par la formule de BFGS inverse (11.10) de l'approximation *à mémoire limitée* de l'inverse de la hessienne, qui suit les principes énoncés aux points 1-3 ci-dessus. L'algorithme correspondant est appelé l'algorithme ℓ-BFGS [458]. La matrice implicite (c.-à-d., non stockée) à l'itération k est toujours notée W_k. Au moment de la mise à jour, on suppose que x_k et x_{k+1} sont connus ; le premnier itéré est noté x_1. Comme mentionné ci-dessus, on utilise en général les couples les plus récents :

$$\{(y_{k-\bar{m}+i}, s_{k-\bar{m}+i}) : i = 1, \dots, \bar{m}\}, \tag{11.22}$$

où

$$\bar{m} := \min(m, k).$$

Partant d'une matrice initiale W_k^0 facilement mémorisable et qu'il faut spécifier, la matrice implicite suivante W_{k+1} est obtenue par

$$\begin{aligned} W_k^i &:= \overline{\text{BFGS}}(W_k^{i-1}, y_{k-\bar{m}+i}, s_{k-\bar{m}+i}), \quad \text{pour } i = 1, \dots, \bar{m} \\ W_{k+1} &= W_k^{\bar{m}}, \end{aligned}$$

où $\overline{\text{BFGS}}$ symbolise la fomule de BFGS inverse (11.10). Il s'agit ici d'une description formelle de W_{k+1}, car cette matrice n'est pas stockée.

Notre but à présent est de montrer que le produit de la matrice implicite W_{k+1}, représentée par la matrice W_k^0 et les couples (11.22), par un vecteur v peut se faire aisément. Dans (11.1), $v = g_{k+1}$, mais ce qui suit est valable pour un vecteur v arbitraire, ce qui pourra être utile pour les problèmes avec contraintes. On note

$$\begin{array}{ll} \text{pour } i = 1, \ldots, \bar{m} : & \rho_i := (y_{k-\bar{m}+i}^{\mathsf{T}} s_{k-\bar{m}+i})^{-1} \quad \text{et} \quad V_i := I - \rho_i y_{k-\bar{m}+i} s_{k-\bar{m}+i}^{\mathsf{T}}, \\ \text{pour } i = 0, \ldots, \bar{m} : & \text{la fonction } v \in \mathbb{R}^n \mapsto \varphi_i(v) := (v^{\mathsf{T}} W_k^i v)/2 \end{array}$$

et on définit par récurrence

$$q_{\bar{m}} = v \qquad \text{et} \qquad q_{i-1} = V_{k-\bar{m}+i} q_i \quad (\text{pour } i = \bar{m}, \ldots, 1).$$

Alors, grâce à (11.10), on a par récurrence

$$\begin{aligned} \varphi_{\bar{m}}(v) &= \varphi_{\bar{m}}(q_{\bar{m}}) \\ &= \frac{1}{2} q_{\bar{m}}^{\mathsf{T}} V_k^{\mathsf{T}} W_k^{\bar{m}-1} V_k q_{\bar{m}} + \frac{\rho_k}{2} (s_k^{\mathsf{T}} q_{\bar{m}})^2 \\ &= \varphi_{\bar{m}-1}(q_{\bar{m}-1}) + \frac{\rho_k}{2} (s_k^{\mathsf{T}} q_{\bar{m}})^2 \\ &= \varphi_0(q_0) + \sum_{i=1}^{\bar{m}} \frac{\rho_{k-\bar{m}+i}}{2} (s_{k-\bar{m}+i}^{\mathsf{T}} q_i)^2. \end{aligned}$$

Cette valeur $\varphi_{\bar{m}}(v)$ se calcule donc par l'algorithme

$$\begin{array}{l} \varphi := 0; \\ q := v; \\ \texttt{for } (i := \bar{m};\ i \geqslant 1;\ i := i-1)\ \{ \\ \quad \alpha_i := s_{k-\bar{m}+i}^{\mathsf{T}} q; \\ \quad \varphi := \varphi + \frac{\rho_{k-\bar{m}+i}}{2} \alpha_i^2; \\ \quad q := q - \rho_{k-\bar{m}+i} \alpha_i y_{k-\bar{m}+i}; \\ \} \\ \varphi := \varphi + \frac{1}{2} q^{\mathsf{T}} W_k^0 q; \end{array} \tag{11.23}$$

Comme $W_{k+1} v$ est le gradient en v de $v \mapsto \varphi_{\bar{m}}(v)$, on peut calculer ce vecteur en exécutant le code adjoint de (11.23) (voir la section 5.5.3 et plus spécialement le code (5.32) et l'exercice 5.9). On note $\overline{\varphi}$, $\overline{q}$ et $\overline{\alpha} := (\overline{\alpha}_1, \ldots, \overline{\alpha}_{\bar{m}})$ les variables adjointes de φ, q et $\alpha := (\alpha_1, \ldots, \alpha_{\bar{m}})$. Ce code adjoint s'écrit alors

$$\begin{array}{l} \overline{\varphi} := 1; \\ \overline{q} := 0; \\ \overline{\alpha} := 0; \\ \overline{q} := W_k^0 q; \\ \texttt{for } (i := 1;\ i \leqslant \bar{m};\ i := i+1)\ \{ \\ \quad \overline{\alpha}_i := \overline{\alpha}_i - \rho_{k-\bar{m}+i} y_{k-\bar{m}+i}^{\mathsf{T}} \overline{q}; \\ \quad \overline{\alpha}_i := \overline{\alpha}_i + \rho_{k-\bar{m}+i} \alpha_i \overline{\varphi}; \\ \quad \overline{q} := \overline{q} + \overline{\alpha}_i s_{k-\bar{m}+i}; \\ \quad \overline{\alpha}_i := 0; \\ \} \\ \overline{v} := \overline{q}; \\ \overline{q} := 0; \\ \overline{\varphi} := 0; \end{array} \tag{11.24}$$

La valeur de $W_{k+1}v$ est obtenu à la sortie de l'algorithme (11.24) dans le vecteur $\overline{v}$. En combinant (11.23) et (11.24), en ne conservant que les instructions qui sont utiles au calcul de $W_{k+1}v$ et en plaçant $\overline{q}$ dans q et tous les $\overline{\alpha}_i$ dans β, on obtient l'algorithme suivant :

$$
\begin{array}{l}
q := v; \\
\texttt{for } (i := \bar{m};\ i \geqslant 1;\ i := i-1)\ \{ \\
\quad \alpha_i := s_{k-\bar{m}+i}^{\mathsf{T}} q; \\
\quad q := q - \rho_{k-\bar{m}+i}\alpha_i y_{k-\bar{m}+i}; \\
\} \\
q := W_k^0 q; \\
\texttt{for } (i := 1;\ i \leqslant \bar{m};\ i := i+1)\ \{ \\
\quad \beta := \rho_{k-\bar{m}+i}(\alpha_i - y_{k-\bar{m}+i}^{\mathsf{T}} q); \\
\quad q := q + \beta s_{k-\bar{m}+i}; \\
\}
\end{array}
\tag{11.25}
$$

La valeur de $W_{k+1}v$ est obtenu à la sortie de l'algorithme (11.25) dans le vecteur q. L'approche que nous avons utilisée pour obtenir l'algorithme de calcul (11.25) montre pourquoi celui est efficace : il est fondé sur l'algorithm compact (11.23) calculant rapidement $\frac{1}{2}v^{\mathsf{T}}W_{k+1}v$ et le mode inverse de différentiation automatique, que l'on sait ne pas dégrader l'efficacité de (11.23).

Le nombre d'opérations pour le calcul de la direction de ℓ-BFGS par (11.25) est approximativement $4mn$ additions et $4mn$ multiplications.

Pour conclure, résumons les instructions de l'algorithme ℓ-BFGS.

Algorithme 11.9 (ℓ-BFGS)

0. On se donne deux constantes ω_1 et ω_2 pour la recherche linéaire de Wolfe : $0 < \omega_1 < \frac{1}{2}$ et $\omega_1 < \omega_2 < 1$.
   Choix d'un itéré initial $x_1 \in \mathbb{R}^n$ et d'un nombre de mises à jour m.
   Initialisation : $k := 1$, $\bar{m} = 0$.
1. *Test d'arrêt* : si $\nabla f(x_k) = 0$, arrêt de l'algorithme.
2. *Calcul de la direction de descente* d_k : on détermine une matrice initiale W_k^0 et on calcule d_k par l'algorithme (11.25) (sauf si $\bar{m} = 0$, auquel cas on prend $d_k = -W_k^0 g_k$).
3. *Recherche linéaire de Wolfe* : trouver un pas $\alpha_k > 0$ tel que l'on ait (6.11a) et (6.11b).
4. $x_{k+1} := x_k + \alpha_k d_k$.
5. Si $k > m$, effacer (y_{k-m}, s_{k-m}) de la mémoire. Stocker (y_k, s_k).
6. Accroître k de 1 et aller en 1.

On résout aujourd'hui couramment des problèmes avec $n \simeq 10^6 \ldots 10^8$ par l'algorithme ℓ-BFGS, en particulier en météorologie [145, 146] ou océanographie.

Notes ▲

L'algorithme de la sécante pour trouver un zéro d'une fonction non linéaire remonte au moins à Newton, probablement vers 1665 [617 ; I, p. 489-491]. [639]

Le théorème 11.6 et son corollaire 11.7 sont dus à Powell [490 ; 1976]. Ce résultat ne peut pas s'étendre aux fonctions non convexes : on a trouvé des contre-exemples lorsque l'algorithme est globalisé par une recherche linéaire exacte [492 ; 1984], par celle de Wolfe [151 ; 2002] ou par celle prenant le pas α_k donnant le minimum exact le long de la direction de recherche et assurant la condition d'Armijo (6.9) [411 ; 2004]. Ce point noir de l'algorithme n'empêche pas sont utilisation avec succès. Certains auteurs ont proposé des modifications de l'algorithme de manière à pouvoir en montrer la convergence sur des fonctions non convexes, mais cette manière de procédé n'a pas l'approbation de tous. Ainsi Li et Fukushima [393 ; 2001] ont obtenu un résultat de convergence globale et superlinéaire avec une modification du vecteur y_k en $y_k + r_k s_k$, où le scalaire $r_k \sim \|g_k\|$. Comme lot de consolation, mentionnons que l'algorithme de DFP, bien moins efficace que l'algorithme de BFGS, converge pour les problèmes avec 2 variables et une recherche linéaire prenant le premier minimum local [496 ; 2000].

Prise en compte de la parcimonie : technique originale due à Toint [592 ; 1977] ; difficultés de non-existence et de non-définie-positivité relevées par Sorensen [564 ; 1982] ; approche de Fletcher [213 ; 1995] utilisant la fonction ψ de Byrd et Nocedal.

L'algorithme ℓ-BFGS de la section 11.2.5 a été proposé par Nocedal [458]. La dérivation de l'algorithme (11.25) présentée, utilisant la différentiation automatique, est reprise de [246]. Dans les problèmes avec contraintes, c'est l'approximation de hessiennes directes dont on a besoin ; on peut en construire des approximations quasi-newtoniennes à mémoire limitée [103, 102]. Le choix du bon nombre m de mises à jour reste aujourd'hui un sujet intriguant. De manière surprenante, il n'est pas vrai que prendre m le plus grand possible améliore les performances de l'algorithme. Le nombre d'itérations pour atteindre un seuil d'optimalité donné peut facilement varier avec un facteur 3 lorsque m s'échelonne, disons de 5 à 50, avec une efficacité optimale pour des valeurs erratiques de m. Boggs et Byrd [68] ont proposé des techniques efficaces pour déterminer m de manière adaptative au cours des itérations. Lin, Harchaoui et Mairal [396 ; 2017] proposent un algorithme fondé sur la technique ℓ-BFGS, destiné à minimiser $f = f_0 + \psi$, où f_0 est convexe $\mathcal{C}^{1,1}$ et ψ est convexe non lisse (par exemple $\psi = \|\cdot\|_1$ ou $\psi = \mathcal{I}_C$), qui peut se voir comme une méthode ℓ-BFGS inexacte sur la régularisée de Moreau-Yosida de f (section 3.7.2).

Des algorithmes de BFGS et de ℓ-BFGS pour la minimisation d'une fonction réelle de variables complexes sont proposés dans [562].

Divers résultats ont aussi été obtenus pour minimiser une fonction non lisse avec l'algorithme de BFGS [392, 637, 296].

Exercices

11.1. *Formule de BFGS par blocs.* Soient Y, $S \in \mathbb{R}^{n\times p}$ avec S injective et $M \in \mathcal{S}^n_{++}$. On considère le problème

$$\begin{cases} \min_{\bar{M}} \ \psi(M^{-1/2}\bar{M}M^{-1/2}) \\ Y = \bar{M}S \\ \bar{M} \in \mathcal{S}^n_{++}, \end{cases}$$

où $\psi : \mathcal{S}^n_{++} \to \mathbb{R}$ est la fonction définie par (11.6). Montrez que ce problème a une solution si, et seulement si, $Y^\mathsf{T}S \in \mathcal{S}^n_{++}$. Montrez que, sous cette condition, la solution $\bar{M}$ du problème est unique et est donnée par l'une des formules suivantes :

$$\bar{M} = M - MS(S^\mathsf{T}MS)^{-1}S^\mathsf{T}M + Y(Y^\mathsf{T}S)^{-1}Y^\mathsf{T}$$

$$\begin{aligned}\bar{W} &= \Big(I - S(Y^\mathsf{T}S)^{-1}Y^\mathsf{T}\Big)W\Big(I - Y(Y^\mathsf{T}S)^{-1}S^\mathsf{T}\Big) + S(Y^\mathsf{T}S)^{-1}S^\mathsf{T}\\ &= W + (S - WY)(Y^\mathsf{T}S)^{-1}S^\mathsf{T} + S(Y^\mathsf{T}S)^{-1}(S - WY)^\mathsf{T}\\ &\quad - S(Y^\mathsf{T}S)^{-1}Y^\mathsf{T}(S - WY)(Y^\mathsf{T}S)^{-1}S^\mathsf{T}.\end{aligned}$$

où on a noté $W := M^{-1}$ et $\bar{W} := \bar{M}^{-1}$.

11.2. *Formule de mise à jour PSB.* Soient y_k et s_k deux vecteurs de $\mathbb{R}^n$ et M_k une matrice d'ordre n symétrique. On considère le problème en $M \in \mathbb{R}^{n\times n}$ suivant

$$\left\{\begin{array}{l}\min\ \|M - M_k\|_F\\ y_k = Ms_k\\ M = M^\mathsf{T},\end{array}\right. \tag{11.26}$$

où $\|\cdot\|_F$ est la norme de Frobenius sur l'ensemble des matrices. Montrez que, si $s_k \neq 0$, le problème (11.26) a une solution unique M_{k+1} et que celle-ci s'écrit

$$M_{k+1} = M_k + \frac{(y_k - M_ks_k)s_k^\mathsf{T} + s_k(y_k - M_ks_k)^\mathsf{T}}{\|s_k\|_2^2} - \frac{(y_k - M_ks_k)^\mathsf{T}s_k}{\|s_k\|_2^4}s_ks_k^\mathsf{T}. \tag{11.27}$$

Si $s_k = 0$ et $y_k \neq 0$, le problème (11.26) n'a pas de solution. Si $s_k = y_k = 0$, le problème (11.26) a comme unique solution $M_{k+1} = M_k$.

Remarque. La formule (11.27) porte le nom de *formule PSB* (Powell-Symétrique-Broyden). En pratique, elle donne généralement de moins bons résultats que la formule de BFGS. On observera que le cas $s_k = 0$ et $y_k \neq 0$ ne peut pas se présenter en optimisation ; quant au cas $s_k = y_k = 0$, il est normalement signe que l'itéré courant est stationnaire.

11.3. *Initialisation de l'algorithme de BFGS.* Montrez que σ_1 et $\bar{\sigma}_1$ donnés dans (11.20) sont respectivement solutions des problèmes en $(\sigma, M) \in \mathbb{R}\times\mathbb{R}^{n\times n}$ et $(\bar{\sigma}, W) \in \mathbb{R}\times\mathbb{R}^{n\times n}$ suivants

$$\left\{\begin{array}{l}\min_{\sigma,M}\ \|\sigma I - M\|_F\\ y_1 = Ms_1\\ M = M^\mathsf{T}\end{array}\right. \quad\text{et}\quad \left\{\begin{array}{l}\min_{\bar{\sigma},W}\ \|\bar{\sigma} I - W\|_F\\ Wy_1 = s_1\\ W = W^\mathsf{T}.\end{array}\right.$$

11.4. *Propriétés de l'algorithme de BFGS.* On considère l'algorithme de BFGS pour minimiser une fonction non linéaire f. On note x_k le k-ième itéré, g_k le gradient de f en x_k, $\mathcal{G}_k := \text{vect}\{g_1, \ldots, g_k\}$, $W_k = M_k^{-1}$ la matrice générée et $\perp_Q$ l'orthogonalité par rapport au produit scalaire associé à une matrice symétrique définie positive Q. Montrez que pour tout $k \geqslant 1$, on a

(*i*) $W_kv \in W_1(\mathcal{G}_k)$, pour tout $v \in \mathcal{G}_k$; en particulier, $W_kg_k \in W_1(\mathcal{G}_k)$;
(*ii*) $W_kv = W_1v$, pour tout $v \perp_{W_1} \mathcal{G}_k$.

De plus, si $W_1 = \sigma I$, avec $\sigma > 0$, on a

(*iii*) $M_kv \in \mathcal{G}_k$, pour tout $v \in \mathcal{G}_k$;
(*iv*) $M_kv = \sigma v$, pour tout $v \perp_I \mathcal{G}_k$.

Remarque. Ce résultat montre que la direction de recherche est dans $W_1(\mathcal{G}_k)$.

Partie III

Méthodes de l'optimisation avec contraintes

12 Projection et activation

Dans ce chapitre, nous présentons quelques méthodes d'optimisation dans lequel le critère est non linéaire et les contraintes sont suffisamment simples. Elles s'inspirent directement des techniques de descente en optimisation sans contrainte, en utilisant la notion de direction de descente et en faisant décroître le critère à chaque itération. Elles ont aussi la caractéristique de maintenir la suite des itérés $\{x_k\}$ dans l'ensemble admissible X. Si l'approche est générale, elle ne trouvera son efficacité que si les contraintes sont affines, si bien que le problème devra pouvoir être mis sous la forme suivante :

$$\begin{cases} \min f(x) \\ Ax \leqslant b, \end{cases} \tag{12.1}$$

où $f : \mathbb{R}^n \to \mathbb{R}$ est une fonction régulière, A est une matrice $m \times n$ et $b \in \mathbb{R}^m$.

Une direction de descente de f en un point qui n'est pas dans l'ensemble admissible X est de peu d'utilité : si elle permet de faire décroître f, elle n'est d'aucune aide pour ramener les itérés dans l'ensemble admissible. Par conséquent, si on désire utiliser des directions de descente de f, on est forcé de générer des suites $\{x_k\}$ *admissibles*, c'est-à-dire contenues dans X. On est alors amené à introduire la notion de *direction de descente admissible* d_k. De petits pas le long de ces directions permettent de faire décroître f tout en restant dans X. Alors on peut par exemple définir la suite $\{x_k\}$ par $x_{k+1} = x_k + \alpha_k d_k$, avec α_k suffisamment petit pour que f décroisse et que x_{k+1} soit dans X.

Cette méthode est attrayante par sa simplicité, mais **elle n'est pas convergente**. On connaît en effet un contre-exemple, dû à Wolfe [625 ; 1972]. Nous n'allons pas le décrire en détail, mais il est instructif de comprendre ce qui s'y passe. Dans celui-ci (voir figure 12.1), on minimise sur l'orthant positif ($X = \{x \in \mathbb{R}^3 : x \geqslant 0\}$) une fonction convexe de 3 variables; on utilise la méthode ci-dessus, en prenant $d_k = -g_k$ comme direction de descente admissible; on prend le pas optimal si celui-ci laisse x_{k+1} dans X, sinon on prend le plus grand pas possible de manière à rester dans X. On constate que l'algorithme construit une suite $\{x_k\}$ oscillant entre deux faces de l'orthant positif et convergeant vers un point *non* stationnaire (!) situé sur l'intersection des deux faces. Que se passe-t-il ? Une façon d'interpréter ce résultat est de dire que les bornes ($x \geqslant 0$) de l'ensemble admissible forcent le pas à être trop petit et provoquent ainsi la « fausse » convergence de la suite. On a vu, en effet, au chapitre 6.3, qu'un des devoirs de la recherche linéaire est d'empêcher le pas d'être trop petit afin d'assurer la convergence des méthodes à directions de descente. La présence des contraintes d'inégalité empêche ici de satisfaire cette propriété.

On peut trouver plusieurs remèdes pour éviter cette fausse convergence. L'un d'eux consiste à poursuivre la recherche du pas au delà du *point d'activation* $x_k + \hat{\alpha}_k d_k$ du

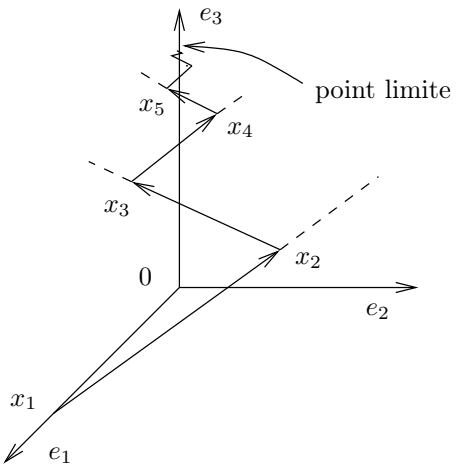


Fig. 12.1. Contre-exemple de Wolfe

chemin $\alpha \mapsto x_k + \alpha d_k$ sur le bord de X. Comme l'on veut que les itérés restent dans X, la recherche se poursuit alors, non pas le long de d_k, mais le long du chemin $\alpha \mapsto x_k + \alpha d_k$ projeté sur X. On parle de *méthodes de projection*. Cette approche sera examinée à la section 12.1. On comprend que pour que celle-ci soit efficace, il faut que la projection sur X soit peu coûteuse. C'est le cas notamment des ensembles X définis par des *contraintes de borne* :

$$X = \{x \in \mathbb{R}^n : a \leqslant x \leqslant b\}, \quad a \text{ et } b \in \mathbb{R}^n. \tag{12.2}$$

Une autre approche consiste à bloquer certaines contraintes pendant quelques itérations (cas où X est donné par des contraintes d'inégalité affines), c'est-à-dire de considérer certaines contraintes d'inégalité comme des contraintes d'égalité. On parle alors de *méthodes différentiation automatique*, aussi appelées *méthodes de pivotage* dans les problèmes de complémentarité [141] (section 12.2). Ces méthodes sont surtout utilisées en *optimisation quadratique*, c'est-à-dire pour la minimisation de fonctions quadratiques en présence de contraintes affines :

$$X = \{x \in \mathbb{R}^n : Ax \leqslant b\}, \quad A : m \times n, \quad b \in \mathbb{R}^m. \tag{12.3}$$

12.1 Méthode de projection

Dans cette section, nous étudions les méthodes dans lesquelles le chemin de recherche de l'itéré suivant est obtenu en projetant sur X le chemin affine $\alpha \mapsto x_k + \alpha d_k$:

$$x_k(\alpha) := \mathrm{P}_X(x_k + \alpha d_k), \tag{12.4}$$

où $d_k \in \mathbb{R}^n$, $\alpha > 0$ et P_X est un opérateur de projection sur X (voir figure 12.2). Par conséquent, $\alpha \mapsto x_k(\alpha)$ est un chemin de recherche admissible et en prenant

$$x_{k+1} = x_k(\alpha_k), \quad \alpha_k > 0, \tag{12.5}$$

on générera une suite $\{x_k\}$ admissible, c'est-à-dire avec $x_k \in X$.

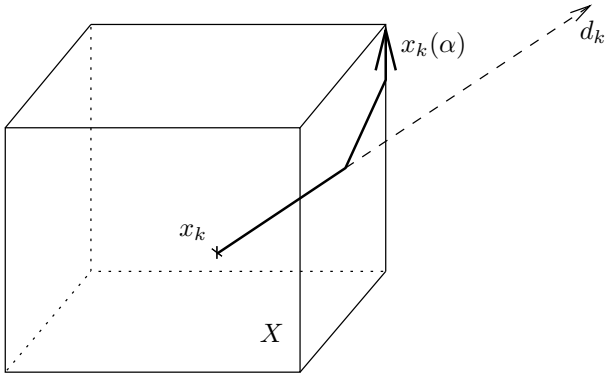


Fig. 12.2. Méthode du chemin projeté

12.1.1 Méthode du gradient projecté

Venons-en maintenant à l'étude d'un algorithme de résolution du problème (12.1), basé sur (12.4)–(12.5). On suppose que X est un convexe fermé non vide de $\mathbb{R}^n$.

La première question qui se pose est de savoir si l'on fait décroître f le long du chemin $x_k(\alpha)$ défini en (12.4). En général, ce n'est pas le cas, tout au moins si d_k est une direction de descente stricte quelconque de f (c.-à-d., une direction vérifiant $g_k^{\mathsf{T}} d_k < 0$, $g_k = \nabla f(x_k)$). On peut en effet construire aisément un exemple où l'on n'a pas

$$f(x_k(\alpha)) < f(x_k), \quad \text{pour } \alpha > 0 \text{ petit.} \tag{12.6}$$

La figure 12.3 illustre une telle situation. La direction d_k est de descente car elle forme avec $-g_k$ un angle plus petit que $\pi/2$ (on suppose que gradient et angle font ici référence au produit scalaire euclidien), ce qui n'est pas le cas du chemin projeté $\alpha \mapsto x_k(\alpha)$. Cependant, si l'on prend $d_k = -g_k$ et donc

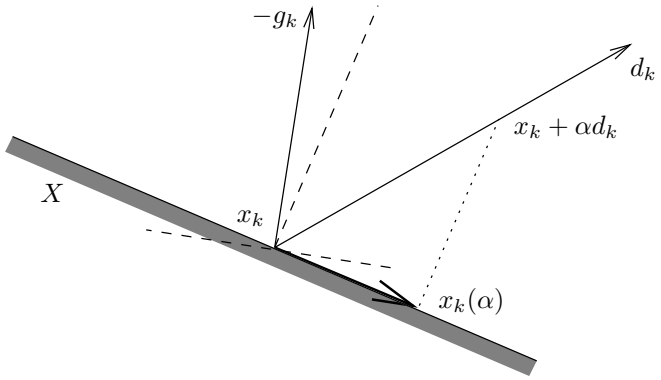


Fig. 12.3. Projection d'une direction de descente d_k quelconque

$$x_k(\alpha) := \mathrm{P}_X(x_k - \alpha g_k), \tag{12.7}$$

la condition (12.6) est vérifiée. Nous allons le voir.

La méthode utilisant le chemin (12.7) comme chemin de descente porte le nom de *méthode du gradient projeté*. Elle a été introduite par Goldstein [262 ; 1964] et Levitin et Polyak [391 ; 1966]. Cette méthode n'est pratiquement utilisable que lorsque la projection sur l'ensemble des contraintes est facile à réaliser, par exemple dans le cas de contraintes de borne (12.2). Les lemmes 12.1, 12.2 et 12.3 donnent quelques propriétés fondamentales de cette méthode.

Lemme 12.1 *Soit X un convexe fermé non vide de $\mathbb{R}^n$. Alors, le point $\bar{x} \in X$ est un point stationnaire du problème (12.1) si, et seulement si,*

$$\bar{x} = \mathrm{P}_X(\bar{x} - \alpha \nabla f(\bar{x})), \tag{12.8}$$

pour un (ou tout) $\alpha > 0$.

Démonstration. D'après la proposition 4.7, lorsque X est convexe, la stationnarité de f en $\bar{x}$ s'exprime par

$$\langle \nabla f(\bar{x}), y - \bar{x} \rangle \geqslant 0, \quad \forall y \in X.$$

Pour $\alpha > 0$, cette condition est équivalente à

$$\langle \bar{x} - (\bar{x} - \alpha \nabla f(\bar{x})), y - \bar{x} \rangle \geqslant 0, \quad \forall y \in X.$$

Comme $\bar{x} \in X$, la caractérisation (2.28a) implique que cette condition est équivalente à (12.8). □

En conséquence de ce lemme, l'itéré courant x_k de l'algorithme (12.5) est un point stationnaire du problème (12.1) si, et seulement si, $x_{k+1} = x_k$.

Lemme 12.2 *On suppose que X est un convexe fermé non vide de $\mathbb{R}^n$ et pour $x_k \in X$, on considère le chemin $\alpha \mapsto x_k(\alpha)$ défini en (12.7). Alors, pour tout $\alpha \in \mathbb{R}$, on a*

$$\alpha \langle g_k, x_k(\alpha) - x_k \rangle \leqslant -\|x_k(\alpha) - x_k\|^2. \tag{12.9}$$

Démonstration. Comme $x_k \in X$ et que $x_k(\alpha)$ est la projection de $x_k - \alpha g_k$ sur X, la caractérisation (2.28a) donne

$$\langle x_k - x_k(\alpha), x_k(\alpha) - x_k + \alpha g_k \rangle \geqslant 0.$$

On en déduit (12.9). □

Ce résultat montre que

$$s_k(\alpha) := x_k(\alpha) - x_k$$

est une direction de descente de f en x_k, quel que soit $\alpha > 0$: l'angle entre $-g_k$ et $s_k(\alpha)$ reste toujours inférieur à $\pi/2$.

Lemme 12.3 *On suppose que X est un convexe fermé non vide de $\mathbb{R}^n$, que f est de classe C^1 et que sa dérivée est lipschitzienne, de module L. Soit $x_k \in X$. Alors, si $0 < \sigma < 1$ et $0 \leqslant \alpha \leqslant 2(1-\sigma)/L$, on a*

$$f(x_k(\alpha)) \leqslant f(x_k) + \sigma \langle g_k, x_k(\alpha) - x_k \rangle. \tag{12.10}$$

Démonstration. On a

$$\begin{aligned} f(x_k(\alpha)) - f(x_k) &= \int_0^1 f'(x_k + t s_k(\alpha)) \cdot s_k(\alpha)\, dt \\ &\leqslant \langle g_k, s_k(\alpha) \rangle + \frac{L}{2} \|s_k(\alpha)\|^2. \end{aligned}$$

En utilisant (12.9), on trouve

$$f(x_k(\alpha)) - f(x_k) \leqslant \left(1 - \frac{L\alpha}{2}\right) \langle g_k, s_k(\alpha) \rangle.$$

En remarquant que, d'après (12.9), $\langle g_k, s_k(\alpha) \rangle$ est négatif lorsque α est positif, on en déduit (12.10) lorsque $0 \leqslant \alpha \leqslant 2(1-\sigma)/L$. □

Dès lors, lorsque x_k n'est pas stationnaire, les inégalités (12.9) et (12.10) montrent que l'on fait décroître f en se déplaçant le long du chemin $\alpha \mapsto x_k(\alpha)$ ($\alpha > 0$ petit).

Les deux résultats suivants donnent des méthodes de détermination du pas α qui assurent la convergence de la méthode du gradient projecté : pas petit (théorème 12.4) et pas d'Armijo (théorème 12.5). Ce sont des extensions naturelles de résultats connus pour les problèmes sans contraintes.

Théorème 12.4 *Soit X un convexe fermé non vide de $\mathbb{R}^n$. Supposons que f soit bornée inférieurement sur X, que f soit de classe C^1 et que sa dérivée soit lipschitzienne de module L. Alors l'algorithme du gradient projecté, avec un pas α_k pris dans un compact de $]0, 2/L[$ génère une suite de points $\{x_k\}$ telle que*

$$\|x_{k+1} - x_k\| \to 0.$$

De plus, tous les points d'adhérence de $\{x_k\}$ sont stationnaires.

Démonstration. Si α_k est dans un compact de $]0, 2/L[$, on peut trouver $\sigma \in]0, 1[$ tel que

$$\alpha_k \in \left[\sigma, \frac{2(1-\sigma)}{L}\right].$$

Alors, d'après les lemmes 12.2 et 12.3, on a

$$\begin{aligned} f(x_{k+1}) &\leqslant f(x_k) - \sigma \frac{\|x_{k+1} - x_k\|^2}{\alpha_k} \\ &\leqslant f(x_k) - \frac{\sigma L}{2(1-\sigma)} \|x_{k+1} - x_k\|^2. \end{aligned}$$

Comme f est bornée inférieurement sur X et que $f(x_k)$ décroît, $f(x_{k+1}) - f(x_k) \to 0$. On en déduit que $\|x_{k+1} - x_k\| \to 0$.

Si $\bar{x}$ est un point d'adhérence de $\{x_k\}$, on peut trouver une sous-suite d'indices $\mathcal{K}$ telle que $x_k \to \bar{x}$, pour $k \to \infty$ dans $\mathcal{K}$. Comme $\|x_{k+1} - x_k\| \to 0$, on a aussi $x_{k+1} \to \bar{x}$, pour $k \to \infty$ dans $\mathcal{K}$. On peut également supposer que $\alpha_k \to \bar{\alpha}$, pour $k \to \infty$ dans $\mathcal{K}$ (au besoin on extrait une nouvelle sous-suite de $\mathcal{K}$). En passant à la limite dans

$$x_{k+1} = \mathrm{P}_X(x_k - \alpha_k \nabla f(x_k))$$

(on se rappelle que P_X est continu, par (2.29)), on a

$$\bar{x} = \mathrm{P}_X(\bar{x} - \bar{\alpha} \nabla f(\bar{x})).$$

Comme $\bar{\alpha} > 0$, le lemme 12.3 montre que $\bar{x}$ est stationnaire. □

En général la constante de Lipschitz de f' n'est pas connue et il n'est pas réaliste (et il est souvent inefficace) de choisir le pas comme dans le théorème 12.4. Il est préférable de déterminer α_k selon la règle suivante qui est une extension de la règle d'Armijo utilisée en optimisation sans contrainte (voir section 6.3.3).

On se donne deux constantes $\sigma \in]0,1[$ et $\beta \in]0,1[$ ainsi qu'une estimation $\alpha_k^0 > 0$ du pas à l'itération k (il est important que $\{\alpha_k^0\}$ et $\{1/\alpha_k^0\}$ forment des suites bornées). On prend $\alpha_k = \beta^{j_k} \alpha_k^0$, j_k étant le plus petit entier tel qu'avec ce α_k, on ait

$$f(x_k(\alpha_k)) \leqslant f(x_k) + \sigma \langle g_k, x_k(\alpha_k) - x_k \rangle. \tag{12.11}$$

Lorsque f est régulière, on peut toujours trouver un tel pas α_k. En effet, d'après le lemme 12.3, cette inégalité est vérifiée lorsque α_k est plus petit que $2(1-\sigma)/L$ (où L est la constante de Lipschitz de f). Ceci montre que par la règle ci-dessus, on a

$$\alpha_k \geqslant \min\left(\alpha_k^0, \frac{2(1-\sigma)\beta}{L}\right). \tag{12.12}$$

On prend ensuite $x_{k+1} = x_k(\alpha_k)$.

Le théorème suivant montre que cette manière de procéder est satisfaisante.

Théorème 12.5 ([50]) *Soit X un convexe fermé non vide de $\mathbb{R}^n$. Supposons que f soit bornée inférieurement sur X, que f soit de classe C^1, avec une dérivée lipschitzienne. Alors, l'algorithme du gradient projeté, avec un pas α_k déterminé par l'extension de la règle d'Armijo (12.11), génère une suite $\{x_k\}$ telle que*

$$\|x_{k+1} - x_k\| \to 0. \tag{12.13}$$

De plus, tous les points d'adhérence de $\{x_k\}$ sont stationnaires.

DÉMONSTRATION. D'après (12.11), le lemme 12.2 et le fait que $\alpha_k \leqslant \alpha_k^0 \leqslant C$, on a

$$f(x_{k+1}) \leqslant f(x_k) - \frac{\sigma}{C}\|x_{k+1} - x_k\|^2.$$

Comme $f(x_k)$ converge, on en déduit (12.13). Si $\bar{x}$ est un point d'adhérence de $\{x_k\}$, on peut trouver une sous-suite d'indices $\mathcal{K}$ telle que

$$x_k \to \bar{x}, \quad \bar{\alpha}_k \to \bar{\alpha}, \quad \text{pour } k \to \infty \text{ dans } \mathcal{K}.$$

D'après (12.12) et le fait que $\{1/\alpha_k^0\}$ est bornée, on a $\bar{\alpha} > 0$. On s'y prend ensuite comme à la fin de la preuve de le théorème 12.4. □

Étant une extension de la méthode de la plus forte pente, la méthode du gradient projeté n'a pas une bonne vitesse de convergence. On peut toutefois utiliser les idées développées ci-dessus pour définir une méthode de Newton projetée : voir Bertsekas [52 ; 1982]. On peut aussi préconditionner la méthode en utilisant des produits scalaires différents pour la définition du gradient de f et pour la projection sur X : voir Gafni et Bertsekas [227 ; 1984].

12.2 Méthodes différentiation automatique ▲

La méthode que nous décrivons dans cette section est fondée sur une idée assez générale. Toutefois, d'un point de vue pratique, elle ne pourra être utilisée avec efficacité qu'en optimisation quadratique. Ces restrictions apparaîtront au cours de l'exposé.

12.2.1 Motivation et schéma des méthodes

Pour construire une suite minimisante $\{x_k\}$ admissible, c'est-à-dire contenue dans X, on peut s'appuyer sur la notion de direction de descente admissible.

Définition 12.6 Soit $x \in X$. On dit que d est une *direction admissible* en x pour X si $x + \alpha d \in X$ pour tout $\alpha > 0$ suffisamment petit (de manière plus précise : il existe $\bar{\alpha} > 0$ tel que pour $0 \leqslant \alpha \leqslant \bar{\alpha}$ on ait $x + \alpha d \in X$). On dit que d est une *direction de descente admissible* en x pour le problème (12.1), s'il existe $\bar{\alpha} > 0$ tel que pour $0 \leqslant \alpha \leqslant \bar{\alpha}$:

$$x + \alpha d \in X \quad \text{et} \quad f(x + \alpha d) \leqslant f(x). \tag{12.14}$$

Il s'agit donc d'une direction de descente (non stricte) de f telle qu'avec un petit déplacement dans cette direction, on reste dans X. □

On ne peut pas toujours trouver une direction vérifiant ces conditions, mais si X est convexe et x_k n'est pas stationnaire, une telle direction existe toujours. En effet, du fait de la convexité de X, les directions de la forme $d_k = y_k - x_k$, où $y_k \in X$ et $y_k \neq x_k$ sont des directions admissibles pour X. Alors, s'il n'existait pas de direction de descente admissible en x_k, cela voudrait dire que, pour $y_k \in X$ fixé, il existerait une suite de $\alpha^i > 0$, $\alpha^i \to 0$, telle que $f(x_k + \alpha^i(y_k - x_k)) > f(x_k)$. En passant à la limite, on trouverait

$$\langle g_k, y_k - x_k \rangle \geqslant 0, \quad \forall y_k \in X.$$

Ceci impliquerait la stationnarité de x_k (proposition 4.7). Nous supposerons donc que X **est convexe** pour pouvoir trouver à chaque étape une direction de descente admissible.

Nous allons utiliser des directions admissibles dans une approche fondé sur la notion d'*activation de contraintes* (en anglais, *Active Set Methods*). L'idée est de bloquer — on dit aussi *activer* — certaines contraintes définissant X, c'est-à-dire que des contraintes d'inégalité seront considérées comme contraintes d'égalité pendant un certain nombre d'itérations (on évite ainsi l'oscillation qui se produit dans le contre-exemple de Wolfe). Ceci n'est concevable que si les contraintes sont affines, sinon la réalisation de $c_i(x_k) = 0$ est trop coûteuse. Pour simplifier l'exposé, nous supposerons qu'il n'y a pas de contraintes d'égalité.

Le problème (P) à résoudre dans cette section est donc

$$(P) \quad \begin{cases} \min f(x) \\ Ax \leqslant b, \end{cases} \tag{12.15}$$

où A est une matrice $m \times n$ et $b \in \mathbb{R}^m$. L'ensemble admissible est convexe (clair). Les équations d'optimalité s'écrivent en $(\bar{x}, \bar{\lambda})$ (les contraintes étant affines, elles sont qualifiées; on peut donc trouver un multiplicateur $\bar{\lambda}$) :

$$\begin{cases} \nabla f(\bar{x}) + A^{\mathsf{T}}\bar{\lambda} = 0, \\ A\bar{x} \leqslant b, \\ \bar{\lambda} \geqslant 0, \\ \bar{\lambda}^{\mathsf{T}}(A\bar{x} - b) = 0. \end{cases} \tag{12.16}$$

On note $A_{(j)}$, la j-ième ligne de A. Si $J \subseteq [1:m]$ est un ensemble d'indices, on note A_J la matrice $|J| \times n$, obtenue en extrayant de A ses lignes $j \in J$. On définit $b_{(j)}$ et b_J de manière analogue. Enfin, l'ensemble actif en x_k, $I^0(x_k)$, se notera ici simplement

$$I_k = \{j : 1 \leqslant j \leqslant m,\ A_{(j)}x_k = b_{(j)}\}.$$

Soit T_k une partie de I_k qu'on appelle *ensemble de travail*. On considère alors le problème

$$(\mathrm{P}_{T_k}) \quad \begin{cases} \min f(x) \\ A_{T_k}x = b_{T_k}, \end{cases} \tag{12.17}$$

dans lequel x est contraint à appartenir à l'*espace actif*

$$E_{T_k} = \{x \in \mathbb{R}^n : A_{T_k}x = b_{T_k}\}.$$

Ce problème est en général beaucoup plus facile à résoudre que le problème (12.15). Ses équations d'optimalité du premier ordre s'écrivent : $\exists \lambda_k \in \mathbb{R}^m$ tel que

$$\begin{cases} \nabla f(x_k) + A_{T_k}^{\mathsf{T}}(\lambda_k)_{T_k} = 0 \\ A_{T_k} x_k = b_{T_k}. \end{cases} \tag{12.18}$$

On a le résultat instructif suivant.

Proposition 12.7 *Soient $\bar{x}$ une solution du problème (P) et x_k une solution du problème (P_{T_k}), dans lequel $T_k = I^0(\bar{x})$.*
(i) Si f est convexe sur E_{T_k} et $x_k \in X$, alors x_k est solution de (P).
(ii) Si f est strictement convexe sur E_{T_k}, alors $x_k = \bar{x}$.

Démonstration. Montrons que

$$f(x_k) = f(\bar{x}).$$

D'une part, $f(x_k) \leqslant f(\bar{x})$, car $\bar{x}$ est solution du problème (P_{T_k}) auquel on ajoute la contrainte d'appartenance à X. D'autre part, si $d_k = x_k - \bar{x}$, la convexité de f sur E_{T_k} permet d'écrire

$$f(\bar{x}) + f'(\bar{x}) \cdot d_k \leqslant f(x_k).$$

Alors, si $f(x_k) < f(\bar{x})$, on aurait $f'(\bar{x}) \cdot d_k < 0$ et d_k serait une direction de descente stricte de f en $\bar{x}$:

$$f(\bar{x} + \alpha d_k) < f(\bar{x}), \quad \text{pour } \alpha > 0 \text{ petit.}$$

Mais comme $d_k \in \mathcal{N}(A_{I^0(\bar{x})})$, on a

$$\bar{x} + \alpha d_k \in X, \quad \text{pour } \alpha > 0 \text{ petit.}$$

Ces deux dernières conclusions contrediraient l'optimalité de $\bar{x}$. Donc $f(x_k) = f(\bar{x})$.

Les conditions (i) et (ii) de l'énoncé se déduisent immédiatement de ce résultat. □

Ce résultat montre que pour minimiser une fonction f strictement convexe sur X, il y a deux choses à faire : identifier $I^0(\bar{x})$, c'est-à-dire trouver les indices des contraintes actives en la solution, et minimiser f sur l'espace actif correspondant.

Donc, si f est strictement convexe et le problème (P) a une solution, on pourrait s'y prendre de la manière suivante. On calcule la solution de (P_{T_k}) pour toutes les combinaisons possibles d'ensemble actif (en nombre fini) — on écarte les problèmes sans solution. Parmi ces solutions, on ne retient que celles qui sont dans X. Parmi ces dernières, celle qui donne la plus petite valeur de f est la solution de (12.15). Cependant, *cette manière de procéder est à proscrire*, du fait du nombre exponentiel de problèmes (P_{T_k}) à résoudre (il y en a 2^m si on résout aussi ceux qui ont plus de n contraintes d'égalité, ce qui n'est pas inutile car la matrice extraite A_{T_k} peut être de rang inférieur à n).

Les méthodes différentiation automatique gardent de cette démarche l'idée de résoudre un nombre fini (et restreint) de problèmes (P_{T_k}), en les sélectionnant de manière qu'on espère appropriée. À chaque itération, on dispose d'un ensemble de travail $T_k \subseteq I_k$ servant à identifier $I^0(\bar{x})$. Celui-ci est mis à jour après une phase de minimisation (éventuellement partielle) de f sur l'espace actif E_{T_k} correspondant. Voici précisément comment ces algorithmes s'y prennent.

Schéma 12.8 (activation de contraintes)

1. Initialisation : choisir $x_1 \in X$ ($Ax_1 \leqslant b$) et $T_1 \subseteq I_1$;
2. Pour $k = 1, 2, \ldots$ faire :
   - 2.1. Si x_k est un point stationnaire de (P_{T_k}) et $(\lambda_k)_{T_k}$ est un multiplicateur associé, alors :
     - 2.1.1. Si $(\lambda_k)_{T_k} \geqslant 0$, x_k est un point stationnaire de (P) et on s'arrête;
     - 2.1.2. Sinon on désactive des contraintes correspondant à des $(\lambda_k)_{(j)} < 0$, $j \in T_k$ (mise à jour de T_k); On prend $x_{k+1} := x_k$; On passe à l'itération suivante;
   - 2.2. Trouver une direction de descente admissible d_k pour (P_{T_k}) en x_k;
   - 2.3. Par une règle de recherche linéaire « convenable », trouver un pas $\alpha_k \geqslant 0$ tel que $x_{k+1} := x_k + \alpha_k d_k \in X$;
   - 2.4. Choisir T_{k+1} tel que $T_k \subseteq T_{k+1} \subseteq I_{k+1}$; Si $\alpha_k = 0$ en 2.3, alors il faut que pour un indice $l \in I_k \setminus T_k$ d'une contrainte sur laquelle on bute en x_k dans la direction d_k, on ait $l \in T_{k+1}$.

Cet algorithme demande quelques explications, notamment en ce qui concerne les affirmations qui y sont faites. Nous allons les donner.

Avant cela, remarquons que tout n'est pas spécifié dans cet algorithme : à l'étape 2.1.2, on ne dit pas si on désactive une ou toutes les contraintes à multiplicateur strictement négatif; à l'étape 2.2, on ne précise pas comment on doit déterminer la direction d_k; à l'étape 2.3, la règle de recherche linéaire n'est pas très explicite; enfin, à l'étape 2.4, un certain choix est laissé à l'utilisateur quant à la détermination de T_{k+1}. Il ne s'agit donc que d'un schéma d'algorithme qu'on adaptera à chaque situation. Remarquons également qu'on ne désactive des contraintes qu'à l'étape 2.1.2, lorsque x_k est solution de (P_{T_k}), et qu'on n'active des contraintes qu'à l'étape 2.4.

L'affirmation de l'étape 2.1.1 est justifiée parce que si x_k est point stationnaire de (P_{T_k}), on a (12.18). On voit alors qu'en prenant $\bar{x} = x_k$ et $\bar{\lambda} \in \mathbb{R}^m$ avec

$$\bar{\lambda}_{(i)} = \begin{cases} (\lambda_k)_{(i)} & \text{si } i \in T_k \\ 0 & \text{sinon,} \end{cases}$$

le couple $(\bar{x}, \bar{\lambda})$ vérifie (12.16). Pour avoir un minimum de (P), il suffirait que x_k soit un minimum de (P_{T_k}) avec $\nabla^2 f(x_k)$ défini positif sur E_{T_k} et que les composantes de $(\lambda_k)_{T_k}$ soient strictement positives (utiliser directement le théorème **??** et la remarque **??**). Ce n'est pas ce que l'algorithme teste; il ne pourra donc trouver que des points stationnaires.

À l'étape 2.1.2, l'intérêt de désactiver des contraintes correspondant à des multiplicateurs strictement négatifs est de pouvoir, *en général*, faire décroître f à l'itération suivante. En effet, supposons que l'on désactive les contraintes d'indices $j_1 \in T_k, \ldots, j_p \in T_k$, correspondant à des multiplicateurs $(\lambda_k)_{(j_1)}, \ldots, (\lambda_k)_{(j_p)}$ strictement négatifs. Supposons également que A_{T_k} soit surjective; elle a donc un inverse à droite que

l'on note $A^-_{T_k}$. Alors, en prenant des coefficients positifs $\alpha_1, \ldots, \alpha_p$, non tous nuls, la direction

$$d = -\sum_{i=1}^{p} \alpha_i A^-_{T_k} e^{j_i}_{T_k}$$

(e^j est le j-ième vecteur de base de $\mathbb{R}^m$ et $e^j_{T_k}$ est le vecteur extrait) vérifie

$$A_{T_k} d = -\sum_{i=1}^{p} \alpha_i e^{j_i}_{T_k}. \tag{12.19}$$

En utilisant l'équation d'optimalité de x_k, on a

$$f'(x_k) \cdot d = -(\lambda_k)^{\mathsf{T}}_{T_k} A_{T_k} d = \sum_{i=1}^{p} \alpha_i (\lambda_k)_{(j_i)} < 0.$$

Ceci montre que d est une direction de descente de f en $x_k = x_{k+1}$. On déduit aussi de (12.19) que $A_{(j)}d = 0$, pour $j \in T_k \setminus \{j_1, \ldots, j_p\} = T_{k+1}$ et donc que d est admissible pour $(\mathrm{P}_{T_{k+1}})$. Mais bien que l'on ait $A_{(j_i)}d = -\alpha_i \leqslant 0$, pour $i = 1, \ldots, p$, il se peut que d ne soit pas admissible pour (P), parce qu'il peut y avoir un indice $j \in I_k \setminus T_k$ pour lequel $A_{(j)}d > 0$. On aurait alors un pas nul à l'étape 2.3 de l'itération suivante avec un danger de *cyclage* (on retourne périodiquement au même ensemble de travail sans bouger en x). Comme notre discussion l'indique, ceci ne se produira pas si $\mathcal{R}(A^{\mathsf{T}}_{T_k}) = \mathcal{R}(A^{\mathsf{T}}_{I_k})$ et si A_{T_k} est surjective. En particulier, si A_{I_k} est surjective, il suffira de prendre $T_k = I_k$. On est donc conduit à la définition suivante.

Définition 12.9 On dit que les contraintes $Ax \leqslant b$ sont *non dégénérées* si $A_{I^0(x)}$ est surjective pour tout x vérifiant $Ax \leqslant b$. □

Comme à l'étape 2.3, x_k n'est pas stationnaire pour (P_{T_k}), on peut trouver une direction de descente admissible pour ce problème.

Il sera utile de calculer le *pas d'activation* $\hat{\alpha}_k$, c'est-à-dire le pas α maximal laissant $x_k + \alpha d_k$ dans X. On doit avoir $\hat{\alpha}_k$ maximal tel que

$$A_{(j)}(x_k + \hat{\alpha}_k d_k) \leqslant b_{(j)}, \quad \forall j.$$

Comme $Ax_k \leqslant b$, il suffit de considérer les indices j tels que $A_{(j)}d_k > 0$. On voit que

$$\hat{\alpha}_k = \min_{j : A_{(j)}d_k > 0} \frac{b_{(j)} - A_{(j)}x_k}{A_{(j)}d_k}. \tag{12.20}$$

Par convention, $\hat{\alpha}_k = +\infty$ si $A_{(j)}d_k \leqslant 0$ pour tout j.

L'algorithme différentiation automatique s'utilise avec plus ou moins de succès pour la minimisation de fonctions régulières *quelconques* en présence de contraintes linéaires, avec une légère modification toutefois. Lorsque la fonction n'est pas quadratique, calculer une solution exacte de (P_{T_k}) est une tâche longue et coûteuse. On se contente alors de tester à l'étape 2.1 si x_k est « presque » stationnaire et si c'est le cas, on désactive des contraintes correspondant à des multiplicateurs négatifs.

Comme on le comprend, la nécessité de devoir désactiver des contraintes afin d'identifier l'espace actif au point optimal – ce qui ne peut se faire en toute sécurité qu'en des points stationnaires d'un problème (P_{T_k}) – fait que la méthode différentiation automatique convient préférentiellement à la minimisation de fonctions quadratiques :

$$f(x) = \frac{1}{2} x^{\mathsf{T}} Q x + q^{\mathsf{T}} x, \tag{12.21}$$

où Q est une matrice symétrique d'ordre n et $q \in \mathbb{R}^n$. Dans la suite nous allons voir comment préciser l'algorithme pour qu'il soit convergent dans les deux cas suivants :

- Q est définie positive (optimisation quadratique strictement convexe),
- Q est semi-définie positive (optimisation quadratique convexe).

Le second cas inclut le premier, mais l'introduction progressive des solutions algorithmiques aidera à la compréhension des méthodes. Lorsque Q est indéfinie le problème peut avoir de nombreuses solutions locales et la recherche d'un minimum local est plus difficile; quant à la recherche du minimum global, c'est un problème très difficile (NP-complet, dans la terminologie de la théorie de la complexité).

Notes

La méthode du gradient projeté a été à l'origine proposée par Goldstein [262 ; 1964] ainsi que Levitin et Polyak [391 ; 1966] pour minimiser une fonction différentiable sur un convexe fermé non vide de $\mathbb{R}^n$. Le théorème 12.4 est repris de ces articles. Le théorème 12.5 est repris de Bertsekas [50 ; 1976].

Pour l'algorithme du gradient projeté, voir Calamai et Moré [104 ; 9187].

Exercices

12.1. *Projection sur un pavé.* Pour $l \in (\mathbb{R} \cup \{-\infty\})^n$ et $u \in (\mathbb{R} \cup \{+\infty\})^n$ tels que $l \leqslant u$, on définit le pavé $[l, u] := \{x \in \mathbb{R}^n : l \leqslant x \leqslant u\}$. Le projeté $\bar{x} = \mathrm{P}_{[l,u]}(x)$ de $x \in \mathbb{R}^n$ sur $[l, u]$, pour le produit scalaire euclidien, est donnée par

$$\bar{x}_i = \begin{cases} l_i & \text{si} \quad x_i < l_i \\ x_i & \text{si} \quad l_i \leqslant x_i \leqslant u_i \\ u_i & \text{si} \quad u_i < x_i. \end{cases}$$

En particulier, le projeté orthogonal de $x \in \mathbb{R}^n$ sur l'orthant positif $\mathbb{R}^n_+$ est $x^+ := \max(0, x)$ (le maximum est pris composante par composante).

12.2. *Gradient projeté.* Soient X une partie *convexe* d'un espace euclidien $\mathbb{E}$ et $f : \mathbb{E} \to \mathbb{R}$ une fonction dérivable en $x \in X$. Le *gradient projeté* de f en x est la projection orthogonale du gradient $\nabla f(x)$ de f en x sur $-\mathrm{T}_x X$. On le note

$$g^{\mathrm{P}}(x) := \mathrm{P}_{-\mathrm{T}_x X} \nabla f(x).$$

1) On a aussi

$$g^{\mathrm{P}}(x) := -\mathrm{P}_{\mathrm{T}_x X}(-\nabla f(x)).$$

2) La condition d'optimalité du premier ordre $\nabla f(x) \in (\mathrm{T}_x\, X)^+$ (théorème 4.6) est équivalente à l'équation $g^{\mathrm{P}}(x) = 0$.
3) S'il est non nul, l'opposé du gradient projeté est une direction de descente de f en x, car on a
$$\langle \nabla f(x), -g^{\mathrm{P}}(x)\rangle \leqslant -\|g^{\mathrm{P}}(x)\|^2.$$
4) On simplifie les notations en introduisant $g := \nabla f(x)$ et $g^{\mathrm{P}} := g^{\mathrm{P}}(x)$. Alors, quel que soit $\alpha \geqslant 0$, on a
$$\mathrm{P}_X(x - \alpha g) = x - \alpha g^{\mathrm{P}} \qquad \Longleftrightarrow \qquad x - \alpha g^{\mathrm{P}} \in X, \tag{12.22}$$
autrement dit, la projection du chemin émanant de x et porté par l'opposé du gradient se confond avec le chemin émanant de x et porté par l'opposé du gradient projeté, tant que ce dernier chemin est dans X. Par ailleurs, lorsque X est polyédrique, les propriétés dans (12.22) ont lieu pour tout $\alpha \geqslant 0$ petit.

Supposons maintenant que $\mathbb{E} = \mathbb{R}^n$, que l'on munit du produit scalaire euclidien. Pour $l \in (\mathbb{R} \cup \{-\infty\})^n$ et $u \in (\mathbb{R} \cup \{+\infty\})^n$ tels que $l < u$, on définit le pavé $[l, u] := \{x \in \mathbb{R}^n : l \leqslant x \leqslant u\}$ et on considère un point $x \in [l, u]$.
5) $\mathrm{T}_x[l, u] = \{d \in \mathbb{R}^n : d_i \geqslant 0 \text{ si } x_i = l_i,\ \ d_i \leqslant 0 \text{ si } x_i = u_i\}$.
6) Pour tout $i \in [1\,{:}\,n]$, on a
$$[g^{\mathrm{P}}(x)]_i = \begin{cases} \min(0, \frac{\partial f}{\partial x_i}(x)) & \text{si } l_i = x_i \\ \frac{\partial f}{\partial x_i}(x) & \text{si } l_i < x_i < u_i \\ \max(0, \frac{\partial f}{\partial x_i}(x)) & \text{si } \quad x_i = u_i. \end{cases}$$
7) Si x minimise f sur $[l, u]$, alors
$$\frac{\partial f}{\partial x_i}(x) \begin{cases} \geqslant 0 & \text{si } l_i = x_i \\ = 0 & \text{si } l_i < x_i < u_i \\ \leqslant 0 & \text{si } \quad x_i = u_i. \end{cases}$$

12.3. *Projection sur l'intersection d'un pavé et d'une contrainte affine.*

12.4. *Projection sur* $\mathbb{R}^n_{\leqslant}$.

13 Pénalisation

Cette théorie et ces phénomènes nous montrent comment on peut amener autrui à modifier ses comportements sans recourir à l'autorité, ni même à quelque stratégie persuasive, mais par des moyens détournés.

J.-L. Beauvois et R.-V. Joule (1987). Petit traité de manipulation à l'usage des honnêtes gens [39].

13.1 Vue d'ensemble

La pénalisation est un concept simple qui permet de transformer un problème d'optimisation avec contraintes en *un* problème ou en *une suite* de problèmes d'optimisation sans contrainte ou avec des contraintes simples. C'est un concept qui a une utilité à la fois théorique et numérique. Ce que le qualificatif « simple » signifie dépendra du contexte, des questions que l'on se pose, de la disponibilité d'algorithmes de résolution.

En analyse, l'approche par pénalisation est parfois utilisée pour étudier un problème d'optimisation dont certaines contraintes sont difficiles à prendre en compte, alors que le problème pénalisant ces contraintes difficiles a des propriétés (l'existence de solution par exemple) mieux comprises ou plus simples à mettre en évidence. Si l'on a de la chance ou si la pénalisation est bien choisie, des *passages à la limite* parfois délicats permettent d'obtenir des propriétés du problème original. Par exemple, on peut obtenir des conditions d'optimalité d'un problème avec contraintes, à partir des conditions d'optimalité des problèmes pénalisés ([207, 99], proposition 13.11). D'autre part, comme nous allons le souligner ci-dessous, la pénalisation est un outil permettant d'étudier les problèmes d'optimisation avec et sans contrainte dans un même formalisme.

D'un point de vue numérique, cette transformation en problèmes sans contrainte (ou avec contraintes simples) permet d'utiliser des algorithmes d'optimisation sans contrainte (ou avec ces contraintes simples) pour obtenir la solution de problèmes dont l'ensemble admissible peut avoir une structure complexe. Cela semble merveilleux, inespéré, de voir que l'on puisse ainsi utiliser des algorithmes qui ne cherchent qu'à minimiser une fonction pour trouver des points qui, en plus d'être optimaux, sont admissibles. Cette approche est de ce fait très souvent utilisée. Elle permet d'obtenir une solution de qualité suffisante rapidement sans avoir à entrer dans l'algorithmique

sophistiquée de l'optimisation avec contraintes. Ce n'est cependant pas une technique universelle, car elle a ses propres inconvénients : non-différentiabilité, nécessité de minimiser une *suite* de fonctions, parfois de plus en plus mal conditionnées, paramétrage délicat. C'est cette approche qui est suivie, avec un raffinement remarquable, dans les méthodes de points intérieurs (chapitres 18, 18 et 20), conduisant ainsi à des algorithmes polynomiaux pour de grandes familles de problèmes d'optimisation.

La pénalisation peut intervenir dans un contexte très général. On désigne par $\mathbb{E}$ un ensemble arbitraire, par X une partie de $\mathbb{E}$ et par $f : \mathbb{E} \to \mathbb{R}$ une fonction. On considère le problème de minimiser f sur X. L'ensemble admissible X pourra être l'intersection de deux ensembles :

$$X := X_r \cap X_s.$$

L'ensemble X_r est celui défini par des contraintes qui seront *relaxées* ou *relâchées* par la pénalisation. Théoriquement, elles ne seront vérifiées que lorsqu'un paramètre de pénalisation atteindra sa limite ; d'une point de vue numérique, elles ne seront vérifiées qu'à la convergence des algorithmes. L'ensemble X_s est celui des contraintes *simples* ou *strictes*, qui seront maintenues sans être relaxées par la pénalisation. Formalisons cela. Le problème considéré s'écrit donc

$$(P_X) \qquad \inf_{x \in X_r \cap X_s} f(x)$$

et les différentes techniques de pénalisation que nous verrons consistent souvent à transformer ce problème par un ou des problème(s) de la forme

$$(P_r) \qquad \inf_{x \in X_s} \Theta_r(x),$$

où $\Theta_r(x)$ est obtenu en ajoutant à $f(x)$ le terme $r\,p(x)$:

$$\Theta_r := f + r\,p. \tag{13.1}$$

Ici, r est un scalaire strictement positif, appelé *facteur de pénalisation*, et $p : \mathbb{E} \to \mathbb{R}$ est une fonction, dénommée *fonction pénalisante*. La locution *fonction de pénalisation* sera réservée à la fonction que l'on minimise dans le problème de pénalisation, qui est Θ_r dans (P_r). Le but de cette fonction pénalisante est de pénaliser la violation des contraintes (on parle alors de *pénalisation extérieure*, section 13.2) ou l'abord de la frontière de l'ensemble admissible (on parle dans ce cas de *pénalisation intérieure*, section 13.3). Parfois, au lieu de pénaliser le critère f, on pénalise le lagrangien du problème (c'est le cas de l'approche du lagrangien augmenté, section 13.4) ; cette technique a des avantages, mais aussi des inconvénients. Numériquement, l'intérêt de (P_r) est de pouvoir être résolu par une méthode d'optimisation sans contrainte (lorsque $X_s = \mathbb{E}$) ou avec contraintes simples ; celles définissant X_s.

La transformation du problème avec contraintes (P_X) en problème(s) (P_r) sans contrainte ou avec contraintes simples soulève deux questions : Est-ce possible ? Quel en est le prix ? La première question renvoie à celle de savoir si en résolvant (P_r) on résout (P_X) et, de manière plus précise, à celle de la détermination du lien entre les ensembles de solutions de (P_X) et (P_r) et du lien entre les valeurs optimales. La seconde question concerne l'efficacité numérique d'une telle transformation. La

réponse à ces deux questions va dépendre du choix de la fonction pénalisante p et du facteur de pénalisation r.

Par exemple, on pourrait choisir la fonction pénalisante p égale à la fonction indicatrice de X_r, $p = \mathcal{I}_{X_r}$, c'est-à-dire

$$p(x) = \begin{cases} 0 & \text{si } x \in X_r \\ +\infty & \text{si } x \notin X_r. \end{cases}$$

Il est clair que dans ce cas, les problèmes (P_X) et (P_r) sont identiques : ils ont les mêmes ensembles de solutions et la même valeur optimale. Cette fonction pénalisante est parfois utilisée dans la théorie, car elle permet de traiter en même temps les problèmes avec ou sans contrainte (voir chapitre 14). Ce choix de p n'est cependant pas très utile numériquement car les méthodes classiques d'optimisation ne peuvent pas être utilisées sur des fonctions qui prennent la valeur $+\infty$ dans des régions visitées par les itérés (dans quelle direction se déplacer pour faire décroître Θ_r si l'itéré courant n'est pas dans X_r ?). Nous allons donc, dans ce chapitre, introduire diverses fonctions pénalisantes p, autres que l'indicatrice de X_r, et en étudier les propriétés théoriques et algorithmiques.

La première question posée ci-dessus conduit à la notion de pénalisation exacte, à laquelle on fera souvent allusion dans ce chapitre.

Définition 13.1 (pénalisation exacte) On dit qu'une fonction de pénalisation Θ_r associée au problème (P_X) est *exacte* si toute « solution » de (P_X) est « solution » de (P_r) ; on dit qu'elle est *inexacte* dans le cas contraire, c'est-à-dire qu'il y a des « solutions » de (P_X) qui ne sont pas « solution » de (P_r). □

Dans cette définition, le terme « solution » est pris dans un sens ambigu et il faudra chaque fois préciser si l'on veut parler de point stationnaire, de minimum local ou de minimum global.

La structure du problème (P_r), dont le critère est la somme pondérée de deux fonctions, permet d'énoncer d'emblée une propriété très générale sur le comportement de chaque terme *en un minimum global* $\bar{x}_r$ de (P_r), lorsque le facteur de pénalisation r varie. Intuitivement, si r augmente, on attache moins d'importance à f et plus d'importance à p, si bien qu'il semble normal que $f(\bar{x}_r)$ croisse et que $p(\bar{x}_r)$ décroisse. La proposition suivante énonce cela de façon rigoureuse. Le résultat est très général puisqu'il ne requiert aucune hypothèse sur l'ensemble X_s, ni a fortiori d'hypothèse de convexité ou de différentiabilité ; seule la structure du critère Θ_r intervient.

Proposition 13.2 (monotonie en pénalisation) *Soient X_s un ensemble non vide, f et $p : X_s \to \mathbb{R}$ deux fonctions, $r \in \mathbb{R}$ et $\Theta_r := f + r\,p$. Si $r_1 < r_2$ sont deux réels et si $\bar{x}_{r_i} \in \arg\min\{\Theta_{r_i}(x) : x \in X_s\}$ $(i = 1, 2)$, alors*

1) $p(\bar{x}_{r_1}) \geqslant p(\bar{x}_{r_2})$,
2) $f(\bar{x}_{r_1}) \leqslant f(\bar{x}_{r_2})$ *si* $r_1 \geqslant 0$,
3) $\Theta_{r_1}(\bar{x}_{r_1}) \leqslant \Theta_{r_2}(\bar{x}_{r_2})$ *si* $p(\bar{x}_{r_2}) \geqslant 0$.

Démonstration. 1) En exprimant que $\bar{x}_{r_i}$ minimise Θ_{r_i} sur X_s, on obtient :

$$\begin{aligned} f(\bar{x}_{r_1}) + r_1 p(\bar{x}_{r_1}) &\leqslant f(\bar{x}_{r_2}) + r_1 p(\bar{x}_{r_2}) \\ f(\bar{x}_{r_2}) + r_2 p(\bar{x}_{r_2}) &\leqslant f(\bar{x}_{r_1}) + r_2 p(\bar{x}_{r_1}). \end{aligned}$$

En sommant, on trouve $(r_2 - r_1)p(\bar{x}_{r_2}) \leqslant (r_2 - r_1)p(\bar{x}_{r_1})$. Alors $r_2 > r_1$ implique alors que $p(\bar{x}_{r_2}) \leqslant p(\bar{x}_{r_1})$.

2) En tenant compte du point 1 et du fait que $r_1 \geqslant 0$, la première inégalité exposée ci-dessus permet d'écrire

$$f(\bar{x}_{r_1}) + r_1 p(\bar{x}_{r_1}) \leqslant f(\bar{x}_{r_2}) + r_1 p(\bar{x}_{r_1}),$$

si bien que $f(\bar{x}_{r_1}) \leqslant f(\bar{x}_{r_2})$.

3) On a $\Theta_{r_1}(\bar{x}_{r_1}) \leqslant \Theta_{r_1}(\bar{x}_{r_2})$ par l'optimalité de $\bar{x}_{r_1}$; puis $\Theta_{r_1}(\bar{x}_{r_2}) \leqslant \Theta_{r_2}(\bar{x}_{r_2})$ par la positivité de $p(\bar{x}_{r_2})$. □

Le même raisonnement montre que l'on a une croissance ou décroissance *stricte* des suites si $\bar{x}_r$ est l'*unique* minimum de (P_r) et si $\bar{x}_r$ change avec r.

Un second résultat général, ne dépendant principalement que de la structure de Θ_r, concerne les points d'adhérence des minimiseurs $\bar{x}_r$ des fonctions de pénalisation Θ_r lorsque $r \downarrow 0$. Dans ce cas, la fonction pénalisante p agit de moins en moins et il est naturel de se demander si les minimiseurs $\bar{x}_r$ de Θ_r ne convergeraient pas vers un minimiseur de f sur X_s, donc sans tenir compte de la contrainte d'appartenance à X_r, à savoir un élément de $S := \arg\min\{f(x) : x \in X_s\}$. La proposition suivante donne des conditions pour qu'il en soit ainsi. Puisque'il est question de convergence des $\bar{x}_r$, on a besoin cette fois d'une topologie sur $\mathbb{E}$.

Proposition 13.3 (point d'adhérence lorsque $r \downarrow 0$) *Soit $\mathbb{E}$ un espace topologique. Supposons que X_s soit fermé dans $\mathbb{E}$, que f et $p : X_s \to \mathbb{R}$ soient s.c.i., que $S := \arg\min\{f(x) : x \in X_s\}$ soit non vide et que, pour une suite de $r \downarrow 0$, (P_r) ait au moins une solution, notée $\bar{x}_r$. Alors tout point d'adhérence de $\{\bar{x}_r\}_{r\downarrow 0}$ est solution de*

$$\inf_{x \in S} p(x).$$

Démonstration. Soient $\hat{x} \in S \neq \varnothing$ et $\bar{x}$ un point d'adhérence de $\{\bar{x}_r\}_{r\downarrow 0}$ (pour une sous-suite de $r \downarrow 0$). Il suffit de montrer que $\bar{x} \in S$ et que $p(\bar{x}) \leqslant p(\hat{x})$.

[*Montrons que $\bar{x} \in S$*] Pour cela, on observe que $\bar{x} \in X_s$, parce que X_s est fermé, que $\bar{x}_r \in X_s$ et que $\bar{x}_r \to \bar{x}$. Il reste à montrer que, pour un point $x \in X_s$ arbitraire, on a $f(\bar{x}) \leqslant f(x)$. L'optimalité de $\bar{x}_r$ permet d'écrire

$$\forall\, x \in X_s : \quad f(\bar{x}_r) + r p(\bar{x}_r) \leqslant f(x) + r p(x). \tag{13.2}$$

On passe à la limite inférieure lorsque $r \downarrow 0$ dans cette inégalité :

$$\begin{aligned} f(\bar{x}) &\leqslant \liminf_{r\downarrow 0} f(\bar{x}_r) \qquad [f \text{ est s.c.i. et } \bar{x}_r \to \bar{x}] \\ &\leqslant \liminf_{r\downarrow 0} f(\bar{x}_r) + \liminf_{r\downarrow 0} rp(\bar{x}_r) \qquad [(\text{A.12}),\ p \text{ est s.c.i. et } p(\bar{x}) \in \mathbb{R}] \\ &\leqslant \liminf_{r\downarrow 0} \big(f(\bar{x}_r) + rp(\bar{x}_r)\big) \qquad [(\text{A.2a})] \\ &\leqslant \liminf_{r\downarrow 0} \big(f(x) + rp(x)\big) \qquad [(13.2)] \\ &= f(x). \end{aligned}$$

[*Montrons que* $p(\bar{x}) \leqslant p(\hat{x})$] Observons d'abord que $f(\bar{x}_r) \in \mathbb{R}$ (car f est à valeurs réelles sur X_s et $\bar{x} \in S \subseteq X_s$). Ensuite, comme $\hat{x} \in S$, on a $f(\hat{x}) \leqslant f(\bar{x}_r)$, si bien que (13.2) en $x = \hat{x} \in X_s$, $f(\bar{x}_r) \in \mathbb{R}$ et $r > 0$ conduisent à

$$p(\bar{x}_r) \leqslant p(\hat{x}).$$

En prenant la limite inférieure lorsque $r \downarrow 0$, on trouve que $p(\bar{x}) \leqslant p(\hat{x})$. □

Nous serons souvent amenés à considérer le problème d'optimisation sous contraintes fonctionnelles suivant :

$$(P_{EI,X}) \qquad \begin{cases} \min\ f(x) \\ c_i(x) = 0, & i \in E \\ c_i(x) \leqslant 0, & i \in I \\ x \in X_s, \end{cases} \tag{13.3}$$

où f et les c_i sont des fonctions définies sur un espace euclidien $\mathbb{E}$ (produit scalaire noté $\langle \cdot, \cdot \rangle$) à valeurs dans $\mathbb{R}$, (E, I) forme une partition de $[1\,{:}\,m]$ et X_s est un fermé de $\mathbb{E}$. Le nombre de contraintes d'égalité et d'inégalité se note m_E et m_I, et leur somme $m = m_E + m_I$. Le lagrangien de ce problème est l'application $\ell : \mathbb{E} \times \mathbb{R}^m \mapsto \mathbb{R}$ définie en $(x, \lambda) \in \mathbb{E} \times \mathbb{R}^m$ par

$$\ell(x, \lambda) := f(x) + \lambda^{\mathsf{T}} c(x).$$

Si $v \in \mathbb{R}^m$, on note $v^{\#} \in \mathbb{R}^m$ le vecteur défini par

$$(v^{\#})_i = \begin{cases} v_i & \text{si } i \in E \\ v_i^+ & \text{si } i \in I. \end{cases} \tag{13.4}$$

Les contraintes de (P_{EI,X_s}) s'écrivent alors simplement $c(x)^{\#} = 0$ (attention : l'application $x \mapsto c(x)^{\#}$ n'est pas différentiable en général ; on n'a donc fait que remplacer la difficulté liée à la présence de contraintes d'inégalité par une autre !). Nous renvoyons le lecteur à la section 4.4 pour d'autres notations.

13.2 Pénalisation extérieure

13.2.1 Définition et exemples

Commençons l'exposé par un exemple, qui résume assez bien le fonctionnement de la pénalisation extérieure.

Exemple 13.4 On considère le problème à une variable et une contrainte suivant

$$\begin{cases} \inf\ 1 - x - \frac{1}{3}x^3 \\ x \leqslant 0, \end{cases} \tag{13.5}$$

auquel on associe la fonction de pénalisation suivante

$$\Theta_r(x) = 1 - x - \frac{1}{3}x^3 + r(x^+)^2.$$

où $x^+ := \max(0, x)$. Cette pénalisation est dite *quadratique*, à cause de l'exposant 2 sur x^+ (section 13.2.4). L'effet de cette pénalisation peut s'observer à la figure 13.1, dont l'abscisse est $c(x) = x$. On voit que le terme $r(x^+)^2$ ne joue un rôle qu'à l'extérieur

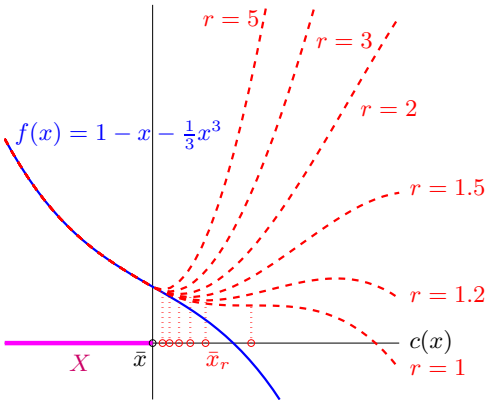


Fig. 13.1. Pénalisation quadratique du problème (13.5), avec $r = 1$, 1.2, 1.5, 2, 3 et 5.

de l'ensemble admissible $\mathbb{R}_-$. C'est la raison pour laquelle on qualifie cette pénalisation d'*extérieure*. D'autre part, on observe que le minimiseur *local* de Θ_r (il n'existe ici que si $r > 1$) est extérieur à l'ensemble admissible. Plus r est grand, plus le minimiseur se rapproche de la solution du problème, qui est ici $\bar{x} = 0$, et la suite de ces minimiseurs converge vers cette solution lorsque r tend vers l'infini. Cependant, plus r est grand, plus le minimum local est accentué (la dérivée seconde de Θ_r si elle existait serait élevée), ce qui pourra être une source de difficultés numériques. □

Dans la section 13.2.2, nous montrerons de manière rigoureuse que ce que nous venons d'observer sur cet exemple simple se produit pour une grande classe de fonctions de pénalisation. Cet exemple est en effet représentatif d'une technique de pénalisation appelée *pénalisation extérieure*, que l'on peut formaliser. Dans celle-ci, la fonction de pénalisation est de la forme $\Theta_r = f + r\,p$, avec une fonction pénalisante p vérifiant les propriétés suivantes :

$$p \text{ est s.c.i. sur } \mathbb{E}, \tag{13.6a}$$

$$\forall x \in \mathbb{E} :\ p(x) \geqslant 0, \tag{13.6b}$$

$$p(x) = 0 \iff x \in X. \tag{13.6c}$$

Observons que ces conditions impliquent que X est un ensemble de sous-niveau de la fonciton s.c.i. p, si bien qu'il est nécessairement fermé. Le qualificatif « *extérieur* » vient de la propriété (13.6c), qui exprime que Θ_r ne modifie f qu'à l'extérieur de l'ensemble admissible.

Le tableau 13.1 donne des exemples de fonctions pénalisantes satisfaisant (13.6),

Contraintes définissant X	Fonction pénalisante	Nom de la pénalisation	Références
$c(x) = 0$	$p(x) = \|c(x)\|_2^2$	quadratique	[143, 144]
$c(x) \leqslant 0$	$p(x) = \|c(x)^+\|_2^2$	quadratique	[2]

Tableau 13.1. Exemples de pénalisation extérieure.

si l'on suppose que c a une propriété de continuité adéquate, conduisant à ce que l'on appelle la *pénalisation quadratique* (section 13.2.4). Pour $u \in \mathbb{R}^p$, on y a noté u^+ le vecteur de $\mathbb{R}^p$ dont la i-ième composante vaut $\max(u_i, 0)$. La norme ℓ_2 y est notée $\|\cdot\|_2$. Les fonctions pénalisantes p du tableau 13.1 font de $\Theta_r = f + r\,p$ une fonction de pénalisation inexacte, puisque l'on trouve pour toute solution $\bar{x}$ de (P_X) :

$$\nabla \Theta_r(\bar{x}) = \nabla f(\bar{x})$$

et que rien n'impose, dans les conditions d'optimalité, que ce vecteur soit nul. Donc $\bar{x}$ n'est généralement pas solution de (P_r).

13.2.2 Propriétés

Nous étudions dans cette section les propriétés asymptotiques des solutions (minima globaux) $\bar{x}_r$ de (P_r), lorsque r tend vers l'infini. La proposition 13.5 énonce des conditions pour que les points d'adhérence de $\{\bar{x}_r\}$ soient solutions du problème original (P_X). La proposition 13.6 montre que, si l'une des fonctions de pénalisation Θ_r est coercive, l'existence d'une suite bornée $\{\bar{x}_r\}$ de minimiseurs est assurée; comme cette suite a alors des points d'adhérence, le résultat de la proposition précédente 13.5 s'y applique.

Proposition 13.5 (optimalité asymptotique) *Soient X un fermé non vide, f une fonction semi-continue inférieurement et $p : \mathbb{E} \to \mathbb{R}$ une fonction vérifiant (13.6). Alors, tout point d'adhérence de la suite $\{\bar{x}_r\}_{r\uparrow\infty}$ est solution de (P_X).*

Démonstration. Soit $\bar{x}$ un point d'adhérence de $\{\bar{x}_r\}_{r\uparrow\infty}$ et $x_{r_i} \to \bar{x}$ pour une sous-suite de facteurs de pénalisation $\{r_i\}_{i\geqslant 0} \to \infty$. Observons que, pour i fixé,

$$\forall\, x \in X : \quad \Theta_{r_i}(\bar{x}_{r_i}) \leqslant \Theta_{r_i}(x) = f(x), \tag{13.7}$$

où la première inégalité provient de l'optimalité de $\bar{x}_{r_i}$ et la seconde de l'admissibilité de x et de (13.6c).

Montrons d'abord que $\bar{x} \in X$ ou encore, par (13.6c), que $p(\bar{x}) = 0$. Comme f est s.c.i., on peut trouver un indice i_0 tel que

$$\forall\, i \geqslant i_0 : \quad f(\bar{x}) - 1 \leqslant f(\bar{x}_{r_i}).$$

En ajoutant $r_i p(\bar{x}_{r_i})$ aux deux membres de cette inégalité et en utilisant (13.7), on obtient

$$\forall\, x \in X \text{ et } i \geqslant i_0 : \quad f(\bar{x}) + r_i p(\bar{x}_{r_i}) - 1 \leqslant \Theta_{r_i}(\bar{x}_{r_i}) \leqslant f(x).$$

Dès lors,

$$\forall\, x \in X \text{ et } i \geqslant i_0 : \quad p(x_{r_i}) \leqslant \frac{f(x) - f(\bar{x}) + 1}{r_i}.$$

En fixant $x \in X$ et en prenant la limite inférieure quand $i \to \infty$, on trouve par (13.6a) et (13.6b) que $0 \leqslant p(\bar{x}) \leqslant \liminf_{i\to\infty} p(x_{r_i}) \leqslant 0$. Donc $p(\bar{x}) = 0$.

Montrons maintenant l'optimalité de $\bar{x}$. En minorant $\Theta_{r_i}(\bar{x}_{r_i})$ par $f(\bar{x}_{r_i})$ dans (13.7), on obtient

$$\forall\, x \in X : \quad f(\bar{x}_{r_i}) \leqslant f(x).$$

Comme f est s.c.i., en prenant la limite inférieure, on a $f(\bar{x}) \leqslant \liminf f(x_{r_i}) \leqslant f(x)$. COmme x est arbitraire dans X et $\bar{x} \in X$, $\bar{x}$ minimise f sur X. □

La proposition suivante suppose la coercivité de Θ_{r_0} pour un certain $r_0 \geqslant 0$, ce qui permet d'assurer l'existence de minimiseurs globaux des Θ_k lorsque $r \geqslant r_0$, ainsi que

Proposition 13.6 (existence d'une suite de minimiseurs) *Soient X un fermé non vide, f une fonction semi-continue inférieurement et $p : \mathbb{E} \to \mathbb{R}$ une fonction vérifiant (13.6). Supposons aussi qu'il existe un $r_0 \geqslant 0$ tel que Θ_{r_0} soit coercive. Alors,*

1) *$\forall r \geqslant r_0$, (P_r) a au moins une solution, que l'on note $\bar{x}_r$,*
2) *la suite $\{\bar{x}_r\}_{r\uparrow\infty}$ est bornée,*
3) *tout point d'adhérence de la suite $\{\bar{x}_r\}_{r\uparrow\infty}$ est solution de (P_X).*

Démonstration. 1) On voit que, pour $r \geqslant r_0$, Θ_r est s.c.i. sur $\mathbb{E}$ (exercice A.3) et est coercive. Par conséquent, (P_r) a au moins une solution $\bar{x}_r$ (corollaire 1.4).

2) Pour $x \in X$ et $r \geqslant r_0$, on a

$$\Theta_{r_0}(\bar{x}_r) \leqslant \Theta_r(\bar{x}_r) \leqslant \Theta_r(x) = f(x). \tag{13.8}$$

La première inégalité vient du fait que $(r - r_0)p(\bar{x}_r) \geqslant 0$ et la seconde de l'optimalité de $\bar{x}_r$. En fixant x dans X, cela montre que $\{\Theta_{r_0}(\bar{x}_r)\}_{r\uparrow\infty}$ est bornée et par la coercivité de Θ_{r_0}, on en déduit que $\{\bar{x}_r\}_{r\uparrow\infty}$ est bornée.

3) C'est une conséquence de la proposition 13.5. □

Les résultats précédents ont une valeur indicative sur le comportement des minima *globaux* de Θ_r. Malheureusement, le même résultat ne tient plus pour les minima locaux. Par exemple, si

$$c(x) = 2x^3 - 3x^2 + 5 = (x+1)(2x^2 - 5x + 5), \tag{13.9}$$

le problème

$$\begin{cases} \inf\ 0 \\ c(x) = 0 \end{cases} \tag{13.10}$$

consiste à chercher l'unique racine réelle $\bar{x} = -1$ de c. Mais le problème pénalisé

$$\inf\ r|c(x)|^2$$

a un minimum local en $\bar{x}_r = 1$ quel que soit $r > 0$. On n'a donc pas la convergence de ces minima locaux vers $\bar{x} = -1$. La proposition 13.11 ci-dessous donne les propriétés des points d'adhérence de la suite des points stationnaires approchés de Θ_r lorsque $r \to +\infty$. Elle est donc complémentaire de la proposition précédente qui ne s'intéresse qu'aux minima globaux.

Une autre limitation du théorème est de supposer que Θ_{r_0} est bornée inférieurement sur $\mathbb{E}$. Si ce n'est pas le cas, il se peut que (P_X) ait une solution mais que (P_r) n'en ait pas. C'est le cas pour le problème suivant

$$\begin{cases} \inf\ x^3 \\ x \geqslant 0 \end{cases}$$

Alors $\Theta_r(x) = x^3 + r(x^-)^2$ n'est pas bornée inférieurement. Dans de pareils cas, on peut rajouter un terme de pénalisation plus fort à l'infini ou introduire des bornes sur les variables.

Remarque 13.7 Pour que la suite de minimiseurs $\{\bar{x}_r\}$ s'approche d'une solution de (P_X), il faut faire croître le facteur de pénalisation r (proposition 13.5). Dans ces conditions, la proposition 13.2 nous apprend que la suite $\{f(\bar{x}_r)\}$ croît. Si la croissance de cette suite est stricte, les points $\bar{x}_r$ ne peuvent être qu'extérieurs à X, sinon les points d'adhérence de $\{\bar{x}_r\}$ ne pourraient pas être solutions de (P_X). En effet, on aurait alors des points de X aussi proches que l'on veut d'une solution $\bar{x}$ en lesquels f prendrait une valeur strictement inférieure à $f(\bar{x})$, ce qui contredirait l'optimalité de $\bar{x}$.

Nous verrons à la remarque 13.12, une autre manière d'aboutir à la même conclusion, après avoir montré comment construire une suite convergeant vers un multiplicateur optimal. □

La propriété suivante montre que la pénalisation extérieure permet d'avoir une borne inférieure du coût optimal de (P_X). Ceci peut être utile pour certaines méthodes de résolution de problèmes d'optimisation avec variables entières, du type *branch-and-bound*. En première analyse, le résultat est un peu magique, car en minimisant la fonction Θ_r qui majore f, on obtient une valeur optimale plus petite que celle de f sur X. La magie s'estombe si l'on se rappelle que (i) la pénalisation n'opère qu'en dehors de l'ensemble admissible, là où le critère peut prendre des valeurs plus basses que $\mathrm{val}(P_X)$ et (ii) la valeur optimale de (P_r) est obtenue par minimisation sur l'espace $\mathbb{E}$ tout entier, alors que celle de (P_X) est le résultat d'une minimisation sur X seulement. La figure 13.1 illustre ces remarques.

Proposition 13.8 (minorant de la valeur optimale)
1) *Si* $p(\cdot) \geqslant 0$, *la suite* $\{\mathrm{val}(P_r)\}$ *croît avec* r.
2) *Si* $p(\cdot) \geqslant 0$ *et* $p(X) = \{0\}$, *la suite* $\{\mathrm{val}(P_r)\}$ *est majorée par* $\mathrm{val}(P)$.

DÉMONSTRATION. 0) Comme au point 3 de la proposition 13.2 (mais ici, on ne suppose pas l'existence de minimiseurs des Θ_r), lorsque $r_1 < r_2$, on a pour tout $x \in \mathbb{E}$:

$$\mathrm{val}(P_{r_1}) \leqslant f(x) + r_1 p(x) \leqslant f(x) + r_2 p(x), \tag{13.11}$$

parce que $r_1 < r_2$ et $p(\cdot) \geqslant 0$.

1) En prenant l'infimum en $x \in \mathbb{E}$ à droite, on obtient $\mathrm{val}(P_{r_1}) \leqslant \mathrm{val}(P_{r_2})$, si bien que $\{\mathrm{val}(P_r)\}$ croît avec r.

2) Si à droite dans (13.11), on prend l'infimum en $x \in X$, on obtient $\mathrm{val}(P_{r_1}) \leqslant \mathrm{val}(P)$, parce que $p(X) = 0$. □

13.2.3 Schéma algorithmique

La proposition 13.5 conduit au schéma algorithmique suivant, qui approche des solutions de (P_X) par des solutions approchées (à ε près) des problèmes pénalisés (P_r), avec $\varepsilon \downarrow 0$ et $r \uparrow \infty$.

Algorithme 13.9 (pénalisation extérieure) Une itération passe de l'itéré courant $(x_k, \varepsilon_k, r_k) \in \mathbb{E} \times \mathbb{R}_{++} \times \mathbb{R}$ à l'itéré suivant $(x_{k+1}, \varepsilon_{k+1}, r_{k+1})$ par les étapes suivantes.

1. *Test d'arrêt.* Arrêt si x_k est satisfaisant.
2. *Nouveaux paramètres.* Choisir $\varepsilon_{k+1} < \varepsilon_k$ (plus de précision) et $r_{k+1} > r_k$ (plus de pénalisation).
3. *Nouvel itéré.* Trouver un minimiseur *approché* (à ε_{k+1} près) x_{k+1} de $\Theta_{r_{k+1}}$ en démarrant les itérations en x_k.

Le schéma algorithmique décrit ci-dessus est simple mais peu précis. Il pose quelques problèmes de mise en œuvre qui méritent d'être discutés.

Remarques 13.10 1) Voici deux critères d'arrêt qui peuvent être utilisés au point 1.

- Pour une fonction pénalisante p vérifiant $p(X) = \{0\}$, $\bar{x}$ est solution de (P_X), si $\bar{x}$ est admissible et minimise Θ_r. En effet, dans ce cas, en plus d'être admissible, $\bar{x}$ vérifie

$$\begin{array}{rcll} f(\bar{x}) & = & \Theta_r(\bar{x}) & [p(X) = \{0\}] \\ & \leqslant & \Theta_r(x), \quad \forall\, x \in \mathbb{E} & [\text{optimalité de } \bar{x}] \\ & = & f(x), \quad \forall\, x \in X & [p(X) = \{0\}], \end{array}$$

si bien que $\bar{x}$ minimise f sur X. On peut donc considérer que si x_k minimise Θ_{r_k} avec suffisamment de précision (i.e., avec un $\varepsilon_k > 0$ suffisamment petit) et si x_k est presque admissible (la vérification de cette propriété peut faire intervenir les fonctions décrivant X), alors x_k est satisfaisant.

- On peut aussi se contenter de l'optimalité au premier ordre, en ne vérifiant que la satisfaction approchée des conditions d'optimalité de KKT (4.33). Il faut alors disposer d'un multiplicateur optimal approché, ce qui est possible pour certaines fonctions pénalisantes (pour la pénalisation quadratique, voir la proposition 13.11).

2) Si l'on veut que les solutions approchées de (P_r) convergent vers une solution de (P_X), il faut nécessairement que $\varepsilon_k \downarrow 0$ et $r_k \uparrow \infty$ (proposition 13.11).

3) À l'étape 3, il n'est pas aisé de donner un critère d'arrêt pour la minimisation de $\Theta_{r_{k+1}}$ qui soit entièrement satisfaisant. Si l'on ne s'intéresse qu'aux points stationnaires, on pourra par exemple décider d'arrêter cette étape si

$$\|\nabla\Theta_{r_{k+1}}(x_{k+1})\| \leqslant \varepsilon_{k+1}.$$

On aimerait en effet ne pas passer trop de temps dans la minimisation de Θ_r si son minimiseur est éloigné de la solution du problème original, parce que r n'est pas assez grand. Nous verrons que, dans l'approche par points intérieurs en optimisation linéaire (chapitre 18), cette question est traitée avec beaucoup plus de précision qu'ici.

4) La manière avec laquelle l'algorithme 13.9 procède, celle de minimiser (approximativement) une suite de fonctions de pénalisation avec des facteurs de pénalisation r croissants, a un aspect étrange. Pourquoi, en effet, ne pas prendre directement r très grand et ne minimiser qu'une seule fonction Θ_r ?

Une première raison est que l'on ne sait pas ce qu'est une valeur de r très grande (cela dépend du problème traité), ce qui donne du sens à cette détermination progressive d'un bon facteur de pénalisation.

Mais il y a une autre raison, plus importante que la première, qui provient du mauvais conditionnement du problème de minimisation de Θ_r pour de grandes valeurs de r (remarque 13.12.2) et des erreurs numériques. Ainsi, si l'on commence à minimiser Θ_r avec une grande valeur de r à partir d'un itéré initial arbitraire, il n'est numériquement pas possible de trouver son minimum, parce que les itérés générés vont surtout se concentrer sur la décroissance de la fonction pénalisante p en ne voyant plus f, dont la contribution à Θ_r est alors marginale. Pour remédier à cette situation, il faut suivre le *chemin des minimiseurs* $r \mapsto \bar{x}_r$ (le tracé progressif rouge à la figure 13.2). Les chances de trouver un point minimisant Θ_r seront d'autant plus grandes que l'itéré initial n'est pas trop éloigné du chemin des minimiseurs et que le minimiseur de Θ_r n'est pas trop éloigné de cet itéré initial. Ces conditions sont remplies dans l'algorithme 13.9 si $r_{k+1} > r_k$ n'est pas pris beaucoup plus grand que r_k, car on y suppose que l'itéré initial x_{r_k} de minimisation de $\Theta_{r_{k+1}}$ est proche de celui minimisant Θ_{r_k}, donc proche du chemin des minimiseurs. □

Les considérations précédentes montrent que l'algorithme 13.9 est conceptuellement simple mais coûteux (il demande beaucoup de temps de calcul), puisqu'il faut

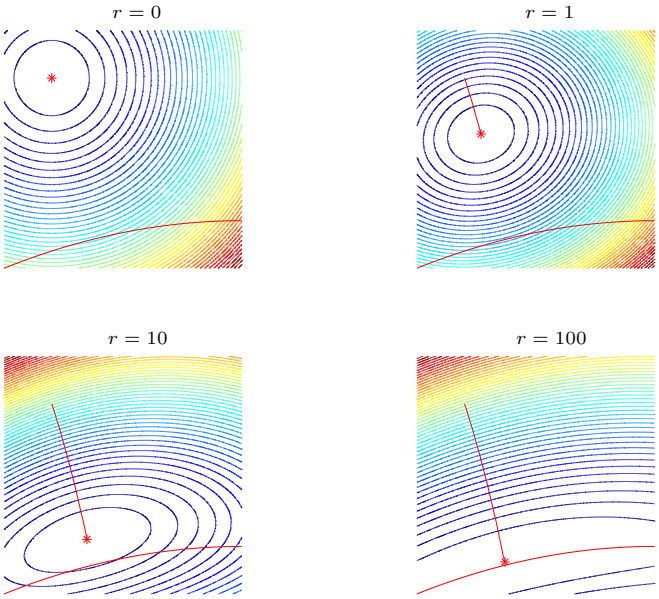


Fig. 13.2. Tracé progressif du *chemin des minimiseurs* dans la pénalisation quadratique du problème $\inf\{\|x - x_0\|_2^2 : c(x) = 0\}$, où $x_0 = (0.2, 0.8)$ et $c(x) = x_2 + 0.2(x_1 - 1)^2 - 0.2$. Celui-ci converge vers la solution du problème et l'on pourra trouver cette solution en suivant ce chemin. Les courbes de niveaux des fonctions de pénalisation, pour des facteurs de pénalisation r pris dans $\{0, 1, 10, 100\}$, montrent la détérioration du conditionnement avec l'augmentation de r (rapprochement progressif de celles-ci sur celles de $\|c(\cdot)\|_2$).

nécessairement résoudre une *suite* de problèmes d'optimisation *non linéaires*, avec l'intérêt de ne jamais devoir prendre en charge explicitement les contraitnes de (P_X).

13.2.4 Pénalisation quadratique

On considère à présent le cas plus concret de la résolution numérique par pénalisation quadratique du problème (P_{EI}), un problème introduit à la section 4.4 et rappelé en (13.3). On parle de *pénalisation quadratique*, lorsque le problème pénalisé associée à (P_{EI}) est le suivant

$$\inf_{x \in \mathbb{E}} \left(\Theta_r(x) := f(x) + \frac{r}{2} \|c(x)^\#\|^2 \right), \tag{13.12}$$

où $v^\# \in \mathbb{R}^m$ est défini en (13.4). Lorsque c est continue, il s'agit d'une pénalisation extérieure, puisque la fonction pénalisante $x \mapsto p(x) := \frac{r}{2}\|c(x)^\#\|^2$ vérifie les conditions (13.6) ; dès lors les résultats obtenus à la section 13.2.2 s'appliquent. Lorsque f et c sont différentiable en $x \in \mathbb{E}$, on a

$$\nabla \Theta_r(x) = \nabla f(x) + \sum_{i \in E \cup I} r[c(x)]_i^\# \, \nabla c_i(x). \tag{13.13}$$

On trouvera à l'exercice 13.1 quelques informations supplémentaires sur cette fonction de pénalisation.

Bien qu'étant une méthode entièrement primale, le schéma algorithmique 13.9 utilisant la pénalisation quadratique permet d'obtenir une estimation des multiplicateurs optimaux associés aux contraintes d'égalité ou d'inégalité de (P_{EI}). Cette estimation peut être précieuse, en particulier, pour estimer l'erreur commise sur les conditions d'optimalité du premier ordre (4.33) et ainsi mesurer la proximité d'un point stationnaire de (P_{EI}). C'est aussi un myen de montrer l'existence de multiplicateur optimaux.

Proposition 13.11 (approximation d'un multiplicateur optimal) *On suppose que les fonctions f et c définissant (P_{EI,X_s}) sont continûment différentiables et que pour une suite de $r \to \infty$, Θ_r a un point stationnaire approché $\bar{x}_r$, dans le sens où*

$$\|\nabla \Theta_r(\bar{x}_r)\| \leqslant \varepsilon_r,$$

avec $\varepsilon_r \to 0$ quand $r \to +\infty$. On suppose aussi que $\bar{x}_r \to \bar{x}$ lorsque $r \to +\infty$, que $\bar{x}$ est un point admissible de (P_{EI}) et que les conditions (QC-MF) ont lieu en $\bar{x}$, Alors

1) *$\{r\, c(\bar{x}_r)^{\#}\}_{r\to\infty}$ est bornée,*
2) *tout point d'adhérence $\bar{\lambda}$ de $\{r\, c(\bar{x}_r)^{\#}\}$ est tel que $(\bar{x}, \bar{\lambda})$ vérifie les conditions d'optimalité (4.33) de (P_{EI,X_s}).*

Démonstration. La différentiabilité supposée de f et c implique celle de Θ_r. Alors, en tenant compte de (13.13), on voit que l'optimalité approchée de $\bar{x}_r$ implique que

$$\nabla f(\bar{x}_r) + \sum_{i\in E\cup I} r[c(\bar{x}_r)]_i^{\#}\, \nabla c_i(\bar{x}_r) \to 0, \quad \text{lorsque } r \to \infty. \tag{13.14}$$

1) Notons $\lambda_r := r[c(\bar{x}_r)]^{\#}$ et montrons par l'absurde que la suite $\{\lambda_r\}_{r\to\infty}$ est bornée. S'il n'en était pas ainsi, on pourrait trouver une sous-suite de $r \to \infty$ telle que

$$\|\lambda_r\| \to \infty \qquad \text{et} \qquad \frac{\lambda_r}{\|\lambda_r\|} \to \mu \neq 0. \tag{13.15}$$

Après division par $\|\lambda_r\|$, la limite (13.14) donne

$$\sum_{i\in E\cup I} \mu_i\, \nabla c_i(\bar{x}) = 0. \tag{13.16}$$

Par la définition de λ_r, on voit que $\mu_I \geqslant 0$. Par ailleurs, la convergence de $\bar{x}_r \to \bar{x}$ et la définition de λ_r montrent que $\mu_{I\setminus I^0(\bar{x})} = 0$. Alors, l'identité (13.16) et (QC-MF) (définition 4.41) impliquent que $\mu = 0$, ce qui est en contradiction avec (13.15).

2) Soit $\bar{\lambda}$ un point d'adhérence de $\{\lambda_r\}$. En passant à la limite dans (13.14), on voit que le gradient du lagrangien s'annule :

$$\nabla f(\bar{x}) + \sum_{i\in E\cup I} \bar{\lambda}_i\, \nabla c_i(\bar{x}) = 0.$$

Par ailleurs, la définition de λ_r montre que $\bar{\lambda}_I \geqslant 0$ et que $\bar{\lambda}_{I\setminus I^0(\bar{x})} = 0$. Le couple $(\bar{x}, \bar{\lambda})$ vérifie donc les conditioins d'optimalité du premier ordre de (P_{EI}). □

Remarques 13.12 1) Accessoirement, la convergence d'une sous-suite convergente de $\{rc(\bar{x}_r)^{\#}\}$ vers un multiplicateur optimal $\bar{\lambda}$ montre que, pour les contraintes d'inégalité c_i associées à des multiplicateurs $\bar{\lambda}_i > 0$, on a $rc_i(\bar{x}_r) > 0$ pour r grand. Ceci veut dire que ces contraintes d'inégalité ne sont pas vérifiées pour r grand. Autrement dit, $\bar{x}_r$ converge vers $\bar{x}$ par l'*extérieur* de l'ensemble défini par ces contraintes.

2) La pénalisation quadratique permet de mettre en évidence l'influence du facteur de pénalisation r sur le *conditionnement du problème* (P_r). Nous entendons par là le conditionnement de la hessienne de Θ_r lorsque celle-ci existe. La formule (13.13) montre que la fonction de pénalisation quadratique Θ_r est différentiable, mais n'est pas nécessairement deux fois différentiable si $I \neq \varnothing$. Pour mettre en évidence par calcul la détérioration du conditionnement du problème (P_r) avec r, considérons alors le cas où il n'y a que des contratintes d'égalité ($I = \varnothing$ et on note $c_E = c$). Alors

$$\begin{aligned} \Theta_r(x) &= f(x) + \frac{r}{2}\|c(x)\|^2, \\ \nabla\Theta_r(x) &= \nabla f(x) + r\, c'(x)^{\mathsf{T}} c(x), \\ \nabla^2\Theta_r(x) &= \nabla^2 f(x) + \sum_{i=1}^{m} r\, c_i(x)\nabla^2 c_i(x) + r\, c'(x)^{\mathsf{T}} c'(x). \end{aligned}$$

Intéressons-nous à la hesienne $\nabla^2\Theta_r(\bar{x}_r)$ en un point stationnaire $\bar{x}_r$ de Θ_r. Sous les hypothèses de la proposition 13.11, les points d'ahérence des facteurs $r\, c_i(\bar{x}_r)$ de la hessienne $\nabla^2 c_i(\bar{x}_r)$ sont des multiplicateurs optimaux, si bien que les deux premiers termes $\nabla^2 f(\bar{x}_r) + \sum_{i=1}^m r\, c_i(\bar{x}_r)\nabla^2 c_i(\bar{x}_r)$ de la hessienne $\nabla^2\Theta_r(\bar{x}_r)$ ont comme point d'adhérence des hessiennes du lagrangien en une solution, si bien que ces termes n'induisent pas un conditionnement problématique. À l'inverse le dernier terme a en général un conditionnement qui explose avec r. En effet, la forme quadratique associée à $r\, c'(\bar{x}_r)^{\mathsf{T}} c'(\bar{x}_r)$, à savoir $v \mapsto r\|c'(\bar{x}_r)v\|^2$, est nulle dans le noyau de $c'(\bar{x}_r)$, alors que sa courbure explose dans l'espace orthogonal (si celui-ci n'est pas de dimension nulle). □

L'application de la pénalisation quadratique à l'optimisation quadratique est examinée à la section **??**.

13.3 Pénalisation intérieure ▲

Dans certains problèmes, le fait que les itérés $\bar{x}_r$ générés par pénalisation extérieure ne soient pas admissibles peut être un inconvénient, par exemple, parce que f n'est pas définie à l'extérieur de X. On peut introduire des méthodes de pénalisation dans lesquelles les itérés $\bar{x}_r$ restent dans X. On parle alors de *pénalisation intérieure*. L'idée est d'utiliser un terme de pénalisation p qui tend vers l'infini lorsque x s'approche de la frontière ∂X de X.

Au problème simple (13.5), on pourra par exemple associé la fonction de pénalisation intérieure, dite logarithmique, suivante

$$\Theta_r(x) = 1 - x - \frac{1}{3}x^3 - r\log(-x).$$

L'effet de cette pénalisation peut s'observer à la figure 13.3.

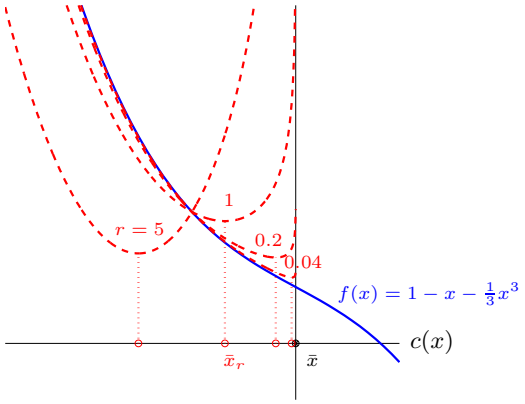


Fig. 13.3. Pénalisation logarithmique de l'exemple (13.5) pour $r = 5,\ 1,\ 0.25,\ 0.05$

On comprend que, dans cette section, il est nécessaire de supposer que l'intérieur X° de X est non vide :

$$X^\circ \neq \varnothing.$$

Cette hypothèse exclut d'emblée la possibilité de prendre en compte directement des contraintes d'égalité. Des artifices permettent toutefois de traiter de telles contraintes.

Les fonctions pénalisantes p considérées dans cette section satisfont les conditions suivantes :

$$p \text{ est continue sur } X^\circ \tag{13.17a}$$

$$p \geqslant 0, \text{ sur } X^\circ \tag{13.17b}$$

$$p(x) \to +\infty, \quad \text{quand } x \in X^\circ \text{ converge vers un point de } \partial X. \tag{13.17c}$$

On considère alors le problème de pénalisation

$$(P_r) \qquad \inf_{x \in X^\circ} \Theta_r(x),$$

où $\Theta_r(x) = f(x) + r\,p(x)$. La condition (13.17c) crée une « barrière » au bord de l'ensemble admissible, si bien que Θ_r porte parfois le nom de *fonction barrière*.

Le tableau 13.2 donne deux exemples de fonctions p satisfaisant (13.17) lorsque l'ensemble admissible s'écrit $X = \{x \in \mathbb{E} : c(x) \leqslant 0\}$, avec $c : \mathbb{E} \to \mathbb{R}^m$. On suppose que $\{x \in \mathbb{E} : c(x) < 0\}$ n'est pas vide. La fonction de *pénalisation intérieure inverse* est due à Carroll [107 ; 1961]. La fonction de *pénalisation logarithmique* est due à l'économétricien norvégien R. Frisch [223 ; 1955]. Cette pénalisation a connu un renouveau avec les *algorithmes de points intérieurs*, que nous étudierons plus en détail au chapitre 18 dans le cadre de l'optimisation linéaire.

Le théorème suivant étudie la suite $\{\bar{x}_r\}$ des solutions des problèmes pénalisés. Contrairement à la pénalisation extérieure, il faut ici faire tendre r vers 0 (et non vers $+\infty$), ce qui a pour effet de diminuer l'influence de la fonction pénalisante, dont le

Contraintes définissant X	Fonction pénalisante	Nom de la pénalisation	Références
$c(x) \leqslant 0$	$p(x) = \sum_{i=1}^{m} \frac{1}{-c_i(x)}$	inverse	[107, 206]
$c(x) \leqslant 0$	$p(x) = -\sum_{i=1}^{m} \log(-c_i(x))$	logarithmique	[223]

Tableau 13.2. Exemples de pénalisation intérieure.

rôle est de repousser les points vers l'intérieur de X, et donc de permettre à $\bar{x}_r$ de se rapprocher de la frontière de l'ensemble admissible, si cela est nécessaire.

Théorème 13.13 (convergence de la pénalisation intérieure) *Supposons que f soit continue sur $\mathbb{E}$ et que l'ensemble admissible X, non vide, vérifie*

$$X = \overline{X^\circ}.$$

On suppose également que soit X est borné, soit $f(x) \to \infty$ quand $\|x\| \to \infty$. Alors, si la fonction pénalisante p vérifie (13.17), on a

1) *$\forall r > 0$, (P_r) a au moins une solution $\bar{x}_r$,*
2) *la suite $\{\bar{x}_r\}_{r\downarrow 0}$ est bornée,*
3) *tout point d'adhérence de $\{\bar{x}_r\}_{r\downarrow 0}$ est solution de (P_X).*

Algorithme 13.14 (pénalisation intérieure) Une itération passe de l'itéré courant $(x_k, \varepsilon_k, r_k) \in \mathbb{E}\times\mathbb{R}_{++}\times\mathbb{R}$ à l'itéré suivant $(x_{k+1}, \varepsilon_{k+1}, r_{k+1})$ par les étapes suivantes.

1. *Test d'arrêt.* Arrêt si x_k est satisfaisant.
2. *Nouveaux paramètres.* Choisir $\varepsilon_{k+1} < \varepsilon_k$ (plus de précision) et $r_{k+1} \in]0, r_k[$ (moins de pénalisation).
3. *Nouvel itéré.* Trouver un minimiseur *approché* (à ε_{k+1} près) x_{k+1} de $\Theta_{r_{k+1}}$ en démarrant les itérations en x_k.

13.4 Le lagrangien augmenté

Since you ask me to mention a gratifying paper, let me pick "A method for nonlinear constraints in minimization problems", because it is regarded as one of the sources of the "augmented Lagrangian method", which is now of fundamental importance in mathematical programming. I have been very fortunate to have played a part in discoveries of this kind.

M.J.D. Powell [497 ; 2003]

Le fait de devoir faire tendre r vers sa valeur limite ($+\infty$ ou 0 suivant le type de pénalisation, extérieure ou intérieure) dans les fonctions de pénalisation précédentes, pour retrouver la solution du problème original, induit, on l'a vu, un mauvais conditionnement des problèmes pénalisés. Il est raisonnable de penser que cet effet numériquement indésirable (si la pénalisation est utilisée comme outil théorique, on se fiche complètement de cette question) sera moins critique si l'on peut construire des fonctions de pénalisation qui, pour une valeur finie du paramètre de pénalisation, ont un minimum en la solution du problème original. C'est ce que l'on a appelé une pénalisation exacte.

Si une fonction de pénalisation Θ_r est exacte et différentiable en une solution $\bar{x}$, on doit avoir $\nabla\Theta_r(\bar{x}) = 0$. C'est bien ce qui manque aux fonctions de la table 13.1, puisqu'elles satisfont $\nabla\Theta_r(\bar{x}) = \nabla f(\bar{x})$, qui n'est pas nul en général. Cela se voit aussi à la figure 13.1 dans laquelle la pente $\Theta_r'(0) = f'(0)$ est non nulle et indépendante de r. Dès lors, avant d'ajouter un terme quadratique à f, il semble judicieux de lui ajouter un terme linéaire de telle sorte que la pénalisation agisse sur une fonction ayant une pente nulle en $\bar{x}$. C'est ce que suggère la figure 13.4, dans laquelle on a

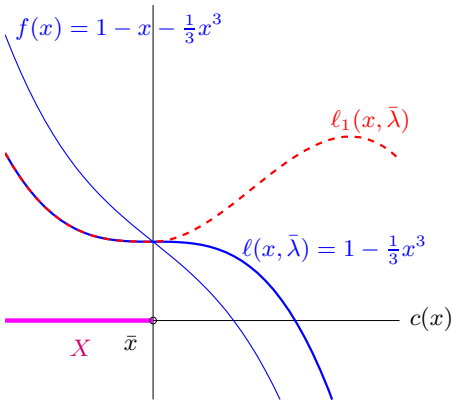


Fig. 13.4. Pénalisation quadratique du lagrangien du problème (13.5)

ajouté la fonction linéaire $x \to x$ au critère $x \mapsto 1 - x - \frac{1}{3}x^3$ de l'exemple 13.4. On voit que la pénalisation quadratique de la fonction résultante admet $\bar{x}$ comme minimiseur local, ici quel que soit le facteur de pénalisation $r > 0$; la fonction de pénalisation obtenue est donc exacte. Que prendre comme terme linéaire ? On peut raisonner de deux manières différentes, qui sont d'ailleurs reliées entre elles. Le plus simple est de dire que l'on connait une fonction dont la pente est nulle en la solution : c'est le lagrangien $\ell(\cdot, \bar{\lambda})$ avec multiplicateur optimal $\bar{\lambda}$, comme nous l'apprennent les conditions d'optimalité (4.33). C'est donc le terme $\bar{\lambda}^\mathsf{T} c(x)$ qu'il faut ajouter à $f(x)$ avant de faire agir la pénalisation. On remarquera que le terme est linéaire en $c(x)$, pas en x (c'est la même chose pour le problème 13.5). On peut aussi raisonner en utilisant l'analyse de sensibilité de la section 4.6.1. La fonction valeur v admet comme dérivée première (sous les hypothèses fortes de la proposition 4.72): $v'(0) = \bar{\lambda}$. Si l'on

veut modifier le problème pour que cette dérivée soit nulle, il faut ajouter $\bar{\lambda}^{\mathsf{T}} c(x)$ au critère, ce qui revient à faire la transformation $f(\cdot) \curvearrowright \ell(\cdot, \bar{\lambda})$ opérée ci-dessus.

13.4.1 Conditions d'exactitude du lagrangien

Le lagrangien du problème (P_{EI,X_s}) est une fonction de pénalisation exacte si les données du problème sont convexes et si le multiplicateur utilisé est optimal. C'est en substance ce qu'affirme le résultat suivant.

Proposition 13.15 (exactitude du lagrangien d'un problème convexe) *Supposons que f et les $\{c_i\}_{i\in I}$ soient convexes et que c_E soit affine. On suppose également que f et c sont différentiables en une solution $\bar{x}$ du problème et qu'il existe un multiplicateur $\bar{\lambda}$ tel que les conditions d'optimalité (KKT) aient lieu. Alors $x \mapsto \ell(x, \bar{\lambda})$ a un minimum en $\bar{x}$.*

Démonstration. Avec les hypothèses de convexité et la positivité de $\bar{\lambda}_I$, l'application $x \in \mathbb{E} \mapsto \ell(x, \bar{\lambda})$ est convexe. Selon les hypothèses, cette fonction est différentiable en $\bar{x}$ et, d'après les conditions d'optimalité (KKT), on a $\nabla_x \ell(\bar{x}, \bar{\lambda}) = 0$. On en déduit que $\ell(\cdot, \bar{\lambda})$ a un minimum (global) en $\bar{x}$. □

Si l'on connaissait un multiplicateur optimal $\bar{\lambda}$, $\ell(\cdot, \bar{\lambda})$ serait une fonction de pénalisation, n'ayant peut-être pas la forme de Θ_r dans (13.1), mais qui pourrait être minimisée pour trouver une solution de (P_{EI,X_s}). Mais $\bar{\lambda}$ est une solution duale, qui doit aussi être trouvée. On appelle *relaxation lagrangienne*, la technique qui recherche une solution de (P_{EI,X_s}) en minimisant $\ell(\cdot, \lambda)$ pour une suite de multiplicateurs λ que l'on fait converger vers un multiplicateur optimal. Il reste à préciser comment on met à jour les multiplicateurs et à étendre cette approche aux problèmes non convexes. C'est ce à quoi nous nous attacherons dans cette section 13.4.

13.4.2 Le lagrangien augmenté de (P_E)

Si le problème n'est pas convexe, le lagrangien n'est plus nécessairement une fonction de pénalisation exacte. Il peut aussi ne pas être borné inférieurement. C'est le cas dans l'exemple (13.5), pour lequel le multiplicateur optimal vaut $\bar{\lambda} = 1$ (d'après la proposition 4.72, c'est $f'(0)$) : $x \mapsto \ell(x, \bar{\lambda}) = 1 + \frac{1}{3}x^3$ n'est pas bornée inférieurement. On comprend qu'en l'absence de convexité, il n'y a plus de sens à minimiser le lagrangien pour obtenir une solution de (P_{EI,X_s}). Le lagrangien augmenté peut être vu comme un moyen de remédier à cet inconvénient.

Commençons par le cas où il n'y a que des contraintes d'égalité, en considérant le problème (P_E) de la page 569. Les conditions d'optimalité du premier et second ordre du théorème 4.17 et de la proposition **??** nous disent que, si l'on restreint h à être voisin de 0 et à appartenir à l'espace tangent aux contraintes en $\bar{x}$, l'application $h \mapsto \ell(\bar{x} + h, \bar{\lambda})$ est convexe et minimale en $h = 0$. Par contre, les conditions d'optimalité ne disent rien sur les valeurs prises par $\ell(\cdot, \bar{\lambda})$ dans l'espace complémentaire au plan tangent. Comme $\nabla_x \ell(\bar{x}, \bar{\lambda}) = 0$, on comprend que l'on pourra convexifier $\ell(\cdot, \bar{\lambda})$ *autour de* $\bar{x}$, et créer ainsi localement une cuvette minimisée en $\bar{x}$, en lui ajoutant un terme positif

qui croît transversalement à la contrainte, mais qui est sans effet longitudinalement. Un terme ayant cette propriété est $\|c(x)\|_2^2$. Il semble donc naturel de considérer la fonction

$$\boxed{\ell_r(x,\mu) = f(x) + \mu^\mathsf{T} c(x) + \frac{r}{2}\|c(x)\|_2^2.} \tag{13.18}$$

Cette fonction, définie pour $(x,\mu) \in \mathbb{E} \times \mathbb{R}^m$, est le *lagrangien augmenté* associé au problème (P_E)[1]. Le facteur de pénalisation $r > 0$ est un paramètre qu'il faudra ajuster, tout comme le multiplicateur μ qui devra tendre vers un multiplicateur optimal.

Examinons la structure du lagrangien augmenté. On vient de le construire comme un lagrangien que l'on a pénalisé par un terme semblable à ceux utilisés en pénalisation extérieure (voir le tableau 13.1). Contrairement aux fonctions de pénalisation de cette section, le lagrangien augmenté sera une fonction de pénalisation exacte lorsque $\mu = \bar\lambda$ et r est assez grand (mais fini !), mais cette propriété ne sera que locale (notion qui sera clarifiée ci-dessous). En ajoutant le terme $\bar\lambda^\mathsf{T} c(x)$ à la fonction de pénalisation inexacte $x \mapsto f(x) + \frac{r}{2}\|c(x)\|_2^2$, on a forcé son exactitude en corrigeant la pente de cette fonction en $\bar x$. En ajoutant le terme $\frac{r}{2}\|c(x)\|_2^2$ au lagrangien, on peut à présent traiter des problèmes non convexes.

L'étude de l'exactitude du lagrangien augmenté (13.18) passe par le lemme B.3 de Finsler. Celui-ci décrit en quelque sorte la version linéarisée (ou « quadratisée ») du mécanisme en jeu dans le lagrangien augmenté (et c'est comme cela que nous l'utiliserons) : M est définie positive dans le noyau de $A^\mathsf{T}A$, mais on n'a pas d'information sur M dans l'espace complémentaire $\mathcal{R}(A^\mathsf{T}A)$ (comparez avec ce que l'on sait sur le lagrangien dans le voisinage d'une solution). En ajoutant un multiple de $A^\mathsf{T}A$ à M, on peut rendre la matrice résultante définie positive. On observera que la matrice ajoutée $rA^\mathsf{T}A$ n'a pas d'effet dans $\mathcal{N}(A)$, puisque $u^\mathsf{T}A^\mathsf{T}Au = 0$ pour tout $u \in \mathcal{N}(A)$.

Théorème 13.16 (exactitude du lagrangien augmenté de (P_E)) *On suppose que f et c sont deux fois dérivables en un minimum local $\bar x$ de (P_E). On suppose également qu'il existe un multiplicateur $\bar\lambda$ tel que $\nabla_x \ell(\bar x, \bar\lambda) = 0$ et tel que la condition suffisante d'optimalité du second ordre (4.28) ait lieu. Alors, il existe un réel $\bar r$ tel que pour tout $r \geqslant \bar r$, le lagrangien augmenté (13.18) a un minimum local strict en $\bar x$.*

Démonstration. On a $\nabla_x \ell_r(x,\bar\lambda) = \nabla_x \ell(x,\bar\lambda) + rA(x)^\mathsf{T} c(x)$, avec $A(x) := c'(x)$. Dès lors $\nabla_x \ell_r(\bar x,\bar\lambda) = 0$. D'autre part, $\nabla^2_{xx}\ell_r(\bar x,\bar\lambda) = \nabla^2_{xx}\ell(\bar x,\bar\lambda) + rA(\bar x)^\mathsf{T}A(\bar x)$. D'après la condition du second ordre (4.28), $\nabla^2_{xx}\ell(\bar x,\bar\lambda)$ est définie positive dans le noyau de $A(\bar x)$. Alors, le lemme B.3 de Finsler nous apprend que $\nabla^2_{xx}\ell_r(\bar x,\bar\lambda)$ est définie positive lorsque r est assez grand. On en déduit que, pour r assez grand, $\ell_r(\cdot,\bar\lambda)$ a un minimum local strict en $\bar x$ (proposition 4.11). □

[1] Ce type de pénalisation porte le nom de lagrangien *modifié* dans la littérature russe.

13.4.3 Le lagrangien augmenté de (P_{EI})

Le lagrangien augmenté associé au problème (P_{EI,X_s}) s'introduit de manière naturelle en utilisant la dualité ; nous le ferons au chapitre 14. Nous allons l'introduire ici par une approche plus intuitive, qui s'appuie sur la formule (13.18) et qui apportera des informations qualitatives intéressantes.

Une idée pourrait être de suivre la structure de (13.18) en ajoutant au lagrangien le terme pénalisant les contraintes utilisé dans (13.12), ce qui donnerait la fonction

$$x \mapsto f(x) + \mu^{\mathsf{T}} c(x) + \frac{r}{2} \|c(x)^{\#}\|_2^2.$$

Celle-ci a l'inconvénient de ne pas toujours être deux fois différentiable en une solution (car $t \mapsto (t^+)^2$) ne l'est pas), quel que soit la régularité de f et c. On préfère donc adopter la démarche suivante, due à Rockafellar [520 ; 1973].

Dans un premier temps, on écrit (P_{EI,X_s}) sous une forme équivalente, en introduisant des *variables d'écart* $s \in \mathbb{R}^{m_I}$:

$$\begin{cases} \inf_{(x,s)} \ f(x) \\ c_E(x) = 0 \\ c_I(x) + s = 0 \\ s \geqslant 0. \end{cases}$$

Ensuite, ce problème est *approché* en utilisant le lagrangien augmenté associé à ses contraintes d'égalité (formule (13.18)) avec un facteur de pénalisation $r > 0$:

$$\inf_x \inf_{s \geqslant 0} \left(f(x) + \mu_E^{\mathsf{T}} c_E(x) + \frac{r}{2} \|c_E(x)\|_2^2 + \mu_I^{\mathsf{T}} (c_I(x)+s) + \frac{r}{2} \|c_I(x)+s\|_2^2 \right).$$

Le lagrangien augmenté du problème (P_{EI,X_s}) est la fonction de (x, μ) définie comme la valeur minimale du problème d'optimisation en $s \geqslant 0$ ci-dessus :

$$\begin{aligned} \ell_r(x,\mu) := \inf_{s \geqslant 0} \Big(& f(x) + \mu_E^{\mathsf{T}} c_E(x) + \frac{r}{2} \|c_E(x)\|_2^2 \\ & + \mu_I^{\mathsf{T}} (c_I(x)+s) + \frac{r}{2} \|c_I(x)+s\|_2^2 \Big). \end{aligned} \tag{13.19}$$

La minimisation en s peut être menée explicitement puisque le problème est quadratique en s avec une hessiene diagonale et que l'on n'a que des contraintes de borne sur s. Plus précisément, comme le critère du problème ci-dessus s'écrit

$$\frac{r}{2} \left\| s + c_I(x) + \frac{\mu_I}{r} \right\|_2^2 + \text{« des termes indépendants de } s \text{ »},$$

il s'agit de projeter $-c_I(x) - \mu_I/r$ sur l'orthant positif. On trouve donc

$$s = \max\left(-c_I(x) - \frac{\mu_I}{r}, 0 \right),$$

si bien que

$$c_I(x) + s = \max\left(-\frac{\mu_I}{r}, c_I(x) \right). \tag{13.20}$$

En remplaçant $c_I(x) + s$ par cette valeur dans le critère du problème d'optimisation ci-dessus on obtient le *lagrangien augmenté* associé aux contraintes $c_E(x) = 0$ et

$c_I(x) \leqslant 0$ du problème (P_{EI,X_s}). C'est la fonction $\ell_r : \mathbb{E} \times \mathbb{R}^m \to \mathbb{R}$, définie pour $(x, \mu) \in \mathbb{E} \times \mathbb{R}^m$ et $r > 0$ par

$$\boxed{\ell_r(x, \mu) = f(x) + \mu^{\mathsf{T}} \tilde{c}_{\mu,r}(x) + \frac{r}{2} \|\tilde{c}_{\mu,r}(x)\|_2^2,} \tag{13.21}$$

où, pour $\lambda \in \mathbb{R}^{m_I}$ et $r > 0$, $\tilde{c}_{\mu,r} : \mathbb{E} \to \mathbb{R}^m$ est définie par

$$(\tilde{c}_{\mu,r}(x))_i = \begin{cases} c_i(x) & \text{si } i \in E \\ \max\left(\frac{-\mu_i}{r}, c_i(x)\right) & \text{si } i \in I. \end{cases} \tag{13.22}$$

Ce lagrangien augmenté a donc une structure tout à fait semblable au lagrangien augmenté associé au problème avec contraintes d'égalité (P_E), pourvu que l'on fasse intervenir la modification non différentiable de c définie par $\tilde{c}_{\mu,r}$ ci-dessus. On notera d'ailleurs que

$$\tilde{c}_{\mu,r}(x) = 0 \qquad \Longleftrightarrow \qquad \begin{cases} c_E(x) = 0 \\ 0 \leqslant \mu_I \perp c_I(x) \leqslant 0, \end{cases}$$

où la notation $0 \leqslant u \perp v \leqslant 0$ signifie $u \geqslant 0$, $v \leqslant 0$ et $u^{\mathsf{T}} v = 0$. La nullité des composantes I de $\tilde{c}_{\mu,r}(x)$ exprime donc à la fois les conditions de signes sur la contrainte et son multiplicateur, ainsi que la complémentarité.

Les deux relations de monotonie du lemme ci-dessous nous seront utiles. La seconde provient de l'expression suivante des termes associés aux contraintes d'inégalité dans la lagrangien augmenté :

$$\mu_I^{\mathsf{T}} (\tilde{c}_{\mu,r}(x))_I + \frac{r}{2} \|(\tilde{c}_{\mu,r}(x))_I\|_2^2 = \frac{1}{2r} \sum_{i \in I} \left[\max(0, \mu_i + r c_i(x))^2 - \mu_i^2\right]. \tag{13.23}$$

Lemme 13.17 1) $\forall\, (x, \mu) \in \mathbb{R}^n \times \mathbb{R}^m,\ \ r \in \mathbb{R}_{++} \to \ell_r(x, \mu)$ *est croissante.*
2) $\forall\, r > 0,\ \ \forall\, \mu_I \in \mathbb{R}^{m_I},\ \ v_I \in \mathbb{R}^{m_I} \to \mu_I^{\mathsf{T}} \max\left(\frac{-\mu_I}{r}, v_I\right) + \frac{r}{2} \left\|\max\left(\frac{-\mu_I}{r}, v_I\right)\right\|^2$ *est croissante (pour l'ordre* $v_I \leqslant v_I' \Leftrightarrow v_i \leqslant v_i'$ *pour tout* $i \in I$*).*

Démonstration. 1) C'est une conséquence de la technique utilisée ci-dessus pour construire le lagrangien augmenté (13.21) : l'argument de l'infimum dans (13.19) croît lorsque r augmente.

2) On utilise l'expression de droite dans (13.23), dont la croissance en $c_I(x)$ est claire, quel que soit le signe de μ_I. □

Malgré l'opérateur max dans (13.22), le lagrangien augmenté est différentiable en (x, λ). Pour le voir, le plus simple est d'utiliser l'expression à droite dans (13.23) pour les termes associés aux contraintes d'inégalité. Comme la dérivée de $(t^+)^2$ est $2t^+$ et que $(a + b)^+ - a = \max(-a, b)$, on obtient

$$\nabla_x \ell_r(x, \mu) = \nabla f(x) + c'(x)^{\mathsf{T}} (\mu + r \tilde{c}_{\mu,r}(x)) \tag{13.24}$$

$$\nabla_\mu \ell_r(x, \mu) = \tilde{c}_{\mu,r}(x). \tag{13.25}$$

L'opérateur max dans (13.22) ne permet pas, en général, d'avoir la différentiabilité seconde de $\ell_r(\cdot,\mu)$: c'est le talon d'Achille de ce lagrangien augmenté. Il a été à la source de nombreux développements. En fait, pour x proche d'une solution $\bar{x}$ de (P_{EI,X_s}) et $\mu=\bar{\lambda}$ (un multiplicateur optimal), on a en utilisant la complémentarité $\bar{\lambda}_I^{\mathsf{T}} c_I(\bar{x})=0$ et la positivité de $\bar{\lambda}_I$:

$$\ell_r(x,\bar{\lambda}) = \ell(x,\bar{\lambda}) + \frac{r}{2}\sum_{i\in E\cup I^{0+}(\bar{x})} c_i(x)^2 + \frac{r}{2}\sum_{i\in I^{00}(\bar{x})} (c_i(x)^+)^2. \tag{13.26}$$

La présence de l'opérateur $(\cdot)^+$ dans (13.26) montre que $\ell_r(\cdot,\bar{\lambda})$ peut ne pas être plus d'une fois différentiable en $\bar{x}$. En cas de complémentarité stricte, $I^{00}(\bar{x})=\varnothing$ et la dernière somme disparaît, si bien que le lagrangien augmenté peut s'écrire (toujours pour x proche de $\bar{x}$):

$$\ell_r(x,\bar{\lambda}) = \ell(x,\bar{\lambda}) + \frac{r}{2}\sum_{i\in E\cup I^{0}(\bar{x})} c_i(x)^2.$$

Localement, les contraintes d'égalité et les contraintes d'inégalité actives sont alors traitées de la même manière et $\ell_r(\cdot,\bar{\lambda})$ est régulière en $\bar{x}$ (pourvu que f et c le soient). Nous résumons ces propriétés de différentiabilité dans la proposition suivante.

Proposition 13.18 (différentiabilité du lagrangien augmenté)
1) *Le lagrangien augmenté ℓ_r, défini en (13.21), est différentiable en μ et son gradient est donné par (13.25).*
2) *Si f et c sont différentiables en x, alors ℓ_r est différentiable en x et son gradient est donné par (13.24).*
3) *Si $(\bar{x},\bar{\lambda})$ est un point stationnaire de (P_{EI,X_s}) vérifiant la complémentarité stricte et si $(f,c_{E\cup I^0(\bar{x})})$ est p fois dérivable ($p\in\mathbb{N}$) dans un voisinage de $\bar{x}$, alors le lagrangien augmenté est p fois dérivable dans un voisinage (éventuellement plus petit) de $\bar{x}$.*

Le résultat suivant est l'analogue du théorème 13.16 pour le lagrangien augmenté (13.21). Il donne des conditions pour que ce lagrangien augmenté soit exact en $\bar{x}$. La condition suffisante d'optimalité semi-forte (4.62) qui y est utilisée est plus forte que la condition faible (4.59), mais plus faible que la condition forte (4.63). Le résultat n'a pas lieu si on ne suppose que la condition faible (4.59).

On a vu au théorème 13.16 que l'exactitude du lagrangien augmenté de (P_E) était fondée sur le lemme de Finsler. Le lecteur perspicace remarquera que la technique utilisée dans la démonstration du théorème ci-dessous est calquée sur celle mise en œuvre en annexe pour démontrer le lemme B.3 de Finsler.

Théorème 13.19 (exactitude du lagrangien augmenté de (P_{EI,X_s})) *On suppose que f et $c_{E\cup I^0(\bar{x})}$ sont deux fois dérivables en un minimum local $\bar{x}$ de (P_{EI,X_s}). On suppose également que les conditions de (KKT) ont lieu et que la*

condition suffisante d'optimalité du second ordre semi-forte (4.62) *a lieu pour un certain multiplicateur optimal* $\bar\lambda$. *Alors, il existe un réel* $\bar r > 0$ *et un voisinage* V *de* $\bar x$ *tels que pour tout* $r \geqslant \bar r$, *le lagrangien augmenté* (13.21), *avec* $\mu = \bar\lambda$, *a un minimum strict en* $\bar x$ *sur* V.

DÉMONSTRATION. Il suffit de montrer qu'il existe un $\bar r > 0$ et un voisinage V de $\bar x$ dans $\mathbb{E}$ tel que

$$\ell_{\bar r}(\bar x, \bar\lambda) < \ell_{\bar r}(x, \bar\lambda), \quad \text{pour tout } x \in V \setminus \{\bar x\}.$$

En effet, si cette affirmation est vraie pour $\bar r$, elle le sera pour tout $r \geqslant \bar r$, avec le même voisinage V. Ceci est dû au fait que $\ell_r(\bar x, \bar\lambda) = f(\bar x)$ est indépendant de r et que, d'après le point (i) du lemme 13.17, $r \mapsto \ell_r(x, \bar\lambda)$ croît avec r.

On démontre cette affirmation par l'absurde en supposant qu'il existe une suite de réels $r_k \to \infty$ et une suite de points $x_k \to \bar x$, tels que pour $k \geqslant 1$: $x_k \neq \bar x$ et

$$\ell_{r_k}(x_k, \bar\lambda) \leqslant \ell_{r_k}(\bar x, \bar\lambda). \tag{13.27}$$

En extrayant une sous-suite au besoin, on peut supposer que pour $k \to \infty$:

$$\frac{x_k - \bar x}{\|x_k - \bar x\|} \to d, \quad \text{avec } \|d\| = 1.$$

En posant $t_k := \|x_k - \bar x\|$, on obtient alors

$$x_k = \bar x + t_k d + o(t_k).$$

Notre but à présent est de montrer que d est une direction critique. Ceci s'obtient en faisant un développement limité des fonctions intervenant dans le membre de gauche de (13.27), exprimé par la formule (13.26) : développement au deuxième ordre du lagrangien et au premier ordre des contraintes dans les deux sommes de (13.26). Pour simplifier les notations, on introduit $\bar L := \nabla^2_{xx}\ell(\bar x, \bar\lambda)$. De la régularité de f et c et de l'optimalité de $(\bar x, \bar\lambda)$, on déduit

$$\begin{aligned} \ell(x_k, \bar\lambda) &= \ell(\bar x, \bar\lambda) + \frac{t_k^2}{2}\, d^{\mathsf T} \bar L d + o(t_k^2), \\ c_i(x_k) &= t_k\, c_i'(\bar x)\cdot d + o(t_k), \quad \text{pour } i \in E \cup I^0(\bar x). \end{aligned}$$

On injecte ces estimations dans (13.27), en utilisant (13.26) et $\ell_{r_k}(\bar x, \bar\lambda) = \ell(\bar x, \bar\lambda)$:

$$\begin{aligned} \frac{t_k^2}{2}\, d^{\mathsf T} \bar L d + o(t_k^2) + \frac{r_k}{2} \sum_{i \in E \cup I^{0+}(\bar x)} \big(t_k c_i'(\bar x)\cdot d + o(t_k)\big)^2 \\ + \frac{r_k}{2} \sum_{i \in I^{00}(\bar x)} \Big([t_k c_i'(\bar x)\cdot d + o(t_k)]^+\Big)^2 \leqslant 0. \end{aligned} \tag{13.28}$$

La limite dans (13.28) quand $k \to \infty$, après avoir divisé par $t_k^2 r_k$, conduit à

$$\begin{aligned} &c_i'(\bar x)\cdot d = 0, \quad \text{si } i \in E \cup I^{0+}(\bar x) \\ &c_i'(\bar x)\cdot d \leqslant 0, \quad \text{si } i \in I^{00}(\bar x). \end{aligned}$$

Dès lors, d est une direction critique non nulle.

D'autre part, (13.28) implique également

$$\frac{t_k^2}{2}\, d^{\mathsf{T}} \bar{L} d + o(t_k^2) \leqslant 0.$$

En prenant la limite après avoir divisé par t_k^2, on obtient $d^{\mathsf{T}} \bar{L} d \leqslant 0$. Cette inégalité est en contradiction avec (4.62), puisque $d \in C(\bar{x}) \setminus \{0\}$. □

13.4.4 Méthode du lagrangien augmenté

Le concept de lagrangien augmenté a été introduit dans un but algorithmique. Il est toujours utilisé aujourd'hui pour résoudre de grands problèmes, surtout lorsque les techniques d'algèbre linéaire requises dans les algorithmes newtoniens comme la *programmation quadratique successive* (voir chapitre 15) ne peuvent pas être utilisées, du fait de la dimension des problèmes. On utilise aussi le lagrangien augmenté pour résoudre des problèmes structurés, comme ceux de l'optimisation quadratique convexe [161, 122 ; 2005-2016].

L'algorithme classique associé au lagrangien augmenté est connu sous le nom de *méthode des multiplicateurs*. Cette méthode s'apparente à la relaxation lagrangienne, dans le sens où l'on y minimise (avec plus ou moins de précision) le lagrangien augmenté pour une suite de multiplicateurs λ_k, que l'on cherche à faire converger vers un multiplicateur optimal, et de facteurs de pénalisation r_k que l'on adapte pour qu'ils soient « suffisamment grands ». L'algorithme est piloté par la recherche du multiplicateur optimal, d'où son nom.

D'un point de vue numérique, l'intérêt principal du lagrangien augmenté, par rapport aux méthodes de pénalisation extérieure et intérieure des sections 13.2 et 13.3, est de ne pas devoir faire tendre le facteur de pénalisation vers une limite qui rend la pénalisation indéfinie ($r \to +\infty$ en pénalisation extérieure) ou inopérante ($r \downarrow 0$ en pénalisation intérieure). Grâce à cette propriété, le lagrangien augmenté conserve un conditionnement raisonnable ; ses courbes de niveau ne s'allongent pas trop. Son inconvénient majeur est de devoir mettre à jour un multiplicateur, en plus du facteur de pénalisation. Heureusement, on dispose d'une formule de mise à jour naturelle, que nous introduirons ici comme une heuristique, mais à laquelle nous donnerons davantage de sens après que les notions de fonction duale et de méthode proximale auront été introduites (voir chapitre 14).

Le théorème 13.19 nous apprend que si l'on connaît un multiplicateur optimal $\bar{\lambda}$ et si l'on prend r assez grand, on a quelques chances de trouver une solution de (P_{EI,X_s}) en minimisant le lagrangien augmenté $\ell_r(.,\bar{\lambda})$. On ne connaît en général aucune de ces deux informations, si bien que c'est algorithmiquement qu'elles doivent être recherchées.

Supposons que l'on dispose au début de l'itération k, d'un facteur de pénalisation $r_k > 0$ et d'un multiplicateur approché $\lambda_k \in \mathbb{R}^m$. La mise à jour de se dernier se fera par une formule qui fait suite aux considérations suivantes (cette formule est vue ici comme une heuristique). Le théorème 13.19 suggère de minimiser $\ell_{r_k}(\cdot, \lambda_k)$. Supposons

que ce problème ait une solution, que l'on note x_k. Par optimalité, $\nabla_x \ell_{r_k}(x_k, \lambda_k) = 0$, qui par (13.24) s'écrit

$$\nabla f(x_k) + \sum_{i \in E \cup I} \Big(\lambda_k + r_k \tilde{c}_{\lambda_k, r_k}(x_k)\Big)_i \nabla c_i(x_k) = 0,$$

où $\tilde{c}_{\lambda_k, r_k}$ est défini en (13.22). Comme, pour résoudre (P_{EI,X_s}), on cherche à annuler le gradient du lagrangien de ce problème, les facteurs de $\nabla c_i(x_k)$ ci-dessus semblent être de bons candidats pour être la nouvelle approximation du multiplicateur optimal. On prend donc

$$\lambda_{k+1} = \lambda_k + r_k \tilde{c}_{\lambda_k, r_k}(x_k).$$

Dans certains cas, comme dans le schéma algorithmique 13.20 ci-dessous, on remplace r_k par un « pas » $\alpha_k > 0$. Donnons une autre expression de $\lambda_k + \alpha_k \tilde{c}_{\lambda_k, \alpha_k}(x_k)$, qui nous informera sur le signe de ses composantes correspondant aux inégalités. Pour $i \in I$ et en laissant tomber l'indice d'itération k, on a $\lambda_i + \alpha \tilde{c}_{\lambda,\alpha}(x)_i = \lambda_i + \alpha \max(-\lambda_i/\alpha, c_i(x)) = \max(0, \lambda_i + \alpha c_i(x))$. On en déduit que

$$\lambda_k + \alpha_k \tilde{c}_{\lambda_k, \alpha_k}(x_k) = (\lambda_k + \alpha_k c(x_k))^{\#}.$$

Les composantes d'indice $i \in I$ de λ_{k+1} sont donc positives.

On manque d'arguments solides pour trouver une règle de mise à jour du facteur de pénalisation r_k qui soit entièrement satisfaisante (le cas des fonctions quadratiques convexes est une exception [161, 122 ; 2005-2016]). On se contente en général d'augmenter r_k si $\tilde{c}_{\lambda_k, r_k}$ ne décroît pas suffisamment vite vers zéro. Cela paraît raisonnable, étant donné que r_k apparaît en facteur de la norme des contraintes dans la lagrangien augmenté (sous la forme (13.19)) et que c'est effectivement son rôle lorsque λ_k reste constant (on est alors proche de la pénalisation extérieure). Mais il ne faut pas oublier que λ_k joue un rôle tout aussi important que r_k pour forcer l'admissibilité des itérés. D'ailleurs dès que r_k est supérieur à un certain seuil, le bon réglage de λ_k devrait suffire pour obtenir l'admissibilité.

Nous donnons ci-dessous un schéma algorithmique inspiré de celui utilisé dans LANCELOT, un code d'optimisation généraliste qui est fondé sur le lagrangien augmenté [137 ; 1992]. Fletcher [212 ; 1987, page 292] propose un algorithme semblable.

Algorithme 13.20 (méthode des multiplicateurs) Une itération passe de l'itéré courant $(\lambda_k, r_k) \in \mathbb{R}^m \times \mathbb{R}_{++}$, vérifiant $(\lambda_k)_I \geqslant 0$, à l'itéré suivant $(\lambda_{k+1}, r_{k+1}) \in \mathbb{R}^m \times \mathbb{R}_{++}$, vérifiant $(\lambda_{k+1})_I \geqslant 0$, par les étapes suivantes.

1. *Nouvel itéré primal.* Avec ℓ_{r_k} défini en (13.21), calculer

$$x_k \in \arg\min \ell_{r_k}(\cdot, \lambda_k).$$

2. *Test de convergence.* Arrêt si $\tilde{c}_{\lambda_k, r_k}(x_k) \simeq 0$, où $\tilde{c}_{\mu,r}$ est défini en (13.22).
3. *Nouveau multiplicateur.* Choisir un pas $\alpha_k > 0$ et prendre

$$\lambda_{k+1} = \lambda_k + \alpha_k \tilde{c}_{\lambda_k, \alpha_k}(x_k) = (\lambda_k + \alpha_k c(x_k))^{\#}.$$

3. *Mise à jour du paramètre de pénalisation* r_k si nécessaire.

Ce schéma algorithmique mérite quelques éclaircissements.

- À l'étape 1, il est rare que l'on minimise complètement le lagrangien augmenté ; de toute façon, un test d'arrêt pour la minimisation de ce problème non linéaire doit être introduit, si le roblème original est lui-même non linéaire. Une possibilité est d'utiliser le principe général selon lequel il ne faut pas être plus exigeant dans la minimisation *interne* (l'étape 1), que ce que l'algorithme a obtenu par la boucle *externe* (ensemble des étapes 1 à 4). Autrement dit, on peut utiliser un test qui compare les normes des quantités qui doivent tendre vers zéro dans les problèmes interne ($\nabla_x \ell_{r_k}(x_k, \lambda_k)$, voir (13.24)) et externe ($\tilde{c}_{\lambda_k, r_k}(x_k)$, voir (13.25)) ; l'algorithme se contente d'un x_k vérifiant

$$\|\nabla_x \ell_{r_k}(x_k, \lambda_k)\| \leqslant \sigma \|\tilde{c}_{\lambda_k, r_k}(x_k)\|, \tag{13.29}$$

  où $\sigma > 0$ est une constante « bien choisie ». À notre connaissance, on n'a pas réussi jusqu'à présent (2021) à démontrer la convergence de l'algorithme avec un critère d'arrêt des itérations internes aussi simple que celui-là, même pour les problèmes convexes, sans pour autant avoir d'argument écartant l'opportunité d'un tel critère. Pour diverses contributions sur ce sujet, on pourra consulter [53, 136, 299, 135, 55, 176, 557, 558, 178, 179, 559, 177, 323, 175, 13, 187, 205, 188 ; 1982-2013].

- Le test d'arrêt de l'étape 2 est fondé sur le fait que l'on a bien optimalité au premier ordre si $\tilde{c}_{\lambda_k, r_k}(x_k) = 0$. En effet, l'optimalité à l'étape 1 et (13.24) impliquent que $\nabla_x \ell(x_k, \lambda_k) = 0$. Par ailleurs, $\tilde{c}_{\lambda_k, r_k}(x_k) = 0$ implique que

$$c_E(x_k) = 0 \qquad \text{et} \qquad \min\left(\frac{(\lambda_k)_I}{r_k}, -c_I(x_k)\right) = 0.$$

  Cette dernière relation est équivalente aux relations de *complémentarité* $0 \leqslant (\lambda_k)_I \perp -c_I(x_k) \geqslant 0$, qui avec $c_E(x_k) = 0$, montrent que les autres conditions du système d'optimalité (4.33) de KKT sont vérifiées.

- À l'étape 3, on prend souvent le pas $\alpha_k = r_k$ comme le suggère la discussion qui précède.

- La seconde partie de la formule de mise à jour de λ_k à l'étape 3 montre que $(\lambda_{k+1})_I \geqslant 0$, si bien que l'algorithme se retrouve au début de l'itération suivante dans les mêmes conditions qu'à l'itération courante.

- La mise-à-jour du facteur de pénalisation r_k à l'étape 4 est une opération délicate, car on manque souvent de connaissance pour en détermniner une valeur correcte, si bien que les concepteurs de solveur utilisent des heuristiques variées. Par exemple, dans [13 ; 2007], r_k n'est mis à jour que lorsqu'au cours de l'itération précédente, une amélioration suffisante de l'admissibiité ou de la complémentarité n'est pas observée. Mentionnons le cas particulier de l'optimisation quadratique convexe, qui est bien mieux compris, pour lequel l'algorithme peut régler la valeur de r_k en fonction de la vitesse de convergence prescrite [161, 122].

Algorithme du lagrangien augmenté pour résoudre un problème convexe non réalisable. Lorsque le problème est convexe, l'algorithme du lagrangien augmenté est identique à l'algorithme proximal sur la fonction duale (proposition 14.39), si bien que son comportement sur des problèmes non réalisables peut se déduire de celui de l'algorithme proximal ; il est décrit dans [97, 510, 567 ; 1977-1987]. On montre que l'algorithme a un comportement maîtrisé :

- il trouve la plus petite translation $\bar{s}$, au sens de la norme ℓ_2, randant le problème réalisable,
- il minimise l'objectif sur les contraintes translatées par ce $\bar{s}$.

On peut être plus précis lorsque le problème est quadratique convexe [218 ; 1982, remarque 5.6] et [159 ; 2006] (pour les contraintes d'égalité seulement) et [122 ; 2016] (pour les contraintes d'inégalité) et montrer un résultat de convergence linéaire globale.

13.4.5 Méthode du lagrangien augmenté à directions alternées ▲

L'algorithme du *lagrangien augmenté à directions alternées*, souvent référencé par son sigle anglo-saxon ADMM abrégeant *Alternating Direction Method of Multipliers*, s'est montré sur le devant de la scène ces dernières années, du fait de son intérêt dans la résolution des *problèmes à données massives* (apprentissage approfondi, acquisition comprimée[2], traitement d'images, *etc*) [300, 85]. Dans ces problèmes aux milliards de variables, on ne cherche pas une solution précise, hors d'atteinte, mais l'accent est mis sur l'obtention *rapide* d'une solution *approchée*. L'algorithme ADMM convient à cette situation. Cet algorithme a été introduit dans les années 1970 pour résoudre certains types d'équations aux dérivées partielles [254, 226]. Il s'agit d'une méthode de lagrangien augmenté, dans laquelle la minimisation de ce dernier, à multiplicateur fixé, se fait de manière très approchée par un unique cycle de Gauss-Seidel (section 7.3.3). L'évocation de ce couple de mathématiciens suggère que le problème puisse être décrit par (au moins) deux groupes de variables, notées $x \in \mathbb{E}$ et $y \in \mathbb{F}$, des groupes parfois introduits artificiellement (par duplication de variables par exemple) de manière à pouvoir appliquer la méthode et bénéficier de ses caractéristiques.

L'algorithme ADMM est en réalité adapté à la résolution de problèmes qui peuvent s'écrire de la manière suivante :

$$\begin{cases} \inf_{(x,y)} \ f(x) + g(y) \\ Ax + By = c, \end{cases} \tag{13.30}$$

où $f : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ et $g : \mathbb{F} \to \mathbb{R} \cup \{+\infty\}$ sont des fonctions définies sur des espaces vectoriels $\mathbb{E}$ et $\mathbb{F}$ pouvant prendre la valeur $+\infty$ (c'est-à-dire pouvant prendre en compte implicitement des contraintes supplémentaires), $A \in \mathcal{L}(\mathbb{E}, \mathbb{G})$ et $B \in \mathcal{L}(\mathbb{F}, \mathbb{G})$ sont des applications linéaires à valeurs dans un espace euclidien $\mathbb{G}$, et $c \in \mathbb{G}$. Le lagrangien augmenté associé est la fonction $\ell_r : \mathbb{E} \times \mathbb{F} \times \mathbb{G} \to \overline{\mathbb{R}}$ définie en $(x, y, \lambda) \in \mathbb{E} \times \mathbb{F} \times \mathbb{G}$ par

[2] Traduction française de *compressed sensing* [35]. Il s'agit de trouver des solutions (de systèmes linéaires sous-déterminés par exemple) avec le plus de zéros. On exprime ce problème en utilisant la norme ℓ_1 : $\inf\{\|x\|_1 : x \in X\}$.

$$\ell_r(x,y,\lambda) = f(x) + g(y) + \langle \lambda, Ax + By - c\rangle + \frac{r}{2}\|Ax + By - c\|^2,$$

où $r > 0$ est le facteur d'augmentation, $\langle\cdot,\cdot\rangle$ est le produit scalaire de $\mathbb{G}$ et $\|\cdot\|$ est la norme associée à ce dernier.

Algorithme 13.21 (ADMM) Une itération passe de l'itéré courant $(y_k, \lambda_k) \in \mathbb{F} \times \mathbb{G}$ à l'itéré suivant (y_{k+1}, λ_{k+1}) par les étapes ci-dessous :

$$x_{k+1} \in \operatorname*{arg\,min}_{x\in\mathbb{E}} \ \ell_r(x, y_k, \lambda_k) \tag{13.31}$$

$$y_{k+1} \in \operatorname*{arg\,min}_{y\in\mathbb{F}} \ \ell_r(x_{k+1}, y, \lambda_k) \tag{13.32}$$

$$\lambda_{k+1} := \lambda_k + r(Ax_{k+1} + Bx_{k+1} - c). \tag{13.33}$$

Comme annoncé, l'algorithme ressemble très fort à l'algorithme du lagrangien augmenté 13.20, si ce n'est que la minimisation complète de $\ell_r(\cdot,\cdot,\lambda_k)$ est remplacée par un cycle gauss-seidelien (section 7.3.3) : une première minimisation en x dans (13.31), avec y fixé à y_k, suivie d'une minimisation en y dans (13.32), avec x fixé à x_{k+1} ; c'est tout, alors que l'algorithme de Gauss-Seidel itérerait ce cycle jusqu'à minimisation complète du lagrangien augmenté. C'est à cette double minimisation que l'on doit le vocable *directions alternées* utilisé dans le nom de l'algorithme. On notera que x_k est une variable auxiliaire qui ne doit pas exister au début de l'itération. Par ailleurs, cet algorithme n'est pas une méthode de dualité (chapitre 14), comme l'est l'algorithme du lagrangien augmenté, car l'itéré contient la variable primale y_k.

On peut montrer que la convergence du coût est en $O(1/k)$ [307, 430]. Des efforts importants ont été faits pour améliorer cette vitesse de convergence du coût, pour qu'elle soit en $O(1/k^2)$, notamment par le schéma de Nesterov [453, 264].

13.5 Pénalisation exacte non différentiable ▲

Le lagrangien augmenté étudié à la section 13.4 est une première manière d'obtenir une pénalisation exacte (voir la définition 13.1), pourvu que l'on connaisse un multiplicateur optimal (comme ce n'est généralement pas le cas, la méthode des multiplicateurs – l'algorithme 13.20 – approche celui-ci de manière itérative). L'idée sous-jacente est de pénaliser quadratiquement une fonction dont la dérivée est nulle en la solution, le lagrangien $\ell(\cdot,\bar\lambda)$.

Une autre manière d'obtenir une pénalisation exacte est de le faire au moyen d'une fonction pénalisante non différentiable. Illustrons cela sur l'exemple 13.4. La figure 13.5 montre que si l'on ajoute le terme $r\,x^+$ à la fonction $1 - x - \frac{1}{3}x^3$, on obtient une fonction de pénalisation exacte dès que $r > 1$. Ce seuil $r = 1$ vient bien sûr ici de la pente de f en zéro et en toute généralité de la pente de la fonction valeur en zéro, si bien que c'est à nouveau le multiplicateur optimal $\bar\lambda$ qui jouera un rôle-clé dans la détermination de ce seuil.

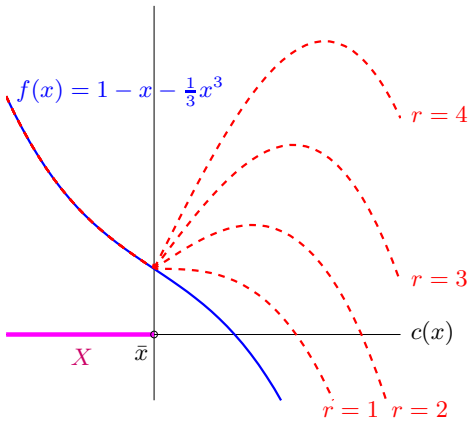


Fig. 13.5. Pénalisation non différentiable du problème (13.5) avec $r = 1$, 2, 3 et 4.

De manière plus générale, si l'on enlève les carrés aux fonctions pénalisantes de la table 13.1, on obtient des fonctions de pénalisation exactes : si r est pris assez grand, les solutions *locales* de (P_X) sont solutions *locales* du problème pénalisé. Cela semble très intéressant, puisque l'on remplace un problème d'optimisation avec contraintes par un *unique* problème d'optimisation sans contrainte. On ne peut cependant pas gagner sur tous les plans : si la fonction de pénalisation est exacte, elle est aussi non différentiable et donc plus délicate à minimiser. La minimisation de ces fonctions se fait en général de manière détournée : on obtient une direction de descente en résolvant un problème auxiliaire et non pas en calculant quelque chose ressemblant à un gradient. Nous verrons au chapitre 15 une utilisation de cette idée.

Dans cette section, on considère le problème d'optimisation avec contraintes explicites

$$(P_{EI}) \quad \begin{cases} \inf\ f(x) \\ c_i(x) = 0, \quad i \in E \\ c_i(x) \leqslant 0, \quad i \in I, \end{cases}$$

décrit dans l'introduction de ce chapitre, auquel on associe la fonction de pénalisation suivante :

$$\Theta_r(x) = f(x) + r\|c(x)^{\#}\|_{\text{P}}, \tag{13.34}$$

où $r > 0$ et $\|\cdot\|_{\text{P}}$ est une norme quelconque. La première pénalisation de ce type a été introduite par Ablow et Brigham [2 ; 1955], avec la norme ℓ_1. Le résultat fondamental est donné à la proposition 13.24, qui énonce des conditions pour que Θ_r soit une fonction de pénalisation exacte.

Sa démonstration utilise les deux lemmes suivants. Le premier étudie la différentiabilité de Θ_r et utilise l'opérateur $P_v : \mathbb{R}^m \to \mathbb{R}^m$, défini pour u et $v \leqslant 0$ dans $\mathbb{R}^m$ par

$$(P_v u)_i = \begin{cases} u_i & \text{si } i \in E \\ u_i^+ & \text{si } i \in I \text{ et } v_i = 0 \\ 0 & \text{si } i \in I \text{ et } v_i < 0. \end{cases}$$

Lemme 13.22 *Si f et c admettent des dérivées directionnelles en un point $x \in \mathbb{E}$, alors Θ_r admet des dérivées directionnelles en x et si $c(x)^\# = 0$, on a*

$$\Theta_r'(x;d) = f'(x;d) + r\|P_{c(x)}c'(x;d)\|_{\mathrm{P}}.$$

Démonstration. Le fait que Θ_r admette des dérivées directionnelles en x vient de ce que $(\cdot)^\#$ et $\|\cdot\|_{\mathrm{P}}$ sont lipschitziennes et admettent, ainsi que f et c, des dérivées directionnelles en x (proposition C.3).

Si $c(x)^\# = 0$, on a

$$\Theta_r'(x;d) = f'(x;d) + r(\|\cdot\|_{\mathrm{P}})'(0;(c^\#)'(x;d)).$$

Mais

$$(\|\cdot\|_{\mathrm{P}})'(0;v) = \lim_{t\downarrow 0}\frac{1}{t}\|tv\|_{\mathrm{P}} = \|v\|_{\mathrm{P}}$$

et

$$(c^\#)'(x;d) = (\cdot^\#)'(c(x);c'(x;d)) = P_{c(x)}c'(x;d).$$

On en déduit le résultat. □

Le second lemme montre qu'avec un facteur de pénalisation suffisamment grand, Θ_r domine le *lagrangien ordinaire* $\ell(\cdot,\bar\lambda)$ sur $\mathbb{E}$. On rappelle de (A.8) que la *norme duale* de $\|\cdot\|_{\mathrm{P}}$ pour le produit scalaire euclidien est la norme $\|\cdot\|_{\mathrm{D}}$ définie par

$$\|y\|_{\mathrm{D}} = \sup_{\|x\|_{\mathrm{P}}\leqslant 1} y^{\mathsf{T}}x.$$

Lemme 13.23 *Si $\lambda \in \mathbb{R}^m$ vérifie $r \geqslant \|\lambda\|_{\mathrm{D}}$ et $\lambda_I \geqslant 0$, alors pour tout $x \in \mathbb{E}$ on a $\ell(x,\lambda) \leqslant \Theta_r(x)$.*

Démonstration. Comme $\lambda_I \geqslant 0$, on a

$$\ell(x,\lambda) \leqslant f(x) + \lambda^{\mathsf{T}}c(x)^\#.$$

On en déduit que

$$\ell(x,\lambda) \leqslant f(x) + \|\lambda\|_{\mathrm{D}}\|c(x)^\#\|_{\mathrm{P}} \leqslant f(x) + r\|c(x)^\#\|_{\mathrm{P}} = \Theta_r(x).$$ □

Voici le résultat annoncé donnant des conditions suffisantes pour que Θ_r soit une fonction de pénalisation exacte en $\bar x$. L'hypothèse (13.35) ne peut avoir lieu que si l'ensemble des multiplicateurs optimaux est borné, ce qui revient à dire que la condition suffisante de qualification des contraintes de Mangasarian-Fromovitz (QC-MF) (définition 4.41) a lieu.

Proposition 13.24 (exactitude de Θ_r) *Soit $\bar{x}$ un minimum local du problème (P_{EI}). On suppose que f et c sont deux fois dérivables en $\bar{x}$ et lipschitziennes dans un voisinage de $\bar{x}$. On suppose également que $\bar{x}$ vérifie les conditions suffisantes d'optimalité du second ordre faible (4.59). On suppose enfin que*

$$r \geqslant \sup_{\bar{\lambda}\in\Lambda(\bar{x})} \|\bar{\lambda}\|_{\mathrm{D}} \qquad \text{et} \qquad r > \|\hat{\lambda}\|_{\mathrm{D}}, \text{ pour un } \hat{\lambda} \in \Lambda(\bar{x}), \tag{13.35}$$

où $\Lambda(\bar{x})$ est l'ensemble non vide des multiplicateurs optimaux associés à $\bar{x}$. Alors, $\bar{x}$ est un minimum local strict de la fonction de pénalisation Θ_r donnée en (13.34).

Démonstration. On raisonne par l'absurde en supposant que $\bar{x}$ n'est pas un minimum local strict de Θ_r. Alors, il existe une suite $\{x_k\}$ telle que $x_k \neq \bar{x}$, $x_k \to \bar{x}$ et

$$\forall k \geqslant 1 : \quad \Theta_r(x_k) \leqslant \Theta_r(\bar{x}). \tag{13.36}$$

1) *Construction d'une direction critique non nulle.* Comme la suite $\{(x_k - \bar{x})/\|x_k - \bar{x}\|\}$ est bornée ($\|\cdot\|$ est une norme arbitraire), on peut en extraire une sous-suite convergente : $(x_k - \bar{x})/\|x_k - \bar{x}\| \to d$, où $\|d\| = 1$. En notant $\alpha_k = \|x_k - \bar{x}\|$, on a

$$x_k = \bar{x} + \alpha_k d + o(\alpha_k).$$

Montrons que d est la direction critique recherchée.

- Comme Θ_r est lipschitzienne dans un voisinage de $\bar{x}$, on a

  $$\Theta_r(x_k) = \Theta_r(\bar{x} + \alpha_k d) + o(\alpha_k).$$

  Alors (13.36) montre que $\Theta'_r(\bar{x}; d) \leqslant 0$. Alors, grâce au lemme 13.22, on peut écrire

  $$f'(\bar{x}) \cdot d + r\|P_{c(\bar{x})}(c'(\bar{x}) \cdot d)\|_{\mathrm{P}} \leqslant 0. \tag{13.37}$$

  On a donc certainement

  $$f'(\bar{x}) \cdot d \leqslant 0. \tag{13.38}$$

- D'autre part, d'après les hypothèses, on peut trouver un multiplicateur optimal $\hat{\lambda}$ tel que $r > \|\hat{\lambda}\|_{\mathrm{D}}$. En utilisant les conditions d'optimalité du premier ordre, notamment la positivité de $\hat{\lambda}_I$ et la complémentarité $\hat{\lambda}_I^{\mathsf{T}} c_I(x_*) = 0$, on a

  $$\begin{aligned} f'(\bar{x}) \cdot d &= -\sum_{i\in E\cup I} \hat{\lambda}_i \, c'_i(\bar{x}) \cdot d \\ &\geqslant -\hat{\lambda}^{\mathsf{T}} P_{c(\bar{x})}(c'(\bar{x}) \cdot d) \\ &\geqslant -\|\hat{\lambda}\|_{\mathrm{D}} \|P_{c(\bar{x})}(c'(\bar{x}) \cdot d)\|_{\mathrm{P}}. \end{aligned}$$

  Alors (13.37) implique que $P_{c(\bar{x})}(c'(\bar{x}) \cdot d) = 0$, c'est-à-dire

  $$\begin{cases} c'_i(\bar{x}) \cdot d = 0 & \text{pour } i \in E \\ c'_i(\bar{x}) \cdot d \leqslant 0 & \text{pour } i \in I^0(\bar{x}). \end{cases}$$

  Ces relations et (13.38) montrent que d est une direction critique non nulle.

Soit à présent $\bar\lambda$ le multiplicateur dépendant de d, déterminé par (4.59). D'après la proposition 4.49, on a

$$d^{\mathsf T}\nabla^2_{xx}\ell(\bar x,\bar\lambda)d>0.$$

Alors le développement suivant (on utilise le fait que $\nabla_x\ell(\bar x,\bar\lambda)=0$)

$$\ell(x_k,\bar\lambda)=\ell(\bar x,\bar\lambda)+\frac{\alpha_k^2}{2}d^{\mathsf T}\nabla^2_{xx}\ell(\bar x,\bar\lambda)d+o(\alpha_k^2)$$

permet de voir que, pour k suffisamment grand, on a

$$\ell(x_k,\bar\lambda)>\ell(\bar x,\bar\lambda). \tag{13.39}$$

On conclut en obtenant une contradiction. Pour k grand, on a

$$\begin{array}{rcll}\Theta_r(x_k) &\leqslant& \Theta_r(\bar x) & \text{[par (13.36)]}\\ &=& f(\bar x)&\\ &=& \ell(\bar x,\bar\lambda)&\\ &<& \ell(x_k,\bar\lambda) & \text{[par (13.39)]}\\ &\leqslant& \Theta_r(x_k) & \text{[par le lemme 13.23]},\end{array}$$

ce qui est absurde. □

Notes

Origine du lagrangien augmenté. Le lagrangien augmenté (13.18) associé aux problèmes avec contraintes d'égalité remonte au moins à Arrow et Solow [24 ; 1958] qui l'utilisent pour localiser des points-selles au moyen d'équations différentielles modifiant les variables primales et duales simultanément. Il a été redécouvert et utilisé indépendamment par Hestenes [311 ; 1969] et Powell [486 ; 1969] dans le même contexte algorithmique qu'ici, celui de la méthode des multiplicateurs du schéma 13.20, qui modifie les variables primales et duales séquentiellement. Le lagrangien augmenté (13.21) permettant de prendre en compte des contraintes d'inégalité est attribué à Rockafellar [518, 520, 521, 524 ; 1971-76] ; voir aussi Buys [100 ; 1972] et Arrow, Gould et Howe [23 ; 1973]. On pourra consulter la revue de Bertsekas [51 ; 1976] pour plus de références bibliographiques avant 1976.

Lagrangiens augmentés plus réguliers. Même en l'absence de complémentarité stricte, on peut obtenir des lagrangiens augmentés plus réguliers que celui (13.19) de la section 13.4.3 (voir ce qu'il en est dit à la proposition 13.18), si le terme de pénalisation quadratique est remplacé par d'autres fonctions (*divergence de Bregman*, *φ-divergence*, *etc*) ; sur ces questions, on pourra consulter [111, 576, 186, 30, 31, 551 ; 1992-2006]. Il n'est pas clair, cependant, qu'une telle fonction plus régulière soit intéressante numériquement [59, 187].

Algorithme du lagrangien augmenté pour résoudre un problème quadratique convexe. La proposition **??** étend un résultat de [252, 253, 218 ; 1976-1982] au cas où H n'est définie positive que dans le noyau de A ; la démonstration est identique. La

convergence linéaire globale de l'algorithme du lagrangien augmenté (ou méthode des multiplicateurs) sur les problèmes quadratiques convexes (éventuellement non réalisables) est étudiée dans [161, 122].

Minimisation inexacte du lagrangien augmenté. On gagne en efficacité en minimisant de manière approchée le lagrangien augmenté à l'étape 1 de la méthode des multiplicateurs (schéma 13.20), avec une précision qui dépend de la proximité de la solution. Pour une étude locale dans le cas de l'optimisation non linéaire, on pourra lire [205].

Pénalisation exacte. Nous avons montré que l'on peut obtenir des fonctions de pénalisation qui sont exactes en des points vérifiant les conditions d'optimalité du second ordre (proposition 13.24). On peut suivre la démarche inverse et retrouver les conditions d'optimalité à partir du concept de pénalisation exacte. Voir [99, 640].

Exercices

13.1. *Pénalisation quadratique.* Soit $\Theta_r : \mathbb{E} \to \mathbb{R}$ la fonction de pénalisation quadratique définie en (13.12).

1) Montrez qu'en $x \in \mathbb{E}$:

$$\Theta_r(x) = \inf_{s \in \mathbb{R}^{m_I}_+} f(x) + \frac{r}{2}\|c_E(x)\|_2^2 + \frac{r}{2}\|c_I(x) + s\|_2^2.$$

Remarque. La fonction Θ_r s'obtient donc en remplaçant les contraintes $c(x)^\# = 0$ de (P_{EI}) par les contraintes équivalentes ($c_E(x) = 0$, $c_I(x) + s = 0$ et $s \geqslant 0$) et en prenant une pénalisation quadratique des contraintes d'égalité tout en gardant explicite les contraintes de positivité $s \geqslant 0$.

2) Montrez que, si f et c définissant (P_{EI}) sont différentiables en x, alors

$$\nabla\Theta_r(x) = \nabla f(x) + \sum_{i \in E \cup I} r[c(x)]_i^\# \, \nabla c_i(x).$$

13.2. *Conditionnement limite de la hessienne du lagrangien augmenté.* Soient H une matrice d'ordre n symétrique et A une matrice de type $m \times n$ telles que $\sigma := \inf\{v^\mathsf{T} H v : v \in \mathcal{N}(A),\ \|v\|_2 = 1\} > 0$. On note $\kappa_2(H_r)$ le conditionnement ℓ_2 de la matrice $H_r := H + rA^\mathsf{T}A$. Montrez que

$$\lim_{r \to +\infty} \frac{\kappa_2(H_r)}{r} = \frac{\|A\|_2^2}{\sigma}.$$

13.3. *Pénalisation exacte ℓ_1.* On suppose que f et c sont régulières dans un voisinage d'une solution $\bar{x}$ de (P_{EI}). On se donne des scalaires positifs r_i, $i \in E \cup I$, et on considère la fonction de pénalisation suivante

$$\Theta_r^1(x) = f(x) + \sum_{i \in E} r_i |c_i(x)| + \sum_{i \in I} r_i c_i(x)^+.$$

On suppose également qu'il existe un multiplicateur $\bar{\lambda} = \{\bar{\lambda}_i\}_{i \in E \cup I}$ tel que $(\bar{x}, \bar{\lambda})$ vérifie les conditions suffisantes d'optimalité du deuxième ordre. Montrez que si

$$r_i > |\bar{\lambda}_i|, \quad \text{pour } i \in E \cup I,$$

alors $\bar{x}$ est un minimum local strict de Θ_r^1.

13.4. *Lagrangien augmenté non différentiable* [70 ; 1989]. On suppose que f et c sont régulières dans un voisinage d'une solution $\bar{x}$ du problème $\inf\{f(x) : c(x) = 0\}$. On suppose également qu'il existe un multiplicateur $\bar{\lambda}$ tel que $(\bar{x}, \bar{\lambda})$ vérifie les conditions suffisantes d'optimalité du deuxième ordre de ce problème. Montrez que

$$\Theta_{\mu,r}(x) = f(x) + \mu^{\mathsf{T}} c(x) + r\|c(x)\|_{\mathrm{P}}$$

est une fonction de pénalisation exacte du problème d'optimisation considéré si $r > \|\bar{\lambda} - \mu\|_{\mathrm{D}}$.

13.5. *Courbes de niveaux de fonctions de pénalisation.* On considère un problème de minimisation sur $\mathbb{R}^2$

$$\begin{cases} \inf\ f(x) \\ x \in \mathbb{R}^2 \\ c(x) = 0, \end{cases}$$

en présence d'une unique contrainte d'égalité. Les fonctions f et $c : \mathbb{R}^2 \to \mathbb{R}$ sont supposées régulières. La figure 13.6 donne les tracés des courbes de niveau de trois

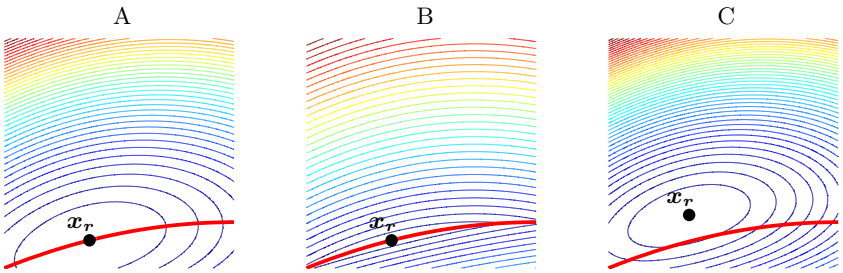


Fig. 13.6. Courbes de niveaux de 3 fonctions de pénalisation

fonctions de pénalisation associées au problème ci-dessus. L'ensemble admissible du problème est représenté dans chaque tracé par la courbe en trait large du bas. Déterminez pour chaque tracé A, B et C, laquelle des fonctions de pénalisation Θ_r^1, Θ_r^2 ou Θ_r^3 données ci-après qui a été utilisée pour dessiner ces courbes de niveau, sachant que toutes les trois ont été utilisées. Dans celles-ci r est un scalaire strictement positif et λ_* est le multiplicateur optimal du problème. Le minimum de ces fonctions est repéré par le point x_r dans chaque dessin.

$$\begin{aligned} \Theta_r^1(x) &= f(x) + \frac{r}{2}\|c(x)\|_2^2 \\ \Theta_r^2(x) &= f(x) + \lambda_*^{\mathsf{T}} c(x) + \frac{r}{2}\|c(x)\|_2^2 \\ \Theta_r^3(x) &= f(x) + r\|c(x)\|_2 \end{aligned}$$

14 Dualité

Les sages lui réservent toujours l'autre moitié du sens que mère nature sagement redoubla.

B. Gracián (1647). Oracle manuel et art de prudence. Traduction de B. Pelegrín [274].

Unus ego et multi in me.

M. Yourcenar. L'Œuvre au noir [636 ; 1968].

On the basis of this duality, close connections between otherwise disparate properties are revealed. [...] *In this way the analysis of a given situation can often be translated into an equivalent yet very different context. This can be a major source of insights as well as a means of unifying seemingly divergent parts of theory.*

R.T. Rockafellar, R.J.-B. Wets [528 ; 1998].

La dualité est une technique magique qui permet de révéler la face cachée de certains problèmes d'optimisation et d'établir des liens entre des problèmes qui peuvent paraître sans point commun au premier abord. Elle permet ainsi de mettre au jour de nouvelles propriétés sur ces problèmes, de mieux les comprendre, et d'introduire des algorithmes de résolution non triviaux. Parfois, la dualité apporte un éclairage original sur des algorithmes connus, conduisant alors à une meilleure conception de ceux-ci, voire à des résultats de convergence.

On sait que les problèmes d'optimisation avec contraintes ont des *variables cachées*, qui n'apparaissent pas dans l'expression du problème original, mais qui sont pourtant extrêmement précieuses : ce sont les multiplicateurs optimaux. On les utilise pour écrire les conditions d'optimalité, pour reconnaître les solutions, mais aussi pour définir des algorithmes de résolution, comme nous allons le voir dans ce chapitre et dans d'autres. Dans cet ouvrage, ces multiplicateurs sont apparus lors de l'obtention des conditions d'optimalité (chapitre 4), par l'intermédiaire du lemme de Farkas (section 2.5.5) : dans le cas du problème (P_{EI}), sachant que le gradient de l'objectif en la solution est dans le cône dual du cône tangent à l'ensemble admissible en ce point (CN1 de Peano-Kantorovitch de la proposition 4.6), qui est alors un cône polyédrique, on en déduit que ce gradient est l'image par une application linéaire d'un vecteur dont les composantes correspondant aux contraintes d'inégalité sont signées ; ce sont les multiplicateurs optimaux (voir la démonstration du théorème 4.32 par

exemple). Ces multiplicateurs forment apparemment un obscur vecteur, dont l'utilité ne semblait a priori que technique (l'écriture des conditions d'optimalité), mais dont la signification s'est manifestée par l'interprétation marginaliste de ses composantes que l'on a pu en faire (section 4.6.1) : les multiplicateurs optimaux expriment la variation au premier ordre de la valeur optimale du problème d'optimisation, lorsqu'on perturbe ses contraintes. Ce chapitre met en évidence une troisième propriété remarquable de ces multiplicateurs optimaux : ils sont parfois solutions d'autres problèmes d'optimisation, que l'on qualifie de *duaux* du problème original. Ce dernier est alors qualifié de *primal*.

En plus de donner une nouvelle interprétation aux multiplicateurs optimaux, la dualité a aussi un intérêt algorithmique. Les *méthodes primales* de résolution d'un problème d'optimisation se concentrent sur la détermination d'une solution optimale $\bar{x}$ en construisant une suite $\{x_k\}$ qui converge vers elle. Les méthodes vues précédemment peuvent être qualifiées de primales : méthodes de projection (section 12.1), méthodes d'activation (section 12.2) et méthodes de pénalisation (chapitre 13). Dans ces approches algorithmiques, un multiplicateur optimal $\bar{\lambda}$ est obtenu comme sous-produit de l'algorithme, permettant en particulier de vérifier l'optimalité de $\bar{x}$ (voir par exemple la proposition 13.11). Dans les *méthodes duales*, celles utilisant la dualité, l'effort est d'abord porté sur la recherche d'un multiplicateur optimal $\bar{\lambda}$ (aussi appelé *variable duale*, d'où le nom donné à ces approches algorithmiques) : on génère une suite de multiplicateurs $\{\lambda_k\}$ convergeant vers $\bar{\lambda}$. Dans les bons cas, la solution primale $\bar{x}$ est obtenue comme sous-produit de l'algorithme. C'est bien sûr le problème dual qui va aider à construire cette suite $\{\lambda_k\}$. Le problème dual n'est pas déterminé de manière unique et on verra une technique permettant d'en générer autant que l'on veut. Ceux-ci sont plus ou moins adaptés à un problème donné et plus ou moins aisés à résoudre. L'intérêt des méthodes duales apparaît lorsqu'on peut trouver un problème dual qui est plus simple à résoudre que le problème original et (cela paraît évident) qui a un rapport avec lui.

Il n'est pas toujours possible de trouver un problème dual dont les solutions sont les multiplicateurs optimaux et qui soit simple à analyser ou à résoudre numériquement. Même si cet objectif n'est pas atteint, les techniques de dualisation introduites dans ce chapitre restent précieuses car les problèmes duaux qu'elles construisent ont la propriété d'avoir une valeur optimale inférieure à celle du problème primal ; c'est la relation dite de *dualité faible* de la proposition 14.2. La différence entre les deux valeurs optimales est appelée le *saut de dualité*. Sur papier, le problème dual permet donc d'obtenir une borne inférieure de la valeur optimale du problème primal, parfois précise et souvent utile (voir l'exercice 19.3 par exemple). Numériquement, si le saut de dualité est faible et si le problème dual est plus simple à résoudre que le problème original, sa résolution permettra de trouver des solutions approchées du problème original dont la pertinence dépendra de chaque cas (comme exemple, on peut citer la relaxation SDP de problèmes non convexes ou combinatoires).

La notion de dualité est très générale et apparaît dans des contextes variés. En optimisation, elle peut s'appliquer à la minimisation de fonctions définies sur un ensemble X non vide quelconque. Celui-ci peut être $\mathbb{R}^n$, un espace fonctionnel, un ensemble discret ou une partie d'un de ces espaces éventuellement définie par des relations fonctionnelles. Nous allons garder cette généralité aussi longtemps que possible.

Les fonctions considérées dans ce chapitre seront à valeurs dans $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$. Elles peuvent donc prendre les valeurs $-\infty$ et $+\infty$. L'intérêt de cette généralisation audacieuse vient de ce que le problème d'optimisation

$$\min_{x \in X} f(x),$$

dans lequel X est un ensemble arbitraire, peut alors représenter un problème d'optimisation avec contrainte. Par exemple, le problème

$$\begin{cases} \min \tilde{f}(x) \\ x \in A, \end{cases}$$

où A est une partie de X et $\tilde{f} : X \to \mathbb{R}$ est à valeurs réelles, pourra s'écrire comme ci-dessus si l'on définit

$$f = \tilde{f} + \mathcal{I}_A,$$

où $\mathcal{I}_A : X \to \mathbb{R} \cup \{+\infty\}$ est la fonction indicatrice de l'ensemble A. On peut donc traiter dans un même cadre les problèmes avec et sans contrainte. On rappelle que le domaine d'une fonction $f : X \to \overline{\mathbb{R}}$ est l'ensemble des points de X où elle ne prend pas la valeur $+\infty$. On le note

$$\operatorname{dom} f := \{x \in X : f(x) < +\infty\}.$$

La fonction f peut y prendre la valeur $-\infty$ cependant.

Il est difficile d'acquérir une bonne maîtrise de la dualité, malgré les nombreuses synthèses qui ont été publiées sur le sujet. C'est en effet une théorie qui peut être abordée avec des points de vue très différents et les liens entre ceux-ci ne sont pas toujours clairs. La dualité peut en effet être présentée sur des problèmes particuliers comme ceux de l'optimisation linéaire ou quadratique [174] ou sur certains problèmes convexes (section 14.3) ; elle peut aussi être introduite à partir des conditions d'optimalité (chapitre 4) des problèmes différentiables [622], à partir de la dualité min-max [572, 406], ou encore à partir de la perturbation de problèmes [236], avec éventuellement l'apport de la conjugaison [515]. Nous essayerons de décrire ici, au moins en partie, ces différents points de vue.

Nous commencerons par introduire la dualité min-max (section 14.1). Pour l'utiliser, on écrit le critère du problème d'optimisation original, qualifié ici de *primal*, comme un supremum, si bien que ce problème primal s'écrit comme l'inf-sup d'une *fonction de couplage* φ. Par définition, le problème *dual* s'obtient alors en inversant l'infimum et le supremum. Il s'écrit donc comme le sup-inf de la même fonction φ. Cette notion de dualité est liée au concept de point-selle. Il y a beaucoup de manières d'écrire une fonction comme un supremum et toutes n'ont pas un intérêt équivalent, parce que le problème dual résultant n'est pas plus simple à résoudre que le problème primal ou n'apporte pas d'information pertinente. À la section 14.2, nous présentons l'approche souvent fructueuse qui consiste à « plonger » le problème original dans une famille de problèmes qui sont des perturbations de celui-ci. Cela donne une procédure pour écrire f comme un supremum et se ramener ainsi à la dualité min-max. Ces notions sont ensuite appliquées au cadre particulier de la dualité de Fenchel (section 14.3) et à la minimisation de fonctions sur $\mathbb{R}^n$ en présence de contraintes d'égalité

et d'inégalité (sections 14.5 et 14.6). Finalement, deux approches numériques de résolution par dualité sont présentées (section 14.7) ; toutes deux recherchent un point-selle de φ : la première approche le fait par l'intermédiaire de la fonction duale, la seconde travaille directement sur φ.

On peut lire ce chapitre de deux manières différentes. Dans la première, l'on se passe de la notion de fonction conjuguée et l'on fait l'impasse sur le lien entre dualité et perturbation de problème d'optimisation (section 14.2). Pour cela, on lit d'abord la section 14.1, puis directement les sections 14.5 et 14.6 en considérant les lagrangiens des formules (14.43) et (14.57) comme des fonctions de couplage φ (celles de la section 14.1) données. Pour avoir une connaissance plus approfondie de la dualité, il faudra passer par la notion de fonction conjuguée, que l'on pourra aborder en seconde lecture. Ce concept permet d'établir les formules des lagrangiens ordinaire (14.43) et augmenté (14.57) à partir de perturbations du problème d'optimisation original.

14.1 Dualité min-max

As far as I can see, there could be no theory of games [...] *without that theorem* [...] *I thought there was nothing worth publishing until the 'Minimax Theorem' was proved.*

J. von Neumann, cité par J.L. Casti [108 ; 1996, p. 19].

La règle du « maximin » nous dit de hiérarchiser les solutions possibles en fonction de leur plus mauvais résultat possible : nous devons choisir la solution dont le plus mauvais résultat est supérieur à chacun des plus mauvais résultats des autres.

J. Rawls (1971). Théorie de la Justice [508].

L'histoire de la dualité min-max a probablement commencé avec le théorème du minimax de von Newmann (1928), auquel il est fait allusion dans la première épigraphe de cette section (il est proposé à l'exercice 17.12). Cette théorie a pris beaucoup d'ampleur et une grande importance, du fait de sa généralité qui permet de l'appliquer à des domaines variés.

14.1.1 Introduction d'un problème dual

Soient X un ensemble et $f : X \to \overline{\mathbb{R}}$ une fonction pouvant prendre des valeurs infinies. On considère le problème d'optimisation

$$(P) \qquad \inf_{x \in X} f(x).$$

Sa valeur optimale (éventuellement infinie) et son ensemble de solution (éventuellement vide) seront respectivement notés

$$\mathrm{val}(P) \qquad \text{et} \qquad \mathrm{Sol}(P).$$

Dans cette section, le problème (P) d'où l'on part est appelé le *problème primal* et ses solutions sont appelées les *solutions primales*.

De manière à introduire un problème dual de (P), on cherchera à représenter $f(x)$ comme un supremum :

$$f(x) = \sup_{y\in Y} \varphi(x,y), \tag{14.1}$$

où Y est un ensemble et $\varphi : X \times Y \to \overline{\mathbb{R}}$ sera appelée la *fonction de couplage* entre variables x et y. Par exemple, si $f \in \mathrm{C}\overline{\mathrm{onv}}(\mathbb{R}^n)$, elle est l'enveloppe supérieure de ses minorantes affines (point 1 de la proposition 3.46). Un autre exemple, plus pratique, est donné ci-dessous (l'exemple 14.1) et nous verrons à la section 14.2, une procédure permettant d'écrire f de cette manière. Lorsque f s'écrit comme ci-dessus, le problème primal devient

$$(P) \qquad \inf_{x\in X} \sup_{y\in Y} \varphi(x,y). \tag{14.2}$$

Insistons sur le fait que cette écriture veut dire que l'on cherche à minimiser la fonction $x \mapsto \sup_{y\in Y} \varphi(x,y)$.

On appelle *problème dual* de (P) relatif à φ le problème noté (D) et défini par

$$(D) \qquad \sup_{y\in Y} \inf_{x\in X} \varphi(x,y). \tag{14.3}$$

On a donc simplement inversé l'ordre dans lequel les extremums en $x \in X$ et en $y \in Y$ sont pris. La valeur optimale du problème (D) (éventuellement infinie) et son ensemble de solution (éventuellement vide) seront respectivement notés

$$\mathrm{val}(D) \qquad \text{et} \qquad \mathrm{Sol}(D).$$

Les solutions de (D) sont appelées les *solutions duales*. Le problème dual consiste donc à minimiser ce que l'on appelle la *fonction duale* δ associée à φ :

$$\delta : y \mapsto \delta(y) := -\inf_{x\in X} \varphi(x,y). \tag{14.4}$$

On gardera à l'esprit que, pour chaque $y \in Y$, il faut résoudre un problème de minimisation pour connaître la valeur $\delta(y)$ de la fonction duale !

Les problèmes

$$\inf_{x\in X} \varphi(x,y) \qquad \text{et} \qquad \sup_{y\in Y} \varphi(x,y)$$

sont appelé respectivement *problème interne primal* en $y \in Y$ et *problème interne dual* en $x \in X$. Dans certains contextes, ils portent parfois le nom de *problèmes de Lagrange*. On note respectivement

$$\bar{X}(y) := \arg\min_{x\in X} \varphi(x,y) \qquad \text{et} \qquad \overline{Y}(x) := \arg\max_{y\in Y} \varphi(x,y) \tag{14.5}$$

leur ensemble de solutions.

On peut souvent représenter f comme en (14.1) au moyen de différentes fonctions φ. À chacune d'elles correspond un problème dual différent. Il n'y a donc pas unicité du problème dual. En voici un exemple de dualisation, dont le principe est généralisable à d'autres contextes.

Exemple 14.1 Considérons le problème d'optimisation avec contraintes d'égalité

$$\inf_{\substack{x\in X\\ c(x)=0}} \quad f(x),$$

dans lequel $c : X \to \mathbb{R}^m$. Ce problème peut s'écrire sous la forme

$$\inf_{x\in X} \ \tilde{f}(x),$$

si on définit $\tilde{f} = f + \mathcal{I}_{X_c}$ où $X_c = \{x \in X : c(x) = 0\}$. Les propriétés (1.1) sont importantes ici et joueront pleinement leur rôle dans tout ce chapitre. L'observation essentielle, à présent, est que l'on peut écrire $\tilde{f}$ comme un supremum :

$$\tilde{f}(x) = \sup_{y\in\mathbb{R}^m} \Big(f(x) + y^{\mathsf{T}} c(x)\Big).$$

Le problème primal s'écrit

$$\inf_{x\in X} \sup_{y\in\mathbb{R}^m} \Big(f(x) + y^{\mathsf{T}} c(x)\Big)$$

et le problème dual associé est

$$\sup_{y\in\mathbb{R}^m} \inf_{x\in X} \Big(f(x) + y^{\mathsf{T}} c(x)\Big).$$

Si $X = \mathbb{R}^n$, la fonction de couplage φ utilisée ici est le lagrangien du problème. Nous reviendrons sur cette dualisation à la section **??**. □

14.1.2 Liens entre problèmes primal et dual, point-selle

Les problèmes primal et dual semblent a priori très différents. Si on met l'accent sur la minimisation en x, le problème primal consiste à minimiser la fonction $x \mapsto \sup_y \varphi(x, y)$ obtenue en maximisant point par point une famille $\mathcal{F}$ de fonctions $x \mapsto \varphi(x, y)$, paramétrées par $y \in Y$. Par ailleurs, le problème dual consiste à chercher dans cette famille $\mathcal{F}$, une fonction $\varphi(\cdot, y)$, $y \in Y$, dont le minimum est le plus élevé possible. Y a-t-il un lien entre ces deux problèmes ? De quelle nature est il ? C'est ce que l'on examine dans cette section, tout en restant à un niveau de généralité élevé.

La proposition suivante donne une relation entre les valeurs optimales primale et duale. Il s'agit d'un résultat très général puisqu'il ne fait d'hypothèse, ni sur les ensembles X et Y, ni sur la fonction φ.

Proposition 14.2 (dualité faible) *On a*

$$\sup_{y\in Y} \inf_{x\in X} \varphi(x, y) \leqslant \inf_{x\in X} \sup_{y\in Y} \varphi(x, y). \tag{14.6}$$

Démonstration. On a bien sûr

$$\forall\, x' \in X,\ \forall\, y' \in Y : \qquad \varphi(x', y') \leqslant \varphi(x', y')$$

et donc certainement

$$\forall\, x' \in X,\ \forall\, y' \in Y : \qquad \inf_{x \in X} \varphi(x, y') \leqslant \varphi(x', y').$$

En fixant $x' \in X$ et en prenant le supremum en $y' \in Y$ dans les deux membres, on obtient

$$\forall\, x' \in X : \qquad \sup_{y \in Y} \inf_{x \in X} \varphi(x, y) \leqslant \sup_{y \in Y} \varphi(x', y).$$

Le membre de gauche est indépendant de x', on peut donc prendre l'infimum en $x' \in X$ à droite et garder l'inégalité. Ceci conduit au résultat. □

Dans le cadre de la dualité min-max, on appelle *saut de dualité* l'écart positif entre les deux membres de (14.6) :

$$\text{Saut de dualité} := \operatorname{val}(P) - \operatorname{val}(D) \geqslant 0,$$

étant sous-entendu que ce saut est nul si $\operatorname{val}(P)$ et $\operatorname{val}(D)$ ont tous deux la même valeur infinie. On dit qu'*il n'y a pas de saut de dualité* si ce dernier est nul. En général, lorsqu'il y a un saut de dualité, les solutions éventuelles des problèmes primal et dual n'ont pas de rapports entre elles. Comme le montre le résultat suivant, l'existence de solutions primale et duale et l'égalité en (14.6) sont étroitement liés à l'existence de point-selle de φ (notion introduite à la définition 3.81).

Théorème 14.3 (caractérisation des points-selles) *Un couple de points $(\bar{x}, \bar{y}) \in X \times Y$ est un point-selle de φ sur $X \times Y$ si, et seulement si, $\bar{x}$ est solution du problème primal (14.2), $\bar{y}$ est solution du problème dual (14.3) et on a*

$$\sup_{y \in Y} \inf_{x \in X} \varphi(x, y) = \inf_{x \in X} \sup_{y \in Y} \varphi(x, y). \tag{14.7}$$

Dans ces conditions, la valeur en (14.7) est $\varphi(\bar{x}, \bar{y})$.

Démonstration. Quel que soit le couple $(\bar{x}, \bar{y}) \in X \times Y$, on a

$$\inf_{x \in X} \varphi(x, \bar{y}) \leqslant \sup_{y \in Y} \inf_{x \in X} \varphi(x, y) \leqslant \inf_{x \in X} \sup_{y \in Y} \varphi(x, y) \leqslant \sup_{y \in Y} \varphi(\bar{x}, y), \tag{14.8}$$

où la première inégalité provient de la définition même du « sup inf », la seconde n'est autre que (14.6) et la troisième provient de la définition même de l'« inf sup ».

[⇒] Si $(\bar{x}, \bar{y}) \in X \times Y$ est un point-selle de φ, alors les membres à l'extrême gauche et à l'extrême droite dans (14.8) sont tous deux égaux à $\varphi(\bar{x}, \bar{y})$. On en déduit que l'on a égalité partout dans (14.8), c'est-à-dire que $\bar{x}$ est solution primale (par l'égalité de droite), que $\bar{y}$ est solution duale (par l'égalité de gauche) et qu'il n'y a pas de saut de dualité (par l'égalité du milieu).

[⇐] Réciproquement, si $\bar{x}$ est solution du problème primal, si $\bar{y}$ est solution du problème dual et s'il n'y a pas de saut de dualité, on a égalité partout dans (14.8). Dès lors (les inégalités à gauche et à droite ci-dessous proviennent de la définition de l'infimum et du supremum)

$$\varphi(\bar{x},\bar{y}) \geqslant \inf_{x\in X} \varphi(x,\bar{y}) = \sup_{y\in Y} \varphi(\bar{x},y) \geqslant \varphi(\bar{x},\bar{y}).$$

On a donc égalité partout dans cette dernière relation. On en déduit (3.72). □

L'ensemble des points-selles d'une fonction $\varphi : X \times Y \to \overline{\mathbb{R}}$ a une structure très particulière, comme le montre le résultat suivant.

Corollaire 14.4 (produit cartésien des points-selles) *Supposons que la fonction $\varphi : X \times Y \to \overline{\mathbb{R}}$ ait un point-selle. Alors*

1) *l'ensemble des points-selles de φ est le produit cartésien* $\mathrm{Sol}(P) \times \mathrm{Sol}(D)$,
2) *la fonction φ prend une valeur constante sur* $\mathrm{Sol}(P) \times \mathrm{Sol}(D)$, *disons $\bar{\varphi}$,*
3) *on a*

$$\begin{aligned} \mathrm{Sol}(P) &= \bigcap_{y\in Y} \{x \in X : \varphi(x,y) \leqslant \bar{\varphi}\}, \\ \mathrm{Sol}(D) &= \bigcap_{x\in X} \{y \in Y : \varphi(x,y) \geqslant \bar{\varphi}\}. \end{aligned}$$

Démonstration. 1) Montrons d'abord que l'ensemble des points-selles est un produit cartésien, disons $X_0 \times Y_0$. Soient (x_1,y_1) et (x_2,y_2) deux points-selles de φ ; il s'agit de montrer que (x_1,y_2) est aussi un point-selle. Par le théorème 14.3, x_1 est solution primale (car (x_1,y_1) est un point-selle), y_2 est solution duale (car (x_2,y_2) est un point-selle) et il n'y a pas de saut de dualité (car φ a un point-selle) ; on utilise alors la condition suffisante du théorème 14.3.

Il reste à montrer que $X_0 \times Y_0 = \mathrm{Sol}(P) \times \mathrm{Sol}(D)$. Comme φ a un point-selle, la condition nécessaire du théorème 14.3 nous dit que $X_0 \subseteq \mathrm{Sol}(P)$ et $Y_0 \subseteq \mathrm{Sol}(D)$, si bien que $X_0 \times Y_0 \subseteq \mathrm{Sol}(P) \times \mathrm{Sol}(D)$. Inversement, si $(\bar{x},\bar{y}) \in \mathrm{Sol}(P) \times \mathrm{Sol}(D)$, l'absence de saut de dualité (qui se déduit de la condition nécessaire du théorème 14.3 et du fait que φ a un point-selle) et la sondition suffisante du théorème 14.3 nous informe que $(\bar{x},\bar{y})$ est un point-sellede φ, donc $(\bar{x},\bar{y}) \in X_0 \times Y_0$.

2) Soient (x_1,y_1) et $(x_2,y_2) in\, \mathrm{Sol}(P)_\times\, \mathrm{Sol}(D)$. Il s'agit de montrer que $\varphi(x_1,y_1) = \varphi(x_2,y_2)$. En utilisant le fait que (x_1,y_2) est un point-selle (inégalité (3.72)), on a

$$\varphi(x_1,y_1) \leqslant \varphi(x_1,y_2) \leqslant \varphi(x_2,y_2).$$

En permutant les indices, on trouve que $\varphi(x_1,y_1) = \varphi(x_2,y_2)$.

3) Soit $\bar{X} := \cap_{y\in Y}\{x \in X : \varphi(x,y) \leqslant \bar{\varphi}\}$. Il s'agit de montrer que $\mathrm{Sol}(P) = \bar{X}$ (l'expression de $\mathrm{Sol}(D)$ s'obtient de la même manière). Si $x_1 \in \mathrm{Sol}(P)$, alors $\varphi(x_1,y) \leqslant \bar{\varphi}$ pour tout $y \in Y$ (c'est l'inégalité de gauche dans (3.72)) ; donc $x_1 \in \bar{X}$. Inversement, si $x_1 \in \bar{X}$, on a $\sup_y \varphi(x_1,y) \leqslant \bar{\varphi} = \inf_x \sup_y \varphi(x,y)$ donc x_1 est solution primale et $x_1 \in \mathrm{Sol}(P)$. □

Lorsque φ a un point-selle, on dispose d'une *approche duale* pour résoudre le problème (P), celle qui consiste à résoudre le problème dual : on cherche $\bar{y} \in Y$ solution de

$$\sup_{y\in Y} \left(\inf_{x\in X} \varphi(x,y) \right).$$

Considérons alors le problème interne primal en une solution duale $\bar{y}$:

$$\inf_{x\in X} \varphi(x,\bar{y}). \tag{14.9}$$

Le corollaire suivant montre que les solutions du problème primal sont solutions de ce problème (mais la réciproque n'est pas nécessairement vraie, voir la section **??**). On rappelle que $\bar{X}(y)$ et $\overline{Y}(x)$ sont définis en (14.5).

Corollaire 14.5 *Si φ a un point-selle $(\bar{x},\bar{y})$, alors*

$$\varnothing \neq \mathrm{Sol}(P) \subseteq \bar{X}(\bar{y}) \qquad \textit{et} \qquad \varnothing \neq \mathrm{Sol}(D) \subseteq \overline{Y}(\bar{x}).$$

Démonstration. D'après le théorème 14.3, $\bar{x}$ est solution primale et $\bar{y}$ est solution duale ; donc $\mathrm{Sol}(P) \neq \varnothing$ et $\mathrm{Sol}(D) \neq \varnothing$. Par ailleurs, si $\hat{x}$ est solution primale, alors $(\hat{x},\bar{y})$ est un point-selle (aussi par le théorème 14.3), et donc $\hat{x} \in \bar{X}(\bar{y})$. De même, si $\hat{y}$ est solution duale, alors $(\bar{x},\hat{y})$ est un point-selle et donc $\hat{y} \in \overline{Y}(\bar{x})$. □

14.1.3 Existence de point-selle ▲

Nous concluons cette section par un résultat d'existence de point-selle qui joue un peu le même rôle que celui de Weierstrass sur la minimisation de fonction (il est aussi de nature topologique), mais qui requiert des hypothèses beaucoup plus fortes. La discussion ci-dessous motive partiellement ces hypothèses.

Observons que la fonction φ a peu de chance d'avoir un point-selle si X ou Y est un ensemble fini. Par exemple si $Y = \{0,1\}$ et si les fonctions $\varphi(\cdot,0)$ et $\varphi(\cdot,1)$ sont comme à gauche dans la figure 14.1, il y aura un saut de dualité (et donc pas de point-selle). À l'inverse, si $\varphi(x,\cdot)$ est affine sur $Y = [0,1]$ avec les mêmes fonctions $\varphi(\cdot,0)$ et $\varphi(\cdot,1)$ que précédemment (ce qui est suggéré par l'ensemble des courbes à droite dans la figure 14.1, correspondant à diverses valeurs de $y \in [0,1]$), on obtient pour un certain $\bar{y} \in [0,1]$ une courbe $\varphi(\cdot,\bar{y})$ ayant $\mathrm{val}(P)$ comme valeur minimale (la convexité des fonctions $\varphi(\cdot,y)$ joue ici un rôle important), ce qui implique qu'il n'y a pas de saut de dualité (voir aussi l'exercice 14.1). Dans le résultat ci-dessous, on suppose que X et Y sont des parties *convexes* d'espaces vectoriels et que la fonction φ est convexe-concave.

La démonstration du résultat ci-dessous est fondée sur l'observation faite dans la remarque précédente.

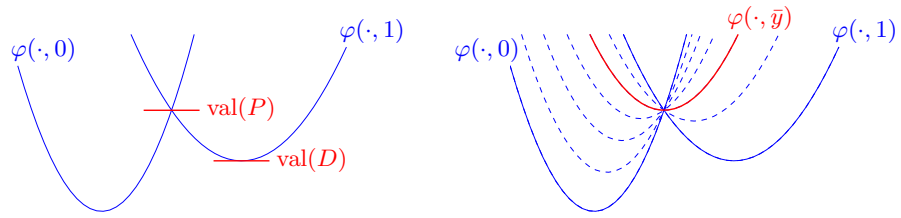


Fig. 14.1. Exemples avec et sans saut de dualité : $\varphi(x,y) = 2(1-y)[(x+1)^2 - 1] + y[(x-1)^2 - 1]$. À gauche, $Y = \{0,1\}$: il y a un saut de dualité car aucune des fonctions $\varphi(\cdot, y)$ n'a pour valeur minimale $\mathrm{val}(P)$. À droite, $Y = [0,1]$: il n'y a pas de saut de dualité car pour $\bar{y} = 2/3$, $\varphi(\cdot, \bar{y})$ a pour valeur minimale $\mathrm{val}(P)$; la solution duale est donc $\bar{y}$.

Théorème 14.6 (Sion, existence de point-selle) *Supposons que X et Y soient des ensembles convexes non vides d'espaces vectoriels topologiques, que X soit compact et que*

(*i*) *pour tout $y \in Y$, $\varphi(\cdot, y)$ est quasi-convexe s.c.i.,*
(*ii*) *pour tout $x \in X$, $\varphi(x, \cdot)$ est quasi-concave s.c.s..*

Alors

$$\sup_{y\in Y} \min_{x\in X} \varphi(x,y) = \min_{x\in X} \sup_{y\in Y} \varphi(x,y). \tag{14.10}$$

Démonstration. Il suffit de montrer l'absence de saut de dualité, car le minimum (au lieu de l'infimum) en $x \in X$ dans (14.10) est une conséquence directe de la compacité de X, du caractère s.c.i. de $\varphi(\cdot, y)$ (quel que soit $y \in Y$) et de $x \mapsto \sup_{y\in Y} \varphi(x,y)$ et du théorème de Weierstrass (théorème 1.2). Comme $\mathrm{val}(D) \leqslant \mathrm{val}(P)$ par la dualité faible, il suffit de montrer que

$$\forall\, \varepsilon > 0, \quad \exists\, y \in Y : \qquad \inf_{x\in X} \varphi(x,y) \geqslant \mathrm{val}(P) - \varepsilon.$$

On procède par étapes, avec un ensemble Y de plus en plus général.

Étape 1 : *on suppose que Y est un intervalle $[y_1, y_2]$.*

Étape 2 : *on suppose que Y est un polytope $\mathrm{co}\{y_1, \ldots, y_p\}$.*

Étape 3 : *cas général.* □

Le théorème 14.6 affirme l'absence de saut de dualité, mais pas l'existence d'un point-selle. Celui-ci sera toutefois bien présent si Y est également compact. En effet, sous les hypothèses du théorème, $y \mapsto \min_{x\in X} \varphi(x,y)$ est s.c.s., si bien que son maximum sur le compact Y est atteint (théorème de Weierstrass) ; ce maximum est une solution duale ; l'existence d'une solution primale-duale et l'absence de saut de dualité assure alors l'existence d'un point-selle.

Par ailleurs, si X n'est pas compact, mais que Y l'est, on peut appliquer le théorème à $-\varphi$, ce qui donne

$$\max_{y\in Y} \inf_{x\in X} \varphi(x,y) = \inf_{x\in X} \max_{y\in Y} \varphi(x,y).$$

14.1.4 Schéma algorithmique

La dualisation min-max conduit au schéma algorithmique (ou principe algorithmique très grossièrement décrit) suivant. Ce schéma se concentre sur la génération d'une suite $\{y_k\} \subseteq Y$ que l'on cherche à faire tendre vers une solution du problème dual (D) défini en (14.3). La suite primale générée, $\{x_k\} \subseteq X$, apparaît ici comme un sous-produit, aidant à la construction de la suite duale.

Schéma algorithmique 14.7 (de dualité) Une itération passe de l'itéré courant $y_k \in Y$ à l'itéré suivant $y_{k+1} \in Y$ par les étapes suivantes.
1. *Calcul de l'itéré primal* :

$$x_k \in \arg\min_{x \in X} \varphi(x, y_k). \tag{14.11}$$

2. *Vérification de l'optimalité* : si (x_k, y_k) est satisfaisant, on s'arrête.
3. *Nouvel itéré dual* : on calcule y_{k+1} en utilisant des informations recueillies en résolvant le problème interne dans (14.11).

Ce schéma est très peu précis. On ne dit pas en effet comment se vérifie l'optimalité, ni comment on calcule y_{k+1} à partir de y_k et des informations obtenues en résolvant le problème interne dans (14.11), mais on ne peut guère en dire davantage à ce niveau de généralité. Nous serons plus précis lorsque les problèmes d'optimisation considérés seront plus palpables, décrits en termes de critère à minimiser et de contraintes fonctionnelles à réaliser. Le schéma algorithmique dual ci-dessus servira alors de canevas permettant d'y broder des méthodes plus précises.

Pour que le schéma algorithmique dual, qui cherche à résoudre (D), ait quelqu'intérêt, il est souhaitable que les conditions suivantes soient remplies.

(D_1) En résolvant (D) on résout aussi (P). Il faut donc que les deux problèmes aient un lien entre eux, ce qui n'est pas toujours le cas. L'existence d'un point-selle de φ sera une propriété précieuse pour obtenir ce lien (voir la section 14.1.2).

(D_2) Il est beaucoup plus facile de résoudre les problèmes internes dans (14.11) que de résoudre (P). Il faut en effet résoudre un tel problème interne en X à chaque itération.

(D_3) Si Y est un espace topologique et si $\bar{y}$ est un point d'adhérence de la suite générée $\{y_k\}$, il est souhaitable que les solutions du problème interne primal $\inf_{x \in X} \varphi(x, \bar{y})$ soient solutions du problème primal, c'est-à-dire que ce problème interne n'ait pas de solutions importunes (voir le corollaire 14.5).

Comme beaucoup dépend de la fonction φ, son choix sera guidé par les trois conditions ci-dessus.

14.2 Dualité par perturbation

As a graduate student at Harvard, I got interested in convexity because I was amazed by linear programming duality and wanted to invent a nonlinear programming duality. That was around 1961. [...] *It was while I was writing up my dissertation—focused then on dual problems stated in terms of polar cones—that I came across Fenchel's conjugate convex functions.* [...] *They turned out to be a wonderful vehicle for expressing nonlinear programming duality, and I adopted them wholeheartedly. Around the time the thesis was nearly finished, I also found out about Moreau's efforts to apply convexity ideas, including duality, to problems in mechanics.*

R.T. Rockafellar sur le site Wikimization.

Dans cette section, nous décrivons une méthode permettant, dans les bons cas, d'écrire f comme le supremum d'une fonction, appelée *lagrangien*. On se ramène ainsi à la notion de dualité min-max introduite en section 14.1. Un intérêt de cette approche est de donner des conditions dans lesquelles il n'y a pas de saut de dualité, sans faire référence à la notion de point-selle comme à la section 14.1.2, mais en examinant les propriétés de la fonction valeur associée aux perturbations.

14.2.1 Perturbation du problème primal

Rappelons la forme du *problème primal*

$$(P) \qquad \inf_{x\in X} f(x),$$

dans lequel l'ensemble X est toujours quelconque. On note comme précédemment par $\mathrm{val}(P)$ la valeur de l'infimum dans (P). L'idée est de plonger le problème primal dans une famille de problèmes qui sont des perturbations de celui-ci. Ces perturbations sont introduites en se donnant une *fonction de perturbation*

$$\Phi : X \times P \to \overline{\mathbb{R}} : (x,p) \mapsto \Phi(x,p),$$

où P est un espace vectoriel de dimension finie. On suppose que l'on retrouve (P) lorsque la perturbation est nulle :

$$\Phi(x,0) = f(x), \quad \forall x \in X. \tag{14.12}$$

On considère alors la famille de problèmes

$$(P_p) \qquad \inf_{x\in X} \Phi(x,p),$$

qui sont des perturbations de (P).

La variation de l'infimum de (P_p) avec p est représentée par la *fonction valeur* $v : P \to \overline{\mathbb{R}}$, qui est définie par

$$v(p) = \inf_{x\in X} \Phi(x,p).$$

Cette fonction joue un rôle important dans toute cette section. On a clairement

$$\boxed{\text{val}(P) = v(0).} \tag{14.13}$$

14.2.2 Le problème dual

Comme on a pris soin de prendre la perturbation p dans un espace vectoriel, on peut considérer la conjuguée v^* de la fonction valeur v. On sait que cette opération ne sera utile que si v est propre avec une minorante affine, c'est-à-dire

$$\operatorname{dom} v \neq \varnothing \quad \text{et} \quad v \text{ a une minorante affine.}$$

Dans ces conditions $v^* \in \text{C}\overline{\text{onv}}(P)$ (proposition 3.43) et il y a un sens à minimiser v^* : d'une part, le caractère fermé de v^* a tendance à assurer l'existence d'une solution (par exemple si v^* est coercive) et, d'autre part, la convexité annule l'ambiguïté entre solution locale et globale. Ceci nous amène à considérer le problème

$$(D) \qquad -\inf_{p^*\in P} v^*(p^*) = \sup_{p^*\in P} -v^*(p^*). \tag{14.14}$$

Le signe « $-$ » est introduit pour se ramener au cadre de la dualité min-max (section 14.1). Ce problème est appelé *problème dual associé aux perturbations* et la fonction

$$\delta = v^* \tag{14.15}$$

est appelée *fonction duale associée aux perturbations* : le problème dual consiste donc à minimiser la fonction duale.

On peut aussi exprimer le problème dual en termes de la fonction de perturbation Φ. En effet,

$$\begin{aligned} v^*(p^*) &= \sup_{p\in P} \Big(\langle p^*, p\rangle - v(p) \Big) \\ &= \sup_{p\in P} \Big(\langle p^*, p\rangle - \inf_{x\in X} \Phi(x,p) \Big) \\ &= \sup_{p\in P} \sup_{x\in X} \Big(\langle p^*, p\rangle - \Phi(x,p) \Big) \qquad (14.16) \\ &= \Phi^*(0,p^*). \qquad (14.17) \end{aligned}$$

Dès lors

$$(D) \qquad -\inf_{p^*\in P} \Phi^*(0,p^*) = \sup_{p^*\in P} -\Phi^*(0,p^*).$$

On note $\text{val}(D)$ la valeur valeur optimale du problème dual (D) et $\text{Sol}(D)$ l'ensemble de ses solutions. On a clairement

$$\boxed{\text{val}(D) = v^{**}(0)} \qquad \text{et} \qquad \boxed{\text{Sol}(D) = \partial v^{**}(0),} \tag{14.18}$$

que l'on comparera avec (14.13). La première identité provient du fait que

$$\mathrm{val}(D) = -\inf_{p^* \in P} v^*(p^*) = \sup_{p^* \in P} -v^*(p^*) = v^{**}(0).$$

Alors la seconde identité de (14.18) provient de ce que $\bar{p}^*$ est solution du problème dual si, et seulement si, $-v^*(\bar{p}^*) = v^{**}(0)$, ce qui s'écrit aussi $0 = v^*(\bar{p}^*) + v^{**}(0)$. D'après la proposition 3.55 (avec $f = v^{**}$ et $f^* = (v^{**})^* = v^*$), cette dernière égalité est équivalente au fait que $\bar{p}^* \in \partial v^{**}(0)$.

Nous montrerons avec la proposition 14.8 que le problème primal a « quelques chances » d'être de la forme inf-sup et que le problème dual est bien de la forme sup-inf.

14.2.3 Le lagrangien associé aux perturbations

En (14.16), le supremum en p dépend essentiellement de la perturbation que l'on se donne. Si cette perturbation est simple, il sera en général possible de donner une expression explicite de la valeur du supremum en p. Cela nous conduit à introduire la notion de *lagrangien associé aux perturbations*, qui est la fonction

$$\ell_\Phi : X \times P \to \overline{\mathbb{R}},$$

définie par

$$\boxed{\ell_\Phi(x, p^*) := -\sup_{p \in P} \Big(\langle p^*, p \rangle - \Phi(x, p) \Big) = -\Phi_x^*(p^*),} \tag{14.19}$$

où on a introduit l'application associée à Φ et à $x \in X$ suivante

$$\Phi_x : p \mapsto \Phi(x, p).$$

Ce lagrangien dépend des perturbations choisies et n'est donc pas nécessairement identique au lagrangien ordinaire (4.34), introduit pour écrire les conditions d'optimalité. Nous verrons plus loin (à la section **??**) que l'on peut choisir des perturbations telles que le lagrangien (14.19) redonne, là où il prend des valeurs finies, le lagrangien ordinaire (4.34).

Par (14.15) et (14.16), la valeur de la fonction duale en p^* se récrit

$$\delta(p^*) = -\inf_{x \in X} \ell_\Phi(x, p^*). \tag{14.20}$$

Cette formule est donc identique à celle définissant la fonction duale par (14.4) en dualité min-max, lorsqu'on prend le lagrangien $(x, p^*) \in X \times P \mapsto \ell_\Phi(x, p^*)$ comme fonction de couplage $(x, y) \in X \times Y \mapsto \varphi(x, y)$.

Remarquons bien que, dans le couple (x, p^*) dont dépend ℓ_Φ, x est une variable primale, appartenant à l'espace sur lequel est défini le problème (P), et p^* est une variable duale, variant dans l'espace dual de l'espace des perturbations.

Proposition 14.8 (dualité faible) *On a*

$$\begin{aligned} \mathrm{val}(D) &= \sup_{p^*\in P}\inf_{x\in X} \ell_\Phi(x,p^*) \\ &\leqslant \inf_{x\in X}\sup_{p^*\in P} \ell_\Phi(x,p^*) \qquad (14.21) \\ &\leqslant \mathrm{val}(P). \end{aligned}$$

Démonstration. L'égalité $(14.21)_a$ vient de la définition de (D) et de ce que

$$-v^*(p^*) = -\sup_{x\in X}(-\ell_\Phi(x,p^*)) = \inf_{x\in X}\ell_\Phi(x,p^*).$$

L'inégalité $(14.21)_b$ découle de la proposition 14.2. Enfin l'inégalité $(14.21)_c$ est obtenue comme suit

$$\sup_{p^*\in P}\ell_\Phi(x,p^*) = \sup_{p^*\in P}\inf_{p\in P}(\Phi(x,p)-\langle p^*,p\rangle) \leqslant \sup_{p^*\in P}\Phi(x,0) = f(x).$$

□

Définition 14.9 On appelle *saut de dualité*, la quantité positive $\mathrm{val}(P)-\mathrm{val}(D)$. On dit qu'*il n'y a pas de saut de dualité* si le saut de dualité est nul, c'est-à-dire si

$$\mathrm{val}(D) = \mathrm{val}(P).$$

□

Grâce à (14.13) et (14.18), l'inégalité $\mathrm{val}(D) \leqslant \mathrm{val}(P)$ s'écrit $v^{**}(0) \leqslant v(0)$, qui n'est autre que l'inégalité bien connue entre une fonction et sa biconjuguée (corollaire 3.47). On a donc, par cette théorie de la dualité par perturbation, transformé la question importante de l'absence de saut de dualité par celle de savoir si la fonction valeur et sa biconjuguée (qui la convexifie) coïncide en 0. C'est donc l'analyse de la fonction valeur v qui peut apporter ici une réponse à cette question.

La proposition 14.10 ci-dessous examine les propriétés du lagrangien ℓ_Φ associé aux perturbations. Le point 2 donne des conditions pour que f s'écrive comme un supremum, à savoir que $\Phi_x^{**}(0) = \Phi_x(0)$ pour tout $x \in X$, auquel cas la dualité par pertubation se ramène à la dualité min-max de la section 14.1.

Proposition 14.10 (propriétés du lagrangien associé aux perturbations)
1) *Pour tout $x \in X$, l'application $p^* \in P \to -\ell_\Phi(x,p^*)$ est convexe et fermée.*
2) *Si, pour $x \in X$, $\Phi_x^{**}(0) = \Phi_x(0)$, alors*

$$f(x) = \sup_{p^*\in P}\ell_\Phi(x,p^*). \qquad (14.22)$$

En particulier, l'identité (14.22) *a lieu si* $\Phi_x \in \mathrm{C}\overline{\mathrm{onv}}(P)$ *ou si* $\Phi_x \in \mathrm{Conv}(P)$ *et* $\partial\Phi_x(0) \neq \varnothing$.

DÉMONSTRATION. 1) Il s'agit en effet de l'enveloppe supérieure de fonctions affines (proposition 3.34).

2) Par (14.12) et (14.19), on a

$$\begin{aligned}\Phi_x(0) &= f(x),\\ \Phi_x^{**}(0) &= \sup_{p^*\in P}\,(-\Phi_x^*(p^*)) = \sup_{p^*\in P}\,\ell_\Phi(x,p^*).\end{aligned}$$

On en déduit l'expression de $f(x)$ si $\Phi_x^{**}(0) = \Phi_x(0)$, ce qui est le cas si $\Phi_x \in \mathrm{C}\overline{\mathrm{onv}}(P)$ (point 3 de la proposition 3.46) ou si $\Phi_x \in \mathrm{Conv}(P)$ et $\partial\Phi_x(0) \neq \varnothing$ (point 2 du corollaire 3.60). □

Les propriétés démontrées au chapitre 3 conduisent aux résultats de la proposition 14.11 ci-dessous qui apportent une réponse à la question de l'absence de saut de dualité et à l'existence de solution primale ou duale à partir des propriétés de la fonction valeur v en 0 (point 1) ou sur P tout entier (point 2). La proposition fait l'hypothèse de convexité de v, mais n'est pas pour autant limitée aux problèmes convexes car des problèmes non convexes peuvent très bien avoir une fonction valeur convexe. Remarquons que, dans le point 1 de la proposition, les conditions données sur v en 0 sont de plus en plus restrictives puisque $0 \in (\mathrm{dom}\, v)^\circ \implies 0 \in (\mathrm{dom}\, v)^{\ominus}$ [clair] $\implies \partial v(0) \neq \varnothing$ [proposition 3.62] $\implies v$ est s.c.i. en 0 [clair par le point (ii) de la proposition 3.55] ; il est donc normal que les conséquences de ces hypothèses soient de plus en plus riches.

Proposition 14.11 (existence de solution duale et saut de dualité)

1) *Si la fonction valeur* $v \in \mathrm{Conv}(P)$ *et* $0 \in \mathrm{dom}\, v$, *on a*
   - *a)* v *est s.c.i. en* $0 \iff \mathrm{val}(P) = \mathrm{val}(D)$,
   - *b)* $\partial v(0) \neq \varnothing \iff \mathrm{val}(P) = \mathrm{val}(D)$ *et* $\mathrm{Sol}(D) \neq \varnothing \implies \mathrm{Sol}(D) = \partial v(0)$,
   - *c)* $0 \in (\mathrm{dom}\, v)^{\ominus} \implies \mathrm{val}(P) = \mathrm{val}(D)$ *et* $\mathrm{Sol}(D) = \partial v(0) \neq \varnothing$,
   - *d)* $0 \in (\mathrm{dom}\, v)^\circ \implies \mathrm{val}(P) = \mathrm{val}(D)$, $\mathrm{Sol}(D) = \partial v(0)$ *est un compact non vide et* v *est lipschitzienne dans un voisinage de* 0.

2) *Si* $v \in \mathrm{C}\overline{\mathrm{onv}}(P)$, *alors* $\mathrm{val}(P) = \mathrm{val}(D) \in \overline{\mathbb{R}}$.

DÉMONSTRATION. 1.*a*) C'est une récriture de la seconde partie du corollaire 3.47.

1.*b*) Si $\partial v(0) \neq \varnothing$, alors $v(0) = v^{**}(0)$ (corollaire 3.60, point 2) et $\partial v(0) = \partial v^{**}(0)$ (corollaire 3.60, point 3). L'égalité $v(0) = v^{**}(0)$ montre qu'il n'y a pas de saut de dualité et le fait que $\partial v(0) \neq \varnothing$ implique alors que $\mathrm{Sol}(D) \equiv \partial v^{**}(0)$ est non vide.

Réciproquement, $v(0) = v^{**}(0)$ (pas de saut de dualité) implique que $\partial v(0) = \partial v^{**}(0)$ (corollaire 3.60, point 3). Comme $\mathrm{Sol}(D) \equiv \partial v^{**}(0)$ est non vide, il en est de même de $\partial v(0)$.

1.*c*) Si $0 \in (\mathrm{dom}\, v)^{\ominus}$, v est sous-différentiable en 0 (proposition 3.62). On applique alors le point 1.*b*.

1.*d*) Si $0 \in (\operatorname{dom} v)^\circ$, alors v est lipschitzienne dans un voisinage de 0 (proposition 3.14) et $\partial v(0)$ est non vide et compact (proposition 3.64).

2) Si $v \in \mathrm{C}\overline{\mathrm{onv}}(P)$, $v^{**} = v$ (point 3 de la proposition 3.46), donc $v(0) = v^{**}(0)$, c'est-à-dire qu'il n'y a pas de saut de dualité (mais les valeurs optimales peuvent être infinies). □

14.2.4 Perturbation du problème dual

Le procédé de dualisation par perturbation du problème primal introduit aux sections 14.2.1 et 14.2.2 peut être inversé : on peut plonger le problème dual dans une famille de problèmes perturbés dont le dual associé redonne le problème primal.

Le problème dual consiste à minimiser la fonction $p^* \mapsto v^*(p^*) = \Phi^*(0, p^*)$. Ce fait suggère de prendre comme fonction de perturbation de (D), la fonction $(x^*, p^*) \mapsto \Phi^*(x^*, p^*)$ et de considérer les problèmes duaux perturbés suivants

$$(D_{x^*}) \qquad \inf_{p^* \in P} \Phi^*(x^*, p^*).$$

La *fonction valeur* associée à ces perturbations est la fonction $w : X \to \overline{\mathbb{R}}$, qui est définie par

$$w(x^*) = \inf_{p^* \in P} \Phi^*(x^*, p^*).$$

Contrairement à v dont la convexité n'est pas assurée, w est toujours convexe, comme fonction marginale de la fonction convexe Φ^*. On note que

$$\boxed{\operatorname{val}(D) = -w(0).} \tag{14.23}$$

La conjuguée de w s'écrit

$$w^*(x) = \Phi^{**}(x, 0) \tag{14.24}$$

et le dual du dual est le problème suivant

$$(DD) \qquad \inf_{x \in X} w^*(x) = \inf_{x \in X} \Phi^{**}(x, 0).$$

Proposition 14.12 (retour au primal) *Si X est un espace vectoriel et si $\Phi \in \mathrm{C}\overline{\mathrm{onv}}(X, P)$, alors $f = w^*$ et le problème (DD) n'est autre que le problème primal (P).*

Démonstration. Clair à partir de l'expression (14.24) de w^* et du point 3 de la proposition 3.46 qui assure que $\Phi^{**} = \Phi$ grâce au fait que $\Phi \in \mathrm{C}\overline{\mathrm{onv}}(X, P)$. □

14.3 Dualité de Fenchel

La dualité de Fenchel regroupe un ensemble de résultats issus de la dualisation d'un problème d'optimisation à la structure particulière, le problème (P_F) ci-dessous. Celui-ci se rencontre fréquemment en optimisation convexe. Nous avons choisi de construire le dual par la technique de perturbation (section 14.2), mais nous aurions pu le faire au moyen de la dualité min-max lagrangienne de la section 14.1 (exercice 14.4). Nous allons un peu plus loin que la construction du dual, cependant, on donnant des conditions suffisantes assurant l'absence de saut de dualité et l'existence de solutions primale et duale.

Le cadre est le suivant. On se donne deux espaces euclidiens $\mathbb{E}$ et $\mathbb{F}$ (produits scalaires tous deux notés $\langle\cdot,\cdot\rangle$), une fonction $f : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$, une application linéaire $A : \mathbb{E} \to \mathbb{F}$ (on note A^* son adjointe) et une fonction $g : \mathbb{F} \to \mathbb{R} \cup \{+\infty\}$. On considère le problème, dit *primal*,

$$(P_F) \quad \inf_{x\in\mathbb{E}} \Big(f(x) + g(Ax) \Big). \tag{14.25}$$

Introduisons un problème dual par perturbation de Ax. On considère le *fonction valeur*

$$v_F : \mathbb{F} \to \overline{\mathbb{R}} : p \mapsto v_F(p) := \inf_{x\in\mathbb{E}} \Big(f(x) + g(Ax + p) \Big). \tag{14.26}$$

Sa conjuguée s'écrit

$$\begin{aligned} v_F^*(\lambda) &= \sup_{x\in\mathbb{E}} \sup_{p\in\mathbb{F}} \Big(\langle\lambda, p\rangle - f(x) - g(Ax + p) \Big) \\ &= \sup_{x\in\mathbb{E}} \Big(-\langle\lambda, Ax\rangle - f(x) + g^*(\lambda) \Big) \\ &= f^*(-A^*\lambda) + g^*(\lambda). \end{aligned}$$

Le problème dual consiste à maximiser $-v_F^*(-\lambda)$ (on fait un changement de variable $\lambda \curvearrowright -\lambda$) :

$$(D_F) \quad \sup_{\lambda\in\mathbb{F}} \Big(-f^*(A^*\lambda) - g^*(-\lambda) \Big). \tag{14.27}$$

Un résultat de dualité faible découle directement de la proposition 14.8.

Proposition 14.13 (dualité de Fenchel faible) *On a* $\mathrm{val}(D_F) \leqslant \mathrm{val}(P_F)$.

Pour obtenir un résultat de dualité forte, assurant en particulier l'absence de saut de dualité, $\mathrm{val}(P_L) = \mathrm{val}(D_F)$, nous allons analyser la régularité de la fonction valeur v_F et utiliser les résultats du chapitre 3. Calculons d'abord le domaine de v_F. Clairement, $p \in \operatorname{dom} v_F$ si, et seulement si, il existe un $x \in \operatorname{dom} f$ tel que $Ax + p \in \operatorname{dom} g$ ou encore $p \in \operatorname{dom} g - A(\operatorname{dom} f)$. Nous avons donc montré que

$$\operatorname{dom} v_F = \operatorname{dom} g - A(\operatorname{dom} f).$$

On comprend donc pourquoi l'ensemble à droite intervient dans le résultat suivant.

Théorème 14.14 (dualité de Fenchel forte) *On suppose que* $f \in \mathrm{Conv}(\mathbb{E})$, *que* $g \in \mathrm{Conv}(\mathbb{F})$, *que* $\mathrm{val}(P_F) \in \mathbb{R}$ *et que*

$$0 \in \Big(\mathrm{dom}\, g - A(\mathrm{dom}\, f)\Big)^{\circ}. \tag{14.28}$$

Alors $\mathrm{val}(P_F) = \mathrm{val}(D_F)$ *et* $\mathrm{Sol}(D_F)$ *est non vide et compact.*

Démonstration. En tant que fonction marginale définie par (14.26), avec $f \in \mathrm{Conv}(\mathbb{E})$ et $g \in \mathrm{Conv}(\mathbb{F})$, v_F est une fonction convexe. Elle est aussi propre car $v_F(0) \in \mathbb{R}$ et, par (14.28), $0 \in (\mathrm{dom}\, v_F)^{\circ}$ (donc v_F ne peut pas prendre la valeur $-\infty$). Le résultat se déduit alors de la proposition 14.11, point 1.*d*. □

Ce résultat a de multiples applications. Nous en donnons deux à titre d'illustration. Le corollaire 14.15 étend des résultats des propositions 3.71 et 3.72 au cas de fonctions pouvant prendre la valeur $+\infty$. Le corollaire 14.16 étudie la dualité de Fenchel pour des problèmes avec contraintes linéaires d'égalité.

Corollaire 14.15 *Dans le cadre défini ci-dessus, avec* f *et* g *convexes, on a*

$$\partial(f + g \circ A)(x) \supseteq \partial f(x) + A^* \partial g(Ax),$$

avec égalité si la condition de stabilité (14.28) *a lieu.*

Démonstration. [$\supseteq$] Soit $x \in \mathrm{dom}\, f$ tel que $\partial f(x) \neq \varnothing$ et $\partial g(Ax) \neq \varnothing$ (sinon, l'inclusion est triviale). Si $x^* \in \partial f(x)$ et $y^* \in \partial g(Ax)$, alors $x \in \mathrm{dom}\, f$, $Ax \in \mathrm{dom}\, g$ et quel que soit $d \in \mathbb{E}$, on a

$$f'(x;d) \geqslant \langle x^*, d\rangle \quad \text{et} \quad g'(Ax;d) \geqslant \langle y^*, d\rangle.$$

Alors

$$\Big(f + g \circ A\Big)'(x;d) = f'(x;d) + g'(Ax;Ad) \geqslant \langle x^* + A^* y^*, d\rangle.$$

Donc $x^* + A^* y^* \in \partial(f + g \circ A)(x)$ et l'inclusion a lieu.

[$\subseteq$] Soient à présent $\bar{x} \in \mathrm{dom}(f + g \circ A)$ et $\bar{s} \in \partial(f + g \circ A)(\bar{x})$, ce qui s'écrit aussi $0 \in \partial(f + g \circ A)(\bar{x}) - \bar{s}$. Alors $\bar{x}$ est solution du problème

$$\inf_{x \in \mathbb{E}} \Big(\tilde{f}(x) + g(Ax)\Big),$$

où $\tilde{f}(x) := f(x) - \langle \bar{s}, x\rangle$. Comme $\tilde{f}^*(x^*) = f^*(x^* + \bar{s})$, le dual de ce problème s'écrit

$$\sup_{\lambda \in \mathbb{F}} \Big(-f^*(A^*\lambda + \bar{s}) - g^*(-\lambda)\Big).$$

Si la condition de stabilité (14.28) a lieu ($\mathrm{dom}\, \tilde{f} = \mathrm{dom}\, f$), les valeurs optimales des problèmes ci-dessus sont égales (d'après le théorème). Comme elles sont finies, le problème dual a une solution, disons $\bar{\lambda}$ et on a

$$f(\bar{x}) + f^*(A^*\bar{\lambda} + \bar{s}) + g(A\bar{x}) + g^*(-\bar{\lambda}) = \langle \bar{s}, \bar{x}\rangle.$$

En utilisant le fait que $f(\bar{x}) + f^*(A^*\bar{\lambda} + \bar{s}) \geqslant \langle A^*\bar{\lambda} + \bar{s}, \bar{x}\rangle$ et $g(A\bar{x}) + g^*(-\bar{\lambda}) \geqslant -\langle A^*\bar{\lambda}, \bar{x}\rangle$, on obtient

$$\langle \bar{s}, \bar{x}\rangle \leqslant f(\bar{x}) + f^*(A^*\bar{\lambda} + \bar{s}) + g(A\bar{x}) + g^*(-\bar{\lambda}).$$

On a donc des égalités dans les trois inégalités précédentes. Celles-ci impliquent que $A^*\bar{\lambda} + \bar{s} \in \partial f(\bar{x})$ et $-\bar{\lambda} \in \partial g(A\bar{x})$. Donc $\bar{s} \in \partial f(\bar{x}) + A^*\partial g(A\bar{x})$, ce qui conclut la démonstration. □

Corollaire 14.16 *On suppose donnés $f : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$, une application linéaire $A : \mathbb{E} \to \mathbb{F}$ et un élément $b \in \mathbb{F}$. Alors*

$$\sup_{y\in\mathbb{F}} \Big(\langle b, y\rangle - f^*(A^*y)\Big) \quad \leqslant \quad \inf_{\substack{x\in\mathbb{E}\\ Ax=b}} f(x)$$

Si f et g sont convexes et si $b \in (A(\operatorname{dom} f))^\circ$, alors on a égalité ci-dessus et, lorsqu'il est fini, le supremum à gauche est atteint.

Démonstration. On considère le problème (14.25) avec $g = \mathcal{I}_{\{b\}}$, la fonction indicatrice du singleton $\{b\}$. On a $g^*(y^*) = \langle b, y^*\rangle$ et le dual de (14.25) s'écrit comme à gauche dans l'inégalité ci-dessus. Enfin la condition de stabilité (14.28) s'écrit $0 \in (b - A(\operatorname{dom} f))^\circ$, c'est-à-dire $b \in (A(\operatorname{dom} f))^\circ$. □

14.4 Dualité non convexe de Toland

La dualité de Toland est une technique permettant de construire un problème d'optimisation dual, non nécessairement convexe. Elle trouve des applications dans la dualisation de problèmes dont le coût s'écrit comme une différence de fonctions convexes (section **??**) et en calcul des variations (en dimension infinie) pour montrer l'existence de solution des équations d'Euler-Lagrange.

La technique ressemble à celle de la dualité par perturbation de Rockafellar (section 14.2), mais les perturbations sont supposées ici « concaves », alors qu'elles étaient naturellement « convexes » dans la dualisation de Rockafellar. Il est remarquable que ce changement radical de dispositif ait quand même permis d'aboutir à une dualisation utile.

14.4.1 Perturbation du problème primal

On note le *problème primal*

$$(P) \qquad \inf_{x\in X} f(x),$$

dans lequel l'ensemble X et la fonction $f : X \to \overline{\mathbb{R}}$ sont quelconques. On note comme d'habitude $\mathrm{Sol}(P)$ l'ensemble des solutions de (P) et $\mathrm{val}(P)$ sa valeur optimale.

Comme à la section 14.2.1, on suppose que l'on peut plonger le problème primal dans une famille de problèmes qui sont des perturbations de celui-ci. Ces perturbations sont introduites en se donnant une *fonction de perturbation*

$$\Phi : X \times P \to \overline{\mathbb{R}} : (x, p) \mapsto \Phi(x, p),$$

où P est un espace vectoriel de dimension finie (si l'on n'a pas besoin de structure sur X et f, on en a besoin sur la perturbation). On suppose que l'on retrouve f lorsque la perturbation est nulle :

$$\forall\, x \in X : \quad \Phi(x, 0) = f(x). \tag{14.29}$$

Comme annoncé dans l'introduction de cette section, pour que la dualisation proposée ci-dessous se passe sans saut, l'on aura besoin que la fonction de perturbation soit *concave* en p à x fixé :

$$\boxed{\forall\, x \in X : \quad -\Phi(x, \cdot) \in \mathrm{Conv}(P).} \tag{14.30}$$

14.4.2 Le problème dual

Le problème dual de Rockafellar (14.14) s'écrit

$$(D_R) \qquad -\inf_{p^* \in P} \sup_{x \in X} \sup_{p \in P} \Big(\langle p^*, p \rangle - \Phi(x, p) \Big),$$

où l'on a utilisé l'expression (14.16) de la conjuguée de la fonction valeur. On note $\mathrm{val}(D_R)$ la valeur optimale de ce problème. Le *problème dual de Toland* s'écrit quant à lui

$$(D_T) \qquad \inf_{p^* \in P} \inf_{x \in X} \sup_{p \in P} \Big(\langle p^*, p \rangle + \Phi(x, p) \Big). \tag{14.31}$$

On note $\mathrm{Sol}(D_T)$ l'ensemble des solutions de (D_T) et $\mathrm{val}(D_T)$ sa valeur optimale.

Le signe « $+$ » devant $\Phi(x, p)$ dans (D_T) se justifie par le fait que l'on souhaite introduire une perturbation $\Phi(x, \cdot)$ concave (voir (14.30) et la discussion introduisant la notion de conjuguée avant (3.33)). Avec cette modification, il est difficile de voir au premier coup d'œil s'il y a un lien entre les valeurs optimales duales $\mathrm{val}(D_R)$ et $\mathrm{val}(D_T)$. Ce lien existe cependant et il sera manifeste après la mise en évidence de l'inégalité de dualité faible suivante.

Proposition 14.17 (dualité faible) $\mathrm{val}(P) \leqslant \mathrm{val}(D_T)$.

Démonstration. Il suffit d'observer qu'en prenant $p = 0$, on a

$$\sup_{p \in P} \Big(\langle p^*, p \rangle + \Phi(x, p) \Big) \geqslant \Phi(x, 0) = f(x).$$

En prenant l'infimum en $(p^*, x) \in P \times X$ de chaque membre de cette inégalité, on obtient le résultat. □

La relation de dualité faible de la dualisation de Rockafellar (proposition 14.8) permet alors d'écrire

$$\mathrm{val}(D_R) \leqslant \mathrm{val}(P) \leqslant \mathrm{val}(D_T).$$

Ces inégalités montrent que la dualisation de Toland n'est pas une dualisation min-max, laquelle aurait conduit à une valeur optimale duale plus petite que la valeur optimale primale (proposition 14.2).

De manière à simplifier les écritures, on introduit deux notations. Ces notations rappellent celles utilisées dans la dualisation de Rockafellar de la section 14.2, mais n'ont cependant pas de lien avec elles.

- On peut faire apparaître une conjuguée avec le supremum de (D_T), en introduisant la fonction opposée Ψ de Φ :

$$\Psi := -\Phi \tag{14.32}$$

et les fonctions partielles, à $x \in X$ fixé :

$$\Psi_x := \Psi(x, \cdot) : P \to \overline{\mathbb{R}}.$$

Le dual se récrit alors

$$(D_T) \qquad \inf_{p^* \in P} \inf_{x \in X} \Psi_x^*(p^*). \tag{14.33}$$

Comme $\Psi_x^* : P \to \overline{\mathbb{R}}$ est convexe, il est normal de la minimiser en p^*, ce qui est fait dans (D_T). Quant à la minimisation en $x \in X$ dans (D_T), elle a permis d'avoir l'inégalité de dualité faible de la proposition 14.17.

- La *fonction duale* $\delta : P \to \overline{\mathbb{R}}$ est alors définie en $p^* \in P$ par

$$\delta(p^*) := \inf_{x \in X} \Psi_x^*(p^*). \tag{14.34}$$

Avec ces notations, le problème dual consiste à *minimiser* la fonction duale :

$$(D_T) \qquad \inf_{p^* \in P} \delta(p^*). \tag{14.35}$$

Proposition 14.18 (dualité forte) *Si*

$$\forall\, x \in X : \quad \Psi_x^{**}(0) = \Psi_x(0), \tag{14.36}$$

alors $\mathrm{val}(P) = \mathrm{val}(D_T)$.

Démonstration. On a

$$\begin{aligned}
\mathrm{val}(P) &= \inf_x \ -\Psi_x(0) \qquad [f(x) = -\Psi(x,0) = -\Psi_x(0)] \\
&= \inf_x \ -\Psi_x^{**}(0) \qquad [\text{par hypothèse}] \\
&= \inf_x \inf_{p^*} \ \Psi_x^*(p^*) \qquad [\text{définition de la biconjuguée}] \\
&= \mathrm{val}(D_T) \qquad [(14.33)].
\end{aligned}$$

□

C'est le moment de se rappeler de la condition (14.30) requérant la concavité de la fonction de perturbation $\Phi(x,\cdot)$, c'est-à-dire la convexité de Ψ_x, quel que soit $x \in X$, et qui était annoncée comme souhaitable pour éviter un saut de dualité. La remarque ci-dessous renforce ce point de vue, avec plus de précision.

Remarque 14.19 *La condition suffisante de dualité forte* (14.36) *est réalisée si l'une des conditions suivantes a lieu :*

$$\forall\, x \in X : \quad \Psi_x \in \mathrm{Conv}(P) \text{ et } \partial\Psi_x(0) \neq \varnothing. \tag{14.37}$$

$$\forall\, x \in X : \quad \Psi_x \in \mathrm{C}\overline{\mathrm{onv}}(P), \tag{14.38}$$

En effet, d'une part, (14.37) implique (14.36) par le point 2 de la proposition 3.60. D'autre part, (14.38) implique (14.36) par le point 3 de la proposition 3.46.

Même si les conditions (14.37) et (14.38) ne sont pas comparables, la seconde posera plus de difficulté du fait de la vacuité éventuelle de $\partial\Psi_x(0)$, qui, on le verra, a « vocation » à contenir les solutions duales, pourvu que x soit une solution primale (proposition 14.22). □

14.4.3 Existence de solutions primale et duale

Le premier résultat donne des conditions pour que l'existence d'une solution duale entraîne celle d'une solution primale. En calcul des variations, en dimension infinie, les conditions (14.39) donnent une forme faible des équations d'Euler-Lagrange et la condition de transversalité naturelle [593].

Proposition 14.20 (existence de solution primale) *Supposons que* (14.30) *et* (14.36) *aient lieu. Si $\bar{p}^* \in \mathrm{Sol}(D_T)$, si $x \mapsto \Psi_x^*(\bar{p}^*)$ est s.c.i. et s'il existe une suite minimisante dans* (14.34) *qui converge vers $\bar{x}$, alors $\bar{x}$ est solution primale, il n'y a pas de saut de dualité et l'on a*

$$f(\bar{x}) = \Psi_{\bar{x}}^*(\bar{p}^*) = \delta(\bar{p}^*). \tag{14.39}$$

Démonstration. Par (14.36) et la proposition 14.18, il n'y a pas de saut de dualité, si bien qu'il suffit de montrer que $f(\bar{x}) \leqslant \Psi_{\bar{x}}^*(\bar{p}^*) \leqslant \mathrm{val}(D_T)$.

Soit $\{x_k\}$ la suite minimisante convergeant vers $\bar{x}$ et telle que $\Psi_{x_k}^*(\bar{p}^*)$ converge vers $\delta(\bar{p}^*)$. On a

$$\begin{aligned}
f(\bar{x}) &= \Psi_{\bar{x}}(0) && [(14.29) \text{ et } (14.32)]\\
&\leqslant \Psi_{\bar{x}}^*(\bar{p}^*) && [\text{conjuguée}]\\
&\leqslant \liminf_{k\to\infty} \Psi_{x_k}^*(\bar{p}^*) && [x \mapsto \Psi_x^*(\bar{p}^*) \text{ est s.c.i.}]\\
&= \delta(\bar{p}^*) && [\Psi_{x_k}^*(\bar{p}^*) \to \delta(\bar{p}^*)]\\
&= \mathrm{val}(D_T) && [\bar{p}^* \in \mathrm{Sol}(D_T)].
\end{aligned}$$

Ce sont les inégalités attendues. □

Remarque 14.21 On verra à la remarque **??**, dans le cas particulier de l'optimisation DC, que la condition (14.39) n'est pas suffisante pour assurer que $\bar{x} \in \mathrm{Sol}(P)$ et $\bar{p}^* \in \mathrm{Sol}(D_T)$. □

Le résultat suivant montre que si $\bar{x}$ est une solution primale et que $\partial\Psi_{\bar{x}}(0)$ est non vide, alors ce dernier ensemble est formé de solutions duales et il n'y a pas de saut de dualité.

Proposition 14.22 (existence de solution duale) *Si (14.30) a lieu, si $\bar{x} \in \mathrm{Sol}(P)$ et si $\bar{p}^* \in \partial\Psi_{\bar{x}}(0)$, alors $\bar{p}^*$ est solution duale, il n'y a pas de saut de dualité et l'on a (14.39).*

Démonstration. Par l'inégalité de dualité faible (proposition 14.17), il suffit de montrer que $\delta(\bar{p}^*) \leqslant \Psi^*_{\bar{x}}(\bar{p}^*) \leqslant \mathrm{val}(P)$. L'hypothèse $\bar{p}^* \in \partial\Psi_{\bar{x}}(0)$ permet d'écrire

$$\forall\, p \in P : \quad \Psi_{\bar{x}}(p) \geqslant \Psi_{\bar{x}}(0) + \langle \bar{p}^*, p\rangle = -\,\mathrm{val}(P) + \langle \bar{p}^*, p\rangle.$$

On en déduit

$$\mathrm{val}(P) \geqslant \sup_{p\in P} \Big(\langle \bar{p}^*, p\rangle - \Psi_{\bar{x}}(p) \Big) = \Psi^*_{\bar{x}}(\bar{p}^*) \geqslant \delta(\bar{p}^*),$$

par la définition (14.34) de la fonction duale. C'est l'inégalité attendue. □

14.4.4 Suites minimisantes primales et duales

Dans cette section, l'on montre comment une suite minimisante duale peut être construite à partir d'une suite minimisante primale. Le premier résultat suppose la condition suffisante d'absence de saut de dualité (14.37) selon laquelle $\partial\Psi_x(0) \neq \varnothing$ quel que soit $x \in X$; la suite minimisante duale est alors construite en choisissant ses éléments p^*_k de manière arbitraire dans $\partial\Psi_{x_k}(0)$. Le second résultat ne suppose que (14.38) ; les sous-différentiels $\partial\Psi_{x_k}(0)$ peuvent alors être vides ; dans ce cas, on construit la suite minimisante duale à partir d'éléments des sous-différentiels $\partial\Psi_{x_k}(p_k)$ pris en des points p_k convergeant vers zéro.

Proposition 14.23 (existence de suite minimisante duale I) *Supposons que (14.37) ait lieu. Si $\{x_k\}$ est une suite minimisante pour (P) et $p^*_k \in \partial\Psi_{x_k}(0)$, alors $\{p^*_k\}$ est une suite minimisante pour (D_T).*

Démonstration. On adapte à la situation présente, la démonstration de la proposition 14.22. L'hypothèse $p^*_k \in \partial\Psi_{x_k}(0)$ permet d'écrire

$$\forall\, p \in P : \quad \Psi_{x_k}(p) \geqslant \Psi_{x_k}(0) + \langle p^*_k, p\rangle = -f(x_k) + \langle p^*_k, p\rangle.$$

On en déduit

$$f(x_k) \geqslant \sup_{p\in P}\Big(\langle p_k^*, p\rangle - \Psi_{x_k}(p)\Big) = \Psi_{x_k}^*(p_k^*) \geqslant \delta(p_k^*) \geqslant \mathrm{val}(D_T),$$

par la définition (14.34) de la fonction duale. Comme $\{x_k\}$ est une suite minimisante primale, $f(x_k) \to \mathrm{val}(P)$. Par ailleurs, il n'y a pas de saut de dualité grâce à l'hypothèse (14.37) (voir la remarque 14.19), si bien que $f(x_k) \to \mathrm{val}(D_T)$. En passant à la limite dans l'inégalité ci-dessus, on voit que $\delta(p_k^*) \to \mathrm{val}(D_T)$, c'est-à-dire que $\{p_k^*\}$ est une suite minimisante duale. □

Proposition 14.24 (existence de suite minimisante duale II) *Supposons que (14.38) ait lieu. Si $\{x_k\}$ est une suite minimisante pour (P), alors il existe une suite $\{p_k\} \to 0$ dans P telle que $\partial\Psi_{x_k}(p_k) \neq \varnothing$ et il existe une suite minimisante $\{p_k^*\}$ pour (D_T) avec des $p_k^* \in \partial\Psi_{x_k}(p_k)$.*

Démonstration. □

14.5 Dualisation lagrangienne

14.5.1 Dualisation de contraintes d'égalité et d'inégalité

Dans cette section, on considère le problème d'optimisation sous contraintes fonctionnelles suivant

$$(P) \equiv (P_{X,EI}) \qquad \begin{cases} \min f(x) \\ x \in X \\ c_i(x) = 0, \quad \forall i \in E \\ c_i(x) \leqslant 0, \quad \forall i \in I, \end{cases} \tag{14.40}$$

où X est toujours un ensemble quelconque et E et I sont des ensembles finis d'indices, contenant respectivement $m_E = |E|$ et $m_I = |I|$ éléments, formant une partition de $[1\,{:}\,m]$. Donc $m = m_E + m_I$ et on note $c_E = \{c_i\}_{i\in E}$ et $c_I = \{c_i\}_{i\in I}$. La fonction f et les fonctions c_i, $i \in E \cup I$, sont définies sur X à valeurs dans $\mathbb{R}$.

Nous introduisons deux problèmes duaux de (14.40) en utilisant deux types de problèmes perturbés. Le lagrangien correspondant aux premières perturbations est très voisin du lagrangien ordinaire (4.34), servant à écrire les conditions d'optimalité. Cette approche est bien adaptée aux problèmes convexes. À l'inverse, lorsque le problème (14.40) n'est pas convexe, c'est le second type de perturbations qui est le plus approprié. Celles-ci conduisent au lagrangien augmenté.

Les problèmes duaux introduits ci-dessous sont obtenus en perturbant les contraintes fonctionnelles de (14.40), sans toucher à la « contrainte » $x \in X$ considérée comme « contrainte dure ». On dit que l'on *dualise* les contraintes fonctionnelles, celles définies au moyen des fonctions c_E et c_I. Comme X est un ensemble quelconque, la démarche suivie englobe le cas où on ne dualiserait qu'une partie des contraintes fonctionnelles. En effet, X peut être un ensemble défini par d'autres contraintes d'égalité

ou d'inégalité et peut donc prendre en compte les contraintes non dualisées. Cette remarque soulève la question du choix des contraintes à dualiser dans un cas concret. Sur ce sujet, ce sont les conditions (D_1)-(D_3) de la page 517 qui servent de gouverne. Comme les contraintes non dualisées (celles prisent en compte par X) se retrouvent dans les problèmes internes primaux (14.11), on cherchera à dualiser les contraintes difficilement traitées par un algorithme de minimisation.

Pour $p = (p_E, p_I) \in \mathbb{R}^m$, on introduit le problème perturbé de $(P_{X,EI})$ suivant et la *fonction valeur* v_0 associée :

$$v_0 : p \in \mathbb{R}^m \mapsto v_0(p) := \inf_{\substack{x\in X \\ c_E(x)+p_E=0 \\ c_I(x)+p_I\leqslant 0}} f(x) \in \overline{\mathbb{R}}. \tag{14.41}$$

La fonction de perturbation Φ_0 est donc définie sur $X \times \mathbb{R}^m$ par

$$\Phi_0(x,p) = \begin{cases} f(x) & \text{si } \ c_E(x) + p_E = 0 \ \text{ et } \ c_I(x) + p_I \leqslant 0 \\ +\infty & \text{sinon.} \end{cases} \tag{14.42}$$

Le lagrangien associé à ces perturbations se calcule au moyen de la formule (14.19). Avec $\lambda = (\lambda_E, \lambda_I) \in \mathbb{R}^m$, on trouve

$$\begin{aligned} \ell_0(x,\lambda) &= -\sup_{p=(p_E,p_I)} \Big(\langle \lambda, p\rangle - \Phi_0(x,p) \Big) \\ &= -\sup_{\substack{p=(p_E,p_I) \\ p_E=-c_E(x) \\ p_I\leqslant -c_I(x)}} \Big(\langle \lambda, p\rangle - f(x) \Big). \end{aligned}$$

Finalement, on obtient comme *lagrangien associé à la fonction de perturbation* (14.42) :

$$\boxed{\ell_0(x,\lambda) = \begin{cases} f(x) + \displaystyle\sum_{i\in E\cup I} \lambda_i c_i(x) & \text{si } \ \lambda_I \geqslant 0 \\ -\infty & \text{sinon.} \end{cases}} \tag{14.43}$$

Pour λ tel que $\lambda_I \geqslant 0$, on retrouve le *lagrangien ordinaire* (4.34), servant à exprimer les conditions d'optimalité. Celui-ci correspond donc aux perturbations (14.41) du problème (14.40).

Le problème dual s'écrit

$$(D_0) \qquad \sup_{\substack{\lambda=(\lambda_E,\lambda_I) \\ \lambda_I\geqslant 0}} \ \inf_{x\in X} \ \ell_0(x,\lambda) \tag{14.44}$$

On notera que les seules contraintes qui portent sur les variables duales sont des contraintes de positivité sur λ_I. Ces contraintes sont simples, surtout si on les compare aux contraintes (éventuellement non linéaires) d'égalité et d'inégalité du problème original (14.40).

Le problème primal (14.40) s'écrit aussi

$$\inf_{x\in X} \ \sup_{\substack{\lambda=(\lambda_E,\lambda_I) \\ \lambda_I\geqslant 0}} \ \ell_0(x,\lambda),$$

ce qui montre que l'on a égalité en $(14.21)_c$. On remarque en effet que cette quantité s'écrit

$$\begin{aligned} & \inf_{x\in X} \sup_{\substack{\lambda \\ \lambda_I \geqslant 0}} \left(f(x) + \sum_{i\in E\cup I} \lambda_i c_i(x) \right) \\ = & \inf_{x\in X} \begin{cases} f(x) & \text{si } c_E(x) = 0 \text{ et } c_I(x) \leqslant 0 \\ +\infty & \text{sinon} \end{cases} \\ = & \operatorname{val}(P_{X,EI}). \end{aligned}$$

La proposition suivante analyse le lien entre les points-selles du lagrangien ordinaire (14.43) et les solutions de (14.40). Schématiquement, les points-selles du lagrangien ordinaire sont des solutions de (14.40) et la réciproque est vraie si le problème est convexe. Le point 2 de ce résultat apporte des précisions sur ce qui est affirmé à la proposition 13.15 : sous des hypothèses de convexité, non seulement $\bar{x}$ minimise le lagrangien $\ell_0(\cdot, \bar{\lambda})$, mais $(\bar{x}, \bar{\lambda})$ est un point-selle de ℓ_0.

Proposition 14.25 (point-selle de ℓ_0 et KKT) *On suppose que $X = \mathbb{R}^n$.*

1) *Si $(\bar{x}, \bar{\lambda})$ est point-selle du lagrangien ordinaire (14.43) sur $\mathbb{R}^n \times \mathbb{R}^m$, alors $\bar{x}$ est solution (globale) de (14.40). Si, de plus, les fonctions f et c sont différentiables, alors les conditions d'optimalité (4.33) soient vérifiées avec $(x_*, \lambda_*) \equiv (\bar{x}, \bar{\lambda})$.*
2) *Réciproquement, supposons que f et les $\{c_i\}_{i\in I}$ soient convexes, que c_E soit affine et que les fonctions f et c soient différentiables en une solution $\bar{x}$ du problème. Supposons également qu'il existe un multiplicateur $\bar{\lambda}$ tel que les conditions d'optimalité de Karush, Kuhn et Tucker (4.33) aient lieu avec $(x_*, \lambda_*) \equiv (\bar{x}, \bar{\lambda})$. Alors, $(\bar{x}, \bar{\lambda})$ est un point-selle du lagrangien ordinaire (14.43) sur $\mathbb{R}^n \times \mathbb{R}^m$.*

Démonstration. 1) Si $(\bar{x}, \bar{\lambda})$ est point-selle du lagrangien (14.43) sur $\mathbb{R}^n \times \mathbb{R}^m$, alors $\bar{x}$ est solution du problème primal (théorème 14.3), qui n'est autre que (14.40). On peut aussi invoquer la proposition 4.74.

Si, de plus, f et c sont différentiables, le fait que $\bar{x}$ minimise $\ell_0(\cdot, \bar{\lambda})$ sur $\mathbb{R}^n$ implique que $\nabla_x \ell_0(\bar{x}, \bar{\lambda}) = 0$. Par ailleurs, le fait que $\bar{\lambda}$ maximise $\ell(\bar{x}, \cdot)$ sur $\mathbb{R}^m$ implique l'admissibilité de $\bar{x}$, ainsi que la complémentarité (implication $(i) \Rightarrow (ii)$ de la proposition 4.73).

2) La minimalité de $\ell_0(\cdot, \bar{\lambda})$ en $\bar{x}$ a été démontrée par la proposition 13.15 et la maximalité de $\ell_0(\bar{x}, \cdot)$ en $\bar{\lambda}$ résulte de l'implication $(ii) \Rightarrow (i)$ de la proposition 4.73. □

D'après le théorème 14.3 et la proposition 14.25, on voit que, pour une grande classe de problèmes *convexes*, il n'y a pas de saut de dualité et que l'on peut trouver $\bar{\lambda}$ en résolvant le problème dual (14.44). Sans convexité, il peut y avoir ou ne pas y avoir de saut de dualité comme le montre l'exemple 1 ci-dessous.

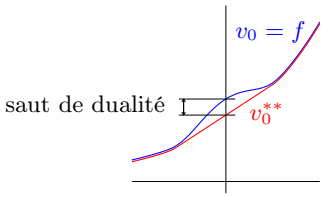
Exemples 14.26 (saut de dualité et existence de solution primale-duale) Il est un cas où la fonction valeur (14.41) associée au problème perturbé se calcule

facilement. C'est celui où le critère $f : \mathbb{R} \to \mathbb{R}$ est une fonction *croissante*, où $X = \mathbb{R}$ et où les contratintes se résument à l'appartenance à $\mathbb{R}_+$:

$$\begin{cases} \inf\ f(x) \\ x \geqslant 0, \end{cases}$$

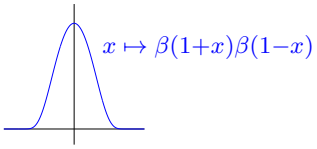
Par la croissance de f, la valeur optimale de ce problème est $f(0)$ et sa fonction valeur s'écrit $p \in \mathbb{R} \mapsto v_0(p) := \inf(f(x) : x \geqslant p\} = f(p)$. La biconjuguée de v_0 est donc l'enveloppe convexe fermée de f, si bien que la saut de dualité $v_0(0) - v_0^{**}(0)$ s'écrit aussi $f(0) - f^{**}(0)$, ce qui peut se voir aisément par l'examen de f. Voici quelques cas de fonctions croissantes f.

1) *Problème non convexe avec ou sans saut de dualité.* Prenons pour f la perturbation suivante de l'exponentielle



$$x \mapsto f(x) = e^x + \alpha\beta(1+x)\beta(1-x), \qquad (14.45)$$

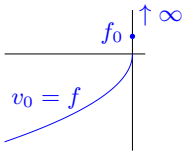
où $\alpha > 0$ est pris assez petit, tandis que la perturbation a pour support le compact $[-1, 1]$ et est construite au moyen de la fonction $\beta : \mathbb{R} \to \mathbb{R}_+$, de classe C^∞, définie par



$$\beta(t) = \begin{cases} e^{-1/t^2} & \text{si } t > 0 \\ 0 & \text{si } t \leqslant 0. \end{cases}$$

Dans ce cas, le saut de dualité $v_0(0) - v_0^{**}(0)$ est non nul (voir la figure ci-dessus, à côté de la formule de f). Si au lieu de f on prenait la fonction $f_t : x \mapsto f_t(x) := f(x + t)$, avec f donnée par (14.45) et $t \notin [-1, 1]$, il n'y aurait pas de saut de dualité, bien que v_0 ne soit pas convexe.

2) *Problème convexe avec ou sans saut de dualité et sans solution duale.* Supposons que f soit définie par



$$f(x) = \begin{cases} -\sqrt{-x} & \text{si } x < 0 \\ f_0 & \text{si } x = 0 \\ +\infty & \text{si } x > 0. \end{cases}$$

Si $f_0 \geqslant 0$, la fonction est convexe. Dans ce cas, la biconjuguée de $v_0 = f$ est la fonction définie par

$$v_0^{**}(x) = f^{**}(x) = \begin{cases} -\sqrt{-x} & \text{si } x \leqslant 0 \\ +\infty & \text{si } x > 0. \end{cases}$$

En accord avec le point 1.a de la proposition 14.11, on voit que le saut de dualité $f_0 - 0$ est nul si, et seulement si, $f_0 = 0$, c'est-à-dire si, et seulement si, $v_0 = f$ est s.c.i. en 0. Par ailleurs, comme v_0^{**} n'est pas sous-différentiable en 0 [par la proposition 3.62 et $(v_0^{**})'(0; -1) = -\infty$], le problème dual n'a pas de solution ; voir (14.18). □

La proposition 14.27 ci-dessous est intéressante pour ses interprétations, qui sont multiples. Donnons-en deux, qui ne sont pas sans rapport.

- Le calcul de la fonction valeur v_0 en une perturbation p donnée requiert la résolution du problème $(P_{X,EI})$ perturbé par p comme dans (14.41), qui est potentiellement difficile à résoudre. La proposition nous apprend que l'on peut évaluer cette fonction en une perturbation p non connue à l'avance en minimisant le lagrangien ordinaire $\ell_0(\cdot,\bar\lambda)$, problème normalement plus simple : si $\bar x$ est un minimiseur de ce lagrangien, alors $v_0(-c(\bar x)) = f(\bar x)$. On comprend par la figure 14.2, qui

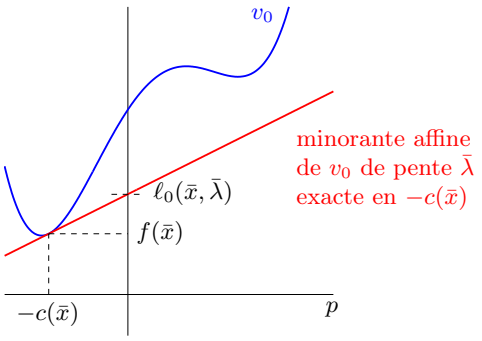


Fig. 14.2. Illustration de la proposition 14.27

illustre la proposition, que toutes les perturbations ne sont pas nécessairement atteignables par cette méthode, en particulier si v_0 n'est pas convexe ou si v_0 est convexe mais n'est pas sous-différentiable. Ainsi, un point $\bar x$ satisfaisant les contraintes ne pourra pas nécessairement être obtenu par cette méthode.

- Le résultat exprime aussi, à sa manière, le *théorème d'Everett* (exercice 14.12) : si $\bar x$ minimise le lagrangien $\ell_0(\cdot,\bar\lambda)$ avec $\bar\lambda_I \geqslant 0$, alors $\bar x$ minimise f sur $X_p := \{x \in X : c_E(x) + p_E = 0,\ c_I(x) + p_I \leqslant 0\}$. En effet, puisque $v_0(-c(\bar x)) = f(\bar x)$, il s'en ensuit que $f(x) \geqslant f(\bar x)$ pour tout $x \in X_p$; de plus, comme $\bar x \in X_p$, $\bar x$ est bien solution du problème perturbé comme dans (14.41) par $p = -c(\bar x)$.

Proposition 14.27 (minorante affine de v_0) *Soient $\bar\lambda \in \mathbb{R}^m$ tel que $\bar\lambda_I \geqslant 0$ et ℓ_0 le lagrangien ordinaire, défini par (14.43). Alors $\bar x \in \arg\min\{\ell_0(x,\bar\lambda) : x \in X\}$ si, et seulement si, la fonction affine*

$$p \in \mathbb{R}^m \mapsto \ell_0(\bar x,\bar\lambda) + \bar\lambda^{\mathsf T} p \tag{14.46}$$

est une minorante de la fonction valeur v_0 définie en (14.41). Dans ce cas, cette minorante affine est exacte en $p = -c(\bar x)$ et on a $v_0(-c(\bar x)) = f(\bar x)$.

Démonstration. [Préliminaire] Pour $p \in \mathbb{R}^m$, on note $X_p := \{x \in X : c_E(x) + p_E = 0,\ c_I(x) + p_I \leqslant 0\}$ l'ensemble admissible du problème perturbé dans (14.41). On observe que $\operatorname{dom} v_0 = \{p \in \mathbb{R}^m : X_p \neq \varnothing\}$.

[⇒] Pour tout $\varepsilon > 0$ et $p \in \operatorname{dom} v_0$, on peut trouver un $x \in X_p$ tel que $f(x) \leqslant v_0(p) + \varepsilon$. Par la propriété de $\bar{x}$ et $\bar{\lambda}_I \geqslant 0$, on a

$$\ell_0(\bar{x}, \bar{\lambda}) + \bar{\lambda}^{\mathsf{T}} p \leqslant f(x) + \bar{\lambda}^{\mathsf{T}}(c(x) + p) \leqslant v_0(p) + \varepsilon.$$

Comme $\varepsilon > 0$ et $p \in \operatorname{dom} v_0$ sont arbitraires, on en déduit que la fonction affine (14.46) est une minorante de v_0 sur $\operatorname{dom} v_0$, donc aussi sur $\mathbb{R}^m$.

[⇐] Quel que soit le couple $(x, p) \in X_p \times \mathbb{R}^m$, on a par hypothèse :

$$\ell_0(\bar{x}, \bar{\lambda}) + \bar{\lambda}^{\mathsf{T}} p \leqslant v_0(p) \leqslant f(x). \tag{14.47}$$

Quel que soit $x \in X$, $(x, p) \in X_p \times \mathbb{R}^m$ en prenant $p = -c(x)$, si bien que l'inégalité entre les membres extrêmes de (14.47) conduit à $\ell_0(\bar{x}, \bar{\lambda}) \leqslant \ell_0(x, \bar{\lambda})$ et montre que $\bar{x} \in \arg\min\{\ell_0(x, \bar{\lambda}) : x \in X\}$.

[Conclusion] Si $p = -c(\bar{x})$, $(\bar{x}, p) \in X_p \times \mathbb{R}^m$ et (14.47) devient $f(\bar{x}) \leqslant v_0(-c(\bar{x})) \leqslant f(\bar{x})$, ce qui montre d'une part l'exactitude de la minorante affine (14.46) en $p = -c(\bar{x})$ et $v_0(-c(\bar{x})) = f(\bar{x})$. □

Voici une propriété reliant la fonction valeur v_0 définie en (14.41) et le problème dual (D_0) défini en (14.44) ; elle donne des conditions pour que le problème dual soit non borné, qui sont que $\operatorname{val}(D_0) \neq +\infty$ et $0 \notin \operatorname{adh}(\operatorname{dom} v_0)$. Dans ce cas, bien sûr, l'infimum de la fonction duale vaut $-\infty$. La démonstration montre que la fonction duale tend en réalité vers $-\infty$ le long du vecteur directeur de l'hyperplan séparant 0 et l'adhérence de $\operatorname{dom} v_0$.

Proposition 14.28 (dual lagrangien non borné) *Lorsque* $\operatorname{val}(D_0) \neq -\infty$, *on a*

$$0 \notin \overline{\operatorname{dom} v_0} \qquad \Longrightarrow \qquad \operatorname{val}(D_0) = +\infty.$$

Démonstration. Soit $\mathscr{D} := \operatorname{dom} v_0$. Si $0 \notin \overline{\mathscr{D}}$, on peut séparer strictement $\{0\}$ et $\overline{\mathscr{D}}$: il existe $\mu \in \mathbb{R}^m$ tel que

$$\sup_{p \in \mathscr{D}} \mu^{\mathsf{T}} p < 0.$$

On déduit deux informations de cette inégalité.

- D'abord $\mu_I \geqslant 0$. En effet, en prenant $x \in X$, $i \in I$ et $t \geqslant 0$, on voit que $p := -c(x) - te^i$ est dans $\mathscr{D}$, si bien que $\mu^{\mathsf{T}} p \leqslant 0$ ou $-\mu^{\mathsf{T}} c(x) - t\mu_i \leqslant 0$ pour tout $t \geqslant 0$, dont on déduit que $\mu_i \geqslant 0$.
- Pour $x \in X$, $p := -c(x) \in \mathscr{D}$ et donc

$$\inf_{x \in X} \mu^{\mathsf{T}} c(x) > 0. \tag{14.48}$$

Par la définition de δ_0 et $X \neq \varnothing$, on a $\delta_0 > -\infty$. Par ailleurs, $\operatorname{val}(D) \neq -\infty$ implique qu'il existe un $\lambda \geqslant 0$ tel que $\delta_0(\lambda) < +\infty$. Pour ce $\lambda \geqslant 0$, on a donc $\delta_0(\lambda) \in \mathbb{R}$.

Avec ce λ et le μ utilisé en (14.48), on a pour tout $t \geqslant 0$: $\lambda + t\mu \geqslant 0$ et donc

$$\begin{aligned}\delta_0(\lambda+t\mu) &= -\inf_{x\in X} f(x)+(\lambda+t\mu)^{\mathsf{T}}c(x)\\ &\leqslant \underbrace{-\inf_{x\in X}\ell(x,\lambda)}_{\delta_0(\lambda)\in\mathbb{R}} -t\quad \underbrace{\inf_{x\in X}\mu^{\mathsf{T}}c(x)}_{>0 \text{ par } (14.48)},\end{aligned}$$

qui tend donc vers $-\infty$ lorsque $t\to+\infty$. On en déduit que $\mathrm{val}(D_0)=+\infty$. □

Voici un corrolaire immédiat qui donnent des conditions pour que $\mathrm{val}(P)\in\mathbb{R}$ (ce qui implique, en particulier, que (P) est réalisable). Une de ces conditions est que $\mathrm{val}(D_0)\in\mathbb{R}$, mais on pourra très bien avoir $\mathrm{val}(D_0)<\mathrm{val}(P)$ (présence d'un saut de dualité).

Corollaire 14.29 *Si* $\mathrm{dom}\, v_0$ *est fermé et* $\mathrm{val}(D_0)\in\mathbb{R}$*, alors* $\mathrm{val}(P)\in\mathbb{R}$*.*

14.5.2 Dualisation de contraintes générales

Dans cette section, on considère un problème avec containtes plus générales que dans la section précédente, comme en (4.5), à savoir

$$(P_{PQ})\quad \begin{cases}\inf\ f(x)\\ x\in P\\ c(x)\in Q,\end{cases}$$

où $f:\mathbb{E}\to\mathbb{R}$ et $c:\mathbb{E}\to\mathbb{F}$ sont des fonctions, $P\subseteq\mathbb{E}$ est un ensemble et $Q\subseteq\mathbb{F}$ est un *convexe fermé* non vide de $\mathbb{F}$ (pas nécessairement un *cône*). Dans cette description et dans toute cette section, $\mathbb{E}$ et $\mathbb{F}$ sont deux espaces euclidiens dont les produits scalaires sont tous deux notés $\langle\cdot,\cdot\rangle$.

Le cas particulier où Q est un *cône* convexe fermé non vide permet parfois d'avoir des propriétés intéressantes supplémentaires. On note alors

$$Q\equiv K$$

pour éviter les confusions, et le problème (P_{PQ}) est alors noté (P_{PK}).

On prend comme *lagrangien* du problème (P_{PQ}) la fonction $\ell:\mathbb{E}\times\mathbb{F}\to\mathbb{R}$ qui est définie en $(x,\lambda)\in\mathbb{E}\times\mathbb{F}$ par

$$\ell(x,\lambda)=f(x)+\langle\lambda,c(x)\rangle.$$

Ce choix de lagrangien indique clairement que l'on cherche à dualiser la contrainte fonctionnelle $c(x)\in Q$, tandis que la contrainte $x\in P$ sera laissée telle quelle.

Évidemment, le problème (P_{PQ}) peut s'écrire

$$\inf_{x\in P} f(x)+\mathcal{I}_Q(c(x)),$$

où $\mathcal{I}_Q$ est l'indicatrice de Q. La dualisation min-max du problème (section 14.1) s'obtient en observant d'abord que $\mathcal{I}_Q(c(x))$ peut s'écrire comme un supremum. On

utilise pour cela le fait que, Q étant un convexe fermé non vide, l'indicatrice $\mathcal{I}_Q$ est incluse dans $\mathrm{C}\overline{\mathrm{onv}}(\mathbb{F})$ (proposition 3.6), si bien que cette dernière est égale à sa biconjuguée (point 3 de la proposition 3.46) et s'écrit donc bien comme un supremum. En utilisant l'identité $\mathcal{I}_Q^* = \sigma_Q$ (point 3 de l'exercice 3.28), dans laquelle σ_Q est la fonction d'appui de Q, on a pour tout $y \in \mathbb{F}$:

$$\mathcal{I}_Q(y) = \mathcal{I}_Q^{**}(y) = \sup_{\lambda \in \mathbb{F}} \langle \lambda, y \rangle - \sigma_Q(\lambda). \tag{14.49}$$

Le raisonnement précédent montre donc que le problème (P_{PQ}) s'écrit aussi

$$\inf_{x \in P} \sup_{\lambda \in \mathbb{F}} \ell(x, \lambda) - \sigma_Q(\lambda).$$

Le dual min-max correspondant à cette formulation s'écrit donc

$$\boxed{(D_{PQ}) \qquad \sup_{\lambda \in \mathbb{F}} \inf_{x \in P} \ell(x, \lambda) - \sigma_Q(\lambda).} \tag{14.50}$$

Dans ce problème dual, la minimisation en x se fait uniquement sur le lagrangien, tandis que la fonction d'appui σ_Q n'intervient que dans la maximisation en λ. Comme cette fonction d'appui peut prendre des valeurs infinies, elle impose à λ d'appartenir à son domaine qui est dans $(Q^\infty)^-$, le cône dual négatif du cône asymtotique de Q (point 5 de la proposition 3.11). Par sonséquent, le problème dual s'écrit aussi

$$\sup_{\lambda \in (Q^\infty)^-} \inf_{x \in P} \ell(x, \lambda) - \sigma_Q(\lambda). \tag{14.51}$$

Dans le cas où $Q \equiv K$ est un *cône* convexe fermé non vide, $\sigma_K = \mathcal{I}_{K^-}$ (point 3 de l'exercice 3.28) et le problème dual se simplifie en

$$\boxed{(D_{PK}) \qquad \sup_{\lambda \in K^-} \inf_{x \in P} \ell(x, \lambda).} \tag{14.52}$$

La proposition suivante établit un lien entre les points-selles de $(x, \lambda) \in P \times \mathbb{F} \mapsto \ell(x, \lambda) - \sigma_Q(\lambda)$ et des conditions d'optimalité *globales* du problème (P_{PQ}) qui, en dehors de leur caractère global (minimisation *globale* du lagrangien), rappellent les conditions du premier ordre (4.73) du théorème 4.56. Il n'y a pas d'hypothèse de convexité dans ce résultat, mais le fait d'avoir un point-selle du lagrangien est rarement vérifié sans convexité.

Proposition 14.30 (point-selle et COs globales) *Dans le cadre décrit au début de cette section, les conditions suivantes sont équivalentes*

(*i*) $(\bar{x}, \bar{\lambda})$ *est point-selle de* $(x, \lambda) \in P \times \mathbb{F} \mapsto \ell(x, \lambda) - \sigma_Q(\lambda)$,
(*ii*) $\bar{x} \in \arg\min \{\ell(x, \bar{\lambda}) : x \in P\}$ *et* $\bar{\lambda} \in \mathrm{N}_Q(c(\bar{x}))$.

Lorsque $Q \equiv K$ est un cône convexe fermé non vide, les conditions d'optimalité du point (ii) s'écrivent

$$\bar{x} \in \arg\min \{\ell(x, \bar{\lambda}) : x \in P\} \qquad et \qquad K^- \ni \bar{\lambda} \perp c(\bar{x}) \in K. \tag{14.53}$$

Démonstration. 1) $[(i) \Rightarrow (ii)]$ Montrons la première affirmation de (ii). Observons d'abord que

$$\sigma_Q(\bar{\lambda}) \in \mathbb{R}. \tag{14.54}$$

En effet $\sigma_Q(\bar{\lambda}) > -\infty$, parce que Q est non vide, et $\sigma_Q(\bar{\lambda}) < +\infty$, parce que $\bar{\lambda}$ maximise $\lambda \mapsto \ell(\bar{x}, \lambda) - \sigma_Q(\lambda)$ qui est fini en $\lambda = 0$. Alors, le fait que $\bar{x}$ minimise $x \mapsto \ell(x, \bar{\lambda}) - \sigma_Q(\bar{\lambda})$ sur P implique que $\bar{x}$ minimise $x \mapsto \ell(x, \bar{\lambda})$ sur P.

La deuxième affirmation de (ii) se démontre en exploitant davantage le fait que $\bar{\lambda}$ maximise $\lambda \mapsto \ell(\bar{x}, \lambda) - \sigma_Q(\lambda)$. Comme $f(\bar{x}) \in \mathbb{R}$, cela s'écrit

$$\sup_{\lambda \in \mathbb{F}} \langle \lambda, c(\bar{x}) \rangle - \sigma_Q(\lambda) = \langle \bar{\lambda}, c(\bar{x}) \rangle - \sigma_Q(\bar{\lambda}). \tag{14.55}$$

Par (14.49), le membre de gauche prend pour valeur $\mathcal{I}_Q(c(\bar{x}))$, qui est fini (car le membre de droite l'est par (14.54)). Donc $c(\bar{x}) \in Q$. Cela implique aussi que chaque membre est nul et donc que $\sigma_Q(\bar{\lambda}) = \langle \bar{\lambda}, c(\bar{x}) \rangle$ ou encore $\langle \bar{\lambda}, y \rangle \leqslant \langle \bar{\lambda}, c(\bar{x}) \rangle$ pour tout $y \in Q$, c'est-à-dire que $\bar{\lambda} \in \mathrm{N}_Q(c(\bar{x}))$.

$[(ii) \Rightarrow (i)]$ On remonte le raisonnement en sens inverse, avec quelques nuances. D'abord, $\ell(x, \bar{\lambda})$ est fini quel que soit $x \in P$; donc, que $\sigma_P(\bar{\lambda})$ soit fini ou pas, il est clair que $\bar{x} \in \arg\min \{\ell(x, \bar{\lambda}) : x \in P\}$ implique que $\bar{x}$ minimise $x \in P \mapsto \ell(x, \bar{\lambda}) - \sigma_P(\bar{\lambda})$, si bien que $\bar{x}$ est solution primale. Pour montrer que $\bar{\lambda}$ est slution duale et qu'il n'y a pas de saut de dualité, il s'agit de démontrer (14.55) ou, en utilisant (14.49) et $c(\bar{x}) \in Q$ (qui provient de la seconde condition dans (ii)), que $\langle \bar{\lambda}, c(\bar{x}) \rangle = \sigma_Q(\bar{\lambda})$ ou encore que

$$\forall\, y \in Q : \qquad \langle \bar{\lambda}, y \rangle \leqslant \langle \bar{\lambda}, c(\bar{x}) \rangle.$$

Lorsque $c(\bar{x}) \in Q$, comme ici, cette condition est équivalente à $\bar{\lambda} \in \mathrm{N}_Q(c(\bar{x}))$, qui est garantie par (ii).

2) Pour obtenir (14.53), on utilise le lemme 4.55. □

14.6 Dualisation lagrangienne augmentée

14.6.1 Dualisation de contraintes d'égalité et d'inégalité

En dualité, le lagrangien augmenté a plusieurs intérêts :

- si le paramètre de pénalisation r est pris assez grand, il n'y a pas de saut de dualité pour des problèmes (14.40) non convexes (proposition 14.31),
- la fonction duale associée est différentiable, c'est la *régularisée de Moreau-Yosida* de la fonction duale associée au lagrangien ordinaire,
- si r est pris assez grand, les problèmes internes primaux n'ont pas de solutions importunes.

Mais il a aussi des inconvénients, celui de détériorer le conditionnement des problèmes internes primaux et celui d'empêcher le découplage des variables des problèmes décomposables.

Le lagrangien augmenté a déjà été introduit dans le chapitre sur la pénalisation, essentiellement au moyen de considérations intuitives, cherchant à créer une cuvette autour d'une solution vérifiant les conditions d'optimalité du second ordre (voir la section 13.4). On l'introduit ici par une technique de perturbation avec un autre objectif, celui de convexifier la fonction valeur dans un voisinage de 0. Si l'on ajoute au critère de (14.41) le terme $(r/2)\|p\|_2^2$, celui-ci se retrouvera ajouter tel quel à la fonction valeur, ce qui aura pour effet de la convexifier près de l'origine.

Soit $r \geqslant 0$. Pour $p = (p_E, p_I) \in \mathbb{R}^m$, on introduit le problème perturbé de $(P_{X,EI})$ suivant et la *fonction valeur* v_r associée :

$$v_r : p \in \mathbb{R}^m \mapsto v_r(p) := \inf_{\substack{x \in X \\ c_E(x)+p_E=0 \\ c_I(x)+p_I \leqslant 0}} \left(f(x) + \frac{r}{2}\|p\|_2^2 \right) = v_0(p) + \frac{r}{2}\|p\|_2^2 \in \overline{\mathbb{R}}. \quad (14.56)$$

On a noté $\|\cdot\|_2$ la norme ℓ_2. La perturbation des contraintes est donc identique à celle utilisée dans (14.41), mais ici une perturbation du critère est aussi introduite. Le problème d'optimisation dans (14.56) équivaut à minimiser en x la fonction définie sur $X \times \mathbb{R}^m$ par

$$\Phi_r(x,p) = \begin{cases} f(x) + \dfrac{r}{2}\|p\|_2^2 & \text{si } \; c_E(x) + p_E = 0 \;\; \text{et} \;\; c_I(x) + p_I \leqslant 0 \\ +\infty & \text{sinon.} \end{cases}$$

Calculons le lagrangien ℓ_r associé aux perturbations introduites. Pour $\lambda = (\lambda_E, \lambda_I)$, il s'écrit

$$\begin{aligned} \ell_r(x,\lambda) &= -\sup_{p=(p_E,p_I)} \Big(\langle \lambda, p \rangle - \Phi_r(x,p) \Big) \\ &= f(x) - \sup_{\substack{p=(p_E,p_I) \\ p_E=-c_E(x) \\ p_I \leqslant -c_I(x)}} \Big(\langle \lambda, p \rangle - \frac{r}{2}\|p\|_2^2 \Big). \end{aligned}$$

Si $r = 0$, on retrouve le lagrangien ordinaire ℓ_0 de la formule (14.43). Si $r > 0$, le supremum en p_I est atteint pour $p_I = \min(\lambda_I/r, -c_I(x))$, si bien que l'on a

$$\boxed{\begin{aligned} \ell_r(x,\lambda) = f(x) &+ \sum_{i \in E} \left[\lambda_i c_i(x) + \frac{r}{2} c_i(x)^2 \right] \\ &+ \sum_{i \in I} \left[\lambda_i \max\left(\frac{-\lambda_i}{r}, c_i(x) \right) + \frac{r}{2} \left(\max\left(\frac{-\lambda_i}{r}, c_i(x) \right) \right)^2 \right]. \end{aligned}} \quad (14.57)$$

On a donc retrouvé le lagrangien augmenté défini en (13.21). Remarquons que, contrairement au lagrangien ordinaire, le lagrangien augmenté ne prend que des valeurs finies. Cette formule est compliquée, surtout du fait de la présence des contraintes d'inégalité. S'il n'y a que des contraintes d'égalité, elle devient

$$\ell_r(x,\lambda) = f(x) + \lambda^{\mathsf{T}} c(x) + \frac{r}{2}\|c(x)\|_2^2,$$

qui est le lagrangien ordinaire avec pénalisation quadratique des contraintes.

Le problème dual s'écrit comme d'habitude

$$\sup_{\lambda=(\lambda_E,\lambda_I)} \inf_{x\in X} \ell_r(x,\lambda)$$

et le problème primal s'écrit

$$\inf_{x\in X} \sup_{\lambda=(\lambda_E,\lambda_I)} \ell_r(x,\lambda).$$

Pour ce dernier, on constate en effet que

$$\begin{aligned}
&\sup_{\lambda=(\lambda_E,\lambda_I)} \ell_r(x,\lambda) \\
&\quad = \quad f(x) + \sum_{i\in E} \sup_{\lambda_i} \left[\lambda_i c_i(x) + \frac{r}{2} c_i(x)^2\right] \qquad (14.58)\\
&\qquad\quad + \sup_{\lambda_I\in\mathbb{R}^{m_I}} \left[- \sum_{\substack{i\in I\\ rc_i(x)+\lambda_i\leqslant 0}} \frac{(\lambda_i)^2}{2r} + \sum_{\substack{i\in I\\ rc_i(x)+\lambda_i>0}} \left(\lambda_i c_i(x) + \frac{r}{2} c_i(x)^2\right)\right].
\end{aligned}$$

Le supremum en λ_i ($i \in E$) du premier crochet vaut $+\infty$ si $c_i(x) \neq 0$ et vaut 0 si $c_i(x) = 0$. Le supremum en λ_i ($i \in I$) du second crochet vaut $+\infty$ si $c_i(x) > 0$ et 0 si $c_i(x) \leqslant 0$ [il est atteint en tout $\lambda_i \geqslant 0$ si $c_i(x) = 0$ et en $\lambda_i = 0$ si $c_i(x) < 0$]. Dès lors, on retrouve le problème primal :

$$\inf_{x\in X} \sup_{\lambda=(\lambda_E,\lambda_I)} \ell_r(x,\lambda) = \inf_{x\in X} \begin{cases} f(x) & \text{si } \ c_E(x) = 0 \ \text{ et } \ c_I(x) \leqslant 0 \\ +\infty & \text{sinon.} \end{cases}$$

On a donc, ici aussi, égalité en $(14.21)_c$.

La proposition suivante donne des conditions pour ne pas avoir de saut de dualité localement. Pour les problèmes non convexes sur $\mathbb{R}^n$, on peut donc avoir des résultats de dualité locaux via le lagrangien augmenté. Ce résultat renforce celui du théorème 13.19 en affirmant que, non seulement $\ell_r(\cdot,\bar\lambda)$ est minimisée localement en $\bar x$, mais que $\ell_r(\bar x,\cdot)$ est maximisée en $\bar\lambda$.

Proposition 14.31 (point-selle de ℓ_r et KKT) *On considère le problème (14.40) avec $X = \mathbb{R}^n$ et on suppose que f et $c_{E\cup I^0(\bar x)}$ sont deux fois dérivables en un minimum local $\bar x$ de ce problème. On suppose également que les conditions de (KKT) ont lieu en $(x_*,\lambda_*) = (\bar x,\bar\lambda) \in \mathbb{R}^n \times \mathbb{R}^m$ et que la condition suffisante d'optimalité du second ordre semi-forte (4.62) a lieu pour un certain multiplicateur optimal $\bar\lambda$. Alors, il existe un réel $\bar r > 0$ et un voisinage V de $\bar x$, tels que pour tout $r \geqslant \bar r$, tout $x \in V \setminus \{\bar x\}$ et tout $\lambda \in \mathbb{R}^m$, on ait*

$$\ell_r(\bar{x},\lambda) \leqslant \ell_r(\bar{x},\bar{\lambda}) < \ell_r(x,\bar{\lambda}). \tag{14.59}$$

Démonstration. La minimalité stricte de $\ell_r(\cdot,\lambda)$ en $\bar{x}$ sur V, pour r au-delà d'un certain seuil $\bar{r}>0$, a été établie au théorème 13.19. Il reste donc à montrer la maximalité de $\ell_r(\bar{x},\cdot)$ en $\bar{\lambda}$. Ici le seuil $\bar{r}>0$ n'est pas nécessaire : un $r>0$ suffit.

Soient $\lambda\in\mathbb{R}^m$ et $r>0$. D'après (14.57), on a

$$\ell_r(\bar{x},\lambda) = f(\bar{x}) + \sum_{\substack{i\in I\\ rc_i(\bar{x})+\lambda_i>0}} \Big(\lambda_i c_i(\bar{x}) + \frac{r}{2}c_i(\bar{x})^2\Big) - \sum_{\substack{i\in I\\ rc_i(\bar{x})+\lambda_i\leqslant 0}} \frac{(\lambda_i)^2}{2r}.$$

Le dernier terme est clairement négatif. Quant au second, en utilisant $rc_i(\bar{x})+\lambda_i>0$ et $c_i(\bar{x})\leqslant 0$, on voit que chaque terme de la somme est $\leqslant -rc_i(\bar{x})^2+\frac{r}{2}c_i(\bar{x})^2\leqslant 0$. Donc

$$\ell_r(\bar{x},\lambda)\leqslant f(\bar{x}).$$

On conclut en observant que $f(\bar{x})=\ell_r(\bar{x},\bar{\lambda})$. En effet, il suffit d'utiliser dans l'expression de $\ell_r(\bar{x},\bar{\lambda})$ ci-dessus le fait que $\bar{\lambda}_i c_i(\bar{x})=0$, que $rc_i(\bar{x})+\bar{\lambda}_i>0$ implique que $\bar{\lambda}_i>0$ et donc $c_i(\bar{x})=0$, et que $rc_i(\bar{x})+\bar{\lambda}_i\leqslant 0$ implique que $\bar{\lambda}_i=0$. □

Les *fonctions duales* associées au lagrangien ordinaire (14.43) et au lagrangien augmenté (14.57) sont définies en $\lambda\in\mathbb{R}^m$ respectivement par

$$\delta_0(\lambda) = -\inf_{x\in X}\ \ell_0(x,\lambda) \qquad \text{et} \qquad \delta_r(\lambda) = -\inf_{x\in X}\ \ell_r(x,\lambda). \tag{14.60}$$

La proposition 14.32 montre que δ_r est une régularisée de Moreau-Yosida de δ_0 sur $\mathbb{R}^m$ [519 ; 1973], une notion de régularisation introduite à la section 3.7.2.

Proposition 14.32 (fonction duale régularisée) *On suppose que f et les $\{c_i\}_{i\in I}$ sont convexes, que c_E est affine, que f et c sont à valeurs réelles et que $r>0$. Alors, les propriétés suivantes ont lieu :*

1) *δ_r est convexe fermée et ne prend pas la valeur $-\infty$,*
2) *δ_r est la régularisée de Moreau-Yosida de δ_0 pour le produit scalaire $(u,v)\mapsto \frac{1}{r}(u^{\mathsf{T}}v)$, c.-à-d., pour tout $\lambda\in\mathbb{R}^m$,*

$$\delta_r(\lambda) = \inf_{\mu\in\mathbb{R}^m}\Big(\delta_0(\mu)+\frac{1}{2r}\|\mu-\lambda\|_2^2\Big), \tag{14.61}$$

3) *si, de plus, $\delta_0\not\equiv+\infty$, alors δ_r prend des valeurs finies (elle est donc dans $\overline{\mathrm{Conv}}(\mathbb{R}^m)$) et est $\mathcal{C}^{1,1}$ sur $\mathbb{R}^m$.*

Démonstration. 1) Par sa définition (14.60) et l'expression (13.19) du lagrangien augmenté, δ_r est l'enveloppe supérieure en $(x,s)\in X\times\mathbb{R}^{m_I}_+$ de fonctions affines, si bien qu'elle est convexe, fermée (proposition 3.34) et ne prend pas la valeur $-\infty$.

2) La formule (14.60) de δ_r, le fait que ℓ_r est le lagrangien associé à la perturbation (14.56) de (P_{EI}) et (14.20) nous informent que $\delta_r = v_r^*$, où v_r est la fonction valeur définie en (14.56). De même $\delta_0 = v_0^*$. Comme v_r et v_0 sont reliés par la formule

$$v_r = v_0 + r\chi,$$

où χ la fonction autoconjuguée $p \mapsto \frac{1}{2}\|p\|_2^2$, on aura un lien entre δ_r et δ_0 en calculant v_r^* comme conjuguée d'une somme de fonctions (proposition 3.54).

Par la convexité de (P_{EI}), la fonction Φ_0 définie par (14.42) est convexe et donc aussi v_0 qui en est la fonction marginale (proposition 3.35). D'autre part $v_0 \not\equiv +\infty$ (quel que soit $x \in \mathbb{R}^n$, $v_0(-c(x)) \leqslant f(x) < +\infty$). Si v_0 prend la valeur $-\infty$, il en est de même de v_r (pour la même valeur de l'argument), auquel cas $\delta_0 = v_0^* \equiv +\infty$, $\delta_r = v_r^* \equiv +\infty$ et la formule (14.61) est vérifiée. Autrement $v_0 \in \mathrm{Conv}(\mathbb{R}^m)$ et en appliquant le point 2 de la proposition 3.54, on trouve également la formule (14.61) :

$$\begin{aligned}\delta_r(\lambda) &= v_r^*(\lambda)\\ &= \Big(v_0^* \boxplus (r\chi)^*\Big)(\lambda)\\ &= \inf_{\mu\in\mathbb{R}^m} \delta_0(\mu) + \frac{1}{2r}\|\mu-\lambda\|_2^2,\end{aligned}$$

où on a utilisé la formule (3.25) de l'inf-convolution et le fait que $(r\chi)^* = \chi/r$.

3) Observons d'abord que $\delta_0 \in \mathrm{C\overline{onv}}(\mathbb{R}^m)$ par le point 1 et par le fait que $\delta_0 \not\equiv +\infty$ par hypothèse. Pour tout λ, le problème de minimisation dans (14.61) a alors une solution unique dans $\mathrm{dom}\,\delta_0$, si bien que δ_r est à valeurs réelles. Comme régularisée de Moreau-Yosida, δ_r est dans $\mathrm{C\overline{onv}}(\mathbb{R}^m)$ et est $\mathcal{C}^{1,1}$ (proposition 3.80). □

Le lagrangien augmenté a également un intérêt pour les *problèmes convexes*, car il a un effet régularisant sur la fonction duale ainsi qu'un effet stabilisateur sur les solutions primales.

La proposition suivante est au lagrangien augmenté ce que la proposition 14.27 est au lagrangien ordinaire. Elle est illustrée par la figure 14.3. La minorante de la fonction valeur v_0 qui se manifeste ici est quadratique par morceaux plutôt qu'affine.

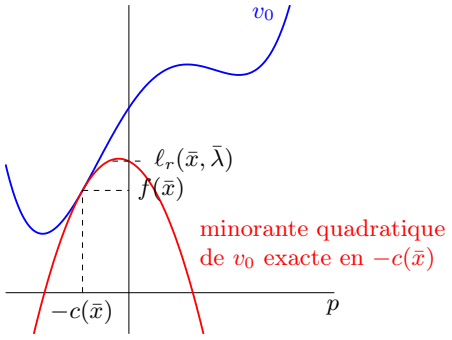


Fig. 14.3. Illustration de la proposition 14.33

Proposition 14.33 (minorante quadratique par morceaux de v_0) *Soient $r > 0$, $\bar\lambda \in \mathbb{R}^m$ et ℓ_r le lagrangien augmenté (14.57). Alors $\bar x \in \arg\min\{\ell_r(x,\bar\lambda) : x \in X\}$ si, et seulement si, la fonction quadratique par morceaux*

$$p \in \mathbb{R}^m \mapsto \ell_r(\bar x,\bar\lambda) + \bar\lambda_E^\mathsf{T} p_E - \frac{r}{2}\|p_E\|_2^2 - \frac{1}{2r}\left(\|(\bar\lambda_I - rp_I)^+\|_2^2 - \|\bar\lambda_I\|_2^2\right) \quad (14.62)$$

est une minorante de la fonction valeur v_0 définie en (14.41). Dans ce cas, cette minorante est exacte en $p = -c(\bar x)$ et on a $v_0(-c(\bar x)) = f(\bar x)$.

Démonstration. [Préliminaire] Soient $r > 0$ et $\bar\lambda \in \mathbb{R}^m$. Pour $p \in \mathbb{R}^m$, on note $X_p := \{x \in X : c_E(x) + p_E = 0,\ c_I(x) + p_I \leqslant 0\}$ l'ensemble admissible du problème perturbé dans (14.56). On note aussi $P := \{p \in \mathbb{R}^m : X_p \neq \varnothing\} = \operatorname{dom} v_0 = \operatorname{dom} v_r$. On note enfin $p \mapsto \chi(p)$ la fonction quadratique par morceaux définie en (14.62).

[$\Rightarrow$] Pour tout $\varepsilon > 0$ et $p \in P$, on peut trouver un $x \in X_p$ tel que $f(x) \leqslant v_0(p) + \varepsilon$. Dès lors

$$\begin{aligned}
\chi(p) &\leqslant \ell_r(x,\bar\lambda) + \bar\lambda_E^\mathsf{T} p_E - \frac{r}{2}\|p_E\|_2^2 - \frac{1}{2r}\left(\|(\bar\lambda_I - rp_I)^+\|_2^2 - \|\bar\lambda_I\|_2^2\right)\\
&\qquad [\bar x \in \arg\min\{\ell_r(x,\bar\lambda) : x \in X\}]\\
&= f(x) + \bar\lambda_E^\mathsf{T}(c_E(x) + p_E) + \frac{1}{2r}\left(\|(\bar\lambda_I + rc_I(x))^+\|_2^2 - \|(\bar\lambda_I - rp_I)^+\|_2^2\right)\\
&\qquad [\text{expressions (13.21) et (13.23) de } \ell_r]\\
&\leqslant f(x) \qquad [x \in X_p \text{ et donc } c_E(x) + p_E = 0 \text{ et } c_I(x) + p_I \leqslant 0]\\
&\leqslant v_0(p) + \varepsilon \qquad [\text{définition de } x].
\end{aligned}$$

Comme $\varepsilon > 0$ et $p \in P$ sont arbitraires, on en déduit que la fonction χ est une minorante de v_0 sur P, donc aussi sur $\mathbb{R}^m$ (v_0 vaut $+\infty$ en dehors de P).

[$\Leftarrow$] Quel que soit le couple $(x,p) \in X_p \times \mathbb{R}^m$, on a par hypothèse :

$$\chi(p) \leqslant v_0(p) \leqslant f(x). \quad (14.63)$$

Quel que soit $x \in X$, $(x,p) \in X_p \times \mathbb{R}^m$ en prenant $p = -c(x)$, si bien que l'inégalité entre les membres extrêmes de (14.63) conduit à $\chi(-c(x)) \leqslant v_0(p) \leqslant f(x)$ ou

$$\begin{aligned}
\ell_r(\bar x,\bar\lambda) &\leqslant f(x) + \bar\lambda_E^\mathsf{T} c_E(x) + \frac{r}{2}\|c_E(x)\|_2^2 + \frac{1}{2r}\left(\|(\bar\lambda_I + rc_I(x))^+\|_2^2 - \|\bar\lambda_I\|_2^2\right)\\
&= \ell_r(x,\bar\lambda),
\end{aligned}$$

par les expressions (13.21) et (13.23) de ℓ_r. Dès lors $\bar x \in \arg\min\{\ell_0(x,\bar\lambda) : x \in X\}$.

[Conclusion] Enfin $(\bar x, p) \in X_p \times \mathbb{R}^m$ si $p = -c(\bar x)$, si bien que (14.63) devient $f(\bar x) = \chi(-c(\bar x)) \leqslant v_0(-c(\bar x)) \leqslant f(\bar x)$, ce qui montre d'une part l'exactitude de χ en $p = -c(\bar x)$ et $v_0(-c(\bar x)) = f(\bar x)$. □

14.6.2 Dualisation de contraintes générales

Voir le syllabus complet.

14.7 Méthodes numériques ▲

On s'intéresse dans cette section à des algorithmes de résolution du problème dual

$$\sup_{y\in Y}\ \inf_{x\in X}\ \varphi(x,y),$$

où $\varphi : X \times Y \to \overline{\mathbb{R}}$. Deux méthodes sont décrites. La première minimise directement la fonction duale, si bien qu'il faut résoudre complètement les problèmes de Lagrange pour chaque valeur prise par la variable duale. Il faut donc que le problème soit bien adapté à cette approche : la fonction φ doit être facile à minimiser en $x \in X$. La seconde est une version simplifiée de la première, dans laquelle le problème de Lagrange est résolu de manière très grossière, puisque l'on ne fait qu'un seul pas de minimisation.

Nous présentons ces algorithmes dans le cadre de la *relaxation lagrangienne*, c'est-à-dire lorsque la fonction φ ci-dessus est le lagrangien d'un problème d'optimisation avec contraintes.

14.7.1 Minimisation de la fonction duale

Nous allons introduire un algorithme minimisant la fonction duale lorsqu'il s'agit de résoudre un problème d'optimisation sous la forme suivante :

$$\begin{cases} \min\ f(x) \\ c(x) \leqslant 0 \\ x \in X, \end{cases} \tag{14.64}$$

où $f : \mathbb{R}^n \to \mathbb{R}$, $c : \mathbb{R}^n \to \mathbb{R}^m$ et X est une partie non vide de $\mathbb{R}^n$ permettant d'exprimer les contraintes aisées à prendre en compte (typiquement des contraintes de borne). On note $\ell(x,\lambda) = f(x) + \lambda^{\mathsf{T}} c(x)$ le lagrangien qui « dualise » la contrainte d'inégalité. Le problème dual s'écrit

$$\inf_{\lambda\geqslant 0}\ \delta(\lambda), \tag{14.65a}$$

où $\delta : \mathbb{R}^m_+ \to \mathbb{R}$ est la *fonction duale*, qui est définie par

$$\delta(\lambda) := -\inf_{x\in X}\ \ell(x,\lambda). \tag{14.65b}$$

On rappelle que le problème de minimisation dans (14.65b) est appelé le *problème de Lagrange*.

Même en l'absence de convexité de f et c, la fonction duale est convexe. Cette propriété remarquable provient de l'infimum dans (14.65b) et de la structure du lagrangien, qui est affine par rapport au multiplicateur λ. La fonction duale n'est pas différentiable en général ; mais lorsque le problème de Lagrange a (au moins) une solution, on dispose d'un sous gradient de δ, dont on peut faire usage dans des algorithmes adaptés à la minimisation de fonctions non différentiables (par exemple la *méthode de faisceaux*, décrite dans [317 ; 1993, chapitres XIV et XV]).

Proposition 14.34 (convexité et sous-gradient de la fonction duale) *Si $\delta \not\equiv +\infty$, alors $\delta \in \mathrm{C}\overline{\mathrm{onv}}(\mathbb{R}^n)$. De plus,*

$$-c\left(\operatorname*{arg\,min}_{x\in X} \ell(x,\lambda)\right) \subseteq \partial\delta(\lambda).$$

Démonstration. Supposons que $\delta \not\equiv +\infty$. Alors δ est propre car elle ne prend pas la valeur $-\infty$ (X est supposé non vide). D'autre part, comme enveloppe supérieure (de la famille paramétrée par $x \in X$) de fonctions affines (donc convexes et fermées) $\lambda \mapsto -\ell(x,\lambda)$, δ est convexe et fermée (proposition 3.34).

Soit à présent $\bar{x}_\lambda$ une solution du problème de Lagrange (14.65b). Pour tout $\mu \in \mathbb{R}^m$, on a

$$\delta(\mu) \geqslant -\ell(\bar{x}_\lambda, \mu) = \delta(\lambda) - (\mu - \lambda)^\mathsf{T} c(\bar{x}_\lambda).$$

Donc $-c(\bar{x}_\lambda) \in \partial\delta(\lambda)$. □

Voici à présent des conditions assurant la différentiabilité de la fonction duale. Dans ce cas et lorsqu'une solution $\bar{x}_\lambda$ du problème de Lagrange existe, d'après le résultat précédent, on doit avoir $\nabla\delta(\lambda) = -c(\bar{x}_\lambda)$.

Proposition 14.35 (différentiabilité de la fonction duale) *Supposons que le problème de Lagrange (14.65b) ait une solution unique, notée $\bar{x}_\lambda$, pour tout λ voisin d'un certain λ_0 et que l'application $\lambda \mapsto c(\bar{x}_\lambda)$ soit continue en λ_0 ; alors δ est Fréchet-différentiable en λ_0 et $\nabla\delta(\lambda_0) = -c(\bar{x}_{\lambda_0})$. Si de plus, $\lambda \mapsto c(\bar{x}_\lambda)$ est continue dans un voisinage de λ_0, alors δ est C^1 dans un voisinage de λ_0.*

Démonstration. Comme dans la démonstration de la proposition 14.34, pour λ et $\lambda' \in \operatorname{dom}\delta$ et pour une solution $\bar{x}_\lambda$ du problème de Lagrange en λ, on a $-\delta(\lambda')+\delta(\lambda) \leqslant (\lambda' - \lambda)^\mathsf{T} c(\bar{x}_\lambda)$. En inversant le rôle de λ et λ', on obtient

$$(\lambda' - \lambda)^\mathsf{T} c(\bar{x}_{\lambda'}) \leqslant -\delta(\lambda') + \delta(\lambda) \leqslant (\lambda' - \lambda)^\mathsf{T} c(\bar{x}_\lambda).$$

Fixons à présent $\lambda = \lambda_0$ et prenons $\lambda' = \lambda_0 + \mu$ avec μ petit, de telle sorte que $\bar{x}_{\lambda_0+\mu}$ existe et que $c(\bar{x}_{\lambda_0+\mu})$ dépende continûment de μ voisin de zéro. Alors

$$\mu^\mathsf{T}[c(\bar{x}_{\lambda_0+\mu}) - c(\bar{x}_{\lambda_0})] \leqslant -\delta(\lambda_0+\mu) + \delta(\lambda_0) - \mu^\mathsf{T} c(\bar{x}_{\lambda_0}) \leqslant 0.$$

Dès lors

$$-\delta(\lambda_0+\mu) + \delta(\lambda_0) - c(\bar{x}_{\lambda_0})^\mathsf{T}\mu = o(\|\mu\|), \quad \text{lorsque } \mu \to 0.$$

Ceci montre que δ est différentiable en λ_0 et que $\nabla\delta(\lambda_0) = -c(\bar{x}(\lambda_0))$. □

Il n'est pas surprenant que le gradient $\nabla\delta(\lambda)$ soit donné par $-c(\bar{x}_\lambda)$, si on prend le point de vue suivant. Supposons en effet des conditions un peu plus régulières que celles de la proposition ci-dessus : la solution $\bar{x}_\lambda \equiv \bar{x}(\lambda)$ du problème de Lagrange

dans (14.65b) est une fonction différentiable de λ et ℓ est différentiable. Alors $\delta(\lambda) = -\ell(\bar{x}(\lambda), \lambda)$ et on a

$$\delta'(\lambda) \cdot \mu = -\ell'_x(\bar{x}(\lambda), \lambda) \cdot (\bar{x}'(\lambda) \cdot \mu) - \ell'_\lambda(\bar{x}(\lambda), \lambda) \cdot \mu = -\ell'_\lambda(\bar{x}(\lambda), \lambda) \cdot \mu,$$

puisque $\ell'_x(\bar{x}(\lambda), \lambda) = 0$ par optimalité de $\bar{x}_\lambda$. On retrouve donc $\nabla\delta(\lambda) = -c(\bar{x}_\lambda)$.

On peut à présent énoncer l'algorithme, qui peut être vu comme une méthode du type gradient projeté (section 12.1) appliquée au problème dual. Le gradient projeté permet de prendre en compte la contrainte de positivité sur le multiplicateur dans (14.65a).

Algorithme 14.36 (minimisation de la fonction duale)

1. Initialisation : choix de $\lambda_1 \geqslant 0$;
2. Pour $k = 1, 2, \ldots$ faire :
   2.1. Trouver x_k solution du problème $\min_{x \in X} \ell(x, \lambda_k)$;
   2.2. Si x_k est satisfaisant, on s'arrête ;
   2.3. $\lambda_{k+1} = (\lambda_k + \alpha_k c(x_k))^+$, où $\alpha_k > 0$ est un pas bien choisi.

On retrouve le déplacement du gradient projeté à l'étape 2.3. On passe pour l'instant sous silence l'étape délicate du choix du pas $\alpha_k > 0$.

Nous allons montrer la convergence de cet algorithme dans le cas des problèmes (fortement) convexes, avec de petits pas α_k. En ce qui concerne la convergence, des contraintes d'égalité affines peuvent être prise en compte par c (on les remplace par deux contraintes d'inégalité opposées).

Proposition 14.37 *On considère le problème* (14.64) *dans lequel on suppose que* $X = \mathbb{R}^n$, *que* $f : \mathbb{R}^n \to \mathbb{R}$ *est différentiable et fortement convexe de module* $\kappa > 0$ *et que* $c : \mathbb{R}^n \to \mathbb{R}^m$ *est lipschitzienne de module* $L > 0$ *et convexe. On suppose également que le lagrangien* ℓ *a un point-selle* $(\bar{x}, \bar{\lambda})$ *sur* $\mathbb{R}^n \times \mathbb{R}^m_+$. *On considère l'algorithme* 14.36 *dans lequel* α_k *pris dans un compact de* $]0, \frac{2\kappa}{L^2}[$. *Alors l'algorithme génère une suite* $\{x_k\}$ *qui converge vers* $\bar{x}$.

Démonstration. Comme f est fortement convexe et l'ensemble admissible du problème (14.64) est convexe, ce dernier a une solution unique, qui n'est autre que $\bar{x}$. En effet $\bar{x}$ est solution du problème primal associé au lagrangien ℓ — théorème 14.3 — qui est le problème (14.64).

L'optimalité de $\bar{x}$ donne

$$\nabla f(\bar{x})^\mathsf{T}(x - \bar{x}) + \bar{\lambda}^\mathsf{T}\Big(c'(\bar{x}) \cdot (x - \bar{x})\Big) = 0, \quad \forall x \in \mathbb{R}^n.$$

En utilisant la convexité de c et $\bar{\lambda} \geqslant 0$, on obtient

$$\nabla f(\bar{x})^\mathsf{T}(x - \bar{x}) + \bar{\lambda}^\mathsf{T}\Big(c(x) - c(\bar{x})\Big) \geqslant 0, \quad \forall x \in \mathbb{R}^n.$$

De même pour l'optimalité en x_k :

$$\nabla f(x_k)^\mathsf{T}(x - x_k) + \lambda_k^\mathsf{T}\Big(c(x) - c(x_k)\Big) \geqslant 0, \quad \forall x \in \mathbb{R}^n.$$

En prenant $x = x_k$ dans l'avant-dernière inégalité, $x = \bar{x}$ dans la dernière et en sommant, on obtient

$$(\nabla f(x_k) - \nabla f(\bar{x}))^\mathsf{T}(x_k - \bar{x}) \leqslant -(\lambda_k - \bar{\lambda})^\mathsf{T}\Big(c(x_k) - c(\bar{x})\Big).$$

On utilise alors la forte convexité de f et on note $\mu_k := \lambda_k - \bar{\lambda}$, ce qui conduit à

$$\kappa\|x_k - \bar{x}\|^2 \leqslant -\mu_k^\mathsf{T}\Big(c(x_k) - c(\bar{x})\Big). \tag{14.66}$$

Il reste à utiliser la récurrence sur λ_k. On a $\nabla_\lambda \ell(\bar{x}, \lambda) = c(\bar{x})$ et on se rappelle que $\ell(\bar{x}, \cdot)$ a un maximum en $\bar{\lambda}$ sur $\mathbb{R}^m_+$. Dès lors, pour tout $\lambda \in \mathbb{R}^m_+$:

$$0 \leqslant -\alpha_k c(\bar{x})^\mathsf{T}(\lambda - \bar{\lambda}) = \Big(\bar{\lambda} - (\bar{\lambda} + \alpha_k c(\bar{x}))\Big)^\mathsf{T}(\lambda - \bar{\lambda}),$$

ce qui s'écrit aussi

$$\bar{\lambda} = P_{\mathbb{R}^m_+}(\bar{\lambda} + \alpha_k c(\bar{x})).$$

D'autre part, la récurrence sur λ_k donne

$$\lambda_{k+1} = P_{\mathbb{R}^m_+}(\lambda_k + \alpha_k c(x_k)).$$

On utilise alors le fait que le projecteur $P_{\mathbb{R}^m_+}$ est 1-lipschitzien, la propriété de Lipschitz de c et (14.66) :

$$\begin{aligned}
\|\mu_{k+1}\| &\leqslant \|\mu_k + \alpha_k(c(x_k) - c(\bar{x}))\| \\
\|\mu_{k+1}\|^2 &\leqslant \|\mu_k\|^2 + 2\alpha_k\mu_k^\mathsf{T}(c(x_k) - c(\bar{x})) + \alpha_k^2\|c(x_k) - c(\bar{x})\|^2 \\
&\leqslant \|\mu_k\|^2 + (\alpha_k L^2 - 2\kappa)\alpha_k\|x_k - \bar{x}\|^2 \\
&\leqslant \|\mu_k\|^2 - \epsilon^2 L^2\|x_k - \bar{x}\|^2,
\end{aligned}$$

si $\alpha_k \in [\epsilon, \frac{2\kappa}{L^2} - \epsilon]$, avec $\epsilon > 0$ petit. Dès lors $\{\|\mu_k\|\}$ est décroissante, donc converge. On en déduit que $x_k \to \bar{x}$. □

Le résultat précédent n'assure pas la convergence de la suite des multiplicateurs $\{\lambda_k\}$, alors que l'algorithme 14.36 s'intéresse principalement à cette suite ! En fait, sous les hypothèses de la proposition, il n'y a pas unicité du multiplicateur optimal $\bar{\lambda}$. Par contre, si l'on suppose que les gradients des contraintes actives sont linéairement indépendants, alors on peut montrer la convergence de λ_k vers l'unique multiplicateur optimal $\bar{\lambda}$ (voir l'exercice 14.14).

Les inconvénients de la relaxation lagrangienne sont les suivants.

- La fonction duale est non différentiable en λ si le problème de Lagrange $\inf_x \ell(x, \lambda)$ n'a pas de solution unique. Il faut donc utiliser des algorithmes non différentiables sophistiqués (souvent lents) pour minimiser la fonction duale.
- Le recouvrement de la solution primale par $\inf_x \ell(x, \bar{\lambda})$ ($\bar{\lambda}$ est une solution duale) n'est pas aisé.

14.7.2 Minimisation de la fonction duale régularisée

On considère le problème $(P_{X,EI})$ défini en (14.40) et on étudie le cas où l'approche duale se fait via le lagrangien augmenté associé (14.57).

L'approche duale considéré ici consiste à minimiser la fonction duale δ_r associée au lagrangien augmenté ℓ_r défini en (14.57), fonction duale dont on rappelle ici la définition :

$$\delta_r(\lambda) = -\inf_{x\in X} \ell_r(x,\lambda).$$

Dans la version la plus simple, l'algorithme prend

$$\lambda_+ = \lambda + r\,\tilde{c}_{\lambda,r}(x_\lambda), \tag{14.67}$$

où $\tilde{c}_{\lambda,r}$ est défini en (13.22) et x_λ est un minimiseur de $\ell_r(\cdot,\lambda)$. De manière plus précise, l'algorithme s'écrit comme suit.

Algorithme 14.38 (du lagrangien augmenté)

1. Initialisation : choix d'un multiplicateur initial $\lambda_1 \in \mathbb{R}^m$ et d'un facteur d'augmentation initial $r_1 > 0$;
2. Pour $k = 1, 2, \ldots$ faire :
   2.1. Trouver x_k solution du problème $\min_{x\in X} \ell_{r_k}(x,\lambda_k)$;
   2.2. Si $\tilde{c}_{\lambda_k,r_k}(x_k) \simeq 0$, on s'arrête ;
   2.3. $\lambda_{k+1} = \lambda_k + r_k\,\tilde{c}_{\lambda_k,r_k}(x_k)$.

La proposition suivante établit le fait remarquable selon lequel, pour les problèmes convexes, l'algorithme du lagrangien augmenté ci-dessus se confond avec l'algorithme proximal sur la fonction duale δ_0 [519 ; 1973], un algorithme étudié à la section 7.2.

Proposition 14.39 (interprétation proximale de l'algorithme du LA) *On suppose que f et les $\{c_i\}_{i\in I}$ sont convexes, que c_E est affine, que f et c sont à valeurs réelles et que $r > 0$. Soit $\lambda \in \mathbb{R}^m$. On suppose que $x_\lambda \in X$ réalise l'infimum dans la définition (14.60) de δ_r (c.-à-d., $\delta_r(\lambda) = -\ell_r(x_\lambda,\lambda)$) et que $\lambda_+ \in \mathbb{R}^m$ réalise l'infimum dans (14.61). Alors*

$$\nabla\delta_r(\lambda) = -\nabla_\lambda \ell_r(x_\lambda,\lambda) = -\tilde{c}_{\lambda,r}(x_\lambda), \tag{14.68}$$

où $\tilde{c}_{\lambda,r}(x_\lambda)$ est défini par (13.22). De plus, λ_+ et x_λ sont reliés par

$$\lambda_+ = \lambda + r\,\tilde{c}_{\lambda,r}(x_\lambda) \qquad \text{et} \qquad -\tilde{c}_{\lambda,r}(x_\lambda) \in \partial\delta_0(\lambda_+). \tag{14.69}$$

Démonstration. Soient $\lambda \in \mathbb{R}^m$, x_λ réalisant l'infimum dans la définition (14.60) de δ_r et $\lambda_+ \in \mathbb{R}^m$ réalisant l'infimum dans (14.61). On se rappelle que $\ell_r(x_\lambda,\cdot)$ est dérivable (proposition 13.18) et concave (par son expression (13.19)). Pour tout $\lambda' \in \mathbb{R}^m$ on a

$$\begin{aligned}\delta_r(\lambda') &\geqslant -\ell_r(x_\lambda,\lambda') \qquad \text{[définition (14.60) de } \delta_r\text{]}\\ &\geqslant -\ell_r(x_\lambda,\lambda) - \nabla_\lambda \ell_r(x_\lambda,\lambda)^{\mathsf{T}}(\lambda'-\lambda) \qquad \text{[concavité de } \ell_r(x_\lambda,\cdot)\text{]}\\ &= \delta_r(\lambda) - \nabla_\lambda \ell_r(x_\lambda,\lambda)^{\mathsf{T}}(\lambda'-\lambda).\end{aligned}$$

Comme δ_r est $\mathcal{C}^{1,1}$ par le point 3 de la proposition 14.32, on déduit de l'inégalité ci-dessus que $\nabla\delta_r(\lambda) = -\nabla_\lambda \ell_r(x_\lambda,\lambda) = -\tilde{c}_{\lambda,r}(x_\lambda)$, par (13.25).

Par ailleurs, comme δ_r est la régularisée de Moreau-Yosida de δ_0 pour le produit scalaire $(u,v) \mapsto \frac{1}{r}(u^{\mathsf{T}}v)$, son gradient en λ pour ce produit scalaire s'écrit $\lambda - \lambda_+$ (point (iv) de la proposition 3.80) et celui pour le produit scalaire euclidien est donc $\frac{1}{r}(\lambda - \lambda_+)$, ce qui grâce à (14.68) donne la première relation de (14.69).

Enfin, par définition de λ_+ comme unique minimiseur du problème à droite dans (14.61), il vient $0 \in \partial\delta_0(\lambda_+) + \frac{1}{r}(\lambda_+ - \lambda)$. Comme $\frac{1}{r}(\lambda_+ - \lambda) = \tilde{c}_{\lambda,r}(x_\lambda)$, on obtient la seconde relation de (14.69). □

Concluons en donnant trois interprétations de l'algorithme 14.38.

1. L'algorithme 14.38 est identique à la *méthode des multiplicateurs* introduite à la section 13.4.4 comme heuristique dans une méthode de pénalisation : voir le schéma algorithmique 13.20.
2. La proposition 14.32 nous apprend que, si $(P_{X,EI})$ est convexe et si $\operatorname{dom}\delta_0 \neq \varnothing$, alors δ_r est différentiable et son gradient s'écrit $\nabla\delta_r(\lambda) = -\tilde{c}_{\lambda,r}(x_\lambda)$. Dès lors (14.67) peut déjà être vu comme une méthode de gradient sur δ_r avec le pas $r > 0$.
3. Enfin, la proposition 14.39 affirme que, si $(P_{X,EI})$ est convexe et si $\operatorname{dom}\delta_0 \neq \varnothing$, alors l'algorithme du lagrangien augmenté est identique à l'algorithme proximal pour minimiser δ_0 (section 7.2).

Le dernier point de vue implique que l'algorithme du lagrangien augmenté est monotone sur δ_0, puisque l'on a

$$\delta_r(\lambda) = \delta_0(\lambda_+) + \frac{1}{2r}\|\lambda_+ - \lambda\|^2 \leqslant \delta_0(\lambda).$$

Comme $\lambda_+ - \lambda = r\,\tilde{c}_{\lambda,r}(x_\lambda)$, on peut estimer la décroissance de δ_0 à chaque itération par la norme de $\tilde{c}_{\lambda,r}(x_\lambda)$:

$$\delta_0(\lambda_+) \leqslant \delta_0(\lambda) - \frac{r}{2}\|\tilde{c}_{\lambda,r}(x_\lambda)\|^2.$$

Si δ_0 est bornée inférieurement, cette inégalité implique la convergence de $\tilde{c}_{\lambda,r}(x_\lambda)$ vers zéro. Elle donne aussi une estimation de la vitesse de décroissance de $\|\tilde{c}_{\lambda,r}(x_\lambda)\|$, qui est en $O(r^{-1/2})$, d'où l'intérêt de prendre r grand. Une grande valeur de r rend par ailleurs la minimisation de ℓ_r souvent plus difficile, ce qui nécessite de trouver un compromis.

14.7.3 L'algorithme d'Arrow-Hurwicz

Algorithme 14.40 (Arrow-Hurwicz)

1. Initialisation : choix de $\lambda_1 \in \Lambda$ et de $x_1 \in X$;

2. Pour $k = 1, 2, \ldots$ faire :
   2.1. $x_{k+1} = P_X\left(x_k - \alpha_k^1 \nabla_x \ell(x_k, \lambda_k)\right)$;
   2.2. $\lambda_{k+1} = P_\Lambda\left(\lambda_k + \alpha_k^2 \nabla_\lambda \ell(x_{k+1}, \lambda_k)\right)$;
   2.3. Si (x_{k+1}, λ_{k+1}) est satisfaisant on s'arrête.

Cet algorithme est très semblable à l'algorithme 14.36 mais on ne fait plus qu'un seul pas de minimisation en x. Dans le cas convexe, on montre que l'on a convergence de la méthode si α_k^1 et α_k^2 sont pris assez petits.

Sans hypothèse de forte convexité-concavité, l'algorithme d'Arrow-Hurwicz a peu de chance de converger. En effet, si l'on cherche à trouver l'unique point-selle $(0,0) \in \mathbb{E}^2$ de la fonction $(x, y) \in \mathbb{E}^2 \mapsto \varphi(x, y) = \langle x, y\rangle$, définie sur un espace euclidien $\mathbb{E}$ de produit scalaire $\langle\cdot,\cdot\rangle$, l'algorithme d'Arrow-Hurwicz avec pas infinitésimal devient l'équation différentielle suivante

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -\alpha\, y \qquad \text{et} \qquad \frac{\mathrm{d}y}{\mathrm{d}t} = \alpha\, x$$

où $\alpha > 0$. Comme $(\|x\|^2 + \|y\|^2)' = 2(\langle x, x'\rangle + \langle y, y'\rangle) = 0$, les trajectoires sont sur des sphères et ne convergent donc pas vers l'unique point-selle [17 ; 1992].

Notes

La dualité min-max (section 14.1) remonte au moins à l'*identité du minimax de von Neumann* [612 ; 1928] (exercice 17.12), qui considère le cas où φ est bilinéaire et les ensembles X et Y sont des simplexes unitaires de dimension finie. Selon [614], ce résultat avait été obtenu en dimensioin deux par E. Borel en 1922. Le théorème 14.6 sur l'existence de point-selle est une généralisation de cette identité (voir Sion [554 ; 1958] pour une version plus générale, démontrée simplement par Komiya [363 ; 1988], et Brézis [90 ; 1973] dont nous avons suivi la démonstration). La dualité par perturbation (section 14.2) a été introduite par Rockafellar [516, 517 ; 1969-1970]. La dualité de Fenchel (section 14.3) a débuté avec le travail de Fenchel [203, 204 ; 1949-1951], puis étendue par Rockafellar [517, 522 ; 1970-1974]. La section 14.4 présentant la dualité de Toland est adaptée de [593 ; 1978].

Sur la dualité en général, on pourra consulter l'ouvrage d'Hiriart-Urruty et Lemaréchal [317 ; chapitre XII] (problèmes convexes) et les monographies de Rockafellar [522], de Walk [614] et Goh et Yang [256] qui considèrent aussi les inéquations variationnelles. La théorie est présentée en dimension infinie par Laurent [382 ; 1972] et Ekeland et Temam [191 ; 1974].

Nous l'avons déjà mentionné au chapitre 13 : le lagrangien augmenté (14.57), identique à (13.21), a été proposé par Rockafellar [518, 519 ; 1971-1973]. Nous avons suivi ici l'approche par perturbation de [519 ; 1973]. La proposition 14.31 est en partie due à Arrow, Gould et Howe [23 ; 1973] qui ont montré que, sous les conditions énoncées, $(\bar{x}, \bar{\lambda})$ est point-selle de $\ell_r(x, \lambda)$ sur $B(\bar{x}, \varepsilon) \times (\mathbb{R}^{m_E} \times \mathbb{R}_+^{m_I})$; la restriction $\lambda_I \geqslant 0$ n'est pas nécessaire comme l'avait déjà observé Rockafellar [519 ; 1973] dans le cas convexe. Comme autre contribution intéressante, citons [420 ; 2008].

Les algorithmes fondés sur la dualité n'ont été ici qu'esquissés. Pour leur utilisation dans la résolution de problèmes quadratiques, on pourra consulter la revue de Lin et Pang [397 ; 1987], [161 ; 2005] (avec le lagrangien augmenté). Pour des problèmes plus généraux, la non-différentiabilité de la fonction duale, lorsqu'elle est présente, complique évidemment beaucoup cette approche et il faut alors recourir à des algorithmes adaptés. La *méthode de faisceaux* en est un exemple ; elle est étudiée en détail dans [317 ; 1993, chapitres XIV et XV]. L'utilisation de la dualité lagrangienne pour résoudre des problèmes variés est passée en revue par Lemaréchal [387 ; 2001]. Une autre possibilité est d'utiliser une fonction de couplage φ autre que le lagrangien, de manière à rendre la fonction duale différentiable. On l'a vu, le lagrangien augmenté a cette propriété (voir aussi [520 ; 1973]). D'autres fonctions partageant cette propriété et ayant divers avantages sont étudiées dans le livre de Golshtein et Tretyakov [266 ; 1996, section 2.5] et dans les articles d'Auslender, Ben-Tiba et Teboulle [30 ; 1999] et d'Auslender et Teboulle [28 ; 2000].

La présentation de la décomposition de Benders à la section **??**, qui remonte à [45 ; 1962], synthétise assez fidèlement [237 ; 1972].

Les techniques de dualité s'utilisent aussi pour résoudre des problèmes de grande taille décomposables. Par exemple l'algorithme du *recouvrement progressif* [527 ; 1991] est adapté aux problèmes d'*optimisation stochastique* discrétisés sur des *arbres de scénarios*.

Exercices

14.1. *Dualité min-max avec Y fini* (adapté de [266 ; lemme 4.5]). On considère le problème min-max (14.2) avec un ensemble Y fini, que l'on prend égal à $[1:p]$:

$$\inf_{x\in X} \max_{i\in[1:p]} f_i(x),$$

où X est un ensemble et les $f_i : X \to \mathbb{R} \cup \{+\infty\}$, $i \in [1:p]$, sont des fonctions propres telles que $D := \cap_{i\in[1:p]} \operatorname{dom} f_i \neq \varnothing$. On sait que l'inversion de l'infimum et du maximum dans ce problème donne le plus souvent lieu à un saut de dualité (voir l'exemple de la figure 14.1). Cependant, si l'on introduit une combinaison convexe, on évite le saut de dualité, dans le sens suivant :

$$\inf_{x\in X} \max_{i\in[1:p]} f_i(x) = \sup_{\alpha\in\Delta_p} \inf_{x\in X} \sum_{\substack{i\in[1:p]\\ \alpha_i\neq 0}} \alpha_i f_i(x), \tag{14.70}$$

où $\Delta_p := \{\alpha \in \mathbb{R}^p_+ : \sum_{i=1}^p \alpha_i = 1\}$ est le simplexe unité de $\mathbb{R}^p$.

14.2. *Dualité forte pour des fonctions séparables*. Soient $p \in \mathbb{N}$, $X := X_1 \times \cdots \times X_p$ le produit cartésien d'ensembles non vides X_i et $\varphi_i : X_i \to \mathbb{R}$, pour $i \in [1:p]$, des fonctions arbitraires. Un $x \in X$ est écrit $x = (x_1, \ldots, x_p)$, avec des $x_i \in X_i$ pour $i \in [1:p]$. Montrez la relation de dualité forte (comparez avec l'exemple de gauche à la figure 14.1 dans lequel il n'y a pas séparabilité et saut de dualité)

$$\inf_{x\in X} \max_{i\in[1:p]} \varphi_i(x_i) = \max_{i\in[1:p]} \inf_{x_i\in X_i} \varphi_i(x_i). \tag{14.71}$$

14.3. *Suite minimaximisante et saut de dualité* [417, 110]. Soient X et Y deux ensembles et $\varphi : X \times Y \to \overline{\mathbb{R}}$ une fonction. On dit que la suite $\{(x_k, y_k)\}_k \subseteq X \times Y$ est *minimaximisante* pour φ sur $X \times Y$ s'il existe une suite $\{\varepsilon_k\} \downarrow 0$ telle que

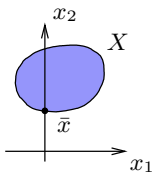
$$\forall\,(x,y)\in X\times Y:\quad \varphi(x_k,y)\leqslant \varepsilon_k+\varphi(x,y_k). \tag{14.72}$$

1) Il existe une suite minimaximisante pour φ si, et seulement si, il n'y a pas de saut de dualité, c'est-à-dire si l'on a l'égalité (14.7).
2) Si $\{(x_k,y_k)\}$ est une suite minimaximisante pour φ, alors $\varphi(x_k,y_k)$ converge vers $\sup_y \inf_x \varphi(x,y)=\inf_x \sup_y \varphi(x,y)$.

14.4. *Dualité de Fenchel et lagrangienne.* Retrouver le dual de Fenchel (D_F) dans (14.27) du problème (P_F) dans (14.25) par la dualité lagrangienne.

14.5. *Dualisation d'un problème dans* $\mathbb{R}^2$ [54]. Soit X une partie quelconque de $\mathbb{R}^2$. On considère le problème :

$$(P)\quad \begin{cases} \inf\ x_2 \\ x\in X \\ x_1=0. \end{cases}$$



1) En dualisant la contrainte d'égalité par le lagrangien ordinaire, montrez que le problème dual consiste à trouver la droite de $\mathbb{R}^2$ qui est en-dessous de X et qui rencontre l'axe des ordonnées le plus haut possible.
2) Montrez que si X est un convexe fermé, si $\{x_2\in\mathbb{R}\colon (0,x_2)\in X\}$ est borné inférieurement et si X a un point d'abscisse <0 et un point d'abscisse >0, alors le lagrangien a un point-selle.
3) En choisissant des ensembles X particuliers, montrez que l'on peut rencontrer les situations suivantes.
   (a) Il y a un saut de dualité et les solutions du problème interne $\inf_{x\in X}(\bar\lambda x_1+x_2)$, où $\bar\lambda$ est une solution duale, ne sont pas solutions de (P).
   (b) Le lagrangien a un point-selle, mais certaines solutions du problème interne $\inf_{x\in X}(\bar\lambda x_1+x_2)$, où $\bar\lambda$ est une solution duale, ne sont pas solutions de (P).
   (c) Tout point de $\mathbb{R}$ est solution duale.
   (d) Le problème primal a une (ou n'a pas de) solution, la valeur optimale primale est finie, le problème dual n'a pas de solution et il n'y a pas de saut de dualité.
4) Dualisez la contrainte « $x_1=0$ » par le lagrangien augmenté et donnez une interprétation géométrique du problème dual. Donnez un exemple d'ensemble X pour lequel il n'y a pas de saut de dualité avec cette dualisation par le lagrangien augmenté, alors qu'il y en aurait un avec la dualisation par le lagrangien ordinaire.

Remarque. Cet exercice décrit bien ce qui peut se passer lors de la dualisation de contraintes d'égalité. C'est alors l'épigraphe de la fonction valeur primale qui tient lieu d'ensemble X.

14.6. *Dualisation lagrangienne de problèmes classiques.* Écrire le dual lagrangien des problèmes suivants.

1) Le problème d'optimisation linéaire en $x\in\mathbb{R}^n$:

$$\begin{cases} \inf\ c^{\mathsf T}x \\ Ax=b \\ x\geqslant 0, \end{cases}$$

où $c\in\mathbb{R}^n$, A est une matrice $m\times n$ et $b\in\mathbb{R}^m$.
2) Un dual du dual du problème linéaire défini au point 1.

3) Le problème généralisant le problème linéaire défini au point 1 : on se donne deux espaces euclidiens $\mathbb{E}$ et $\mathbb{F}$, $c \in \mathbb{E}$, un cône non vide K de $\mathbb{E}$, $b \in \mathbb{F}$, une application linéaire $A : \mathbb{E} \to \mathbb{F}$ et on considère le problème $\inf\{\langle c, x\rangle : x \in \mathbb{R}^n,\ Ax = b,\ x \in K\}$. On dualisera soit la contrainte d'égalité, soit les deux contraintes (on suppose alors que K est aussi convexe et fermé), pour retrouver le même problème dual.
4) Les problèmes d'optimisation quadratique

$$\inf\{g^{\mathsf{T}}x + \tfrac{1}{2}x^{\mathsf{T}}Hx : x \in \mathbb{R}^n,\ Ax = b\},$$
$$\inf\{g^{\mathsf{T}}x + \tfrac{1}{2}x^{\mathsf{T}}Hx : x \in \mathbb{R}^n,\ Ax = b,\ x \geqslant 0\},$$

où $g \in \mathbb{R}^n$, H est une matrice d'ordre n symétrique semi-définie positive, A est une matrice $m \times n$ et $b \in \mathbb{R}^m$.

14.7. *Dualisation lagrangienne de la projection sur un polyèdre convexe* (inspiré de [394]). On considère le problème de la projection sur un polyèdre convexe, qui peut s'écrire de la manière suivante

$$\begin{cases} \inf_{x\in\mathbb{R}^n} \ \frac{1}{2}\|x - z\|^2 \\ Ax = b \\ x \geqslant 0, \end{cases} \tag{14.73}$$

où $z \in \mathbb{R}^n$ est donné, $\|\cdot\|$ est la norme euclidienne, $A \in \mathbb{R}^{m\times n}$ et $b \in \mathbb{R}^m$. On suppose que l'ensemble admissible du problème est non vide : il existe un $x_0 \geqslant 0$ tel que $Ax_0 = b$.

1) Dans quel sens peut on dire que le dual lagrangien du problème (14.73) est le problème suivant:

$$\sup_{y\in\mathbb{R}^m} \left(-\frac{1}{2}\|(A^{\mathsf{T}}y + z)^+\|^2 + b^{\mathsf{T}}y + \frac{1}{2}\|z\|^2\right). \tag{14.74}$$

2) Montrez qu'il n'y a pas de saut de dualité entre (14.73) et (14.74).
3) Montrez que les propriétés suivantes sont équivalentes :
   (a) l'ensemble des solutions de (14.74) est borné,
   (b) $\forall\, p \neq 0$ tel que $A^{\mathsf{T}}p \leqslant 0$, on a $b^{\mathsf{T}}p < 0$,
   (c) A est surjective et il existe un $x > 0$ tel que $Ax = b$.
4) Le but de ce numéro est d'examiner un dual de (14.73) de la forme (14.74) mais avec y restreint à l'image de A.
   4.1) Montrez que l'on peut aussi prendre comme dual de (14.73) le problème

$$\sup_{y\in\mathcal{R}(A)} \left(-\frac{1}{2}\|(A^{\mathsf{T}}y + z)^+\|^2 + b^{\mathsf{T}}y + \frac{1}{2}\|z\|^2\right). \tag{14.75}$$

   4.2) Montrez qu'il n'y a pas non plus de saut de dualité entre (14.73) et (14.75).
   4.3) Montrez que les propriétés suivantes sont équivalentes :
      (a) l'ensemble des solutions de (14.75) est borné,
      (b) $\forall\, p \in \mathcal{R}(A) \setminus \{0\}$ tel que $A^{\mathsf{T}}p \leqslant 0$, on a $b^{\mathsf{T}}p < 0$,
      (c) il existe un $x > 0$ tel que $Ax = b$.

14.8. *Dualité lagrangienne en optimisation conique.* Soient $\mathbb{E}$ et $\mathbb{F}$ deux espaces euclidiens, dont on note $\langle\cdot,\cdot\rangle$ les produits scalaires. Soit K un cône non vide de $\mathbb{E}$, $c \in \mathbb{E}$, $b \in \mathbb{F}$ et A une application linéaire de $\mathbb{E}$ dans $\mathbb{F}$. On note K^+ le cône dual de K et $A^* : \mathbb{F} \to \mathbb{E}$ l'application linéaire adjointe de A. On considère le problème en $x \in \mathbb{E}$ suivant :

$$\begin{cases} \inf \langle c, x\rangle \\ Ax = b \\ x \in K. \end{cases} \tag{14.76}$$

1) *Dualisation de la contrainte d'égalité.* En dualisant la contrainte d'égalité avec le lagrangien classique $\ell(x,y) = \langle c,x\rangle - \langle y, Ax-b\rangle$, montrez que l'on obtient comme problème dual

$$\begin{cases} \sup \langle b,y\rangle \\ A^*y + s = c \\ s \in K^+, \end{cases} \tag{14.77}$$

où $K^+ := \{s \in \mathbb{R}^n : \langle s,x\rangle \geqslant 0,\ \forall x \in K\}$ est le cône dual de K.

2) *Dualisation de la contrainte d'appartenance au cône.* Supposons que K soit un cône convexe fermé non vide. Montrez que l'on retrouve le dual (14.77) si l'on dualise la contrainte linéaire *et* la contrainte d'appartenance au cône.

3) On suppose à présent que $\mathbb{E} = \mathbb{R}^3$, $\mathbb{F} = \mathbb{R}$, $A(x) = x_1 - 1$ et $K = \mathbb{R}^3_\triangledown := \{x \in \mathbb{R}^3 : x_1^2 + x_2^2 \leqslant x_3^2,\ x_3 \geqslant 0\}$ (voir l'exercice 2.35). Montrez que dans ce cas (14.76) n'a pas de solution, que son dual (14.77) en a une et qu'il n'y a pas de saut de dualité [512].

14.9. *Dualisation lagrangienne du problème de région de confiance.* On considère le problème de région de confiance (9.8), c'est-à-dire $\min\{g^\mathsf{T}x + \frac{1}{2}x^\mathsf{T}Hx : x \in \mathbb{R}^n,\ \|x\|_2 \leqslant \Delta\}$, dans lequel $g \in \mathbb{R}^n$, la matrice symétrique $H \in \mathcal{S}^n$ n'est pas nécessairement semi-définie positive (le problème peut être non convexe) et le rayon de confiance $\Delta > 0$. Écrivez le dual lagrangien de ce problème et montrez que

$$\min_{\|x\|\leqslant\Delta} \left(g^\mathsf{T}x + \frac{1}{2}x^\mathsf{T}Hx\right) = \max_{\substack{\lambda\geqslant 0 \\ H+\lambda I\succcurlyeq 0 \\ g\in\mathcal{R}(H+\lambda I)}} -\left(g^\mathsf{T}(H+\lambda I)^\dagger g + \frac{\Delta}{2}\lambda\right). \tag{14.78}$$

Observez que cette identité n'est plus nécessairement vraie si $\Delta = 0$.

14.10. *Dualisation wolfienne d'un problème convexe différentiable* [622]. On considère les problèmes (P) et (D) suivants :

$$(P)\quad \begin{cases} \inf_{x\in\mathbb{R}^n} f(x) \\ c(x) \leqslant 0 \end{cases} \qquad \text{et} \qquad (D)\quad \begin{cases} \sup_{(x,\lambda)\in\mathbb{R}^n\times\mathbb{R}^m} \ell(x,\lambda) \\ \lambda \geqslant 0 \\ \nabla_x \ell(x,\lambda) = 0, \end{cases}$$

dans lesquels $f : \mathbb{R}^n \to \mathbb{R}$ est une fonction convexe différentiable, les composantes de $c : \mathbb{R}^n \to \mathbb{R}^m$ sont convexes différentiables et $\ell : \mathbb{R}^n \times \mathbb{R}^m : (x,\lambda) \mapsto \ell(x,\lambda) = f(x) + \lambda^\mathsf{T}c(x)$ est le lagrangien du problème (P). On note $\mathrm{val}(P)$ et $\mathrm{val}(D)$ les valeurs optimales de (P) et de (D), respectivement.

1) *Dualité faible.* Montrez que $\mathrm{val}(D) \leqslant \mathrm{val}(P)$.
2) *Dualité forte I.* Montrez que si la contrainte de (P) est qualifiée au sens de la définition 4.30 en une solution x_* de (P), alors
   - il existe $\lambda_* \in \mathbb{R}^m$ tel que (x_*, λ_*) est solution de (D),
   - $\mathrm{val}(D) = \mathrm{val}(P)$.
3) *Dualité forte II.* Montrez que si la contrainte c de (P) est affine mais non réalisable (donc $\mathrm{val}(P) = +\infty$) et si les contraintes de (D) sont réalisables (donc $\mathrm{val}(D) > -\infty$), alors $\mathrm{val}(D) = +\infty$.
4) *Certificat d'inconsistance* [82]. Montrez que si l'on peut trouver un couple $(x,\lambda) \in \mathbb{R}^n \times \mathbb{R}^m$ tel que $\lambda \geqslant 0$, $\lambda^\mathsf{T}c(x) > 0$ et $c'(x)^\mathsf{T}\lambda = 0$, alors il n'existe pas de x tel que $c(x) \leqslant 0$.

Remarques.

1) Le dual de Wolfe (D) se distingue du dual de Lagrange (14.65) par un critère à minimiser explicite (alors que la fonction duale δ peut n'être connue que

numériquement et évaluable algorithmiquement, pas analytiquement avec évaluation rapide), mais avec une contrainte d'égalité non linéaire « $\nabla_x \ell(x,\lambda) = 0$ », qui peut être compliquée à prendre en compte numériquement (elle est non linéaire, non convexe, et sa dérivée fait intervenir les dérivées secondes de f et c). Il se peut d'ailleurs que (P) soit plus simple à résoudre numériquement que (D).

2) Cette dualisation permet d'avoir une *borne inférieure* sur $\mathrm{val}(P)$, puisqu'elle nous apprend que, *pour les problèmes convexes*, $\ell(x,\lambda)$ est une telle borne inférieure lorsque $\lambda \geqslant 0$ et $\nabla_x \ell(x,\lambda) = 0$.
3) En présence de contraintes de borne additionnelles dans (P), on peut éliminer la contrainte d'égalité du problème dual et obtenir ainsi une borne inférieure sur la valeur optimale primal par simple évaluation de l'objectif dual. Cette borne inférieure de $\mathrm{val}(P)$ peut toutefois aussi s'obtenir en utilisant les bornes sur x et des minorantes affines exactes de f. [82]

14.11. *Dualisation lagrangienne d'un problème homogène* (inspiré de [378]). On rappelle qu'une fonction $\varphi : \mathbb{R}^n \to \mathbb{R}^m$ est dite *positivement homogène* de degré $p \in \mathbb{R}$ si pour tout $x \in \mathbb{R}^n$ et pour tout réel $t \geqslant 0$, on a $\varphi(tx) = t^p \varphi(x)$. On considère le problème d'optimisation en $x \in \mathbb{R}^n$ suivant :

$$(P) \quad \begin{cases} \inf\ f(x) \\ c(x) \leqslant b, \end{cases}$$

où $b \in \mathbb{R}^m$ et les fonctions $f : \mathbb{R}^n \to \mathbb{R}$ et $c : \mathbb{R}^n \to \mathbb{R}^m$ sont différentiables et homogènes de degré $p > 0$ (le même degré strictement positif pour les deux fonctions). On note $(x,\lambda) \in \mathbb{R}^n \times \mathbb{R}^m \mapsto \ell(x,\lambda) := f(x) + \lambda^{\mathsf{T}}(c(x) - b)$ le lagrangien du problème. Montrez que le dual lagrangien de (P) est le problème en $\lambda \in \mathbb{R}^m$ *convexe* suivant :

$$(D) \quad \begin{cases} \sup\ -b^{\mathsf{T}}\lambda \\ \lambda \geqslant 0 \\ \ell_h(x,\lambda) \geqslant 0, \quad \forall x \in \mathbb{R}^n, \end{cases}$$

où $(x,\lambda) \mapsto \ell_h(x,\lambda) = f(x) + \lambda^{\mathsf{T}} c(x)$ est la partie homogène en x de ℓ.

14.12. *Théorème d'Everett.* On considère le problème

$$(P_{EI}) \quad \begin{cases} \min\ f(x) \\ c_E(x) = 0 \\ c_I(x) \leqslant 0, \end{cases}$$

où les fonctions $f : \mathbb{R}^n \to \mathbb{R}$, $c_E : \mathbb{R}^n \to \mathbb{R}^{m_E}$ et $c_I : \mathbb{R}^n \to \mathbb{R}^{m_I}$ n'ont pas de propriété particulière (de différentiabilité ou de convexité). On note $c(x) := (c_E(x), c_I(x)) \in \mathbb{R}^m$ $(m := m_E + m_I)$, $(x,\lambda) \mapsto \ell(x,\lambda) = f(x) + \lambda^{\mathsf{T}} c(x)$ le lagrangien,

$$\lambda \mapsto \delta(\lambda) := -\inf_{x \in \mathbb{R}^n} \ell(x,\lambda) \tag{14.79}$$

la fonction duale et $\inf\{\delta(\lambda) : \lambda \in \mathbb{R}^m,\ \lambda_I \geqslant 0\}$ le problème dual. Le problème à droite dans (14.79) est appelé problème de Lagrange. Démontrez les affirmations suivantes.

1) Si $\lambda_I \geqslant 0$ et si x_λ est solution du problème de Lagrange, alors x_λ est solution du problème (P_{EI}) perturbé suivant

$$\begin{cases} \min\ f(x) \\ c_E(x) = c_E(x_\lambda) \\ c_i(x) \leqslant c_i(x_\lambda), \quad \forall i \in I \text{ tel que } \lambda_i > 0. \end{cases} \tag{14.80}$$

2) Si $\bar{x}$ est solution du problème de Lagrange avec un $\lambda = \bar{\lambda}$ et si $c_E(\bar{x}) = 0$ et $0 \leqslant \bar{\lambda}_I \perp (-c_I(\bar{x})) \geqslant 0$, alors $\bar{x}$ est solution de (P_{EI}).

14.13. *La distance à un ensemble convexe est une valeur-selle* [98]. Soient $\mathbb{E}$ un espace euclidien dont le produit scalaire est noté $\langle \cdot, \cdot \rangle$, $\| \cdot \|$ une norme sur $\mathbb{E}$ qui n'est pas nécessairement celle associée au produit scalaire et C un convexe fermé non vide de $\mathbb{E}$. On note d_C la distance à C pour la norme $\| \cdot \|$ et $\bar{B}_{\mathrm{D}}$ la boule-unité fermée pour la norme duale. Montrez que

1) pour tout $x \in \mathbb{E}$, on a

$$\mathrm{d}_C = \sup_{x^* \in \bar{B}_{\mathrm{D}}} \inf_{y \in C} \langle x^*, x - y \rangle = \inf_{y \in C} \sup_{x^* \in \bar{B}_{\mathrm{D}}} \langle x^*, x - y \rangle,$$

2) l'ensemble des solutions du problème sup-inf ci-dessus est $\partial\, \mathrm{d}_C(x)$ et l'ensemble des solutions du problème inf-sup ci-dessus est $\arg\min\{\|x - y\| : y \in C\}$.

14.14. *Convergence des multiplicateurs dans l'algorithme 14.36.* On reprend le cadre défini par la proposition 14.37 en supposant en plus que les gradients des contraintes actives en $\bar{x}$ sont linéairement indépendants. Alors il y a un unique point-selle $(\bar{x}, \bar{\lambda})$ du lagrangien et la suite $\{\lambda_k\}$ générée par l'algorithme 14.36 converge vers $\bar{\lambda}$.

14.15. *Norme nucléaire et rang* [509]. On considère l'espace vectoriel $\mathbb{R}^{m \times n}$ des matrices réelles de type $m \times n$, muni de différentes normes :

- la norme de Frobenius (B.9), notée $\| \cdot \|_F$, qui est associée au produit scalaire (B.8), noté $\langle \cdot, \cdot \rangle$,
- la norme d'opérateur de $(\mathbb{R}^n, \| \cdot \|_2)$ dans $(\mathbb{R}^m, \| \cdot \|_2)$, où $\| \cdot \|_2$ est la norme euclidienne, dont la valeur en $A \in \mathbb{R}^{m \times n}$ est notée ici

$$\|A\| := \sup_{\|x\|_2 \leqslant 1} \|Ax\|_2,$$

- la *norme nucléaire*, qui est la norme duale de la norme d'opérateur pour le produit scalaire (B.8) de $\mathbb{R}^{m \times n}$ et dont la valeur en $A \in \mathbb{R}^{m \times n}$ est notée et définie par

$$\|A\|_* := \sup_{\|B\| \leqslant 1} \langle A, B \rangle. \tag{14.81}$$

On note $\sigma(A)$ le vecteur formé des valeurs singulières de A (section B.5.4), rangées par ordre décroissant : $\sigma_1(A) \geqslant \sigma_2(A) \geqslant \cdots \geqslant 0$.

1) Montrez que

$$\|A\| = \|\sigma(A)\|_\infty, \qquad \|A\|_F = \|\sigma(A)\|_2 \qquad \text{et} \qquad \|A\|_* = \|\sigma(A)\|_1. \tag{14.82}$$

2) Montrez que

$$\|A\| \leqslant \|A\|_F \leqslant \|A\|_* \leqslant \mathrm{rg}(A)^{1/2} \|A\|_F \leqslant \mathrm{rg}(A) \|A\|. \tag{14.83}$$

L'inégalité $\|A\|_* \leqslant \mathrm{rg}(A)\|A\|$ du point 2 montre que le rang de A est minoré par $\|A\|_*$ sur la boule unité $\mathcal{B} := \{A \in \mathbb{R}^{m \times n} : \|A\| \leqslant 1\}$. Nous allons montrer que $\| \cdot \|_*$ est en réalité la plus grande fonction convexe fermée minorant le rang sur $\mathcal{B}$. On est obligé de restreindre le rang à une partie de $\mathbb{R}^{m \times n}$, car sinon sa biconjuguée est nulle et donc de peu d'intérêt.

3) Montrez que $\mathrm{rg}^{**} = 0$.

On introduit donc la fonction $f = \mathrm{rg} + \mathcal{I}_{\mathcal{B}} : \mathbb{R}^{m \times n} \to \overline{\mathbb{R}}$, où $\mathcal{I}_{\mathcal{B}}$ est l'indicatrice de $\mathcal{B}$.

4) Montrez que $f^*(A^*) = \|(\sigma(A^*) - e)^+\|_1$.

5) Montrez que $f^{**} = \| \cdot \|_* + \mathcal{I}_{\mathcal{B}}$.

14.16. *Poursuite de base ou recouvrement ℓ_1* [117]. La *poursuite de base* ou problème de *recouvrement ℓ_1* consiste à recouvrer un vecteur $x \in \mathbb{R}^n$ satisfaisant une contrainte affine et de *norme ℓ_1* minimale. Il s'écrit donc

$$\begin{cases} \inf_{x\in\mathbb{R}^n} \ \|x\|_1 \\ Ax = b, \end{cases} \tag{14.84}$$

où $A : \mathbb{R}^n \to \mathbb{R}^m$ est linéaire et $b \in \mathbb{R}^m$. On munit $\mathbb{R}^n$ et $\mathbb{R}^m$ du produit scalaire euclidien. Montrez que le dual lagrangien de ce problème s'écrit

$$\begin{cases} \sup_{y\in\mathbb{R}^m} \ b^{\mathsf{T}} y \\ \|A^{\mathsf{T}} y\|_\infty \leqslant 1. \end{cases} \tag{14.85}$$

14.17. *Recouvrement nucléaire* [114]. Le problème de *recouvrement nucléaire* consiste à recouvrer une matrice $X \in \mathbb{R}^{m\times n}$ satisfaisant une contrainte affine et de *norme nucléaire* minimale. Il s'écrit donc

$$\begin{cases} \inf_{X\in\mathbb{R}^{m\times n}} \ \|X\|_* \\ \mathcal{A}(X) = b, \end{cases} \tag{14.86}$$

où $\|\cdot\|_*$ est la norme nucléaire (14.81), $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^p$ est linéaire et $b \in \mathbb{R}^p$. Montrez que le dual lagrangien de ce problème s'écrit

$$\begin{cases} \sup_{y\in\mathbb{R}^p} \ \langle b, y\rangle \\ \|\mathcal{A}^*(y)\| \leqslant 1, \end{cases} \tag{14.87}$$

où $\mathcal{A}^*$ est l'adjointe de $\mathcal{A}$ lorsque $\mathbb{R}^{m\times n}$ est muni du produit scalaire (B.8) et $\mathbb{R}^p$ est muni du produit scalaire aussi noté $\langle\cdot,\cdot\rangle$ dans (14.87).

15 Optimisation quadratique successive

Dans ce chapitre, nous étudions une famille d'*algorithmes newtoniens*, qui procèdent donc par linéarisation de fonctions, pour résoudre les problèmes d'optimisation avec contraintes d'égalité et d'inégalité. Les algorithmes requièrent la résolution de problèmes d'optimisation quadratique à chaque itération, de problèmes non linéaires donc, mais qui sont plus simples que le problème original. On s'écarte ainsi des méthodes de Newton en résolution de systèmes d'équations différentiables ou de problèmes d'optimisation, décritent au chapitre 10, qui ne requièrent que la résolution d'un système linéaire à chaque itération.

L'approche étudiée est suffisamment générale pour pouvoir considérer les problèmes qui s'écrivent sous la forme suivante

$$(P_{EI}) \quad \begin{cases} \min\ f(x) \\ c_i(x) = 0, & i \in E \\ c_i(x) \leqslant 0, & i \in I, \end{cases}$$

où les ensembles d'indices E et I forment une partition de $[1:m] = E \cup I$ ($E \cap I = \varnothing$). Les paramètres x à optimiser sont dans un espace euclidien $\mathbb{E}$ et les fonctions $f : \mathbb{E} \to \mathbb{R}$ et $c_i : \mathbb{E} \to \mathbb{R}$ définissant le critère et les contraintes sont supposées régulières. Les versions quasi-newtoniennes seront aussi étudiées (section 15.4).

Les notations sont celles déjà introduites à la section 4. On rassemble les contraintes d'égalité et d'inégalité en une seule fonction $c : \mathbb{E} \to \mathbb{R}^m$. Si $v \in \mathbb{R}^m$, on note v_E (resp. v_I) le vecteur de $\mathbb{R}^{|E|}$ (resp. $\mathbb{R}^{|I|}$) formé des composantes v_i de v avec $i \in E$ (resp. $i \in I$). Les fonctions définissant les contraintes d'égalité et d'inégalité seront donc notées c_E et c_I, respectivement. À un vecteur $v \in \mathbb{R}^m$, on associe le vecteur $v^\# \in \mathbb{R}^m$, défini par

$$(v^\#)_i = \begin{cases} v_i & \text{si } i \in E \\ v_i^+ & \text{si } i \in I, \end{cases}$$

où $v_i^+ = \max(0, v_i)$. Avec cette notation, les contraintes de (P_{EI}) s'écrivent $c(x)^\# = 0$, qui n'a d'intérêt que par sa compacité, car la fonction $x \mapsto c(x)^\#$ est en général non différentiable.

Si la présence de contraintes d'égalité dans (P_{EI}) rend ce problème plus difficile à résoudre qu'un problème sans contrainte. Le saut de complexité apporté par la présence des contraintes d'inégalité est incomparablement plus important que celui dû aux contraintes d'égalité.

15.1 L'algorithme OQS et sa convergence locale

Nous présentons l'algorithme OQS comme un algorithme de Josephy-Newton sur le système d'optimalité, lequel peut en effet s'écrire comme un problème d'inclusion, et même de complémentarité non linéaire (section 15.1.1). Sa convergence locale peut alors se déduire de celle de l'algorithme de Josephy-Newton (théorème **??**), pourvu que le point stationnaire considéré soit semi-stable et hémi-stable. On montre que c'est effectivement le cas lorsque celui-ci correspond à un minimum local de (P_{EI}) vérifiant les conditions d'optimalité du second ordre semi-fortes (section 15.1.2).

15.1.1 L'algorithme OQS

Par définition, un couple stationnaire de (P_{EI}) est un couple (x, λ) qui vérifie le système d'optimalité (4.33), c'est-à-dire

$$\begin{cases} \nabla f(x) + c'(x)^*\lambda = 0 \\ c_E(x) = 0 \\ 0 \leqslant \lambda \perp c_I(x) \leqslant 0. \end{cases} \tag{15.1}$$

On peut récrire ce système comme l'inclusion ou le problème de complémentarité en $z := (x, \lambda)$ ci-dessous

$$F(z) + \mathrm{N}_K(z) \ni 0 \qquad \text{ou} \qquad K \ni z \perp F(z) \in K^+, \tag{15.2a}$$

avec

$$F(z) = \begin{pmatrix} \nabla f(x) + c'(x)^*\lambda \\ -c(x) \end{pmatrix} \qquad \text{et} \qquad K = \mathbb{E} \times (\mathbb{R}^{m_E} \times \mathbb{R}^{m_I}_+). \tag{15.2b}$$

En effet, alors $K^+ = \{0_{\mathbb{E}}\} \times (\{0_{\mathbb{R}^{m_E}}\} \times \mathbb{R}^{m_I}_+)$ et

- $(x, \lambda) \in K$ s'écrit $\lambda_I \geqslant 0$,
- $F(x, \lambda) \in K^+$ s'écrit $\nabla_x \ell(x, \lambda) = 0$, $c_E(x) = 0$ et $c_I(x) \leqslant 0$,
- $(x, \lambda) \perp F(x, \lambda)$ se ramène alors à la relation de complémentarité $\lambda_I \perp c_I(x)$.

On retrouve donc bien (15.1).

L'algorithme de Josephy-Newton (**??**) appliqué à l'inclusion ou au problème de complémentarité (15.2) consiste à déterminer l'itéré suivant $z_{k+1} := (x_{k+1}, \lambda_{k+1})$ à partir de l'itéré courant $z_k := (x_k, \lambda_k)$ en résolvant l'inclusion linéarisée en $z := (x, \lambda)$ ci-dessous

$$F(z_k) + F'(z_k)(z - z_k) + \mathrm{N}_K(z) \ni 0 \tag{15.3a}$$

ou son problème de complémentarité équivalent

$$K \ni z \perp \big(F(z_k) + F'(z_k)(z - z_k)\big) \in K^+. \tag{15.3b}$$

Voyons comment s'écrit cet algorithme lorsque (F, K) est donné par (15.2b). Observons que

$$F'(x, \lambda) = \begin{pmatrix} L(x, \lambda) & c'(x)^* \\ -c'(x) & 0 \end{pmatrix}, \tag{15.4}$$

où l'on a simplifié l'écriture en introduisant

$$L(x,\lambda) := \nabla^2_{xx}\ell(x,\lambda).$$

La condition $(x,\lambda) \in K$ s'écrit comme précédemment

$$\lambda_I \geqslant 0. \tag{15.5a}$$

La condition $F(z_k) + F'(z_k)(z - z_k) \in K^+$ se traduit par

$$\nabla f(x_k) + c'(x_k)^*\lambda_k + L(x_k,\lambda_k)(x - x_k) + c'(x_k)^*(\lambda - \lambda_k) = 0, \tag{15.5b}$$
$$c_E(x_k) + c'_E(x_k)(x - x_k) = 0, \tag{15.5c}$$
$$c_I(x_k) + c'_I(x_k)(x - x_k) \leqslant 0. \tag{15.5d}$$

Enfin, la relation d'orthogonaité $(x,\lambda) \perp (F(z_k) + F'(z_k)(z - z_k))$ s'exprime alors par

$$\lambda_I \perp [c_I(x_k) + c'_I(x_k)(x - x_k)]. \tag{15.5e}$$

Si l'on élimine λ_k de (15.5b), ce système peut se récrire

$$\begin{cases} \nabla f(x_k) + L(x_k,\lambda_k)(x - x_k) + c'(x_k)^*\lambda = 0, \\ c_E(x_k) + c'_E(x_k)(x - x_k) = 0, \\ 0 \leqslant \lambda_I \perp [c_I(x_k) + c'_I(x_k)(x - x_k)] \leqslant 0. \end{cases} \tag{15.6}$$

Le point important est maintenant de constater que le système (15.6) est formé des conditions d'optimalité de ce que l'on appelle le *problème quadratique osculateur* de (P_{EI}), à savoir

$$\boxed{\begin{cases} \min_x \ \nabla f(x_k)^{\mathsf{T}}(x - x_k) + \frac{1}{2}(x - x_k)^{\mathsf{T}}\nabla^2_{xx}(x_k,\lambda)(x - x_k) \\ c_E(x_k) + c'(x_k)_E(x - x_k) = 0 \\ c_I(x_k) + c'(x_k)_I(x - x_k) \leqslant 0. \end{cases}} \tag{15.7}$$

On peut à présent préciser l'itération de l'algorithme OQS local (c'est-à-dire sans technique de globalisation de la convergence).

Algorithme 15.1 (OQS local) Une itération passe de l'itéré courant $(x_k,\lambda_k) \in \mathbb{E} \times \mathbb{R}^m$ à l'itéré suivant (x_{k+1},λ_{k+1}) par les étapes suivantes :
1. *Test d'arrêt* : si le couple (x_k,λ_k) est satisfaisant, on s'arrête.
2. *Nouvel itéré* : On prend comme nouvel itéré (x_{k+1},λ_{k+1}) une solution primale duale du problème quadratique osculateur (15.7).

L'algorithme OQS décompose donc la réslution du problème d'optimisation non-linéaire (P_{EI}) en une suite de problème d'optimisation quadratique, plus imples à résoudre que le problème original.

15.1.2 Convergence locale

Définition 15.2 On dit qu'un couple stationnaire (x_*,λ_*) de (P_{EI}) est *semi-stable* (resp. *hémi-stable*) si ce couple est une solution semi-stable (resp. hémi-stable) de l'inclusion dans (15.2a) avec F et K donnés par (15.2b).

Si (x_*, λ_*) vérifie les conditions de KKT, on a

$$F(x_*, \lambda_*) = \begin{pmatrix} 0_{\mathbb{E}} \\ 0_{\mathbb{R}^{m_E}} \\ -c_I(x_*) \end{pmatrix}, \tag{15.8}$$

$$F(x_*, \lambda_*) + F'(x_*, \lambda_*)(d, \mu) = \begin{pmatrix} L_* d + c'(x_*)^* \mu \\ -c'_E(x_*) d \\ -c_I(x_*) - c'_I(x_*) d \end{pmatrix}. \tag{15.9}$$

où l'on a utilisé (15.4) et posé $L_* := L(x_*, \lambda_*) := \nabla^2_{xx} \ell(x_*, \lambda_*)$. La condition $F(x_*, \lambda_*) + F'(x_*, \lambda_*)(d, \mu) + \mathrm{N}_K(x_*, \lambda_*) \ni 0$ s'écrit

$$\begin{gathered} L_* d + c'(x_*)^* \mu = 0, \\ c'_{E \cup I^{0+}_*}(x_*) d = 0, \quad c'_{I^{00}_*}(x_*) d \leqslant 0, \quad c_{I \setminus I^0_*}(x_*) + c'_{I \setminus I^0_*}(x_*) d \leqslant 0. \end{gathered}$$

Le résultat suivant analyse la semi-stabilité d'un couple stationnaire pour cette inclusion.

Proposition 15.3 (semi-stabilité d'une solution primale-duale) *Si $z_* := (x_*, \lambda_*)$ est une solution primale-duale de (P_{EI}), les propriétés suivantes sont équivalentes :*
(i) z_* est une solution primale-duale semi-stable,
(ii) $(d, \mu) = 0$ est l'unique solution du système

$$L_* d + c'(x_*)^* \mu = 0, \tag{15.10a}$$

$$c'_{E \cup I^{0+}_*}(x_*) d = 0, \tag{15.10b}$$

$$\mu_{I^{0+}_*} \geqslant 0, \quad 0 \leqslant \mu_{I^{00}_*} \perp c'_{I^{00}_*}(x_*) d \leqslant 0 \quad \text{et} \quad \mu_{I \setminus I^0_*} = 0. \tag{15.10c}$$

Démonstration. Par l'équivalence $(i) \Leftrightarrow (ii)$ de la proposition **??**, la semi-stabilité de (x_*, λ_*) est équivalente à

$$\begin{gathered} (d, \mu) = 0 \text{ est solution isolée de} \\ L_* d + c'(x_*)^* \mu = 0 \\ c'_E(x_*) d = 0 \\ c_I(x_*) + c'_I(x_*) d \in \mathrm{N}_{\mathbb{R}^{m_I}_+}(\lambda_* + \mu), \end{gathered} \tag{15.11}$$

où l'on a tenu compte du fait que $\nabla_x \ell(x_*, \lambda_*) = 0$ et $c_E(x_*) = 0$. Il revient au même de dire que les *petites* solutions du système ci-dessus sont nulles. Observons que la dernière condition ci-dessus, celle qu'il faut exprimer autrement, s'écrit aussi

$$0 \leqslant \big(\lambda_* + \mu\big) \perp \big(c_I(x_*) + c'_I(x_*) d\big) \leqslant 0. \tag{15.12}$$

Exploitons cette condition.

- Si $i \in I^{0+}_*$, $c_i(x_*) = 0$ et $(\lambda_*)_i > 0$, si bien que $(\lambda_* + \mu)_i > 0$ pour μ petit et la complémentarité dans (15.12) implique que $c'_i(x_*) d = 0$

- Pour les indices dans I_*^{00}, $c_i(x_*) = 0$ et $(\lambda_*)_i = 0$, si bien que (15.12) implique que $0 \leqslant \mu_{I_*^{00}} \perp c'_{I_*^{00}}(x_*)d \leqslant 0$.
- Si $i \in I \setminus I_*^0$, $c_i(x_*) < 0$ et $(\lambda_*)_i = 0$, si bien que pour d suffisamment petit $c_i(x_*) + c'_i(x_*)d < 0$ et donc la complémentarité dans (15.12) implique que $\mu_i = 0$.

On a donc montré qu'une *petite* solution du système dans (15.11) est solution de (15.10). On montre de la même manière qu'une *petite* solution de (15.10) est solution du système dans (15.11). Dès lors (15.11) revient à dire que $(d, \mu) = 0$ est solution isolée de (15.10).

On conclut en observant que l'ensemble des solutions de (15.10) est un cône, si bien qu'il revient au même de dire que $(d, \mu) = 0$ est solution isolée de (15.10) ou que $(d, \mu) = 0$ est l'unique solution de (15.10). □

On rappelle que le cône critique C_* en une solution locale de (P_{EI}) a été défini en (4.45a) et que L_* est une écriture simplifiée pour la hessienne $\nabla^2_{xx}\ell(x_*, \lambda_*)$ du lagrangien ℓ évaluée en $(x_*, \lambda_*) \in \mathbb{E} \times \mathbb{R}^m$.

Proposition 15.4 (semi-stabilité d'un minimum local) *Si x_* est un minimum local de (P_{EI}) et λ_* est un multiplicateur optimal associé, alors les propriétés suivantes sont équivalentes :*

(i) (x_*, λ_*) est semi-stable,

(ii) λ_* est l'unique multiplicateur associé à x_* et les conditions suffisantes du second ordre sont satisfaites, c'est-à-dire, $\forall\, d \in C_* \setminus \{0\}$, on a $\langle L_* d, d\rangle > 0$.

Démonstration. 1) On exploite l'équivalence $(i) \Leftrightarrow (iii)$ de la proposition **??** pour obtenir une autre expression de la semi-stabilité de (x_*, λ_*). Dans ce but, on pose

$$(d, \mu) := z - z_* = (x, \lambda) - (x_*, \lambda_*)$$

et on observe qu'avec F définie en (15.2b), (15.8) et (15.9), on a

$$\begin{aligned} \langle F'(x_*, \lambda_*)(d, \mu), (d, \mu)\rangle &= \langle L_* d, d\rangle, \\ \langle F(x_*, \lambda_*), (d, \mu)\rangle = 0 \quad &\Longleftrightarrow \quad \mu_{I \setminus I_*^0} = 0. \end{aligned}$$

Par l'équivalence $(i) \Leftrightarrow (iii)$ de la proposition **??**, la semi-stabilité de (x_*, λ_*) est équivalente à

$$\text{on a } \langle L_* d, d\rangle > 0 \text{ pour tout } (d, \mu) \in \mathbb{E} \times \mathbb{R}^m \text{ non nul tel que} \tag{15.13a}$$

$$(\mu + \lambda_*)_I \geqslant 0, \quad \mu_{I \setminus I_*^0} = 0, \tag{15.13b}$$

$$L_* d + c'_{E \cup I_*^0}(x_*)^* \mu = 0, \tag{15.13c}$$

$$d \in C_*, \quad c_{I \setminus I_*^0}(x_*) + c'_{I \setminus I_*^0}(x_*)d \leqslant 0. \tag{15.13d}$$

où $C_* := \{d \in \mathbb{E} : c'_{E \cup I_*^{0+}}(x_*)d = 0,\ c'_{I_*^{00}}(x_*)d \leqslant 0\}$ est le cône critique.

2) [$(i) \Rightarrow (ii)$] Comme (x_*, λ_*) est semi-stable, c'est une solution *isolée* du système d'optimalité 4.33 (proposition **??**). Par ailleurs, l'ensemble des multiplicateurs associé à x_* est convexe, λ_* est nécessairement l'unique multiplicateur associé à x_*.

L'ensemble des multiplicateurs associés à x_* étant borné (c'est un singleton), les conditions de qualification (QC-MF) ont lieu. On déduit alors de l'optimalité locale de x_* que $\langle L_* d, d\rangle \geqslant 0$ pour toute direction critique $d \in C_*$ (théorème 4.48). Donc si la conclusion n'a pas lieu, il existe une direction $d_1 \in C_* \setminus \{0\}$ telle que $\langle L_* d_1, d_1\rangle = 0$. Cette direction d_1 est donc solution de

$$\begin{cases} \min \langle L_* d, d\rangle \\ c'_{E\cup I_*^{0+}}(x_*)d = 0 \\ c'_{I_*^{00}}(x_*)d \leqslant 0, \end{cases}$$

problème dont les contraintes définissent les directions critiques. Les conditions d'optimalité de ce problème, aux contraintes qualifiées par (QC-A), affirment l'existence d'un multiplicateur $\mu_1 \in \mathbb{R}^m$ tel que l'on ait

$$\begin{gathered} (\mu_1)_{I\setminus I_*^0} = 0, \\ L_* d_1 + c'(x_*)^* \mu_1 = 0, \\ c'_{E\cup I_*^{0+}}(x_*)d_1 = 0, \\ 0 \leqslant (\mu_1)_{I_*^{00}} \perp c'_{I_*^{00}}(x_*)d_1 \leqslant 0. \end{gathered}$$

Alors $(d, \mu) = t(d_1, \mu_1)$, avec $t > 0$ suffisamment petit, vérifie (15.13b)-(15.13d) mais pas la conclusion $\langle L_* d, d\rangle > 0$ dans (15.13a). Cette contradiction montre que les CS2 sont vérifiées.

3) [$(i) \Leftarrow (ii)$] Évidemment, si les CS2 ont lieu, x_* est un minimum local de (P_{EI}) (théorème 4.49).

Pour montrer la semi-stabilité de (x_*, λ_*), on montre que (15.13) a lieu. Soit (d, μ) non nul vérifiant (15.13b)-(15.13d). Il suffit de montrer que $d \neq 0$, car alors $d \in C_* \setminus \{0\}$ et la conclusion $\langle L_* d, d\rangle > 0$ de (15.13a) s'obtient par la CS2 supposée dans (ii). Raisonons par l'absurde en supposant que $d = 0$. Alors $\mu \neq 0$ et il est facile de voir que $\lambda_* + \mu$ serait un autre multiplicateur optimal associé à x_*, ce qui contredirait l'unicité supposée de celui-ci. □

Proposition 15.5 (condition suffisante d'hémi-stabilité) *Si x_* est un minimum local de (P_{EI}) et si (x_*, λ_*) est une solution semi-stable du système d'optimalité (4.33), alors (x_*, λ_*) est hémi-stable.*

Démonstration. Il s'agit de montrer que, pour tout $\alpha > 0$ donné, on peut trouver une constante $\beta > 0$, telle que, quel que soit $(x_0, \lambda_0) \in \bar{B}((x_*, \lambda_*), \beta)$, l'inclusion en (x, λ) suivante

$$\begin{pmatrix} \nabla f(x_0) + c'(x_0)^* \lambda_0 \\ -c(x_0) \end{pmatrix} + \begin{pmatrix} L(x_0, \lambda_0) & c'(x_0)^* \\ -c'(x_0) & 0 \end{pmatrix} \begin{pmatrix} x - x_0 \\ \lambda - \lambda_0 \end{pmatrix} + \mathrm{N}_K(x, \lambda) \ni 0$$

a une solution dans $\bar{B}((x_*, \lambda_*), \alpha)$. Cette inclusion est le système d'optimalité du premier ordre du problème quadratique en $x \in \mathbb{E}$ suivant

$$\begin{cases} \min \langle \nabla f(x_0), x - x_0 \rangle + \frac{1}{2} \langle L(x_0, \lambda_0)(x - x_0), x - x_0 \rangle \\ c_E(x_0) + c'_E(x_0)(x - x_0) = 0 \\ c_I(x_0) + c'_I(x_0)(x - x_0) \leqslant 0. \end{cases} \tag{15.14}$$

Celui-ci est une perturbation du problème quadratique que l'on obtient en prenant $(x_0, \lambda_0) = (x_*, \lambda_*)$, à savoir

$$\begin{cases} \min \langle \nabla f(x_*), x - x_* \rangle + \frac{1}{2} \langle L_*(x - x_*), x - x_* \rangle \\ c_E(x_*) + c'_E(x_*)(x - x_*) = 0 \\ c_I(x_*) + c'_I(x_*)(x - x_*) \leqslant 0. \end{cases} \tag{15.15}$$

Observons que le problème non perturbé (15.15) admet (x_*, λ_*) comme solution primale-duale, parce que son système d'optimalité du premier ordre en $x = x_*$ n'est autre que celui de (P_{EI}), vérifié par (x_*, λ_*), et parce que ses conditions d'optimalité du deuxième ordre sont également celles de (P_{EI}), qui sont satisfaites par hypothèses. Par ailleurs, la semi-stabilité entraîne l'unicité du multiplicateur (convexité de l'ensemble des multiplicateurs optimaux et remarque **??**(3)) et donc aussi les conditions de qualification de Mangasarian-Fromovitz (QC-MF) (proposition 4.45). On sait alors [514 ; théorème 4.2 et corollaire 4.3] que l'on peut trouver une solution du système perturbé (15.14) dont l'écart à (x_*, λ_*) est majoré par une constante fois la grandeur de la perturbation. □

Théorème 15.6 (convergence de l'algorithme SQP) *Si f et c sont $C^{2,1}$ dans la voisinage d'un minimum local x_* de (P_{EI}), s'il existe un unique multiplicateur optimal associé à x_*, et si les conditions suffisantes du second ordre sont vérifiées, alors il existe un voisinage V de (x_*, λ_*) tel que si le premier itéré $(x_1, \lambda_1) \in V$,*
1) *l'algorithme SQP peut générer une suite $\{(x_k, \lambda_k)\}$ dans V,*
2) *$\{(x_k, \lambda_k)\}$ converge quadratiquement vers (x_*, λ_*).*

Démonstration. Par la proposition 15.4, l'unicité du multiplicateur optimal et les conditions suffisantes du deuxième ordre, (x_*, λ_*) est une solution semi-stable de (15.2). Par la proposition 15.5, c'est aussi une solution hémi-stable. On peut alors appliquer le théorème **??** qui donne le résultat. □

15.2 L'algorithme local

15.2.1 Un algorithme non convergent

Considérons d'abord le problème avec contraintes d'égalité seulement

$$(P_E) \quad \begin{cases} \min f(x) \\ c(x) = 0, \end{cases}$$

où $f : \mathbb{E} \to \mathbb{R}$ et $c : \mathbb{E} \to \mathbb{R}^m$.

On a vu au chapitre 10 comment on pouvait introduire la méthode de Newton pour résoudre des équations non linéaires (voir (10.3)) et pour minimiser une fonction (voir (10.7)). Pour résoudre le problème (P_E), on pourrait donc être tenté d'obtenir le déplacement d_k en x_k en prenant une approximation quadratique du critère et en linéarisant les contraintes en x_k. Avec cette méthode, on calculerait d_k comme solution du problème quadratique

$$\begin{cases} \min \nabla f(x_k)^\mathsf{T} d + \frac{1}{2} d^\mathsf{T} \nabla^2 f(x_k) d \\ c_E(x_k) + c'_E(x_k) \cdot d = 0 \end{cases} \tag{15.16}$$

et on prendrait $x_{k+1} = x_k + d_k$. Attention, **cet algorithme n'est pas nécessairement convergent** ! On connaît des exemples (voir l'exercice 15.1) dans lesquels la solution est répulsive pour cet algorithme : on peut trouver des itérés aussi proches que l'on veut de la solution (mais différents de la solution), tels que l'itéré suivant est plus éloigné de la solution que l'itéré en question. En fait, si on écrit l'algorithme comme un processus de point fixe, $x_{k+1} = \Phi(x_k)$ pour une certaine fonction Φ, l'application Φ n'est pas *strictement contractante* proche de la solution. Hélas, on voit encore parfois cette approche utilisée de nos jours !

Pourtant il doit bien exister une *méthode newtonienne*, c'est-à-dire qui procède par linéarisations, pour résoudre les problèmes avec contraintes ! Les chercheurs en optimisation numérique se sont posés la question bien longtemps et ce n'est qu'au milieu des années 1970, soit 30 ans après l'invention de l'algorithme du simplexe, que la situation s'est éclaircie. Ce qui ne fonctionne pas dans l'approche qui a conduit à (15.16), c'est qu'en traitant séparément les problèmes contradictoires « $\min f(x)$ » et « $c(x) = 0$ » (le minimum de f ne vérifie en général pas la contrainte $c(x) = 0$, sinon pourquoi la spécifier), on a mis en place deux algorithmes qui ne se concertent pas et tirent à hue et à dia. Cela paraît à présent évident, la bonne démarche est de résoudre un seul système qui détermine les solutions de (P_E) et cela par un unique algorithme. Ce système est celui formé par les conditions d'optimalité de (P_E) et l'algorithme est celui de Newton. Ce système comporte en effet $n+m$ équations et $n+m$ inconnues (x_*, λ_*), ce qui en fait un candidat convenable pour être résolu par des itérations de Newton. Nous n'allons pas écrire cet algorithme, qui est en fait un cas particulier de celui que nous introduisons à la section suivante.

15.2.2 L'algorithme OQS

La discussion de la section 15.2.1 nous a montré qu'il était judicieux de faire des itérations de Newton sur le système d'optimalité de (P_{EI}). Rappelons (théorème 4.32) que celui-ci détermine les points stationnaires (x_*, λ_*) du problème comme solution du système de Karush, Kuhn et Tucker :

$$\text{(KKT)} \quad \begin{cases} (a) & \nabla f(x_*) + A(x_*)^\mathsf{T} \lambda_* = 0 \\ (b) & c_E(x_*) = 0, \quad c_I(x_*) \leqslant 0 \\ (c) & (\lambda_*)_I \geqslant 0 \\ (d) & (\lambda_*)_I^\mathsf{T} c_I(x_*) = 0. \end{cases} \tag{15.17}$$

Travailler sur ce système d'optimalité n'allait pas de soi et l'on peut dire qu'il a fallu franchir un saut conceptuel pour y arriver.

- D'abord, ces relations forment un système en (x_*, λ_*), pas seulement en x_*. Il faut donc les linéariser par rapport à (x, λ), pas seulement par rapport à x comme dans l'algorithme qui a conduit à (15.16). L'algorithme de Newton qui en résultera sera donc une *méthode primale-duale*, générant à la fois des itérés primaux x_k et duaux λ_k.
- Une autre difficulté provient de la présence d'inégalités dans (15.17). La linéarisation de l'inégalité $F(x) \leqslant 0$ en x, pour un accroissement d, se fera par $F(x) + F'(x) \cdot d \leqslant 0$. On verra par les résultats de convergence locale obtenus qu'il s'agit d'un bon choix.

Soit $(x, \lambda) \in \mathbb{E} \times \mathbb{R}^m$ le point courant auquel on linéarise (15.17) et $(d, \mu) \in \mathbb{E} \times \mathbb{R}^m$ l'accroissement, la correction, que l'on désire apporter au point courant. L'itéré suivant (x_+, λ_+) s'obtiendra donc par

$$x_+ := x + d \qquad \text{et} \qquad \lambda_+ := \lambda + \mu.$$

On obtient le système suivant :

$$\begin{cases} L(x,\lambda)d + A(x)^{\mathsf{T}}\mu = -\nabla_x \ell(x,\lambda) \\ (c(x) + A(x)d)^{\#} = 0 \\ (\lambda + \mu)_I \geqslant 0 \\ (\lambda + \mu)_I^{\mathsf{T}} c_I(x) + \lambda_I^{\mathsf{T}} (A(x)d)_I = 0. \end{cases} \tag{15.18}$$

On a noté $A(x) := c'(x)$ la jacobienne des contraintes et

$$L(x, \lambda) := \nabla^2_{xx} \ell(x, \lambda)$$

la hessienne du lagrangien. Bien que linéaire, le système (15.18) a l'inconvénient d'être très difficile à résoudre. On y trouve en effet des égalités *et* des inégalités. De plus, on ne voit plus le problème d'optimisation dont il est issu. On se rappelle, en effet, que dans le cas de l'optimisation sans contrainte, le pas de Newton pouvait se voir comme un point stationnaire d'un problème d'optimisation quadratique sans contrainte. On aimerait pouvoir faire de même ici.

Dans ce but, on modifie le système (15.18), de manière à ce qu'il soit le système d'optimalité d'un problème quadratique avec contraintes linéaires. Si on ajoute le terme $\mu^{\mathsf{T}} A(x)d$ à la dernière équation, celle-ci ressemble davantage à des conditions de complémentarité (c'est alors un produit de deux facteurs). On observera par ailleurs, que dans le voisinage d'une solution primale-duale, le terme $\mu^{\mathsf{T}} A(x)d$ est formé d'un produit de deux grandeurs très petites, les accroissements μ et d. L'ajout du terme $\mu^{\mathsf{T}} A(x)d$ apporte donc une perturbation insignifiante au système (15.18), qui ne devrait pas modifier les propriétés de convergence rapide de l'algorithme (cette intuition s'avérera correcte). Le système devient alors

$$\begin{cases} L(x,\lambda)d + A(x)^{\mathsf{T}}\mu = -\nabla_x \ell(x,\lambda) \\ (c(x) + A(x)d)^{\#} = 0 \\ (\lambda + \mu)_I \geqslant 0 \\ (\lambda + \mu)_I^{\mathsf{T}} (c(x) + A(x)d)_I = 0. \end{cases} \tag{15.19}$$

On vérifie aisément le fait remarquable suivant : (15.19) est formé des conditions de KKT du problème d'optimisation quadratique

$$\begin{cases} \min_d \nabla f(x)^\mathsf{T} d + \frac{1}{2} d^\mathsf{T} L(x,\lambda) d \\ c_E(x) + A_E(x) d = 0 \\ c_I(x) + A_I(x) d \leqslant 0. \end{cases} \tag{15.20}$$

Plus précisément, le lien entre l'accroissement μ de λ et le multiplicateur λ^{PQ} associé aux contraintes de (15.20) s'écrit $\lambda^{\text{PQ}} = \lambda + \mu$. Dès lors la mise à jour du couple (x, λ) se fait par

$$x_+ := x + d \qquad \text{et} \qquad \lambda_+ := \lambda^{\text{PQ}}. \tag{15.21}$$

L'algorithme qui met à jour (x, λ) par ces formules est appelé l'*algorithme OQS* (pour Sequential Quadratic Programming). Le problème (15.20) que l'on doit résoudre à chaque itération pour déterminer (d, λ^{PQ}) est appelé le *problème quadratique osculateur* (PQO) au problème (P_{EI}) en (x, λ). On le déduit aisément de (P_{EI}). Ses contraintes sont celles de (P_{EI}), linéarisées en x. Son critère est hybride, avec $\nabla f(x)$ dans la partie linéaire et la hessienne du lagrangien dans la partie quadratique. Ce qui manque au problème quadratique (15.16) est maintenant manifeste : ce problème ne fait pas intervenir la courbure des contraintes (les dérivées secondes de c sont absentes). Dans le problème quadratique osculateur (15.20), cette courbure intervient dans la hessienne du lagrangien, pas par une approximation quadratique des contraintes.

On peut à présent résumer l'algorithme OQS local.

Algorithme 15.7 (OQS local — une itération)

On suppose que l'on dispose au début de l'itération d'un couple $(x, \lambda) \in \mathbb{E} \times \mathbb{R}^m$.

1. *Test d'arrêt* : si (x, λ) vérifie les conditions d'optimalité (15.17), arrêt de l'algorithme.
2. *Déplacement* : calculer (d, λ^{PQ}) comme point stationnaire du problème quadratique osculateur (15.20).
3. *Mise à jour des variables* : le nouvel itéré (x_+, λ_+) est donné par (15.21).

Le coût d'une itération est essentiellement dû à la résolution du problème quadratique osculateur. On a donc reporté la *combinatoire* du problème (P_{EI}) sur des problèmes quadratiques, dans lesquels elle est plus facilement prise en compte. Cependant, plusieurs difficultés peuvent se présenter :

- La hessienne du lagrangien n'est pas nécessairement semi-définie positive, si bien que le PQO est NP-ardu (voir le chapitre **??**) et peut présenter des *solutions importunes* (indésirables car trop grandes, nous reviendrons sur cette question à la section 15.2.3). Pour cette raison, les implémentations de cet algorithmes utilisent souvent une modification ou une approximation définie positive M_k de la hessienne de $L(x_k, \lambda_k)$. S'il est réalisable, c'est-à-dire si ses contraintes sont compatibles, le PQO peut alors être résolu en un nombre polynomial d'itérations et n'a plus de solutions indésirables.
- Le PQO peut ne pas être borné (donc ne pas avoir de solution). L'introduction de l'algorithme nous a montré qu'un point stationnaire suffirait, mais il n'y a pas

d'algorithme évident pour trouver un point stationnaire d'un problème quadratique autre qu'un minimum local.

- Les contraintes linéarisées du PQO peuvent être *incompatibles*. Diverses techniques permettent de faire face à cette difficulté : relaxation des contraintes linéarisées dans la globalisation de la convergence par recherche linéaire; la globalisation de la convergence par régions de confiance prend directement en compte cette difficulté.

15.2.3 Convergence locale ▲

Nous présentons ci-dessous le résultat de convergence le plus simple, celui qui se démontre en se ramenant au cas des problèmes avec contraintes d'égalité. Pour un résultat plus fin, nous renvoyons le lecteur à [71, 72, 73].

Commençons donc par considérer le problème où il n'y a que des contraintes d'égalité :

$$(P_E) \quad \begin{cases} \min f(x) \\ c(x) = 0. \end{cases}$$

Dans ce cas, l'algorithme OQS est l'algorithme de Newton (10.2)–(10.3) appliqué au système d'optimalité

$$F(z) = 0, \quad \text{où} \quad z = (x, \lambda) \quad \text{et} \quad F(z) = \begin{pmatrix} \nabla_x \ell(x, \lambda) \\ c(x) \end{pmatrix}.$$

On a noté $(x, \lambda) \mapsto \ell(x, \lambda) = f(x) + \lambda^{\mathsf{T}} c(x)$ le lagrangien associé au problème (P_E). Le résultat de convergence locale est alors un corrolaire immédiat du théorème 10.2, qui requiert une hypothèse de différentiabilité de F et d'inversibilité de $F'(x_*, \lambda_*)$, ce qui conduit à la définition suivante.

Définition 15.8 (point stationnaire régulier) Un point stationnaire (x_*, λ_*) de (P_E) est dit *régulier* si la matrice d'ordre $n + m$

$$\begin{pmatrix} \nabla^2_{xx} \ell(x_*, \lambda_*) & c'(x_*)^* \\ c'(x_*) & 0 \end{pmatrix} \tag{15.22}$$

est inversible.

Un point stationnaire (x_*, λ_*) de (P_E) vérifiant les conditions d'optimalité du second ordre du théorème 4.24 et tel que $c'(x_*)$ est surjective est régulier. En effet, il suffit de montrer que la matrice carrée (15.22) est injective pour en conclure qu'elle est inversible. Or si (d, μ) est dans son noyau, alors $d \in \mathcal{N}(c'(x_*))$ et $d^{\mathsf{T}} \nabla^2_{xx} \ell(x_*, \lambda_*) d = 0$. Donc $d = 0$ par les conditions du second ordre. Ensuite $\mu = 0$ par l'injectivité de $c'(x_*)^*$.

Théorème 15.9 (convergence quadratique locale de l'algorithme de Newton pour (P_E)) *Supposons que f et c soient de classe C^2 dans un voisi-*

nage d'un point stationnaire régulier x_* de (P_E), avec multiplicateur associé λ_*. Alors, il existe un voisinage V de (x_*, λ_*) tel que, si le premier itéré $(x_1, \lambda_1) \in V$, l'algorithme de Newton 15.7 est bien défini et génère une suite $\{(x_k, \lambda_k)\}$ convergeant superlinéairement vers (x_*, λ_*). Si f'' et c'' est lipschitzienne dans un voisinage de x_*, la convergence de la suite est quadratique.

DÉMONSTRATION. On applique le théorème 10.2 avec F définie ci-dessus, $z = (x, \lambda)$ et $z_* = (x_*, \lambda_*)$. La jacobienne $F'(z_*)$ est inversible par la régularité de z_*. La convergence superlinéaire de $\{(x_k, \lambda_k)\}$ vers (x_*, λ_*) a lieu si (x_1, λ_1) est suffisamment proche de (x_*, λ_*). Si f'' et c'' son lipschitzienne proche de x_*, il en est de même de F' proche de z_*, et le convergence quadratique de $\{(x_k, \lambda_k)\}$ s'en suit. □

Une propriété minimale d'un algorithme convergent est de générer un déplacement nul lorsque l'itéré courant est une solution. Cette propriété élémentaire n'a pourtant pas nécessairement lieu lorsque le déplacement de l'algorithme OQS est une solution *arbitraire* du PQO, comme le montre l'exemple suivant.

Exemple 15.10 On cherche à minimiser le logarithme de $(1 + x)$ lorsque x est restreint à l'intervalle $[0, 3]$:

$$\begin{cases} \min_x \log(1 + x) \\ 0 \leqslant x \leqslant 3. \end{cases}$$

Le logarithme a été utilisé de manière à introduire de la non-convexité dans le problème (par la monotonicité du logarithmie, il aurait été équivalent de minimiser $(1+x)$ et le problème serait devenu linéaire). On vérifie aisément que le problème a une unique solution primale-duale $(x_*, \lambda_*) = (0, (1, 0))$, qui satisfait les conditions suffisantes d'optimalité, la complémentarité stricte et la qualification des contraintes (QC-IL). On peut donc arguer qu'il s'agit d'une « bonne » solution. Cependant, le problème quadratique osculateur (15.20) en cette solution s'écrit

$$\begin{cases} \min_d d - \frac{1}{2}d^2 \\ 0 \leqslant d \leqslant 3. \end{cases}$$

Ce PQO a trois points stationnaires primaux-duaux (d, λ) : à savoir un minimum local $(0, (1, 0))$, un maximum $(1, (0, 0))$ et un minimum global $(3, (0, 2))$. Parmi ces points stationnaires, seul le premier est satisfaisant puisqu'il donne un déplacement nul et le multiplicateur optimal. Les deux autres points stationnaires sont dits *importuns*, indésirables. □

Le phénomène qui se produit dans cet exemple ne peut avoir lieu que si $L(x, \lambda)$ n'est pas définie positive. Sinon, le PQO est strictement convexe et a donc une unique solution dès que ses contraintes sont compatibles. Les résultats de convergence de l'algorithme OQS doivent donc faire une hypothèse sur la solution du PQO qui est sélectionnée par l'algorithme à chaque itération. Dans le résultat donné ci-dessous, il est supposé que d est une *solution de norme minimale*.

The next lemma is useful for proving theorem 15.12. We use the notation that is suitable for the present framework. Its proof is proposed in exercise **??**.

Lemme 15.11 (propriété de la surjectivité) *Let A be an $m \times n$ surjective matrix and $\mu \in \mathbb{R}^m$. Then, $\forall \varepsilon > 0$, $\exists \delta > 0$, such that when the $m \times n$ matrix $\tilde{A}$ and the point $\tilde{\mu} \in \mathbb{R}^m$ satisfy $\|\tilde{A} - A\| \leqslant \delta$ and $\|\tilde{A}^{\mathsf{T}}\tilde{\mu} - A^{\mathsf{T}}\mu\| \leqslant \delta$, there must hold $\|\tilde{\mu} - \mu\| \leqslant \varepsilon$.*

Théorème 15.12 (convergence quadratique primale-duale de l'algorithme OQS) *Suppose that f and c are of class C^2 in a neighborhood of a stationary point x_* of (P_{EI}), with associated multiplier λ_*. Suppose also that strict complementarity holds and that $(x_*, (\lambda_*)_{E \cup I_*^0})$ is a regular stationary point of the equality constrained problem*

$$\begin{cases} \min_x \ f(x) \\ c_i(x) = 0, \quad \text{for } i \in E \cup I_*^0, \end{cases} \tag{15.23}$$

*in the sense of definition **??**. Consider the OQS algorithm, in which d_k is a minimum-norm stationary point of the osculating quadratic problem (15.20). Then there is a neighborhood V of (x_*, λ_*) such that, if the first iterate $(x_1, \lambda_1) \in V$:*

(i) the OQS algorithm is well defined and generates a sequence $\{(x_k, \lambda_k)\}$ that converges superlinearly to (x_*, λ_*);

(ii) the active constraints of the osculating quadratic problem (15.20) are those of problem (P_{EI});

(iii) if, in addition, f and c are of class $C^{2,1}$ in a neighborhood of x_*, the convergence of $\{(x_k, \lambda_k)\}$ is quadratic.

Démonstration. The idea of the proof is to show that, close to (x_*, λ_*), the selected minimum-norm stationary point of the osculating quadratic problem (15.20) is actually the primal-dual Newton step on (15.23). The result then follows from theorem 15.9.

Suppose that (x, λ) is close enough to (x_*, λ_*). Since $(x_*, (\lambda_*)_{E \cup I_*^0})$ is a regular stationary point of (15.23), $c'_{E \cup I_*^0}(x_*)$ is surjective and the osculating QP associated with (15.23), namely

$$\begin{cases} \min_{\mathrm{d}t} \ \nabla f(x)^{\mathsf{T}} \mathrm{d}t + \frac{1}{2} \mathrm{d}t^{\mathsf{T}} L(x, \lambda) \mathrm{d}t \\ c_i(x) + c'_i(x) \cdot \mathrm{d}t = 0, \quad \text{for } i \in E \cup I_*^0, \end{cases} \tag{15.24}$$

has a unique primal-dual stationary point. We denoted it by $(\mathrm{d}t, \tilde{\ell}_{E \cup I_*^0})$ and form with $\tilde{\ell}_{E \cup I_*^0}$ a vector $\tilde{\ell} \in \mathbb{R}^m$, by setting $\tilde{\ell}_i = 0$ for $i \in I \setminus I_*^0$.

Let us show that $(\mathrm{d}t, \tilde{\ell})$ is a stationary point of the osculating quadratic problem (15.20), if $(x, \lambda) := (x_k, \lambda_k)$ is in some neighborhood of (x_*, λ_*); this will imply that, for (x, λ) close to (x_*, λ_*), the OQS algorithm is well defined and that d is small (it is a minimum-norm stationary point and $\mathrm{d}t$ is small by theorem 15.9). We only need to show that

$$c_i(x) + c'_i(x) \cdot \mathrm{d}t \leqslant 0 \text{ for } i \in I \setminus I_*^0, \tag{15.25}$$

$$\lambda_i \geqslant 0 \text{ for } i \in I_*^0. \tag{15.26}$$

From theorem 15.9, $(x + \mathrm{d}t, \tilde{\ell})$ is close to (x_*, λ_*), when (x, λ) is close to (x_*, λ_*), so that $\mathrm{d}t$ is small. Now (15.25) follows when (x, λ) close enough to (x_*, λ_*), since $c_i(x_*) < 0$ for $i \in I \setminus I_*^0$. On the other hand, to prove (15.26), observe that, for $i \in I_*^0$, $(\lambda_*)_i > 0$ by strict complementarity. Therefore $\tilde{\ell}_i \geqslant 0$ since, by theorem 15.9, $\tilde{\ell}$ is close to λ_* when (x, λ) is close to (x_*, λ_*).

Let us now show that the pair $(d, \lambda^{\mathrm{PQ}}) := (d_k, \lambda_k^{\mathrm{PQ}})$ formed of the selected minimum-norm stationary point of the osculating QP and its associated multiplier is in fact $(\mathrm{d}t, \tilde{\ell})$, if (x, λ) is in some neighborhood of (x_*, λ_*). Since (15.24) has a single stationary point, for (x, λ) close to (x_*, λ_*), we have to show that $(d, \lambda^{\mathrm{PQ}})$ satisfies its optimality conditions. Knowing that $(d, \lambda^{\mathrm{PQ}})$ satisfies the optimality conditions of the osculating quadratic problem (15.20), we just have to show that

$$\lambda_i = 0 \text{ for } i \in I \setminus I_*^0, \tag{15.27}$$

$$c_i(x) + c_i'(x) \cdot d = 0 \text{ for } i \in I_*^0. \tag{15.28}$$

Condition (15.27) is a consequence of the complementarity in (15.20) and of the fact that, for (x, λ) close to (x_*, λ_*) and $i \in I \setminus I_*^0$, $c_i(x) + c_i'(x) \cdot d < 0$ (since $c_i(x_*) < 0$ and d is small). Condition (15.28) results from the complementarity in (15.20) and the fact that $\lambda_i > 0$ for $i \in I_*^0$. For the latter claim, observe indeed that

$$\nabla f(x) + L(x, \lambda)d + \sum_{i \in E \cup I_*^0} \lambda_i \nabla c_i(x) = 0.$$

Since $\nabla f(x) + L(x, \lambda)d$ is close to $\nabla f(x_*)$ and the gradients $\{\nabla c_i(x) : i \in E \cup I_*^0\}$ are close to the linearly independent gradients $\{\nabla c_i(x_*) : i \in E \cup I_*^0\}$, lemma 15.11 implies that $\lambda_{I_*^0}$ is close to $(\lambda_*)_{I_*^0} > 0$.

Because of (15.28) and the fact that $c_i(x) + c_i'(x) \cdot d < 0$ for $i \in I \setminus I_*^0$ (see the reasonning after (15.28)), we have also proven the claim (ii) of the theorem. □

15.3 Globalisation de la convergence par recherche linéaire

On écrit le problème quadratique osculateur en l'itéré (x_k, λ_k) de l'itération k comme suit

$$\begin{cases} \min_d \ g_k^{\mathsf{T}} d + \frac{1}{2} d^{\mathsf{T}} M_k d \\ (c_k + A_k d)^{\#} = 0. \end{cases} \tag{15.29}$$

Nous avons abrégé les écritures en posant

$$g_k = \nabla f(x_k), \quad c_k = c(x_k) \quad \text{et} \quad A_k = A(x_k) = c'(x_k).$$

La matrice M_k est soit $L(x_k, \lambda_k)$ ou une approximation de cette hessienne (version quasi-newtonienne). Un point stationnaire d_k de ce problème vérifie, pour un certain multiplicateur $\lambda_k^{\mathrm{PQ}} \in \mathbb{R}^m$, les conditions d'optimalité suivantes

$$\begin{cases} (a) & g_k + M_k d_k + A_k^{\mathsf{T}} \lambda_k^{\mathrm{PQ}} = 0 \\ (b) & (c_k + A_k d_k)^{\#} = 0 \\ (c) & (\lambda_k^{\mathrm{PQ}})_I \geqslant 0 \\ (d) & (\lambda_k^{\mathrm{PQ}})_I^{\mathsf{T}} (c_k + A_k d_k)_I = 0. \end{cases} \tag{15.30}$$

15.3.1 Fonction de mérite

Afin de globaliser la méthode de Newton, on introduit la fonction de mérite (ou fonction de pénalisation) suivante

$$\Theta_\sigma(x) = f(x) + \sigma\|c(x)\|_{\mathrm{P}}, \tag{15.31}$$

où $\|\cdot\|_{\mathrm{P}}$ est une norme sur $\mathbb{R}^m$ et $\sigma > 0$ est un paramètre. On supposera que la norme $\|\cdot\|_{\mathrm{P}}$ vérifie

$$v \mapsto \|v^{\#}\|_{\mathrm{P}} \text{ est convexe.} \tag{15.32}$$

C'est le cas des normes ℓ_p définies par (A.5), avec $p \in [1, +\infty]$. On note $\|\cdot\|_{\mathrm{D}}$, la norme duale de $\|\cdot\|_{\mathrm{P}}$ pour le produit scalaire euclidien (voir (A.8)).

15.3.2 Condition de décroissance de la fonction de mérite

Proposition 15.13 (décroissance de Θ_σ le long de d_k) *Si $(d_k, \lambda_k^{\mathrm{PQ}})$ vérifie les conditions d'optimalité (15.30) et si $\|\cdot\|_{\mathrm{P}}$ vérifie (15.32), alors*

$$\Theta'_\sigma(x_k; d_k) \leqslant g_k^{\mathsf{T}} d_k - \sigma\|c_k^{\#}\|_{\mathrm{P}} = -d_k^{\mathsf{T}} M_k d_k + (\lambda_k^{\mathrm{PQ}})^{\mathsf{T}} c_k - \sigma\|c_k^{\#}\|_{\mathrm{P}}. \tag{15.33}$$

Si, de plus,

$$\sigma \geqslant \|\lambda_k^{\mathrm{PQ}}\|_{\mathrm{D}}, \tag{15.34}$$

alors

$$\Theta'_\sigma(x_k; d_k) \leqslant -d_k^{\mathsf{T}} M_k d_k. \tag{15.35}$$

Donc $\Theta'_\sigma(x_k; d_k) < 0$, si $\sigma \geqslant \|\lambda_k^{\mathrm{PQ}}\|_{\mathrm{D}}$, si M_k est définie positive et si x_k n'est pas un point stationnaire de (P_{EI}).

Démonstration. Puisqu'une norme a des dérivées directionnelles (elle est convexe, donc la proposition 3.15 s'applique) et est lipschitzienne (on utilise sa convexité ou l'inégalité triangulaire), la fonction $v \to \|v^{\#}\|_{\mathrm{P}}$ a des dérivées directionnelles. Pour les calculer, on observe d'abord que, par (15.32) et (15.30)-(b), on a pour $t \in]0, 1[$:

$$\begin{aligned}
\left\|\left[c_k + tA_k d_k\right]^{\#}\right\|_{\mathrm{P}} &= \left\|\left[(1-t)c_k + t(c_k + A_k d_k)\right]^{\#}\right\|_{\mathrm{P}} \\
&\leqslant (1-t)\|c_k^{\#}\|_{\mathrm{P}} + t\left\|\left[c_k + A_k d_k\right]^{\#}\right\|_{\mathrm{P}} \\
&= (1-t)\|c_k^{\#}\|_{\mathrm{P}}.
\end{aligned}$$

Dès lors

$$(\|\cdot^{\#}\|_{\mathrm{P}})'(c_k; A_k d_k) = \lim_{t \downarrow 0} \frac{1}{t}\left(\left\|\left[c_k + tA_k d_k\right]^{\#}\right\|_{\mathrm{P}} - \|c_k^{\#}\|_{\mathrm{P}}\right) \leqslant -\|c_k^{\#}\|_{\mathrm{P}}.$$

Alors, en utilisant successivement (15.30)-(a), (15.30)-(b) et (15.30)-(d), on montre (15.33) :

$$\begin{aligned}\Theta'_\sigma(x_k;d_k) &\leqslant g_k^{\mathsf{T}} d_k - \sigma\|c_k^{\#}\|_{\mathrm{P}} \\ &= -d_k^{\mathsf{T}} M_k d_k - (\lambda_k^{\mathrm{PQ}})^{\mathsf{T}} A_k d_k - \sigma\|c_k^{\#}\|_{\mathrm{P}} \\ &= -d_k^{\mathsf{T}} M_k d_k + (\lambda_k^{\mathrm{PQ}})^{\mathsf{T}} c_k - \sigma\|c_k^{\#}\|_{\mathrm{P}}.\end{aligned}$$

Si $\sigma \geqslant \|\lambda_k^{\mathrm{PQ}}\|_{\mathrm{D}}$, en utilisant (15.30)-$(c)$ et l'inégalité de Cauchy-Schwarz généralisée (A.9), on trouve

$$(\lambda_k^{\mathrm{PQ}})^{\mathsf{T}} c_k - \sigma\|c_k^{\#}\|_{\mathrm{P}} \leqslant (\lambda_k^{\mathrm{PQ}})^{\mathsf{T}} c_k^{\#} - \sigma\|c_k^{\#}\|_{\mathrm{P}} \leqslant (\|\lambda_k^{\mathrm{PQ}}\|_{\mathrm{D}} - \sigma)\|c_k^{\#}\|_{\mathrm{P}} \leqslant 0.$$

Alors, (15.35) se déduit de (15.33). Si $\Theta'_\sigma(x_k;d_k) = 0$ et M_k est définie positive, alors $d_k = 0$. De (15.30), on déduit la stationnarité de x_k, avec λ_k^{PQ} comme multiplicateur associé. □

Le seuil sur σ dans (15.34) rappelle celui qui fait de Θ_σ une fonction de pénalisation exacte (proposition 13.24). Pour satisfaire l'inégalité (15.34), il va falloir modifier σ à certaines itérations (on n'est pas maître de λ_k^{PQ}). On note alors σ_k la valeur de σ à l'itération k et *avant* de faire de la recherche linéaire (RL) on modifie éventuellement l'ancien σ $(= \sigma_{k-1})$ de manière à avoir

$$\sigma_k \geqslant \|\lambda_k^{\mathrm{PQ}}\|_{\mathrm{D}} + \bar{\sigma}, \tag{15.36}$$

où $\bar{\sigma} > 0$ est une « petite » constante (choix heuristique). On utilisera les règles suivantes.

- Première itération. On prend

  $$\bar{\sigma} = \max(\sqrt{\texttt{eps}}, \|\lambda_1\|_{\mathrm{D}}/100) \quad \text{et} \quad \sigma_1 = \|\lambda_1\|_{\mathrm{D}} + \bar{\sigma},$$

  où `eps` est l'epsilon-machine.

- Itérations suivantes. On augmente éventuellement σ de manière à réaliser (15.36), mais on peut aussi le faire décroître s'il s'avère être « beaucoup » trop grand (afin d'éviter la troncature du pas).

  si $\sigma_{k-1} < \|\lambda_k^{\mathrm{PQ}}\|_{\mathrm{D}} + \bar{\sigma}$;
  alors $\sigma_k = \max(1.5\,\sigma_{k-1}, \|\lambda_k^{\mathrm{PQ}}\|_{\mathrm{D}} + \bar{\sigma})$;
  sinon
  si $\sigma_{k-1} > 1.1(\|\lambda_k^{\mathrm{PQ}}\|_{\mathrm{D}} + \bar{\sigma})$;
  alors $\sigma_k = (\sigma_{k-1} + \|\lambda_k^{\mathrm{PQ}}\|_{\mathrm{D}} + \bar{\sigma})/2$;
  sinon $\sigma_k = \sigma_{k-1}$.

Plus rien ne garantit la convergence « théorique » de l'algorithme si l'on fait décroître σ_k comme dans la règle ci-dessus, mais en pratique une telle heuristique peut améliorer grandement l'efficacité de l'algorithme. On ne s'en prive donc pas. Aux analystes d'en trouver une qui n'empêche pas la convergence !

15.4 Versions quasi-newtoniennes ▲

Dans les méthodes de quasi-Newton, on remplace la hessienne du lagrangien L qui intervient dans le problème quadratique osculateur (15.20) par une matrice M,

symétrique *définie positive*, mise à jour par l'algorithme. Dans le cadre de OQS, cette approche a au moins deux intérêts : il ne faut pas calculer de dérivées secondes et le problème quadratique (PQ) est toujours « mieux » posé (il a au plus une solution). Bien que la convergence soit plus lente qu'avec la méthode de Newton (elle n'est plus que superlinéaire) et malgré les difficultés conceptuelles rencontrées par cette approche (voir plus loin), le second avantage cité ci-dessous fait que c'est essentiellement l'approche utilisée dans les codes commerciaux ou de recherche.

La version quasi-newtonienne de OQS génère donc une suite primale-duale $\{(x_k, \lambda_k)\}$ et une suite de matrices symétriques définies positives $\{M_k\}$ de la manière suivante. À l'étape k, on calcule d'abord *la* solution primale-duale $(d_k, \lambda_k^{\text{PQ}})$ du problème quadratique osculateur (en espérant qu'elle existe...)

$$\left\{ \begin{array}{l} \min_{d\in\mathbb{E}} \; g_k^{\mathsf{T}} d + \frac{1}{2} d^{\mathsf{T}} M_k d \\ (c_k + A_k d)^{\#} = 0, \end{array} \right. \tag{15.37}$$

dans lequel M_k joue le rôle de $L_k = L(x_k, \lambda_k) = \nabla^2_{xx}\ell(x_k, \lambda_k)$, la hessienne du lagrangien ℓ, $A_E(x_k)$ et $A_I(x_k)$ sont les jacobiennes des contraintes d'égalité c_E et c_I. Ensuite on prend $x_{k+1} = x_k + d_k$, $\lambda_{k+1} = \lambda_k^{\text{PQ}}$ et on met à jour M_k par la formule de BFGS (c'est la formule la plus utilisée).

La formule de BFGS s'écrit :

$$M_{k+1} = M_k - \frac{M_k \delta_k \delta_k^{\mathsf{T}} M_k}{\delta_k^{\mathsf{T}} M_k \delta_k} + \frac{\gamma_k \gamma_k^{\mathsf{T}}}{\gamma_k^{\mathsf{T}} \delta_k}.$$

Les vecteurs γ_k et δ_k sont déterminés de manière à forcer M_{k+1} à être définie positive (on suppose que M_k l'est) et à être proche de L_{k+1}, ce qui, dans certains cas, peut être contradictoire. Pour cela, on prend pour δ_k le déplacement en x :

$$\delta_k = x_{k+1} - x_k.$$

Le vecteur γ_k devrait idéalement être la variation du gradient du lagrangien

$$\gamma_k^{\ell} = \nabla_x \ell(x_{k+1}, \lambda_{k+1}) - \nabla_x \ell(x_k, \lambda_{k+1}).$$

Mais, pour conserver la définie positivité de M_k, on doit avoir $\gamma_k^{\mathsf{T}} \delta_k > 0$, ce qui n'est pas garanti avec $\gamma_k = \gamma_k^{\ell}$. Si bien que l'on utilisera la *correction de Powell* qui consiste à prendre

$$\gamma_k = \theta_k \gamma_k^{\ell} + (1 - \theta_k) M_k \delta_k,$$

où θ_k est pris maximal dans $]0, 1]$ de manière à avoir $\gamma_k^{\mathsf{T}} \delta_k \geqslant 0.2\, \delta_k^{\mathsf{T}} M_k \delta_k$. On trouve

$$\theta_k = \left\{ \begin{array}{ll} 0.8 \, \dfrac{\delta_k^{\mathsf{T}} M_k \delta_k}{\delta_k^{\mathsf{T}} M_k \delta_k - (\gamma_k^{\ell})^{\mathsf{T}} \delta_k} & \text{si } (\gamma_k^{\ell})^{\mathsf{T}} \delta_k < 0.2 \delta_k^{\mathsf{T}} M_k \delta_k \\ 1 & \text{sinon.} \end{array} \right.$$

Il reste à spécifier la matrice initiale. On prendra $M_1 = I$ (matrice identité) à la première itération. Mais, après le calcul de x_2 et avant le calcul de M_2, on modifie la valeur de M_1 en ηI, où η a une valeur reflétant l'échelle du problème (ou la « valeur moyenne » de L_1) :

$$\eta = \frac{\|\gamma_1\|_2^2}{\gamma_1^{\mathsf{T}} \delta_1}.$$

Il faut en effet attendre que la première itération soit terminée pour évaluer cette grandeur. Puis on calcule M_2 par la formule de BFGS.

15.5 Le diable se cache dans les détails ▲

Si l'on veut qu'un algorithme fonctionne bien, il faut soigner les « détails », qui n'en sont pas en réalité, mais qui ont été passé sous silence dans les développements précédents. De nombreuses questions d'apparence anodine doivent en effet être traitées avec le plus grand soin si l'on veut avoir une méthode numérique efficace, donnant de bons résultats sur des bans d'essai de problèmes-tests.

15.5.1 Incompatibilité des contraintes

Un problème d'optimisation non linéaire peut avoir des contraintes incompatibles, donc ne pas avoir de points admissibles, auquel cas le problème n'a évidemment pas de solution. Il est souhaitable toutefois que l'algorithme puisse détecter une telle situation. Dans l'algorithme OQS, cette situation se manifestera par l'intermédiaire du problème quadratique osculateur (PQO) qui pourra lui aussi présenter des contraintes linéaires incompatibles. Le lien entre l'incompatibilité des contraintes du problème non linéaire et celle des PQOs est toutefois complexe à étudier et on se contentera ici de présenter les méthodes qui ont été proposées pour traiter les PQOs inconsistants.

Méthodes possibles :

- La *pénalisation exacte des contraintes linéarisées* [211 ; 1982] consiste à remplacer le PQO par sa version pénalisée exacte suivante

$$\begin{cases} \min_d \nabla f(x)^\mathsf{T} d + \frac{1}{2} d^\mathsf{T} H d + \sigma \, \|(c(x) + A(x)d)^\# \|_1 \\ \|d\|_\infty \leqslant \Delta, \end{cases} \tag{15.38}$$

  dans laquelle $H \simeq \nabla^2_{xx} \ell(x, \lambda)$, $\sigma > 0$ est pris assez grand et la contrainte joue le rôle de région de confiance. Ce problème a toujours une solution, mais il est *non lisse* ; il peut se récrire sous la forme d'un problème quadratique standard, pourvu que l'on introduise des variables auxiliaires, en faisant passer le terme normé du critère en contrainte (exercice 1.8). Si $H \succ 0$, le direction d est de descente pour la fonction de mérite (15.31).

  – Avantage de l'approche : robustesse.

  – Inconvénient de l'approche : pas de solveur disponible pour résoudre un problème sous la forme (15.38); la transformation de (15.38) en problème quadratique standard introduit des variables auxiliaires.

- Le *mode élastique* [250 ; 2002] consiste à relaxer le PQO comme suit

$$\begin{cases} \min_{(d,v,w)} \nabla f(x)^\mathsf{T} d + \frac{1}{2} d^\mathsf{T} H d + \sigma \sum_{i \in E \cup I} (v_i + w_i) \\ c_E(x) + A_E(x)d - v_E + w_E = 0 \\ l \leqslant c_I(x) + A_I(x)d - v_I + w_I \leqslant u \\ v_I \geqslant 0 \\ w_I \geqslant 0, \end{cases}$$

  dans lequel $v = (v_E, v_I)$ et $w = (w_E, w_I)$ assurent la compatibilité des nouvelles contraintes. Par le terme de pénalisation linéaire $\sigma \sum_i (v_i + w_i)$, on essaye d'annuler (v, w) pour que, si possible, d soit une solution du PQO initial.

- Certaines approches algorithmiques de résolution du PQO, comme celle du lagrangien augmenté (section 13.4), calculent une solution de

$$\begin{cases} \min_d \nabla f(x)^\mathsf{T} d + \frac{1}{2} d^\mathsf{T} H d \\ c_E(x) + A_E(x)d + \bar{s}_E(x) = 0 \\ l \leqslant c_I(x) + A_I(x)d + \bar{s}_I(x) \leqslant u, \end{cases}$$

où $\bar{s}(x) = (\bar{s}_E(x), \bar{s}_I(x))$ est le vecteur de norme euclidienne minimale rendant les contraintes du PQO compatibles [122 ; 2016]. La direction d ainsi calculée est de descente pour Θ_σ, pourvu que σ soit assez grand et $H \succcurlyeq 0$.

15.5.2 Troncature du pas : l'effet Maratos

15.5.3 Problèmes de commande optimale

Du point de vue de l'optimisation, un problème de commande optimale discrétisé se présente sous la forme

$$(P_{SI}) \quad \begin{cases} \min\ f(x) \\ c_S(x) = 0 \\ c_I(x) \leqslant 0. \end{cases}$$

Il ressemble très fort au problème (P_{EI}), si ce n'est que la fonction $c_S : \mathbb{R} \to \mathbb{R}^{m_S}$, qui remplace la fonction c_E, a des propriétés particulières. Il peut d'ailleurs y avoir en plus des contraintes d'égalité additionnelles $c_E(x) = 0$ sans ces propriétés particulières; nous les avons omises pour alléger l'exposé. Voici la structure apportée par l'équation $c_S(x) = 0$, dite *équation d'état*.

La variable $x = (y, u)$ est partitionnée en *variable d'état* $y \in \mathbb{R}^{m_S}$ et en *variable de commande* $u \in \mathbb{R}^{n-m_S}$. On partitionne de la même manière la jacobienne de c_S :

$$c'_S(x) = \begin{pmatrix} B(x) & N(x) \end{pmatrix},$$

avec une matrice $B(x)$ carrée d'ordre m_S. L'hypothèse-clé est de supposer que $B(x)$ est inversible en tout point rencontré, en particulier en la solution (et donc dans son voisinage). Cette hypothèse permet, dans certaines approches algorithmiques, de représenter l'état comme fonction implicite de la commande et donc d'éliminer l'état du problème, ce qui peut représenter une réduction importante de la dimension du problème (lorsque $n - m_S \ll n$). Ce n'est cependant pas ce point de vue que nous allons présenter ici. Notre but est de montrer que l'algorithme OQS, dans une version dite *réduite*, permet d'avoir le même gain en dimension et donc en nombre d'opérations, tout en évitant la nécessité souvent coûteuse d'avoir des itérés qui satisfont l'équation d'état. *L'approche est surtout avantageuse lorsqu'il n'y a que des contraintes d'inégalité sur la commande*, autrement dit lorsque $\partial c_I / \partial y \equiv 0$.

Le problème quadratique osculateur (PQO) associé à (P_{SI}) en (x_k, λ_k) s'écrit

$$\begin{cases} \min_d\ g_k^\mathsf{T} d + \frac{1}{2} d^\mathsf{T} L_k d \\ c_S(x_k) + c'_S(x_k)d = 0 \\ c_I(x_k) + c'_I(x_k)d \leqslant 0. \end{cases} \tag{15.39}$$

L'hypothèse d'inversibilité de $B_k = B(x_k)$ permet d'introduire un inverse à droite A_k^- de $A_k \equiv c'_S(x_k)$ et une matrice Z_k^- dont les colonnes forment une base du noyau de A_k :

$$A_k^- = \begin{pmatrix} B_k^{-1} \\ 0 \end{pmatrix} \quad \text{et} \quad Z_k^- = \begin{pmatrix} -B_k^{-1} N_k \\ I_{n-m} \end{pmatrix}.$$

Ces matrices ne peuvent en général pas être calculées explicitement, mais on s'autorise à les appliquer à un vecteur, ce qui requiert à chaque fois la résolution d'un système linéaire. Alors toute solution de l'équation d'état linéarisée est de la forme

$$d_k = r_k + Z_k^- h_k,$$

où

$$r_k = -A_k^- c_S(x_k) \in \mathbb{E}$$

est un pas de restauration de l'équation d'état à commande fixée et $h_k \in \mathbb{R}^{n-m_S}$ est à déterminer. Si on reporte cette structure de d_k dans (15.39), le PQO devient

$$\begin{cases} \min_h \ (g_k + L_k r_k)^\mathsf{T} Z_k^- h + \frac{1}{2} h^\mathsf{T} Z_k^{-\mathsf{T}} L_k Z_k^- h \\ c_I(x_k) + c_I'(x_k) r_k + c_I'(x_k) Z_k^- h \leqslant 0. \end{cases}$$

Ce problème se simplifie considérablement si l'on peut faire disparaître la matrice Z_k^-. Dans ce but, on suit les étapes suivantes :

- on approche $Z_k^{-\mathsf{T}} L_k Z_k^-$ par une matrice M_k générée par une technique quasi-newtonienne,
- ne pouvant plus calculer le terme $L_k r_k$ dans la partie linéaire du critère (car L_k n'est ni calculé ni approché), on le néglige,
- il faut par ailleurs supposer que la matrice $c_I'(x_k) Z_k^-$ est simple à calculer.

Les deux premières étapes définissent ce que l'on appelle les *méthodes de quasi-Newton réduites*. L'abandon du terme $L_k r_k$ fait perdre la convergence superlinéaire, mais on garde toutefois la *convergence superlinéaire en* 2 *pas*, c'est-à-dire

$$\frac{\|x_{k+2} - x_*\|}{\|x_k - x_*\|} \to 0.$$

Le troisième point ci-dessus sera certainement satisfait s'il y a peu de contraintes d'inégalité ou si celles-ci portent uniquement sur les variables de commande (alors $c_I'(x_k) Z_k^- = (\partial c_I / \partial u)(x_k)$). Dans ce dernier cas, le PQO devient particulièrement simple

$$\begin{cases} \min_h \ g_k^\mathsf{T} Z_k^- h + \frac{1}{2} h^\mathsf{T} M_k h \\ c_I(x_k) + c_I'(x_k) r + (\partial c_I / \partial u)(x_k) h \leqslant 0. \end{cases}$$

Le vecteur $Z_k^{-\mathsf{T}} g_k = -N_k^\mathsf{T} B_k^{-\mathsf{T}} \nabla_y f(x_k) + \nabla_u f(x_k)$ est appelé le *gradient réduit*. On y reconnaît l'*état adjoint* $B_k^{-\mathsf{T}} \nabla_y f(x_k)$.

Notes

La section **??** ne fait que guigner sur les inclusions fonctionnelles en espérant inciter le lecteur à davantage s'intéresser davantage à ce vaste sujet [173].

On a mis longtemps à mettre au point l'algorithme de Newton pour résoudre les problèmes d'optimisation avec contraintes d'égalité et d'inégalité. Ce n'est, en

effet, qu'au milieu des années 1970 que cette recherche a abouti, soit près de 30 ans après l'invention de l'algorithme du simplexe (chapitre 17). C'est d'ailleurs ce dernier que l'on a d'abord essayé de généraliser à l'optimisation non linéaire [621], mais la voie n'était pas directe. Les numériciens ont ensuite développé les méthodes de pénalisation (chapitre 13), en particulier l'approche par lagrangien augmenté. La prise de conscience de l'existence d'une méthode newtonienne directe n'est venue qu'ensuite [500, 302, 303, 491].

Pour un état de l'art sur l'algorithme OQS, on pourra consulter par exemple la section 5 de [272 ; 2005], la partie III de [73 ; 2006], Fletcher [214 ; 2010] et l'ouvrage technique et approfondi d'Izmailov et Solodov [330 ; 2014]. Voir aussi Izmailov, Kurennoy et Solodov [329 ; 2012]. La dérivation du résultat de convergence locale de l'algorithme OQS (section 15.1) à partir de celui de Josephy-Newton est reprise de Bonnans [72 ; 1994]. Pour l'établissement de conditions kantorovitchéennes de convergence, on pourra consulter Argyros et Hilout [19 ; 2010].

Exercices

15.1. *Non convergence locale de l'algorithme* (15.16). Écrire l'algorithme dont les itérés sont calculés par la récurrence $x_{k+1} = x_k + d_k$, avec d_k solution de (15.16), sous la forme $x_{k+1} = \Phi(x_k)$, où $\Phi : \mathbb{E} \to \mathbb{E}$. On supposera que x_k est voisin d'une solution x_*, que $c'(x_*)$ est surjective et que $\nabla^2 f(x_*)$ est inversible. Montrez que le spectre de $\Phi'(x_*)$ n'est pas dans la boule unité ouverte pour l'exemple en $x = (x_1, x_2) \in \mathbb{R}^2$ suivant :

$$\begin{cases} \min_x \ -ax_1^2 + 2x_2 \\ x_1^2 + x_2^2 = 1, \end{cases}$$

dans lequel $a \in]0, 1[$.

16 Méthodes de points intérieurs

At the same time, we find that the nature of the [*interior-point*] *methods, is in fact, independent of the specific properties of linear programming problems, so that these methods can be extended onto more general convex programs.*

Y.E. Nesterov, A.S. Nemirovskii (1994). [456].

16.1 Méthodes de points intérieurs en optimisation convexe

16.1.1 Vue d'ensemble

Soit $\mathbb{E}$ un espace vectoriel de dimension finie ($\geqslant 1$), muni d'un produit scalaire $\langle \cdot, \cdot \rangle$, dit de référence, et de sa norme associée $\| \cdot \|$. Dans cette section, nous nous intéressons à la résolution numérique de problèmes d'optimisation convexes, que l'on supposera écrits sous la forme

$$(P) \quad \begin{cases} \inf \langle c, x \rangle \\ x \in X. \end{cases} \tag{16.1}$$

Ici $c \in \mathbb{E}$ et X est un convexe de $\mathbb{E}$. Le fait que le critère soit linéaire n'introduit pas de perte de généralité, car on peut toujours *faire passer en contrainte* un critère, qui serait ici convexe, comme à la proposition 1.13.

La théorie que nous présentons ci-dessous, principalement développée par Nesterov et Nemirovskii [455 ; 1993], peut être vue comme une généralisation des méthodes de points intérieurs que nous verrons sous des formes un peu différentes dans les cas particuliers de l'optimisation linéaire (chapitre 18) et de l'optimisation semi-définie positive (chapitre 20). Chronologiquement, elle a été construite après la découverte de ces dernières, dans un but évident de généralisation et d'extension de son domaine d'application. C'est aussi pour cette raison que nous la présentons dans cette partie décrivant les méthodes générales de l'optimisation avec contraintes. Le lecteur pourra d'ailleurs étudier les chapitres 18 et 20, avant d'aborder cette section, mais rien n'impose un tel ordre de lecture.

Cette théorie est remarquable, puisqu'elle conduit au résultat impressionnant selon lequel tout problème d'optimisation convexe, écrit sous la forme ci-dessus, peut être résolu avec une précision arbitraire donnée en un nombre polynomial d'itérations ne

dépendant de la précision demandée que par son logarithme. La dépendance polynomiale ne se fait pas directement par rapport à la dimension de $\mathbb{E}$, car la structure de l'ensemble admissible X intervient également, mais par l'intermédiaire d'un module de complexité ϑ_f, dont la dépendance par rapport à $\dim \mathbb{E}$ pourra parfois être explicitée. Cela paraît très attrayant. Il faut toutefois que l'on dispose d'une barrière autoconcordante f dont le domaine est l'intérieur de l'ensemble admissible X pour que ce résultat ait lieu. Même si l'on peut montrer qu'une telle fonction existe toujours, on ne dispose pas nécessairement d'une forme analytique de celle-ci, si bien que le résultat peut ne pas avoir une utilité pratique immédiate. On connaît cependant de grandes classes de problèmes auxquelles sont associées des fonctions autoconcordantes : l'optimisation linéaire (chapitre 18) et l'optimisation SDP (chapitre 20). Mais on connait aussi des problèmes convexes qui sont NP-ardus, les problèmes d'*optimisation copositive* [69 ; 2012] ; donc pour lesquels l'évaluation d'une barrière auto-concoirdante ne pourra se faire en temps polynomial que si $\mathsf{P} = \mathsf{NP}$.

Nous introduisons d'abord, à la section 16.1.2, la notion de fonction auto-concordante et en donnons les propriétés utiles à l'étude des algorithmes décrits. Nous présentons ensuite, à la section 16.1.3, les barrières autoconcordantes, qui sont des fonctions autoconcordantes pour lesquelles la direction de Newton est bornée (pour une métrique particulière). Ce sont ces fonctions qui permettent d'obtenir des résultats de complexité itérative polynomiale des méthodes de points intérieurs. Nous concluons la section par trois algorithmes typiques et en démontrons leur complexité polynomiale (section 16.1.4). On peut accéder à l'étude d'un algorithme polynomialement convergent avec un minimum d'effort préalable : il suffit de définir les notions de fonction autoconcordante (définition 16.1 et définitions équivalentes de la proposition 16.5) et de barrière autoconcordante (définition 16.18), d'étudier les propositions 16.8 et 16.11 et d'établir l'inégalité de droite dans (16.15) et celle de la proposition 16.28; on a alors toute l'information nécessaire à la compréhension et à l'analyse du premier algorithme présenté à la section 16.1.4.

La seconde partie de ce chapitre s'intéresse au cas des problèmes non linéaires et non convexes. Il n'y a alors plus de résultat de complexité polynomiale. Après avoir présenté un schéma algorithmique parfois implémenté, nous en étudions le comportement asymptotique.

16.1.2 Fonction autoconcordante

On verra aux chapitres 18 et 20, que la fonction logarithmique joue un rôle important dans la conception des algorithmes de points intérieurs et dans l'étude de leur convergence polynomiale. Les fonctions autoconcordantes sont celles qui ont les propriétés du logarithme qui y jouent un rôle-clé, avec l'intérêt supplémentaire de pouvoir être définie sur des ouverts convexes plus généraux que l'intérieur de l'orthant positif (exemple 16.2) ou du cône des matrices semi-définies positives (exemple 16.3).

Les fonctions autoconcordantes $f : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$, dont l'adhérence du domaine définira l'ensemble admissible de (P), devront au moins vérifier les propriétés suivantes :

$$\begin{cases} \operatorname{dom} f \text{ est un ouvert convexe non vide,} \\ f \text{ est } C^2 \text{ sur son domaine et} \\ \nabla^2 f(x) \succ 0 \text{ pour tout } x \in \operatorname{dom} f. \end{cases} \tag{16.2}$$

Évidemment, de telles fonctions sont strictement convexes. On note

$$g(x) := \nabla f(x) \qquad \text{et} \qquad H(x) := \nabla^2 f(x)$$

le gradient et hessienne de f en $x \in \operatorname{dom} f$ pour le produit scalaire de référence.

Métrique riemannienne associée à une fonction vérifiant (16.2)

On peut introduire sur le domaine d'une fonction $f : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ vérifiant (16.2) une *métrique riemannienne* en se donnant en tout point $x \in \operatorname{dom} f$ un produit scalaire et sa norme associée

$$\langle u, v\rangle_x := f''(x) \cdot (u, v) = \langle H(x)u, v\rangle \qquad \text{et} \qquad \|v\|_x := \langle v, v\rangle_x^{1/2}.$$

Cette métrique ne dépend pas du produit scalaire de référence, puisqu'elle est définie à partir de la donnée de f seule. On utilise aussi la notation $\|\cdot\|_x$ pour la norme d'un opérateur linéaire sur $\mathbb{E}$, subordonnée à la norme vectorielle définie ci-dessus. On note

$$B_x(y, 1) := \{y + v \in \mathbb{E} : \langle H(x)v, v\rangle < 1\}$$
$$\bar{B}_x(y, 1) := \{y + v \in \mathbb{E} : \langle H(x)v, v\rangle \leqslant 1\}$$

les boules ouverte et fermée de centre $y \in \mathbb{E}$ et de rayon 1 pour la métrique $\langle\cdot,\cdot\rangle_x$. On peut les voir comme des *ellipsoïdes* centrés en y.

Les gradient et hessienne de f pour la métrique $\langle\cdot,\cdot\rangle_x$ sont notés g_x et H_x et valent en $y \in \operatorname{dom} f$:

$$g_x(y) = H(x)^{-1}g(y) \qquad \text{et} \qquad H_x(y) = H(x)^{-1}H(y).$$

En particulier $g_x(x) = -d(x)$ est l'opposé de la direction de Newton en x pour minimiser f et $H_x(x) = I$. On notera que $H_x(y)$ est auto-adjoint pour le produit scalaire $\langle\cdot,\cdot\rangle_x$.

Il est clair que, pour v fixé, $\|v\|_x$ varie continûment avec x (car f est C^2), mais on a un peu mieux : *pour tout* $\varepsilon \in]0, 1[$, *il existe un voisinage* V *de* $x \in \operatorname{dom} f$, *tel que pour tout* $y \in V$ *et tout* $v \in \mathbb{E}$ *non nul, on a*

$$1 - \varepsilon \leqslant \frac{\|v\|_y}{\|v\|_x} \leqslant \frac{1}{1-\varepsilon}. \tag{16.3}$$

En effet, pour V suffisamment petit, on a

$$\begin{aligned}
\|v\|_y^2 &= \|v\|_x^2 + \langle (H(x)^{-1/2}H(y)H(x)^{-1/2} - I)H(x)^{1/2}v, H(x)^{1/2}v\rangle \\
&\leqslant \|v\|_x^2 + \|H(x)^{-1/2}H(y)H(x)^{-1/2} - I\|\,\|v\|_x^2 \\
&\leqslant (1+\varepsilon)\|v\|_x^2 \\
&\leqslant (1-\varepsilon)^{-2}\|v\|_x^2,
\end{aligned}$$

car $1 + \varepsilon \leqslant (1 + \varepsilon)^2 \leqslant (1 - \varepsilon)^{-2}$. On en déduit l'inégalité de droite dans (16.3). En inversant x et y, on a

$$\|v\|_x^2 \leqslant \|v\|_y^2 + \|H(y)^{-1/2}H(x)H(y)^{-1/2} - I\|\,\|v\|_y^2 \leqslant (1-\varepsilon)^{-2}\|v\|_y^2.$$

On en déduit l'inégalité de gauche dans (16.3).

On aurait pu mettre $1 + \varepsilon$ à droite dans (16.3) au lieu de $(1 - \varepsilon)^{-1}$. Nous avons choisi ce dernier majorant car cela nous permet d'introduire de façon un peu plus naturelle la notion de fonction autoconcordante, laquelle vérifie cette propriété avec $V = B_x(x, 1)$ et $\varepsilon = \|y - x\|_x$.

Fonction autoconcordante

Définition 16.1 On dit qu'une fonction $f : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ est *autoconcordante*, si elle vérifie les conditions (16.2) et si les deux propriétés suivantes ont lieu :

(AC1) $\forall x \in \operatorname{dom} f$, $B_x(x,1) \subseteq \operatorname{dom} f$,
(AC2) $\forall x \in \operatorname{dom} f$, $\forall y \in B_x(x,1)$, $\forall v \in \mathbb{E} \setminus \{0\}$, on a

$$1 - \|y-x\|_x \leqslant \frac{\|v\|_y}{\|v\|_x} \leqslant \frac{1}{1-\|y-x\|_x}. \tag{16.4}$$

On note $\mathrm{AC}(\mathbb{E})$ l'ensemble des fonctions autoconcordantes sur $\mathbb{E}$. □

Les inégalités (16.4) sont remarquables si on les compare à (16.3) : aucune constante indéterminée n'est utilisée et le voisinage $B_x(x,1)$ sur lequel elles ont lieu est bien précisé. Cette absence de constante inconnue est une caractéristique que l'on s'attachera à préserver dans toutes les propriétés des fonctions autoconcordantes et qui contribuera à l'obtention des résultats de polynomialité.

La condition (AC1) donne à une fonction autoconcordante f une allure proche de la frontière de son domaine qui rappelle celle de l'opposé du logarithme au voisinage de 0 (proposition 16.4). En particulier, $f(x)$ tend vers $+\infty$ lorsque x tend vers un point de $\partial(\operatorname{dom} f)$. La condition (AC2) exprime à sa manière le caractère lipschitzien de la hessienne de f (proposition 16.5). Cette propriété permet en particulier de comparer deux directions de Newton successives (proposition 16.11), ce qui sera important dans le contrôle des itérés dans les méthodes de points intérieurs.

Les fonctions autoconcordantes existent ! En voici deux exemples emblématiques, qui seront respectivement utilisées en optimisation linéaire (chapitre 18) et en optimisation semi-définie positive (chapitre 20). Le logarithme y joue un rôle important. On peut d'ailleurs voir l'autoconcordante comme un moyen d'introduire sur un ensemble convexe quelconque une fonction ayant les propriétés particulières de l'opposé du logarithme sur $\mathbb{R}_{++}$, qui sont décisives pour l'obtention des résultats de polynomialité.

Exemple 16.2 La *fonction log-barrière* $\mathrm{lb} : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ définie en $x \in \mathbb{R}^n$ par

$$\mathrm{lb}(x) = \begin{cases} -\sum_{i=1}^n \log x_i & \text{si } x > 0 \\ +\infty & \text{sinon} \end{cases} \tag{16.5}$$

est autoconcordante (voir l'exercice 16.1 pour une définition géométrique de cette fonction).

Démonstration. On munit $\mathbb{R}^n$ du produit scalaire euclidien et on note $X = \mathrm{Diag}(x_1, \ldots, x_n)$ et e le vecteur dont les composantes valent 1. Alors, pour $x \in \operatorname{dom} \mathrm{lb} = \mathbb{R}^n_{++}$, $g(x) = -X^{-1}e$ et $H(x) = X^{-2}$. La fonction lb vérifie donc bien les hypothèses (16.2).

Ensuite, si $x > 0$ et $y \in B_x(x,1)$, on a $\sum_i (y_i - x_i)^2/x_i^2 < 1$, ce qui implique que $|y_i - x_i| < x_i$ pour tout i, ou encore $y_i > 0$; donc $y \in \operatorname{dom} \mathrm{lb}$.

D'autre part, pour x et $y \in \operatorname{dom} \mathrm{lb}$,

$$\|v\|_y^2 = \sum_{i=1}^n \left(\frac{v_i}{y_i}\right)^2 = \sum_{i=1}^n \left(\frac{v_i}{x_i}\right)^2 \left(\frac{x_i}{y_i}\right)^2 \leqslant \|v\|_x^2 \max_{1\leqslant i\leqslant n} \left(\frac{x_i}{y_i}\right)^2.$$

Mais

$$\frac{y_i}{x_i} \geqslant 1 - \left|\frac{y_i}{x_i} - 1\right| \geqslant 1 - \|y - x\|_x > 0.$$

On en déduit l'inégalité à droite dans (16.4). Pour obtenir l'inégalité de gauche, on écrit

$$\|v\|_y^2 \geqslant \|v\|_x^2 \min_{1 \leqslant i \leqslant n} \left(\frac{x_i}{y_i}\right)^2$$

et on observe que

$$\frac{y_i}{x_i} \leqslant 1 + \left|\frac{y_i}{x_i} - 1\right| \leqslant 1 + \|y - x\|_x \leqslant \frac{1}{1 - \|y - x\|_x}.$$

□

Exemple 16.3 La *fonction log-déterminant* $\mathrm{ld} : \mathcal{S}^n \to \mathbb{R} \cup \{+\infty\}$ définie en $X \in \mathcal{S}^n$ par

$$\mathrm{ld}(X) = \begin{cases} -\log\det X & \text{si } X \in \mathcal{S}^n_{++} \\ +\infty & \text{sinon} \end{cases} \tag{16.6}$$

est autoconcordante.

Démonstration. On munit $\mathcal{S}^n$ du produit scalaire $\langle A, B\rangle = \operatorname{tr} AB$. Le domaine de ld est $\mathcal{S}^n_{++}$. C'est bien un ouvert et sur celui-ci, ld est C^∞. En se rappelant qu'en X inversible $\det'(X) \cdot H = (\det X)(\operatorname{tr} X^{-1}H)$, on trouve que $\mathrm{ld}'(X) \cdot H = -\langle X^{-1}, H\rangle$ et que $\mathrm{ld}''(X) \cdot H^2 = \langle X^{-1}HX^{-1}, H\rangle = \|X^{-1/2}HX^{-1/2}\|^2$, qui est strictement positif pour H non nul. La fonction ld vérifie donc bien les hypothèses (16.2).

Soient à présent $X \succ 0$ et $Y \in B_X(X, 1)$. Alors

$$\begin{aligned} \|Y - X\|_X^2 &= \mathrm{ld}''(X) \cdot (Y - X)^2 \\ &= \operatorname{tr} X^{-1}(Y - X)X^{-1}(Y - X) \\ &= \operatorname{tr}(X^{-1/2}YX^{-1/2} - I)(X^{-1/2}YX^{-1/2} - I) \\ &= \sum_{i=1}^n (1 - \lambda_i)^2, \end{aligned}$$

où les λ_i sont les valeurs propres de $X^{-1/2}YX^{-1/2}$. Comme $\|Y - X\|_X < 1$, tous les $(1 - \lambda_i)^2 < 1$ et donc toutes les valeurs propres $\lambda_i > 0$. Dès lors $X^{-1/2}YX^{-1/2} \succ 0$ ou encore $Y \succ 0$.

Il reste à montrer (16.4). Soient $X \succ 0$, $Y \in B_X(X, 1)$ et $V \in \mathcal{S}^n$. On note

$$\bar{Y} := X^{-1/2}YX^{-1/2} \quad \text{et} \quad \bar{V} := X^{-1/2}VX^{-1/2}.$$

Les valeurs propres de $\bar{Y}$ sont notées $0 < \lambda_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \lambda_n$ (on a vu que $Y \succ 0$ et que $\bar{Y} \succ 0$) et sa factorisation spectrale est de la forme $\bar{Y} = Q\Lambda Q^\mathsf{T}$, où Q est orthogonale et $\Lambda := \operatorname{Diag}(\lambda_1, \ldots, \lambda_n)$. On a vu que

$$\|V\|_X^2 = \|\bar{V}\|^2$$

et on calcule facilement

$$\begin{aligned}
\|V\|_Y^2 &= \operatorname{tr} Y^{-1}VY^{-1}V \\
&= \operatorname{tr}(X^{1/2}Y^{-1}X^{1/2})(X^{-1/2}VX^{-1/2})(X^{1/2}Y^{-1}X^{1/2})(X^{-1/2}VX^{-1/2}) \\
&= \operatorname{tr}\bar{Y}^{-1}\bar{V}\bar{Y}^{-1}\bar{V} \\
&= \operatorname{tr}(\bar{Y}^{-1/2}\bar{V}\bar{Y}^{-1/2})^2 \\
&= \operatorname{tr}(\bar{\Lambda}^{-1/2}Q^{\mathsf{T}}\bar{V}Q\bar{\Lambda}^{-1/2})^2 \\
&= \sum_{i,j} \frac{1}{\lambda_i}(Q^{\mathsf{T}}\bar{V}Q)_{ij}^2 \frac{1}{\lambda_j} \qquad (16.7) \\
&\leqslant \frac{1}{\lambda_1^2}\|Q^{\mathsf{T}}\bar{V}Q\|^2 \\
&= \frac{1}{\lambda_1^2}\|\bar{V}\|^2 \\
&= \frac{1}{\lambda_1^2}\|V\|_X^2
\end{aligned}$$

On obtient l'inégalité de droite dans (16.4) en observant que

$$\lambda_1 \geqslant 1 - |\lambda_1 - 1| \geqslant 1 - \left(\textstyle\sum_i (\lambda_i - 1)^2\right)^{1/2} = 1 - \|Y - X\|_X.$$

L'inégalité de gauche dans (16.4) s'obtient de la même manière. On minore $\|V\|_Y^2$ ci-dessus à partir de (16.7) pour obtenir $\|Y\|_Y^2 \geqslant \|V\|_X^2/\lambda_n^2$ et on utilise

$$\lambda_n \leqslant 1 + |\lambda_n - 1| \leqslant 1 + \left(\textstyle\sum_i (\lambda_i - 1)^2\right)^{1/2} = 1 + \|Y - X\|_X \leqslant \frac{1}{1 - \|Y - X\|_X}.$$

□

L'hypothèse (AC1) porte sur la courbure de f qui doit être assez importante en un point x proche de $\partial(\operatorname{dom} f)$, la frontière du domaine de f. En exploitant cette propriété, on peut déduire le résultat de la proposition suivante selon laquelle, le long d'une droite rencontrant $\partial(\operatorname{dom} f)$, une fonction autoconcordante croît au moins autant que l'opposé du logarithme dans le voisinage de 0. On en déduit que $f(x)$ tend vers l'infini lorsque x tend vers un point de $\partial(\operatorname{dom} f)$.

Proposition 16.4 (croissance proche de la frontière du domaine) *Soient $x \in \operatorname{dom} f$, $\bar{x} \in \partial(\operatorname{dom} f)$ et $x_t = (1-t)\bar{x} + tx$ avec $t \in]0,1]$. Alors*

$$f(x_t) \geqslant f(x) + f'(x) \cdot (x_t - x) - \log t + t - 1. \qquad (16.8)$$

Dès lors, si une suite $\{x_k\} \subseteq \operatorname{dom} f$ converge vers un point de $\partial(\operatorname{dom} f)$, alors $f(x_k) \to +\infty$.

Démonstration. Comme $\bar{x} = x_t + (\bar{x} - x_t) \notin \operatorname{dom} f$, on a $\|\bar{x} - x_t\|_{x_t} \geqslant 1$ (par la propriété (AC1)), ce qui s'écrit $f''(x_t) \cdot (\bar{x} - x_t)^2 \geqslant 1$ ou encore

$$f''(x_t) \cdot (\bar{x} - x)^2 \geqslant \frac{1}{t^2}. \qquad (16.9)$$

Pour $t \in]0,1]$, le développement de Taylor d'ordre 2 avec reste intégral s'écrit

$$f(x_t) = f(x) + f'(x)\cdot(x_t{-}x) + \int_0^1 (1{-}\alpha)\, f''(x+\alpha(x_t{-}x))\cdot(x_t{-}x)^2\,\mathrm{d}\alpha.$$

En utilisant (16.9), on a

$$f''(x+\alpha(x_t{-}x))\cdot(x_t{-}x)^2 = (1{-}t)^2\, f''(x_{1-\alpha(1-t)})\cdot(\bar{x}-x)^2 \geqslant \frac{(1{-}t)^2}{(1-\alpha(1{-}t))^2}.$$

On a aussi en posant $\beta = 1-\alpha(1{-}t)$:

$$\int_0^1 \frac{(1{-}t)^2(1{-}\alpha)}{(1-\alpha(1{-}t))^2}\,\mathrm{d}\alpha = \int_1^t \frac{t-\beta}{\beta^2}\,\mathrm{d}\beta = -1 + t - \log t.$$

En regroupant les trois dernières relations, on trouve (16.8).

Venons-en à la seconde partie de la proposition : une suite $\{x_k\} \subseteq \operatorname{dom} f$ converge vers un point de $\partial(\operatorname{dom} f)$. Soit $x \in \operatorname{dom} f$, qui est un ouvert. On peut donc trouver $r > 0$ tel que $B(x,r) \subseteq \operatorname{dom} f$. Dès que $x_k \in B(\bar{x},r)$, $x + t(x_k{-}x) \notin \operatorname{dom} f$ pour $t > 0$ assez grand. En effet, dans le cas contraire, par convexité du domaine de f, on aurait $B(x_k,r) \subseteq \operatorname{dom} f$, ce qui n'est pas possible puisque $\bar{x}$ appartiendrait à cette boule sans appartenir à $\operatorname{dom} f$. Pour k assez grand, on peut donc trouver un $\bar{x}_k \in \partial(\operatorname{dom} f)$ tel que $x_k = (1{-}t_k)\bar{x}_k + t_k x$, avec $0 < t_k \leqslant 1$. Alors, par convexité de $\operatorname{dom} f$, $B(x_k, t_k r) \subseteq \operatorname{dom} f$. Donc $t_k \leqslant \|x_k - \bar{x}\|/r$ et $t_k \downarrow 0$ lorsque $k \to \infty$. En utilisant l'inégalité (16.8), on trouve

$$f(x_k) \geqslant f(x) + f'(x)\cdot(x_k{-}x) - 1 + t_k - \log t_k.$$

Comme $x_k \to \bar{x}$ et $t_k \downarrow 0$, on trouve que $f(x_k) \to +\infty$ lorsque $k \to \infty$. □

Ce résultat implique qu'une fonction autoconcordante est fermée :

$$\mathrm{AC}(\mathbb{E}) \subseteq \mathrm{C}\overline{\mathrm{onv}}(\mathbb{E}). \tag{16.10}$$

En particulier, f a un minimum si son domaine est borné.

Si on fixe v dans les inégalités de (16.4), on voit que cette propriété essentielle des fonctions autoconcordantes exprime à sa manière que la hessienne $H(x)$ ne varie pas trop vite avec x. La proposition ci-dessous donne une autre expression de cette propriété en faisant disparaître v. Ainsi, en se rappelant que $H_x(x) = I$, on voit que la première inégalité dans (16.12) traduit une condition de Lipschitz radiale locale sur la hessienne (x est fixé et y doit être voisin de x), si bien qu'une fonction autoconcordante est un peu plus régulière que C^2.

Proposition 16.5 (hessienne lipschitzienne) *Dans la définition d'une fonction autoconcordante, on peut remplacer* (AC2) *par l'une des conditions équivalentes suivantes :*

$$\forall x \in \operatorname{dom} f, \quad \forall y \in B_x(x,1), \\ \|H_x(y)\|_x,\ \|H_x(y)^{-1}\|_x \leqslant \frac{1}{(1-\|y-x\|_x)^2} \tag{16.11}$$

ou

$$\forall x \in \operatorname{dom} f, \quad \forall y \in B_x(x,1), \\ \|H_x(y)-I\|_x,\ \|H_x(y)^{-1}-I\|_x \leqslant \frac{1}{(1-\|y-x\|_x)^2}-1. \tag{16.12}$$

DÉMONSTRATION. On note $0<\lambda_1\leqslant\cdots\leqslant\lambda_n$ les valeurs propres de $H_x(y)$.

[(AC2) $\Leftrightarrow$ (16.11)] Les inégalités dans (16.11) s'obtiennent en prenant le supremum et l'infimum en $v\neq 0$ du quotient $\|v\|_y/\|v\|_x$ dans (16.4). En utilisant le fait que $H_x(y)$ est auto-adjointe et définie positive pour le produit scalaire local en x, on obtient (exercice 4.12) :

$$\max_{v\neq0}\frac{\|v\|_y^2}{\|v\|_x^2}=\max_{v\neq0}\frac{\langle H(y)v,v\rangle}{\|v\|_x^2}=\max_{v\neq0}\frac{\langle H_x(y)v,v\rangle_x}{\|v\|_x^2}=\lambda_n=\|H_x(y)\|_x,$$

$$\min_{v\neq0}\frac{\|v\|_y^2}{\|v\|_x^2}=\min_{v\neq0}\frac{\langle H(y)v,v\rangle}{\|v\|_x^2}=\min_{v\neq0}\frac{\langle H_x(y)v,v\rangle_x}{\|v\|_x^2}=\lambda_1=\frac{1}{\|H_x(y)^{-1}\|_x}.$$

L'équivalence entre (AC2) et (16.11) s'en déduit.

[(16.11) $\Leftrightarrow$ (16.12)] Les valeurs propres de $H_x(y)-I$ sont λ_i-1. D'autre part, $\lambda_1>0$ et $(\lambda_1-1)^2\geqslant 0$ permettent d'écrire $1-\lambda_1\leqslant\frac{1}{\lambda_1}-1$. Dès lors (exercice 4.12)

$$\begin{aligned}\|H_x(y)-I\|_x &= \max_i|\lambda_i-1|\\ &= \max\{\lambda_n-1,1-\lambda_1\}\\ &\leqslant \max\{\lambda_n-1,\tfrac{1}{\lambda_1}-1\}\\ &= \max\{\|H_x(y)\|_x-1,\|H_x(y)^{-1}\|_x-1\}.\end{aligned}$$

On démontre de la même manière la seconde inégalité de (16.12). Réciproquement les inégalités de (16.12) impliquent directement celles de (16.11), car $\|H_x(y)-I\|_x\geqslant\|H_x(y)\|_x-1$ et $\|H_x(y)^{-1}-I\|_x\geqslant\|H_x(y)^{-1}\|_x-1$. □

Règles de calcul

Nous énonçons ci-après quelques opérations qui préservent l'autoconcordance.

Proposition 16.6 (restriction à un sous-espace vectoriel) *Si $f\in\mathrm{AC}(\mathbb{E})$ et $\mathbb{E}_0$ est un sous-espace vectoriel de $\mathbb{E}$ qui rencontre $\operatorname{dom} f$, alors $f|_{\mathbb{E}_0}\in\mathrm{AC}(\mathbb{E}_0)$.*

DÉMONSTRATION. Il est clair que le domaine de $f|_{\mathbb{E}_0}$ $(=\mathbb{E}_0\cap\operatorname{dom} f)$ est bien un ouvert non vide de $\mathbb{E}_0$, que $f|_{\mathbb{E}_0}$ est C^2 sur $\mathbb{E}_0$ et que $(f|_{\mathbb{E}_0})''(x)\cdot v^2=f''(x)\cdot v^2>0$,

pour tout $x \in \operatorname{dom} f|_{\mathbb{E}_0}$ et tout v non nul dans $\mathbb{E}_0$. D'autre part, les normes $\|\cdot\|_x$ associées à $f|_{\mathbb{E}_0}$ et f sont identiques, si bien que les autres propriétés requises pour que $f|_{\mathbb{E}_0}$ soit autoconcordante sont remplies. □

Proposition 16.7 (somme) *Si f_1 et $f_2 \in \mathrm{AC}(\mathbb{E})$ et si $\operatorname{dom} f_1 \cap \operatorname{dom} f_2 \neq \varnothing$, alors $f_1 + f_2 \in \mathrm{AC}(\mathbb{E})$.*

Démonstration. Les conditions (16.2) sont clairement vérifiées. D'autre part, si on note $B_{x,i}(x,1)$ la boule unité de centre x pour la métrique $\langle u,v\rangle_{x,i} = f_i''(x)\cdot(u,v)$, la définie positivité des hessiennes $H_i(x)$ implique clairement que $B_x(x,1) \subseteq B_{x,1}(x,1) \cap B_{x,2}(x,1) \subseteq \operatorname{dom} f_1 \cap \operatorname{dom} f_2$ [car les $f_i \in \mathrm{AC}(\mathbb{E})$] $= \operatorname{dom} f$. Enfin, en notant que pour des nombres réels strictement positifs a, b, c et d, on a

$$\min\left(\frac{a}{c},\frac{b}{d}\right) \leqslant \frac{a+b}{c+d} \leqslant \max\left(\frac{a}{c},\frac{b}{d}\right),$$

on obtient pour un vecteur $v \neq 0$:

$$\min\left(\frac{\|v\|_{y,1}^2}{\|v\|_{x,1}^2},\frac{\|v\|_{y,2}^2}{\|v\|_{x,2}^2}\right) \leqslant \frac{\|v\|_y^2}{\|v\|_x^2} \leqslant \max\left(\frac{\|v\|_{y,1}^2}{\|v\|_{x,1}^2},\frac{\|v\|_{y,2}^2}{\|v\|_{x,2}^2}\right).$$

En utilisant le fait que les $f_i \in \mathrm{AC}(\mathbb{E})$, on obtient

$$\min\bigl(1-\|y-x\|_{x,1}, 1-\|y-x\|_{x,2}\bigr) \leqslant \frac{\|v\|_y}{\|v\|_x} \leqslant \max\left(\frac{1}{1-\|y-x\|_{x,1}},\frac{1}{1-\|y-x\|_{x,2}}\right).$$

Il reste à observer que, quel que soit l'indice $i = 1,2$, on a $1-\|y-x\|_x \leqslant 1-\|y-x\|_{x,i}$ et que $(1-\|y-x\|_{x,i})^{-1} \leqslant (1-\|y-x\|_x)^{-1}$, établissant ainsi (16.4). □

Une fonction linéaire n'est pas une fonction autoconcordante (elle n'est pas strictement convexe), mais si on l'ajoute a une fonction autoconcordante, on ne détruit pas l'autoconcordance de cette dernière.

Proposition 16.8 (ajout d'une fonction linéaire) *Si $f \in \mathrm{AC}(\mathbb{E})$ et $c \in \mathbb{E}$, alors $f(\cdot) + \langle c, \cdot\rangle \in \mathrm{AC}(\mathbb{E})$.*

Démonstration. Clair car $\operatorname{dom}(f+c) = \operatorname{dom} f$ est ouvert, $f + c$ est C^2 sur son domaine et $\nabla^2(f+c)(x) = \nabla^2 f(x)$ pour $x \in \operatorname{dom} f$. Les autres propriétés requises s'en suivent. □

Proposition 16.9 (précomposition par une application affine) *Si $\mathbb{F}$ est un autre espace vectoriel de dimension finie, si $a : \mathbb{E} \to \mathbb{F}$ est affine et injective, si $f \in \mathrm{AC}(\mathbb{F})$ et si $\operatorname{dom} f \cap a(\mathbb{E}) \neq \varnothing$, alors $f \circ a \in \mathrm{AC}(\mathbb{E})$.*

Démonstration. On peut écrire $a(x) = Ax - b$, avec $A : \mathbb{E} \to \mathbb{F}$ linéaire injective et $b \in \mathbb{F}$. On note $f^a := f \circ a$ et $\|\cdot\|_x^a = \|A \cdot\|_{a(x)}$ la norme locale en x associée à f^a. Le $\operatorname{dom} f^a = \{x \in \mathbb{E} : a(x) \in \operatorname{dom} f\}$ est un ouvert non vide, sur lequel f^a est C^2 avec $(f^a)''(x) \cdot v^2 = f''(a(x)) \cdot (Av)^2 > 0$ pour tout v non nul, grâce à l'injectivité de A. Dès lors (16.2) a lieu.

Soit à présent $x \in \operatorname{dom} f^a$ et $y \in \mathbb{E}$ tel que $\|y - x\|_x^a < 1$. Alors, $\|A(y-x)\|_{a(x)} < 1$ ou encore $\|a(y) - a(x)\|_{a(x)} < 1$, ce qui implique que $a(y) \in \operatorname{dom} f$ (puisque $a(x) \in \operatorname{dom} f$ et $f \in \mathrm{AC}(\mathbb{F})$) ou encore $y \in \operatorname{dom} f^a$.

Enfin, si $x \in \operatorname{dom} f^a$ et $y \in \mathbb{E}$ est tel que $\|y - x\|_x^a < 1$, on a

$$\frac{\|v\|_y^a}{\|v\|_x^a} = \frac{\|Av\|_{a(y)}}{\|Av\|_{a(x)}} \leqslant \frac{1}{1 - \|a(y) - a(x)\|_{a(x)}} = \frac{1}{1 - \|y - x\|_x^a}.$$

L'inégalité de gauche dans (16.4) se démontre de la même manière. □

Exemple 16.10 En composant la fonction lb sur $\mathbb{R}^m$ et la fonction affine $a : \mathbb{R}^n \to \mathbb{R}^m$ définie par $a(x) = Ax - b$ (on suppose que A est injective et que $\{x : Ax > b\}$ est non vide), on obtient une fonction autoconcordante

$$\begin{cases} -\sum_{i=1}^m \log(A_i x - b_i) & \text{si } Ax > b \\ +\infty & \text{sinon,} \end{cases}$$

dont l'adhérence du domaine est le polyèdre convexe $\{x \in \mathbb{R}^n : Ax \geqslant b\}$. L'hypothèse d'injectivité de A requiert que l'on ait au moins n contraintes linéaires linéairement indépendantes. □

Itérations de Newton sur une fonction autoconcordante

Soit

$$d(x) := -H(x)^{-1} g(x) = -g_x(x)$$

la direction de Newton en $x \in \operatorname{dom} f$ sur une fonction $f \in \mathrm{AC}(\mathbb{E})$. On note

$$x_+ := x + d(x) \tag{16.13}$$

l'itéré calculé par une itération de Newton en x.

La proposition ci-dessous permet de comparer, dans leurs normes locales, les directions de Newton $d(x)$ et $d(x_+)$ en deux points successifs x et $x_+ = x + d(x)$.

Proposition 16.11 (comparaison de deux directions de Newton successives) *Si $f \in \mathrm{AC}(\mathbb{E})$, $x \in \operatorname{dom} f$ et $\|d(x)\|_x < 1$, alors $x_+ \in \operatorname{dom} f$ et*

$$\|d(x_+)\|_{x_+} \leqslant \left(\frac{\|d(x)\|_x}{1-\|d(x)\|_x}\right)^2.$$

Démonstration. On observe d'abord que si $x \in \operatorname{dom} f$ et $\|x_+ - x\|_x = \|d(x)\|_x < 1$, alors $x_+ \in \operatorname{dom} f$ car $B_x(x,1) \subseteq \operatorname{dom} f$ pour les fonction autoconcordante. Ensuite, comme la direction de Newton ne dépend pas du produit scalaire utilisé, on peut écrire $d(x_+) = -H_x(x_+)^{-1} g_x(x_+)$. La norme $\|H_x(x_+)^{-1}\|_x$ pourra se majorer grâce à l'autoconcordance de f, voir (16.11), tandis que le lien entre $g_x(x_+)$ et $g_x(x) = -d(x)$ s'obtiendra par un développement de Taylor.

On a

$$\|d(x_+)\|_{x_+}^2 = \langle g(x_+), H(x_+)^{-1} g(x_+)\rangle = \langle g_x(x_+), H(x_+)^{-1} H(x) g_x(x_+)\rangle_x.$$

Mais $H(x_+)^{-1}H(x) = H_x(x_+)^{-1}$, si bien que

$$\|d(x_+)\|_{x_+}^2 = \langle g_x(x_+), H_x(x_+)^{-1} g_x(x_+)\rangle_x \leqslant \|H_x(x_+)^{-1}\|_x \, \|g_x(x_+)\|_x^2.$$

On note $\delta := \|d(x)\|_x$. Alors (16.11) implique que $\|H_x(x_+)^{-1}\|_x \leqslant (1-\delta)^{-2}$ et il reste donc à montrer que $\|g_x(x_+)\|_x \leqslant \delta^2(1-\delta)^{-1}$. On utilise pour cela un développement de Taylor de $g(x_+)$ autour de x :

$$\begin{aligned} g_x(x_+) &= H(x)^{-1} g(x_+) \\ &= H(x)^{-1}\left[g(x) + \int_0^1 H(x{+}td(x))\, d(x)\, \mathrm{d}t\right] \\ &= \int_0^1 \left[H_x(x{+}td(x)) - I\right] d(x)\, \mathrm{d}t. \end{aligned}$$

En prenant la norme locale en x et en utilisant (16.12) :

$$\begin{aligned} \|g_x(x_+)\|_x &\leqslant \left(\int_0^1 \|H_x(x{+}td(x)) - I\|_x \, \mathrm{d}t\right)\delta \\ &\leqslant \left(\int_0^1 \left[\frac{1}{(1-\delta t)^2} - 1\right] \mathrm{d}t\right)\delta \\ &= \frac{\delta^2}{1-\delta}. \end{aligned}$$

□

Soit q_x le modèle quadratique de f en $x \in \operatorname{dom} f$, c'est-à-dire son développement de Taylor à l'ordre 2. On a pour tout $y \in \operatorname{dom} f$:

$$q_x(y) = f(x) - \langle d(x), y - x\rangle_x + \frac{1}{2}\|y-x\|_x^2.$$

La proposition suivante donne une estimation de l'erreur entre f et son modèle quadratique q_x. On ne s'étonnera pas que l'erreur est en $O(\|y-x\|_x^3)$ puisque la dérivée seconde d'une fonction autoconcordante vérifie une condition de Lipschitz radiale locale; et on appréciera que, pour y dans le voisinage $B_x(x,1)$ de x, aucune constante inconnue n'intervient dans l'estimation.

Proposition 16.12 (erreur du modèle quadratique) *Soient $f \in \mathrm{AC}(\mathbb{E})$, $x \in \operatorname{dom} f$ et $y \in B_x(x,1)$. Alors*

$$\big|f(y) - q_x(y)\big| \leqslant \frac{\|y-x\|_x^3}{3\big(1-\|y-x\|_x\big)}.$$

Démonstration. En utilisant le développement

$$\varphi(1) = \varphi(0) + \varphi'(0) + \frac{1}{2}\varphi''(0) + \int_0^1\int_0^t \big(\varphi''(s) - \varphi''(0)\big)\,\mathrm{d}s\,\mathrm{d}t$$

sur la fonction $\varphi(t) = f(x + t(y-x))$, on obtient

$$f(y) = q_x(y) + \int_0^1\int_0^t \langle [H_x(x+s(y-x)) - I](y-x), y-x\rangle_x\,\mathrm{d}s\,\mathrm{d}t.$$

En prenant les normes, puis en utilisant (16.12), ce qui est licite puisque $y \in B_x(x,1)$, on en déduit

$$\begin{aligned}
\big|f(y) - q_x(y)\big| &\leqslant \|y-x\|_x^2 \int_0^1\int_0^t \|H_x(x+s(y-x)) - I\|_x\,\mathrm{d}s\,\mathrm{d}t \\
&\leqslant \|y-x\|_x^2 \int_0^1\int_0^t \left(\frac{1}{\big(1-s\|y-x\|_x\big)^2} - 1\right)\mathrm{d}s\,\mathrm{d}t \\
&= \|y-x\|_x^3 \int_0^1 \frac{t^2}{1-t\|y-x\|_x}\,\mathrm{d}t \\
&\leqslant \frac{\|y-x\|_x^3}{1-\|y-x\|_x}\int_0^1 t^2\,\mathrm{d}t \\
&= \frac{\|y-x\|_x^3}{3\big(1-\|y-x\|_x\big)}.
\end{aligned}$$

□

La proposition suivante donne, en cas d'existence d'un minimum x_*, une estimation du progrès vers celui-ci que réalise une itération de Newton.

Proposition 16.13 (progrès vers le minimum) *Supposons que $f \in \mathrm{AC}(\mathbb{E})$ ait un minimum en $x_* \in \operatorname{dom} f$.*

1) *Si* $x \in \operatorname{dom} f$ *est tel que* $x_* \in B_x(x,1)$, *alors* x_+ *vérifie*

$$\|x_+ - x_*\|_x \leqslant \frac{\|x - x_*\|_x^2}{1 - \|x - x_*\|_x}.$$

2) *Si* $x \in \operatorname{dom} f$ *est tel que* $x \in B_{x_*}(x_*, \frac{1}{4})$, *alors* x_+ *vérifie*

$$\|x_+ - x_*\|_{x_*} < 4\|x - x_*\|_{x_*}^2.$$

Démonstration. 1) En retranchant x dans chaque membre de (16.13) et en faisant un développement de Taylor avec reste intégral de $g(x)$ autour de x_*, on obtient classiquement

$$x_+ - x_* = \int_0^1 \Big[I - H_x\big(x_* + t(x - x_*)\big)\Big](x - x_*)\,\mathrm{d}t.$$

En prenant les normes, puis en utilisant (16.12), ce qui est légitime puisque $x_* + t(x - x_*) \in B_x(x,1)$, on trouve avec $\delta_+ := \|x_+ - x_*\|_x$ et $\delta := \|x - x_*\|_x$:

$$\delta_+ \leqslant \int_0^1 \Big\|I - H_x\big(x_* + t(x - x_*)\big)\Big\|_x \,\delta\,\mathrm{d}t \leqslant \int_0^1 \left(\frac{\delta}{(1 - (1-t)\delta)^2} - \delta\right)\mathrm{d}t = \frac{\delta^2}{1-\delta}.$$

2) Afin de récrire l'inégalité obtenue au point 1 en termes de $\|x_+ - x_*\|_{x_*}$ et $\|x - x_*\|_{x_*}$, on utilise les inégalités suivantes, déduites de (16.4) :

$$\|x_+ - x_*\|_{x_*}\big(1 - \|x - x_*\|_{x_*}\big) \leqslant \|x_+ - x_*\|_x,$$
$$\|x - x_*\|_x \leqslant \frac{\|x - x_*\|_{x_*}}{1 - \|x - x_*\|_{x_*}}.$$

Lorsque $x \in B_{x_*}(x_*, \frac{1}{2})$, l'inégalité du point 1 devient alors

$$\|x_+ - x_*\|_{x_*} \leqslant \frac{\|x - x_*\|_{x_*}^2}{\big(1 - \|x - x_*\|_{x_*}\big)^2\big(1 - 2\|x - x_*\|_{x_*}\big)}.$$

On obtient alors l'inégalité du point 2 lorsque, par exemple, $x \in B_{x_*}(x_*, \frac{1}{4})$. □

Proposition 16.14 (existence d'un minimum se déduisant d'une direction de Newton petite) *Si* $f \in \mathrm{AC}(\mathbb{E})$ *et si* $\|d(x)\|_x \leqslant \frac{1}{4}$ *pour un* $x \in \operatorname{dom} f$, *alors* f *a un minimum en un point* $x_* \in \operatorname{dom} f$,

$$\|x_+ - x_*\|_x \leqslant \frac{3\|d(x)\|_x^2}{\big(1 - \|d(x)\|_x\big)^3} \quad \text{et} \quad \|x - x_*\|_x \leqslant \|d(x)\|_x + \frac{3\|d(x)\|_x^2}{\big(1 - \|d(x)\|_x\big)^3}.$$

Démonstration. □

Autres propriétés

Proposition 16.15 (existence d'un minimum) *Si $f \in \mathrm{AC}(\mathbb{E})$ est minorée par une constante (par exemple parce que $\operatorname{dom} f$ est borné), alors f a un minimum.*

Démonstration. □

Proposition 16.16 (comportement frontière) *Si $f \in \mathrm{AC}(\mathbb{E})$ et si $\{x_k\} \subseteq \operatorname{dom} f$ est une suite convergeant vers un point $\bar{x} \in \partial(\operatorname{dom} f)$, alors $f(x_k) \to +\infty$ et $\|g(x_k)\| \to +\infty$. En particulier, $f \in \mathrm{C}\overline{\mathrm{onv}}(\mathbb{E})$.*

Démonstration. □

Dans le voisinage d'un point $x \in \operatorname{dom} f$, l'application $g(\cdot)$ est inversible (on peut utiliser le théorème de la fonction inverse, car $g'(x) = H(x)$ est inversible) : pour v petit, on peut écrire $g(x) + v = g(x + u)$ pour un certain $u \in \mathbb{E}$. La proposition suivante quantifie la petitesse de v pour que $g(x) + v$ soit dans l'image de g.

Proposition 16.17 (inversion du gradient) *Soient $f \in \mathrm{AC}(\mathbb{E})$, $x \in \operatorname{dom} f$ et $r \leqslant \frac{1}{4}$. Si $\|v\|_x \leqslant r$, il existe $u \in \bar{B}_x(v, \frac{3r^2}{(1-r)^3})$ tel que $g_x(x) + v = g_x(x+u)$.*

Démonstration. On considère $\tilde{f} \in \mathrm{AC}(\mathbb{E})$, définie en $y \in \mathbb{E}$ par

$$\tilde{f}(y) = f(y) - \langle g_x(x) + v, y\rangle_x.$$

Le minimum de $\tilde{f}$, s'il existe, est le point $x + u$ recherché.

On observera que $\tilde{f} \in \mathrm{AC}(\mathbb{E})$ (proposition 16.8). Le déplacement de Newton en x sur $\tilde{f}$ s'écrit

$$-H(x)^{-1}(g(x) - H(x)g_x(x) - H(x)v) = v.$$

Sa norme locale étant plus petite que $\frac{1}{4}$, $\tilde{f}$ a un minimum (proposition 16.14) ; notons le $x+u \in \operatorname{dom} f$. Selon la même proposition, $\|u-v\|_x \leqslant 3\|v\|_x^2/(1-\|v\|_x)^3 \leqslant 3r^2/(1-r)^3$. □

16.1.3 Barrière autoconcordante

Pour être autoconcordante, une fonction doit vérifier les propriétés globales (16.2) et les propriétés locales (AC1) et (AC2). Pour être une *barrière* autoconcordance, il faut qu'une propriété globale supplémentaire ait lieu.

Définition 16.18 On dit qu'une fonction $f : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ est une (fonction) *barrière autoconcordante*, si $f \in \mathrm{AC}(\mathbb{E})$ et si

$$\vartheta_f := \sup_{x \in \operatorname{dom} f} \|g_x(x)\|_x^2 \tag{16.14}$$

est fini. Le scalaire ϑ_f est appelé le *module de complexité* de f. On note $\mathrm{BAC}(\mathbb{E})$ l'ensemble des fonctions barrières autoconcordantes sur $\mathbb{E}$. □

Comme la direction de Newton s'écrit $d(x) = -g_x(x)$, une barrière auto-concordante a ses directions de Newton bornées en norme locale. On observe que ni $g_x(x)$, ni $\|\cdot\|_x$ ne dépendent du produit scalaire de référence; il en est donc de même de ϑ_f. Enfin, il est clair que $\vartheta_f > 0$ (sinon $g_x(x) = 0$ pour tout $x \in \operatorname{dom} f$, ce qui ne serait possible que si $g(x) = 0$ pour tout $x \in \operatorname{dom} f$ ou encore que si f était constante, en contradiction évidente avec la stricte convexité d'une fonction autoconcordante). Nous verrons que l'on a en fait $\vartheta_f \geqslant 1$ (proposition 16.28), ce qui peut paraître surprenant au premier abord.

Exemple 16.19 La fonction $\mathrm{lb} : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ définie en (16.5) est une barrière autoconcordante avec $\vartheta_{\mathrm{lb}} = n$.

Démonstration. On sait que $\mathrm{lb} \in \mathrm{AC}(\mathbb{R}^n)$ (exemple 16.2). D'autre part, si on note $X = \mathrm{Diag}(x_1, \ldots, x_n)$ et e le vecteur dont les composantes valent 1, le gradient de lb pour la métrique locale s'écrit $g_x(x) = H(x)^{-1} g(x) = (X^{-2})^{-1}(-X^{-1}e) = -x$. Dès lors $\vartheta_{\mathrm{lb}} = \sup_{x>0} \|x\|_x^2 = \sup_x x^{\mathsf{T}} X^{-2} x = n$. □

Exemple 16.20 La fonction $\mathrm{ld} : \mathcal{S}^n \to \mathbb{R} \cup \{+\infty\}$ définie en (16.6) est une barrière autoconcordante avec $\vartheta_{\mathrm{ld}} = n$.

Démonstration. On sait que $\mathrm{ld} \in \mathrm{AC}(\mathcal{S}^n)$ (exemple 16.3). Le gradient $g(X)$ et la hessienne $H(X)$ de ld en $X \succ 0$ vérifient $g(X) = -X^{-1}$ et $H(X)M = X^{-1}MX^{-1}$ pour tout $M \in \mathcal{S}^n$. Alors $-X^{-1} = g(X) = H(X)g_X(X) = X^{-1}g_X(X)X^{-1}$, si bien que $g_X(X) = -X$. Finalement $\|g_X(X)\|_X^2 = \langle g(X), g_X(X)\rangle = \operatorname{tr} I = n$. □

Calcul autoconcordant

Proposition 16.21 (restriction à un sous-espace vectoriel) *Si* $f \in \mathrm{BAC}(\mathbb{E})$ *et si* $\mathbb{E}_0$ *est un sous-espace vectoriel de* $\mathbb{E}$ *qui rencontre* $\operatorname{dom} f$*, alors* $f|_{\mathbb{E}_0} \in \mathrm{BAC}(\mathbb{E}_0)$ *et* $\vartheta_{f|_{\mathbb{E}_0}} \leqslant \vartheta_f$.

Démonstration. En effet, $f|_{\mathbb{E}_0} \in \mathrm{AC}(\mathbb{E}_0)$ (proposition 16.6). Par ailleurs, si $P_{\mathbb{E}_0,x}$ est le projecteur orthogonal sur $\mathbb{E}_0$ pour la métrique riemannienne en x et si $g|_{\mathbb{E}_0,x}(x)$ est le gradient de $f|_{\mathbb{E}_0}$ en x pour cette même métrique, on a $g|_{\mathbb{E}_0,x}(x) = P_{\mathbb{E}_0,x} g_x(x)$ et $\|P_{\mathbb{E}_0,x} v\|_x \leqslant \|v\|_x$ (quel que soit $v \in \mathbb{E}$). Dès lors

$$\vartheta_{f|_{\mathbb{E}_0}} = \sup_{x\in\mathbb{E}_0\cap\operatorname{dom} f} \|g|_{\mathbb{E}_0,x}(x)\|_x^2 \leqslant \sup_{x\in\mathbb{E}_0\cap\operatorname{dom} f} \|g_x(x)\|_x^2 \leqslant \sup_{x\in\operatorname{dom} f} \|g_x(x)\|_x^2 = \vartheta_f.$$

□

Proposition 16.22 (somme) *Si f_1 et $f_2 \in \mathrm{BAC}(\mathbb{E})$ et si $\operatorname{dom} f_1 \cap \operatorname{dom} f_2 \neq \varnothing$, alors $f_1 + f_2 \in \mathrm{BAC}(\mathbb{E})$ et $\vartheta_{f_1+f_2} \leqslant \vartheta_{f_1} + \vartheta_{f_2}$.*

Démonstration. On sait déjà que $f := f_1 + f_2 \in \mathrm{AC}(\mathbb{E})$ (proposition 16.7). On simplifie les écritures en notant $g_i := g_i(x)$, $H_i := H_i(x)$, $g = g_1 + g_2$ et $H = H_1 + H_2$.

Il faut estimer $\|g_x(x)\|_x$, mais comme cette quantité ne dépend pas du produit scalaire de référence, on peut supposer que celui-ci est tel que $H(x) = I$. Alors on a

$$\|g_x(x)\|_x^2 = f''(x) \cdot \big(H(x)^{-1} g(x)\big)^2 = \|g\|^2 = \|g_1 + g_2\|^2.$$

On développe le carré et on introduit $v_i := H_i^{-1/2} g_i$:

$$\|g_x(x)\|_x^2 = \|g_1\|^2 + \|g_2\|^2 + 2\langle g_1, g_2\rangle = \langle H_1 v_1, v_1\rangle + \langle H_2 v_2, v_2\rangle + 2\langle H_1^{1/2} v_1, H_2^{1/2} v_2\rangle.$$

On utilise alors $H_1 + H_2 = H = I$:

$$\begin{aligned} \|g_x(x)\|_x^2 &= \langle (I - H_2)v_1, v_1\rangle + \langle (I - H_1)v_2, v_2\rangle + 2\langle H_1^{1/2} v_1, H_2^{1/2} v_2\rangle \\ &= \|v_1\|^2 + \|v_2\|^2 - \|H_2^{1/2} v_1\|^2 - \|H_1^{1/2} v_2\|^2 + 2\langle H_1^{1/2} v_1, H_2^{1/2} v_2\rangle. \end{aligned}$$

Mais H_1 et H_2 commutent (car $H_1 H_2 = H_1(I - H_1) = (I - H_1)H_1 = H_2 H_1$) et donc aussi $H_1^{1/2}$ et $H_2^{1/2}$. Alors la somme des trois derniers termes ci-dessus est la quantité négative $-\|H_2^{1/2} v_1 - H_1^{1/2} v_2\|^2$. On en déduit

$$\|g_x(x)\|_x^2 \leqslant \|v_1\|^2 + \|v_2\|^2 = \langle H_1^{-1} g_1, g_1\rangle + \langle H_2^{-2} g_2, g_2\rangle \leqslant \vartheta_{f_1} + \vartheta_{f_2}.$$

En prenant le supremum en $x \in \operatorname{dom} f$ à gauche on obtient $\vartheta_f \leqslant \vartheta_{f_1} + \vartheta_{f_2}$. □

La proposition 16.8 ne s'étend pas aux barrières autoconcordantes.

Contre-exemple 16.23 *La fonction $x \in \mathbb{R} \mapsto x - \log x$ n'est pas une barrière autoconcordante.* En effet, $g(x) = 1 - 1/x$, $H(x) = 1/x^2$, $g_x(x) = x^2 - x$, $\|g_x(x)\|_x^2 = (x^2 - x)^2/x^2 = (x - 1)^2$ qui n'est pas borné sur $\mathbb{R}_{++}$. □

Proposition 16.24 (précomposition par une application affine) *Si $\mathbb{F}$ est un autre espace euclidien, si $a : \mathbb{E} \to \mathbb{F}$ est affine et injective, si $f \in \mathrm{BAC}(\mathbb{F})$ et si $\operatorname{dom} f \cap a(\mathbb{E}) \neq \varnothing$, alors $f \circ a \in \mathrm{BAC}(\mathbb{E})$ et $\vartheta_{f\circ a} \leqslant \vartheta_f$.*

Démonstration. On sait déjà que $\tilde{f} := f \circ a \in \mathrm{AC}(\mathbb{E})$ (proposition 16.9). Soient $x \in \mathbb{E}$ tel que $y := a(x) \in \operatorname{dom} f$. On prend $\langle \cdot, \cdot\rangle_y$ comme produit scalaire de référence sur $\mathbb{F}$ et on suppose que $a(x) = Ax + b$ avec A linéaire injective.

Calculons le gradient et la hessienne de $\tilde{f}$ en x. On a $\tilde{f}'(x)\cdot u = f'(y)\cdot(Au) = \langle A^*g(y), u\rangle$ et $\tilde{f}''(x)\cdot(u,v) = f''(y)\cdot(Au,Av) = \langle A^*Au, v\rangle$ (avec le produit scalaire choisi sur $\mathbb{F}$, $H(y)=I$). Dès lors

$$\tilde{g}(x) = A^*g(y) \qquad \text{et} \qquad \tilde{H}(x) = A^*A.$$

On peut alors évaluer

$$\begin{aligned}
\|\tilde{g}_x(x)\|_x^2 &= \langle \tilde{H}(x)^{-1}\tilde{g}(x), \tilde{g}(x)\rangle \\
&= \langle (A^*A)^{-1}A^*g(y), A^*g(y)\rangle \\
&= \langle A(A^*A)^{-1}A^*g(y), g(y)\rangle_y \\
&\leqslant \|A(A^*A)^{-1}A^*\|_y \, \|g_y(y)\|_y^2 \\
&\leqslant \vartheta_f,
\end{aligned}$$

car $A(A^*A)^{-1}A^*$ étant un projecteur orthogonal pour le produit scalaire $\langle\cdot,\cdot\rangle_y$ (il est idempotent et auto-adjoint), il est de norme $\|\cdot\|_y$ inférieure à 1. □

Autres propriétés

Beaucoup de propriétés des barrières autoconcordantes se déduisent des bornes que l'on peut obtenir sur ses dérivées directionnelles. C'est ce sur quoi se concentrent les deux propositions suivantes. On y fait usage de l'inverse de la fonction de Minkowski de $\operatorname{dom} f$, de pôle $x \in \operatorname{adh}(\operatorname{dom} f)$, qui est défini en $y \in \mathbb{E}$ par

$$\alpha_x(y) = \sup\{\alpha \geqslant 0 : x + \alpha(y-x) \in \operatorname{dom} f\}.$$

La *fonction de Minkowski* est l'application $y \mapsto 1/\alpha_x(y)$.

Proposition 16.25 (dérivée directionnelle première) *Soient $f \in \mathrm{BAC}(\mathbb{E})$ et $x \in \operatorname{dom} f$.*

1) *Si $y \in \operatorname{dom} f$, alors*

$$\frac{-\vartheta_f}{\alpha_y(x)-1} \leqslant f'(x)\cdot(y-x) < \vartheta_f. \tag{16.15}$$

2) *Si $y \in \partial(\operatorname{dom} f)$ et $\alpha_y(x) \geqslant \left(1+\vartheta_f^{1/2}\right)^2$, alors*

$$1 - \frac{\left(1+\vartheta_f^{1/2}\right)^2}{\alpha_y(x)} \leqslant f'(x)\cdot(y-x). \tag{16.16}$$

Démonstration. 1) Démontrons d'abord l'inégalité de droite dans (16.15). On examine f le long de la droite portée par x et $y \in \operatorname{dom} f$, en introduisant la fonction φ définie par $\varphi(t) := f(x+t(y-x))$. Il s'agit donc de démontrer que $\varphi'(0) < \vartheta_f$ ou $1/\varphi'(0) > 1/\vartheta_f$, car on peut supposer que $\varphi'(0) > 0$ (sinon le résultat est trivial). Si

$\varphi'(0) > 0$, alors, par convexité, $\varphi'(t) > 0$ pour $t \geqslant 0$ dans le domaine de φ. Comme 1 est dans le $\operatorname{dom}\varphi$, il suffit en fait de démontrer l'inégalité un peu plus forte

$$\frac{1}{\varphi'(0)} - \frac{1}{\varphi'(1)} \geqslant \frac{1}{\vartheta_f}. \tag{16.17}$$

Celle-ci suggère de regarder comment varie $1/\varphi'(t)$ ou encore d'intégrer la dérivée de ce quotient entre 0 et 1. On trouve

$$\left(-\frac{1}{\varphi'(t)}\right)' = \frac{\varphi''(t)}{\varphi'(t)^2} \geqslant \frac{1}{\vartheta_f}, \tag{16.18}$$

où l'inégalité à droite vient des estimations suivantes, dans lesquelles on a noté $x_t := x + t(y-x)$:

$$|\varphi'(t)| = |\langle g_{x_t}(x_t), y - x\rangle_{x_t}| \leqslant \|g_{x_t}(x_t)\|_{x_t}\, \|y-x\|_{x_t} \leqslant \vartheta_f^{1/2}\, \|y-x\|_{x_t},$$
$$\varphi''(t) = \|y-x\|_{x_t}^2.$$

Après intégration de (16.18) entre 0 et 1, on obtient (16.17).

Pour établir l'inégalité de gauche dans (16.15), on introduit la fonction φ définie par $\varphi(t) := f(y + t(x-y))$. Il s'agit donc de majorer $\varphi'(1) = -f'(x)\cdot(y-x)$, que l'on peut supposer strictement positif (sinon le résultat est trivial). Par convexité, on a alors $\varphi'(t) > 0$ pour $t \geqslant 1$ dans le domaine de φ. Comme ci-dessus, $\varphi'(t)$ vérifie (16.18). Après intégration entre 1 et t, on obtient cette fois

$$\vartheta_f - (t-1)\varphi'(1) \geqslant \frac{\vartheta_f \varphi'(1)}{\varphi'(t)}.$$

Le membre de droite reste strictement positif pour tout $t \geqslant 1$ dans $\operatorname{dom}\varphi$; il en est donc de même du membre de gauche. En prenant le plus grand t possible, c'est-à-dire $t = \alpha_y(x)$, on voit que $\varphi'(1) \leqslant \vartheta_f/(\alpha_y(x) - 1)$, qui est l'inégalité recherchée.

2) On note $T := \alpha_y(x)$ si cette quantité est finie; sinon T est un nombre arbitrairement grand. On définit à présent $\varphi \in \mathrm{AC}(\mathbb{R})$ par $\varphi(t) := f(y + t(x-y))$. Il s'agit de majorer $\varphi'(1)$. Le domaine de φ est l'intervalle $]0, T[$. Comme φ est autoconcordante et $0 \notin \operatorname{dom}\varphi$, on a pour tout $t \in]0, T[$:

$$\varphi''(t)t^2 \geqslant 1.$$

D'autre part, en appliquant à φ l'inégalité de gauche dans (16.15), qui est encore vérifiée pour $y \in \partial(\operatorname{dom} f)$, on trouve $\varphi'(t)(-t) \geqslant -\vartheta_f/(\alpha_0(t)-1)$ (proposition 16.21). Comme $\alpha_0(t) = T/t$, on obtient pour tout $t \in]0, T[$:

$$\varphi'(t) \leqslant \frac{\vartheta_f}{T-t}.$$

On intègre alors $\varphi''(t)$ de 1 à $t \in]t, T[$ et on tient compte des deux dernières inégalités :

$$\varphi'(1) - \frac{1}{t} + 1 \leqslant \varphi'(1) + \int_1^t \varphi''(s)\mathrm{d}s = \varphi'(t) \leqslant \frac{\vartheta_f}{T-t}$$

ou encore

$$\varphi'(1) \leqslant \frac{\vartheta_f}{T-t} + \frac{1}{t} - 1. \tag{16.19}$$

Si $T := \alpha_y(x) < +\infty$, on peut prendre $t = T/(1+\vartheta_f^{1/2})$, qui est dans $]1, T[$, par l'hypothèse sur $\alpha_y(x)$. On trouve

$$\varphi'(1) \leqslant \frac{\vartheta_f(1+\vartheta_f^{1/2})}{T\vartheta_f^{1/2}} + \frac{1+\vartheta_f^{1/2}}{T} - 1 = \frac{(1+\vartheta_f^{1/2})^2}{T} - 1,$$

qui est bien (16.16). Si $\alpha_y(x) = +\infty$, on peut prendre T et t arbitrairement grand dans (16.19), si bien que l'on trouve $\varphi'(1) \leqslant -1$ et donc aussi (16.16). □

Les dérivées directionnelles secondes peuvent s'exprimer en utilisant la norme locale. C'est ce que nous faisons dans la proposition suivante.

Proposition 16.26 (dérivée directionnelle seconde) *Soient $f \in \mathrm{BAC}(\mathbb{E})$ et x, $y \in \operatorname{dom} f$ vérifiant $f'(x)\cdot(y-x) \geqslant 0$. Alors $\|y-x\|_x \leqslant 4\vartheta_f + 1$.*

Démonstration. □

Les propositions précédentes ont de nombreuses conséquences. En voici déjà quelques-unes, toutes immédiates, apportant un éclairage sur l'allure des barrières autoconcordantes.

- *Si d fait un angle plus petit que $\pi/2$ avec un gradient $g(x) \neq 0$ d'une barrière autoconcordante, alors la demi-droite $\{x + td : t \geqslant 0\}$ rencontre la frontière de son domaine.*

  Cet énoncé suppose que l'on utilise le même produit scalaire pour définir l'angle et le gradient. Il se déduit de l'inégalité de droite dans (16.15). Si on utilise aussi la proposition 16.26, la zone où doit se trouver le domaine de f est représenté à la figure 16.1.

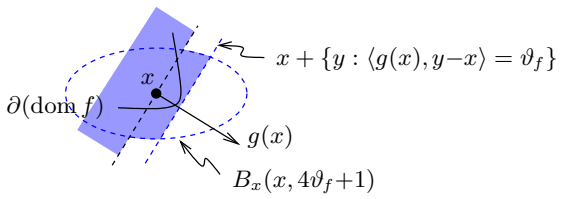


Fig. 16.1. Localisation du domaine d'une barrière autoconcordante

- *Le domaine d'une barrière autoconcordante ne contient pas de droite. En particulier, il ne peut être $\mathbb{E}$ tout entier.*

En effet, supposons que $\{x_0+\alpha d : \alpha \in \mathbb{R}\} \subseteq \operatorname{dom} f$. Alors, pour tout $x = x_0+\alpha d$, on aurait $f'(x)\cdot d = 0$ (prendre $y = x+td$ avec $t \to \pm\infty$ dans l'inégalité de droite dans (16.15)). On en déduirait que $f''(x)\cdot d^2 = 0$, ce qui ne serait pas compatible avec la stricte convexité de f.

- *Si* $x \in \operatorname{dom} f$ *et* $d \in (\operatorname{dom} f)^\infty(x)$, *alors* $f'(x)\cdot d \leqslant -1/\alpha_x(x-d) < 0$.

  Comme $-d \notin (\operatorname{dom} f)^\infty(x)$ (car $\operatorname{dom} f$ ne contient pas de droite), $\alpha := \alpha_x(x-d)$ est fini et $y := x-\alpha d$ est sur la frontière de $\operatorname{dom} f$. D'autre part $\alpha_y(x) = \infty$ et, d'après (16.16), $f'(x)\cdot(y-x) \geqslant 1$.

Le résultat suivant montre que, dans le cas où f a un minimum, celui-ci est bien centré dans le domaine de f. De plus, les ellipsoïdes inscrit $B_{x_*}(x_*, 1)$ et circonscrit $B_{x_*}(x_*, 4\vartheta_f+1)$ permettent d'avoir une bonne estimation de son volume.

Proposition 16.27 (existence d'un minimum) *Une fonction* $f \in \mathrm{BAC}(\mathbb{E})$ *a un minimum* x_* *si, et seulement si, son domaine est borné. Dans ce cas*

$$B_{x_*}(x_*, 1) \subseteq \operatorname{dom} f \subseteq B_{x_*}(x_*, 4\vartheta_f+1).$$

Démonstration. Si $\operatorname{dom} f$ est borné, f a un minimum (proposition 16.15). Inversement si f a un minimum $x_* \in \operatorname{dom} f$, on a $B_{x_*}(x_*, 1) \subseteq \operatorname{dom} f$ par la définition de l'autoconcordance et $\operatorname{dom} f \subseteq B_{x_*}(x_*, 4\vartheta_f+1)$ par la proposition 16.26 (appliquée à $x = x_*$, avec $g(x_*) = 0$). □

Estimation du module de complexité

Mieux vaut disposer d'une barrière autoconcordante avec un module de complexité petit car, comme nous le verrons à la section 16.1.4, la complexité itérative des algorithmes sera proportionnelle à ϑ_f^ω, avec un certain $\omega > 0$. Le module de complexité ne peut toutefois pas être arbitrairement petit. D'une part, il doit être supérieur à 1 (proposition 16.28), ce que ne suggère pas sa définition.

Proposition 16.28 (minorant de ϑ_f) *Si* $f \in \mathrm{BAC}(\mathbb{E})$, *alors* $\vartheta_f \geqslant 1$.

Démonstration. On peut trouver un $x \in \operatorname{dom} f$ tel que $g_x(x) \neq 0$ car une fonction autoconcordante ne peut être constante (elle a une hessienne définie positive). Alors $y := x + tg_x(x)/\|g_x(x)\|_x$ est dans le domaine de f pour tout $t \in [0,1[$ (on sait en effet que $B_x(x,1) \subseteq \operatorname{dom} f$ lorsque f est autoconcordante). En utilisant ce y dans l'inégalité de droite dans (16.15), on obtient

$$\vartheta_f > t\,\frac{\langle g(x), g_x(x)\rangle}{\|g_x(x)\|_x} = t\,\|g_x(x)\|_x.$$

Le supremum sur les $x \in \operatorname{dom} f$ tels que $g_x(x) \neq 0$ et sur les $t \in [0,1[$ à droite conduit à $\vartheta_f \geqslant \vartheta_f^{1/2}$ ou $\vartheta_f \geqslant 1$. □

Existence de barrière autoconcordante

Nous admettrons le théorème fondamental suivant [455 ; Theorem 2.5.1]. Une *droite* fait référence ici à un sous-espace affine de dimension 1.

Théorème 16.29 (fondamental) *Il existe une constante $\nu > 0$ telle que pour tout entier $n \geqslant 1$ et pour tout ouvert convexe Ω de $\mathbb{R}^n$ ne contenant pas de droite, il existe une fonction $f \in \mathrm{BAC}(\mathbb{R}^n)$ telle que $\operatorname{dom} f = \Omega$ et $\vartheta_f \leqslant \nu n$.*

16.1.4 Trois approches algorithmiques ▲

Problème et méthodes

On considère le problème

$$(P) \quad \begin{cases} \inf \langle c, x \rangle \\ x \in X, \end{cases}$$

dans lequel $c \in \mathbb{E}$ et $X \subseteq \mathbb{E}$ est l'adhérence du domaine d'une fonction $f \in \mathrm{BAC}(\mathbb{E})$. On note

$$\mathrm{val}(P)$$

la valeur optimale de (P). On cherche à résoudre ce problème par des méthodes de points intérieurs, générant donc des itérés dans l'intérieur supposé non vide de X.

Comme l'ensemble admissible X n'est pas donné par des contraintes fonctionnelles, les méthodes de points intérieurs pour résoudre (P) sont nécessairement primales, dans le sens où elles ne s'intéressent qu'à la variable primale x (à la différence des méthodes de points intérieurs primales-duales des chapitres 18 et 20, qui génèrent conjointement des itérés primaux et des multiplicateurs).

Définition 16.30 On appelle *chemin central* du problème (P), noté $\mathscr{C}$, l'image de l'application $(\mu > 0) \mapsto x(\mu) \in \operatorname{dom} f$, où $x(\mu)$ est le minimum (s'il existe) de la fonction f_μ définie par

$$f_\mu(x) = \frac{1}{\mu}\langle c, x \rangle + f(x).$$

□

S'il existe, le minimum de la fonction strictement convexe f_μ est nécessairement unique. Comme nous avons supposé que $f \in \mathrm{BAC}(\mathbb{E})$, $f_\mu \in \mathrm{AC}(\mathbb{E})$ (proposition 16.8), mais f_μ peut ne pas être dans $\mathrm{BAC}(\mathbb{E})$ (contre-exemple 16.23).

Les algorithmes de points intérieurs se servent du chemin central comme d'un guide qui va conduire les itérés vers une solution du problème. Cette stratégie est motivée par le fait que la valeur du critère dans (P) décroît vers la valeur optimale le long du chemin central (lorsque $\mu \downarrow 0$). La décroissance de la valeur optimale de f_μ est assurée par la proposition 13.2, tandis que sa convergence vers la valeur optimale du problème (P) se déduit clairement de l'estimation suivante :

$$\boxed{\langle c, x(\mu) \rangle \leqslant \mathrm{val}(P) + \vartheta_f\, \mu.} \tag{16.20}$$

Démontrons cette inégalité. Par optimalité, $\nabla f_\mu(x(\mu)) = 0$, ce qui s'écrit

$$-\frac{1}{\mu}\, c = g(x(\mu)).$$

En prenant le produit scalaire avec $y - x(\mu)$ pour $y \in \operatorname{dom} f$, on a grâce à la proposition 16.25 :

$$-\frac{1}{\mu}\langle c, y - x(\mu)\rangle = \langle g(x(\mu)), y - x(\mu)\rangle < \vartheta_f.$$

Dès lors, $\langle c, x(\mu)\rangle < \langle c, y\rangle + \vartheta_f\, \mu$, pour tout $y \in \operatorname{dom} f$. On en déduit (16.20).

C'est la direction de Newton en x sur f_μ, pour des paramètres μ décroissant vers zéro, qui joue un rôle important dans les algorithmes décrits ci-après. Elle est notée et définie par

$$d_\mu(x) := -H(x)^{-1}\left(\frac{1}{\mu}\, c + g(x)\right). \tag{16.21}$$

On aura aussi besoin d'imposer aux itérés x de se trouver suffisamment proche d'un point central $x(\mu)$. Le contrôle de cette proximité se fera au moyen d'un voisinage qu'il est agréable de définir à partir de la direction de Newton $d_\mu(x)$, parce que celle-ci est calculée dans les algorithmes. Pour $\mu > 0$ et $\alpha > 0$, on définit le voisinage suivant de $x(\mu)$:

$$V_\mu(\alpha) := \{x \in \operatorname{dom} f : \|d_\mu(x)\|_x \leqslant \alpha\}. \tag{16.22}$$

Le voisinage $V(\alpha)$ du chemin central est alors la réunion de ces $V_\mu(\alpha)$ pour tous les $\mu > 0$.

Si f_μ n'a pas de minimum, le voisinage $V_\mu(\alpha)$ peut être vide. En fait, d'après la proposition 16.14, supposer $V_\mu(\alpha) \neq \varnothing$, avec $\alpha \in \left]0, \frac{1}{4}\right[$, assure du même coup l'existence de $x(\mu)$. Cette hypothèse sera implicite dans les algorithmes décrits ci-après.

Algorithme des petits déplacements

L'algorithme des petits déplacements relève du principe de la *poursuite d'un objectif fuyant*, ce que l'on a schématisé à la figure 16.2 : en l'itéré courant x, supposé

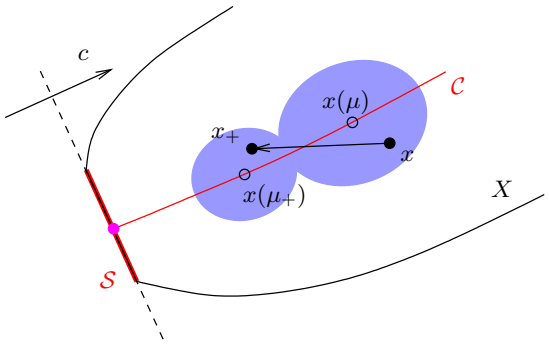


Fig. 16.2. Algorithme des petits déplacements

proche d'un point central $x(\mu)$, on se fixe pour objectif un autre point central $x(\mu_+)$ avec $\mu_+ < \mu$, et l'on fait un déplacement de Newton dans sa direction (donc sur f_{μ_+}), ce qui conduit au nouvel itéré x_+. On n'a pas besoin de connaître $x(\mu_+)$; seuls μ_+ et f_{μ_+} sont utiles au calcul du déplacement de Newton. Le processus se répète jusqu'à ce que μ soit assez petit devant ϑ_f ou devant un majorant de ϑ_f, que l'on suppose connu. L'inégalité (16.20) permet d'avoir une estimation de l'erreur sur le coût sans avoir besoin de connaître la valeur optimale $\mathrm{val}(P)$, pourvu que ϑ_f soit connu. La proximité du point central $x(\mu)$ se fait au moyen du voisinage $V_\mu(\alpha)$ avec $\alpha = 1/9$. Quant à la décroissance de μ, elle se fait simplement par la règle $\mu_+ = \sigma\mu$, avec la constante $\sigma \in]0,1[$ bien choisie.

On peut à présent décrire de manière précise, l'algorithme dont on montrera la polynomialité.

Algorithme 16.31 (PIC petits déplacements — une itération)

On se donne $\mu > 0$ et $\alpha := 1/9$. On suppose que l'itéré courant $x \in V_\mu(\alpha)$. L'itéré suivant x_+ est calculé comme suit.

1. Mise à jour du paramètre $\mu_+ = \sigma\mu$, avec

$$\sigma = \frac{8\,\vartheta_f^{1/2}}{1 + 8\,\vartheta_f^{1/2}}. \tag{16.23}$$

2. Calcul de la direction de Newton $d_{\mu_+}(x)$.
3. Nouvel itéré : $x_+ = x + d_{\mu_+}(x)$.

Cet algorithme fait usage de ϑ_f, qu'il faut donc connaître. On pourrait aussi utiliser un majorant θ de ϑ_f, ce qui aurait pour effet d'augmenter σ, mais alors la complexité itérative de l'algorithme serait en $O(\theta^{1/2})$ au lieu de $O(\vartheta_f^{1/2})$.

Théorème 16.32 (convergence et complexité de l'algorithme des petits déplacements) *L'algorithme 16.31 est bien défini, converge ($\mu_k \to 0$ q-linéairement) et pour tout $\varepsilon \in]0,1[$, il existe un indice $K := \lceil 9\,\vartheta_f^{1/2} \log \varepsilon^{-1} \rceil$ tel que $\mu_k \leqslant \varepsilon\mu_0$, dès que $k \geqslant K$.*

Démonstration. L'algorithme sera bien défini si on montre que $x_+ \in V_{\mu_+}(\alpha)$, c'est-à-dire si $\|d_{\mu_+}(x_+)\|_{x_+} \leqslant \alpha$. La proposition 16.11 permet d'estimer la norme précédente en fonction de $\|d_{\mu_+}(x)\|_x$. Commençons donc par estimer cette dernière en fonction de $\|d_\mu(x)\|_x$ que l'on sait être $\leqslant \alpha$, par récurrence. On a

$$d_{\mu_+}(x) = -H(x)^{-1}\left(\frac{1}{\mu_+}\,c + g(x)\right) = \frac{\mu}{\mu_+} d_\mu(x) + \left(\frac{\mu}{\mu_+} - 1\right) g_x(x).$$

En utilisant $\mu_+ = \sigma\mu$, $\|d_\mu(x)\|_x \leqslant \alpha$, $\|g_x(x)\|_x \leqslant \vartheta_f^{1/2}$, les valeurs de $\alpha = 1/9$ et de σ utilisées par l'algorithme et $\vartheta_f \geqslant 1$ (proposition 16.28), on obtient

$$\|d_{\mu_+}(x)\|_x \leqslant \frac{\alpha}{\sigma} + \left(\frac{1}{\sigma} - 1\right) \vartheta_f^{1/2} \leqslant \frac{1}{9}\left(\frac{1}{8\,\vartheta_f^{1/2}} + 1\right) + \frac{1}{8} \leqslant \frac{1}{4}.$$

Comme $f_{\mu_+} \in \mathrm{AC}(\mathbb{E})$ (proposition 16.8) et $\|d_{\mu_+}(x)\|_x < 1$, on peut utiliser la proposition 16.11, qui conduit à

$$\|d_{\mu_+}(x_+)\|_{x_+} \leqslant \left(\frac{\|d_{\mu_+}(x)\|_x}{1 - \|d_{\mu_+}(x)\|_x}\right)^2 \leqslant \frac{1}{9} = \alpha.$$

L'algorithme est donc bien défini et $\mu_{k+1} = \sigma\mu_k$, avec $\sigma < 1$, montre que $\mu_k \to 0$ linéairement.

Il reste à démontrer la polynomialité. On a

$$\begin{aligned}
\log \mu_k &= \log \mu_{k-1} - \log\left(1 + \frac{1}{8\,\vartheta_f^{1/2}}\right) \\
&\leqslant \log \mu_{k-1} - \frac{1/(8\,\vartheta_f^{1/2})}{1 + 1/(8\,\vartheta_f^{1/2})} \qquad [\text{car } \log(1+t) \geqslant \tfrac{t}{1+t}] \\
&\leqslant \log \mu_0 - \frac{k}{1 + 8\,\vartheta_f^{1/2}} \qquad [\text{par récurrence}].
\end{aligned}$$

Dès lors, $\log(\mu_k/\mu_0) \leqslant \log \varepsilon$, dès que $k \geqslant K := \lceil 9\vartheta_f^{1/2} \log \varepsilon^{-1} \rceil \geqslant (1 + 8\,\vartheta_f^{1/2}) \log \varepsilon^{-1}$ (on se rappelle que $\vartheta_f \geqslant 1$). □

Algorithme des grands déplacements

L'inconvénient de l'algorithme des petits déplacements est de réduire μ très lentement, puisqu'on le multiplie à chaque itération par le facteur σ donné par (16.23), qui est proche de 1 lorsque ϑ_f est grand. Dans l'algorithme des grands déplacements, on décide de réduire μ par un facteur σ arbitrairement petit à chaque étape de l'algorithme : $\mu_+ := \sigma\mu$. Fort bien. Mais il n'est alors plus garanti que $x + d_{\mu_+}(x)$ soit dans le voisinage $V_{\mu_+}(\alpha)$ de $x(\mu_+)$ lorsque $x \in V_\mu(\alpha)$. Or on a besoin que le nouvel itéré x_+ soit dans $V_{\mu_+}(\alpha)$ pour qu'au début de l'étape suivante on se retrouve dans les mêmes conditions qu'à l'étape courante. Dans l'algorithme des grands déplacements, on obtient $x_+ \in V_{\mu_+}(\alpha)$ en faisant plusieurs itérations de Newton amorties, avec des pas déterminés par recherche linéaire sur f_{μ_+}. L'algorithme est schématisé à la figure 16.3. Cet algorithme génère donc des itérés intermédiaires x_i. On appellera *itération* l'ensemble des opérations menant d'un itéré intermédiaire au suivant et *étape* l'ensemble des itérations conduisant de x à x_+ et correspondant à une réduction de μ en $\mu_+ = \sigma\mu$. Voici une description précise d'une étape de l'algorithme.

Algorithme 16.33 (PIC grands déplacements — une étape)

Soient $\alpha = 1/4$ et $\sigma \in]0, 1[$ deux constantes indépendantes de l'étape considérée. On suppose que l'itéré courant $x \in V_\mu(\alpha)$, pour un $\mu > 0$. L'itéré suivant $x_+ \in V_{\mu_+}(\alpha)$ est calculé comme suit.

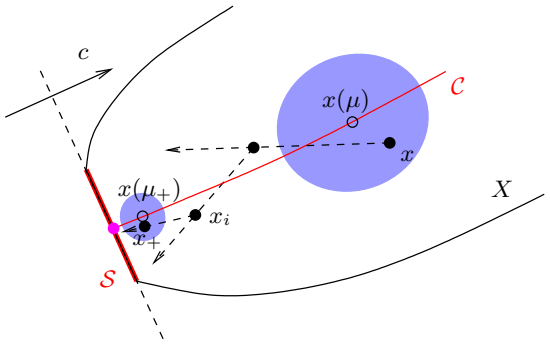


Fig. 16.3. Algorithme des grands déplacements

1. Mise à jour du paramètre : $\mu_+ = \sigma\mu$.
2. Itérations de Newton avec recherche linéaire.
   2.1. $x_0 := x$, $i := 0$.
   2.2. Tant que $x_i \notin V_{\mu_+}(\alpha)$:
   - calcul du pas $t_i > 0$ qui minimise $t \mapsto f_{\mu_+}(x_i + t d_{\mu_+}(x_i))$,
   - $x_{i+1} := x_i + t_i d_{\mu_+}(x_i)$,
   - $i := i + 1$.
3. Nouvel itéré : $x_+ := x_i$, $i(x) := i$.

La difficulté principale de l'analyse de l'algorithme est de montrer que le nombre $i(x)$ d'itérés intermédiaires à chaque étape est borné (par une constante ne dépendant que de ϑ_f et de σ). La complexité polynomiale de l'algorithme sera une conséquence aisée de ce fait.

Théorème 16.34 (convergence et complexité de l'algorithme des grands déplacements) *L'algorithme 16.33 est bien défini, converge ($\mu_k \to 0$ q-linéairement) et pour tout $\varepsilon \in]0,1[$, on a $\mu_k \leqslant \varepsilon\mu_0$, dès que le nombre d'itérations de Newton est supérieur à $(80\,\vartheta_f \log_\sigma \varepsilon)/\sigma$.*

Démonstration. 1) *Première étape : la décroissance maximale que l'on peut obtenir sur f_{μ_+} à partir de $x \in V_\mu(\alpha)$ est bornée :*

$$f_{\mu_+}(x) - f_{\mu_+}(x(\mu_+)) \leqslant \frac{2\vartheta_f}{\sigma}. \tag{16.24}$$

Cette inégalité s'obtient en deux temps, en estimant la décroissance de f_{μ_+} de x à $x(\mu)$ puis de $x(\mu)$ à $x(\mu_+)$:

$$f_{\mu_+}(x) - f_{\mu_+}(x(\mu)) \leqslant \frac{\sqrt{\vartheta_f}}{\sigma} \tag{16.25}$$

$$f_{\mu_+}(x(\mu)) - f_{\mu_+}(x(\mu_+)) \leqslant \left(\frac{1}{\sigma} - 1\right)\vartheta_f. \tag{16.26}$$

En sommant ces deux inégalités et en tenant compte du fait que $\vartheta_f \geqslant 1$ (proposition 16.28), on obtient (16.24).

Pour établir (16.25), on commence par exprimer la convexité de f_{μ_+} :

$$f_{\mu_+}(x) - f_{\mu_+}(x(\mu)) \leqslant f'_{\mu_+}(x) \cdot (x - x(\mu)) = \langle d_{\mu_+}(x), x(\mu) - x \rangle_x.$$

Pour estimer le membre de droite, on note que $\|x - x(\mu)\|_x \leqslant 3/4$ lorsque $x \in V_\mu(1/4)$ (proposition 16.14) et on exprime $d_{\mu_+}(x)$ en fonction de $d_\mu(x)$, dont on connaît la grandeur :

$$d_{\mu_+}(x) = \frac{\mu}{\mu_+} d_\mu(x) + \left(\frac{\mu}{\mu_+} - 1\right) g_x(x) = \frac{1}{\sigma} d_\mu(x) + \left(\frac{1}{\sigma} - 1\right) g_x(x). \tag{16.27}$$

On obtient

$$\begin{aligned} f_{\mu_+}(x) - f_{\mu_+}(x(\mu)) &\leqslant \|d_\mu(x)\|_x \, \|x - x(\mu)\|_x \\ &\leqslant \left(\frac{1}{\sigma}\|d_\mu(x)\|_x + \left(\frac{1}{\sigma} - 1\right) \|g_x(x)\|_x\right) \frac{3}{4} \\ &\leqslant \left(\frac{1}{\sigma}\frac{1}{4} + \left(\frac{1}{\sigma} - 1\right) \sqrt{\vartheta_f}\right) \frac{3}{4} \\ &\leqslant \frac{1}{\sigma}\frac{1}{4}\sqrt{\vartheta_f} + \frac{1}{\sigma}\frac{3}{4}\sqrt{\vartheta_f} \qquad [\text{car } \vartheta_f \geqslant 1] \\ &= \frac{\sqrt{\vartheta_f}}{\sigma}. \end{aligned}$$

Pour établir (16.26), on commence également par exprimer la convexité de f_{μ_+} :

$$\begin{aligned} f_{\mu_+}(x(\mu)) - f_{\mu_+}(x(\mu_+)) &\leqslant f'_{\mu_+}(x(\mu)) \cdot (x(\mu) - x(\mu_+)) \\ &= \langle d_{\mu_+}(x(\mu)), x(\mu_+) - x(\mu) \rangle_{x(\mu)}. \end{aligned}$$

Comme précédemment, la direction de Newton $d_{\mu_+}(x(\mu))$ peut s'exprimer en fonction de $d_\mu(x(\mu))$, qui est nulle car $x(\mu)$ minimise f_μ, grâce à (16.27) avec x remplacé par $x(\mu)$:

$$d_{\mu_+}(x(\mu)) = \left(\frac{1}{\sigma} - 1\right) g_{x(\mu)}(x(\mu)).$$

En utilisant l'inégalité de droite dans (16.15), on trouve bien (16.26) :

$$f_{\mu_+}(x(\mu)) - f_{\mu_+}(x(\mu_+)) \leqslant \left(\frac{1}{\sigma} - 1\right) \langle g_{x(\mu)}(x(\mu)), x(\mu_+) - x(\mu) \rangle_{x(\mu)} \leqslant \left(\frac{1}{\sigma} - 1\right) \vartheta_f.$$

2) *Deuxième étape : tant que $x_i \notin V_{\mu_+}(\alpha)$, f_{μ_+} décroît au moins de 1/40 lors de chaque itération de Newton :*

$$x_i \notin V_{\mu_+}(1/4) \quad \Longrightarrow \quad f_{\mu_+}(x_{i+1}) \leqslant f_{\mu_+}(x_i) - \frac{1}{40}.$$

En effet, $f_{\mu_+}(x_{i+1}) \leqslant f_{\mu_+}(x_i + \tilde{t}_i d_{\mu_+}(x_i))$ avec le pas particulier $\tilde{t}_i := (5\|d_{\mu_+}(x_i)\|_{x_i})^{-1}$ puisque le pas t_i minimise $t \mapsto f_{\mu_+}(x_i + t d_{\mu_+}(x_i))$ et que $\tilde{t}_i \|d_{\mu_+}(x_i)\|_{x_i} < 1$ implique

que $x_i+\tilde{t}_i d_{\mu_+}(x_i) \in \operatorname{dom} f$. La proposition 16.12 permet alors de développer $f_{\mu_+}(x_i+\tilde{t}_i d_{\mu_+}(x_i))$:

$$\begin{aligned}
f(x_{i+1}) &\leqslant f_{\mu_+}(x_i+\tilde{t}_i d_{\mu_+}(x_i)) \\
&\leqslant f_{\mu_+}(x_i) - \tilde{t}_i \|d_{\mu_+}(x_i)\|_{x_i}^2 + \frac{1}{2}\tilde{t}_i^2 \|d_{\mu_+}(x_i)\|_{x_i}^2 + \frac{\tilde{t}_i^3 \|d_{\mu_+}(x_i)\|_{x_i}^3}{3(1-\tilde{t}_i\|d_{\mu_+}(x_i)\|_{x_i})}. \\
&= f_{\mu_+}(x_i) - \frac{1}{5}\|d_{\mu_+}(x_i)\|_{x_i} + \frac{1}{50} + \frac{1}{300} \\
&\leqslant f_{\mu_+}(x_i) - \frac{1}{20} + \frac{1}{50} + \frac{1}{300} \qquad [\|d_{\mu_+}(x_i)\|_{x_i} \geqslant \tfrac{1}{4} \text{ lorsque } x_i \notin V_{\mu_+}(1/4)] \\
&\leqslant f_{\mu_+}(x_i) - \frac{1}{40}.
\end{aligned}$$

3) *Troisième étape : le nombre d'itérations de Newton par étape de l'algorithme est borné.* En effet, en utilisant les résultats obtenus dans les deux premières étapes de la démonstration, on voit que, tant que $x_{i-1} \notin V_{\mu_+}(1/4)$, on a

$$\frac{i}{40} \leqslant f_{\mu_+}(x) - f_{\mu_+}(x_i) \leqslant f_{\mu_+}(x) - f_{\mu_+}(x(\mu_+)) \leqslant \frac{2\vartheta_f}{\sigma}.$$

Dès lors $i(x) \leqslant 80\,\vartheta_f/\sigma$.

4) *Conclusion.* Puisque $x_+ \in V_{\mu_+}(\alpha)$ en un nombre fini d'étapes, l'algorithme est bien défini. De plus, si on note μ_k ($k \geqslant 0$) les paramètres utilisés à chaque étape de l'algorithme on a évidemment $\mu_k = \sigma\mu_{k-1} = \sigma^k\mu_0$, qui converge donc q-linéairement vers zéro. De plus $\mu_k/\mu_0 \leqslant \varepsilon$ dès que $\sigma^k \leqslant \varepsilon$, ce qui s'écrit

$$k \geqslant \log_\sigma \varepsilon \quad \Longrightarrow \quad \mu_k \leqslant \varepsilon\mu_0$$

($\sigma \in \,]0,1[$, donc $x \mapsto \log_\sigma x$ est décroissante). Par ailleurs, le nombre total n_k d'itérations de Newton pour terminer les k premières étapes vérifie

$$n_k := \sum_{l=0}^{k-1} i(x_l) \leqslant \frac{80\,\vartheta_f}{\sigma}\, k.$$

On a donc les implications

$$n_k \geqslant \frac{80\,\vartheta_f}{\sigma}\,\log_\sigma \varepsilon \quad \Longrightarrow \quad k \geqslant \log_\sigma \varepsilon \quad \Longrightarrow \quad \mu_k \leqslant \varepsilon\mu_0.$$

□

La complexité itérative polynomiale de l'algorithme des grands déplacements est donc en $O(\vartheta_f \log_\sigma \varepsilon)$, ce qui est moins bien que celle de l'algorithme des petits déplacements, lequel est en $O(\sqrt{\vartheta_f}\log\varepsilon^{-1})$. Lorsqu'on le particularise à des cas concrets (voir les chapitres 18 et 20), cet algorithme est en pratique plus rapide que l'algorithme des petits déplacements. Il y a donc un écart entre la pratique et la théorie, qui n'est pas comblé à ce jour.

Algorithme prédicteur-correcteur ▲

16.2 Méthodes de points intérieurs en optimisation non linéaire ▲

On considère, pour simplifier, le problème avec contraintes d'inégalité seulement, celles pour lesquelles les algorithmes de points intérieurs (PI) ont été conçus. On l'écrit sous la forme

$$(P_I) \quad \begin{cases} \min\ f(x) \\ c(x) \geqslant 0, \end{cases}$$

où $f : \mathbb{R}^n \to \mathbb{R}$ et $c : \mathbb{R}^n \to \mathbb{R}^m$ sont des fonctions régulières. Le problème peut ne pas être convexe (on ne demande pas que f et les $-c_i$ soient convexes). On note $\lambda \in \mathbb{R}^m$ le multiplicateur associé à la contrainte dans le lagrangien, lequel est la fonction $\ell : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ définie par

$$\ell(x, \lambda) = f(x) - \lambda^{\mathsf{T}} c(x),$$

où le signe « $-$ » tient compte de la positivité requise des contraintes.

Pourvu que les contraintes soient qualifiées en une solution particulière x_*, il existe un multiplicateur optimal λ_* tel que l'on ait les conditions d'optimalité (KKT) suivantes

$$\begin{cases} \nabla_x \ell(x_*, \lambda_*) = 0 \\ C(x_*)\lambda_* = 0 \\ c(x_*) \geqslant 0 \quad \text{et} \quad \lambda_* \geqslant 0. \end{cases} \tag{16.28}$$

On a noté $C(x) := \mathrm{Diag}(c_1(x), \ldots, c_m(x))$. On note $L_* := \nabla^2_{xx} \ell(x_*, \lambda_*)$ la hessienne du lagrangien en $z_* := (x_*, \lambda_*)$ et

$$I^0 := \{i : c_i(x_*) = 0\} \qquad \text{et} \qquad I^- := \{i : c_i(x_*) < 0\}$$

l'ensemble des indices des contraintes actives et inactives en x_*. Évidemment ces ensembles d'indices forment une partition de $[1 : m]$.

Le premier intérêt des méthodes de points intérieurs est leur propension à mieux gérer l'*aspect combinatoire* des problèmes d'optimisation avec contraintes d'inégalité, c'est-à-dire de trouver la bonne manière d'activer celles-ci parmi les 2^m possibilités. Plus exactement, elles remplacent la combinatoire par du *mauvais conditionnement*, celui des systèmes linéaires à résoudre. La gestion de ce mauvais conditionnement relève de l'algèbre linéaire et est en général plus simple à prendre en compte.

16.2.1 Un algorithme de PI primal-dual

Schéma algorithmique

Le problème barrière associé à (P_I) s'écrit, pour un paramètre de pénalisation $\mu > 0$:

$$\min \left(f_\mu(x) := f(x) - \mu \sum_{i=1}^{m} \log c_i(x) \right).$$

Dans celui-ci la contrainte $c(x) > 0$ est implicite (le critère est considéré comme prenant la valeur $+\infty$ si $c(x) \not> 0$). Ce problème n'a donc de sens que s'il existe un

point strictement admissible, ce qui sera le cas sous les hypothèses 16.36 ci-dessous. Un point stationnaire x de f_μ est strictement admissible et vérifie, pour un $\lambda \in \mathbb{R}^m$, les conditions suivantes :

$$\begin{cases} \nabla_x \ell(x,\lambda) = 0 \\ C(x)\lambda = \mu e. \end{cases} \tag{16.29}$$

Ce sont les conditions d'optimalité (16.28), dans lesquelles les conditions de complémentarité ont été perturbées et les contraintes $c(x) > 0$ et $\lambda > 0$ sont implicites. L'ensemble

$$\mathscr{C} := \{z : F(z,\mu) = 0 \text{ pour un certain } \mu > 0\}$$

est appelé le *chemin central primal-dual*.

Même si les fonctions f et $-c_i$ sont convexes, la fonction f_μ ci-dessus n'est en général pas une fonction autoconcordante. Cela n'empêche pas de l'utiliser en pratique pour concevoir des algorithmes efficaces. Cependant, si l'on ne suppose que la continuité et la convexité de f et des $-c_i$, le chemin central pourra avoir une allure surprenante, parfois indésirable. Ainsi, dans $\mathbb{R}^2$, avec une fonction f convexe continue et la contrainte $x_2 \geqslant 0$, on peut avoir les chemins centraux de la figure 16.4. Les deux

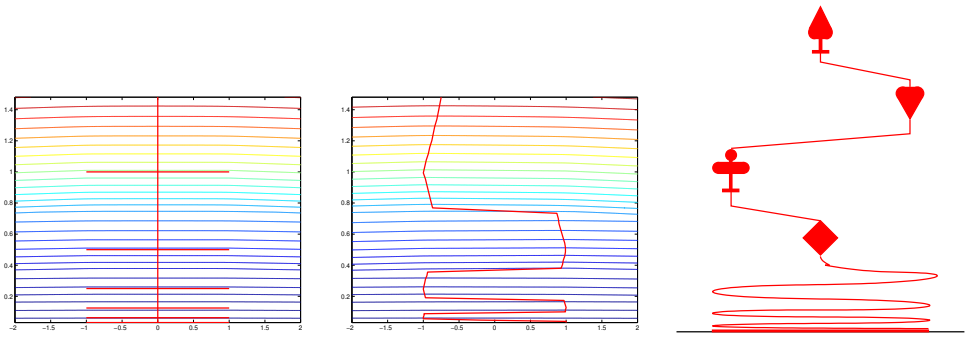


Fig. 16.4. Exemples de chemins centraux indésirables

premiers ont été dessinés avec une fonction f convexe continue, mais peuvent aussi se présenter si f est C^∞ [241]. Dans celui de gauche en forme d'antenne, il y a un nombre infini de segments horizontaux qui s'accumulent sur la droite $x_2 = 0$. Celui du milieu en zigzag a un nombre infini de tournants. De tels chemins centraux ne peuvent pas se présenter si les fonctions f et $-c_i$ sont convexes et analytiques [134, 433] ou si le lagrangien est fortement convexe pour un jeu de multiplicateurs [20]. Si la fonction f n'est que convexe continue, on peut avoir un chemin central comme dans le tracé de droite [33]. Ces exemples suggèrent que l'approche par points intérieurs n'est pas toujours appropriée.

On fera souvent référence à l'application $F : \mathbb{R}^{n+m} \times \mathbb{R} \to \mathbb{R}^{n+m}$ associée à (16.29) et définie en $z = (x,\lambda)$ et μ par

$$F(z,\mu) = \begin{pmatrix} \nabla_x \ell(x,\lambda) \\ C(x)\lambda - \mu e \end{pmatrix}.$$

Alors (16.29) se récrit $F(z,\mu) = 0$ tandis que z_* vérifie $F(z_*,0) = 0$. Observons que

$$F'_z(z,0) = F'_z(z,\mu) = \begin{pmatrix} \nabla^2_{xx}\ell(x,\lambda) & -c'(x)^{\mathsf{T}} \\ \Lambda c'(x) & C(x) \end{pmatrix}.$$

L'algorithme que nous allons étudier utilise des directions de Newton en $z = (x,\lambda)$ sur l'équation $F(z,\mu) = 0$. On la note $d_\mu(z)$ et elle est définie par le système linéaire

$$F'_z(z,\mu)\, d_\mu(z) = -F(z,\mu).$$

D'autre part, il génère des itérés z dans le voisinage $V := \cup_{\mu>0} V_\mu$ du chemin central, où pour $\mu > 0$ fixé

$$V_\mu := \Big\{(x,\lambda) \in \mathbb{R}^{n+m} : \quad c(x) > 0,\ \lambda > 0, \\ \|\nabla_x \ell(x,\lambda)\|_2 \leqslant \varepsilon^{\mathrm{D}}(\mu),\ \|C(x)\lambda - \mu e\|_2 \leqslant \varepsilon^{\mathrm{C}}(\mu)\Big\}. \tag{16.30}$$

La dépendance en μ des seuils $\varepsilon^{\mathrm{D}}(\cdot) > 0$ et $\varepsilon^{\mathrm{C}}(\cdot) > 0$ sera déterminée au cours de l'analyse pour que, asymptotiquement, l'algorithme soit bien posé et génère une suite convergeant rapidement.

Précisons à présent une itération de l'algorithme.

Algorithme 16.35 (PINL asymptotique — une itération)

Asymptotiquement, une itération de l'algorithme, de $z := (x,\lambda)$ à $z_+ := (x_+,\lambda_+)$ se déroule comme suit. On suppose au départ que $z \in V_\mu$.

1. Mise à jour : $\mu_+ := \phi(\mu)$.
2. Calcul de la direction de Newton $d_{\mu_+}(z)$.
3. $z_+ := z + d_{\mu_+}(z)$.

Cet algorithme fonctionne donc de manière assez semblable à l'algorithme 16.31 des petits déplacements et la figure 16.2 peut aussi le schématiser (sauf qu'ici les hypothèses 16.36 impliquent que l'ensemble $\mathcal{S}$ des solutions est *localement* réduit au point x_*). La différence essentielle vient de la définition de la direction de Newton qui est ici primale-duale (calculée sur $F(z,\mu_+) = 0$), alors qu'elle est primale dans l'algorithme des petits déplacements (calculée sur $\min f_{\mu_+}$). L'analyse et l'expérience ont montré qu'une direction primale-duale donnait de meilleurs résultats (théoriques et pratiques) qu'une direction primale, mais on n'avait pas d'autre choix pour résoudre le problème général (16.1), qui ne présente pas de multiplicateur de manière naturelle. Une autre différence porte sur la définition des voisinages V_μ, qui se fait ici à partir des conditions d'optimalité plutôt que de la direction de Newton. Cette différence n'est pas essentielle car, comme on le verra au lemme 16.40, ces deux quantités sont semblables ($\sim$) dans le voisinage d'une solution z_* vérifiant les hypothèses 16.36. Dans le cas non linéaire non convexe qui nous occupe ici, loin de z_*, on utilise souvent plusieurs itérations de Newton avec μ fixé ou les règles de mise à jour de ce paramètre sont telles que celui-ci varie peu lorsque l'itéré courant est éloigné du chemin central.

16.2.2 Analyse asymptotique

Dans cette section, le problème (P_I) et sa solution primale-duale $z_* = (x_*, \lambda_*)$ considérée sont supposés vérifier les hypothèses fortes suivantes.

Hypothèses 16.36
(H_1) Régularité : les fonctions f et c sont C^2 dans un voisinage de la solution x_*.
(H_2) (QC-IL) : $c'_{I^0}(x_*)$ est surjective,
(H_3) Complémentarité stricte : pour tout i, $c_i(x_*) + (\lambda_*)_i > 0$,
(H_4) (CS2) : $u^{\mathsf{T}} L_* u > 0$ pour toute direction non nulle u dans le noyau de $c'_{I^0}(x_*)$.

Dans l'analyse asymptotique de cette section, il sera nécessaire de comparer entre elles des grandeurs dépendant de $\mu > 0$, disons $u \equiv u(\mu)$ et $v \equiv v(\mu)$, lorsque $\mu \to 0$ (il s'agit ici d'une analyse asymptotique). On notera $u = O(v)$ s'il existe une constante C, indépendante de μ, telle que $\|u(\mu)\| \leqslant C\|v(\mu)\|$ pour tout $\mu > 0$ petit. On notera $u \sim v$ si $u = O(v)$ et $v = O(u)$. Enfin, on notera $u = o(v)$, s'il existe une fonction $\mu \mapsto \varepsilon(\mu)$ telle que $\varepsilon(\mu) \to 0$ et $\|u(\mu)\| \leqslant \varepsilon(\mu)\|v(\mu)\|$ lorsque $\mu \to 0$.

Comme le montre la proposition ci-dessous, sous les hypothèses 16.36, le chemin central se présente dans le voisinage de z_* comme une courbe régulière, que l'on note $\mu \mapsto z(\mu)$, définie pour μ voisin de zéro. En particulier, pour $\mu > 0$, la courbe est formée de points strictement admissibles. Par contre, pour de grand $\mu > 0$, rien n'assure que $F(x, \mu) = 0$ ait une unique solution (il peut ne pas y en avoir ou y en avoir plusieurs).

Proposition 16.37 (le chemin central) *Si les hypothèses 16.36 ont lieu, alors $F'_z(z_*, 0)$ est inversible et, dans un voisinage de z_*, le chemin central est une courbe de classe C^1. De plus*

$$(z(\mu) - z_*) \sim \mu.$$

Démonstration. On a

$$F'_z(z_*, 0) = \begin{pmatrix} L_* & -A_*^{\mathsf{T}} \\ \Lambda_* A_* & C_* \end{pmatrix},$$

où $A_* = c'(x_*)$, $\Lambda_* = \mathrm{Diag}(\lambda_*)$ et $C_* = \mathrm{Diag}(c(x_*))$. La première partie de la proposition sera une conséquence du théorème des fonctions implicites (théorème C.14) si l'on montre que $F'_z(z_*, 0)$ est inversible.

Soit (d, ν) dans le noyau de $F'_z(z_*, 0)$. La composante $i \in I^0$ de la seconde équation s'écrit $(\lambda_*)_i c'_i(x_*)d = 0$. Comme $(\lambda_*)_{I^0} > 0$ par complémentarité stricte, on voit que $d \in \mathcal{N}(c'_{I^0}(x_*))$. Les composantes $i \in I^-$ de la seconde équation impliquent par ailleurs que $\nu_{I^-} = 0$, si bien que la première équation devient $L_* d - c'_{I^0}(x_*)^{\mathsf{T}} \nu_{I^0} = 0$. En multipliant cette dernière par d, on trouve $d^{\mathsf{T}} L_* d = 0$ et donc $d = 0$ par la CS2. Alors, cette même équation et (QC-IL) montrent que $\nu_{I^0} = 0$, donc $\nu = 0$.

Venons-en à la seconde partie de la proposition. On note

$$\tilde{e} = (0, e) \in \mathbb{R}^n \times \mathbb{R}^m,$$

avec $e = (1, \ldots, 1) \in \mathbb{R}^m$. Alors, $\mu\tilde{e} = F(z(\mu), 0) - F(z(\mu), \mu) = F(z(\mu), 0)$, car $F(z(\mu), \mu) = 0$. En faisant un développement de Taylor de $F(z(\mu), 0)$ autour de z_*, on obtient

$$\mu\tilde{e} = F'_z(z_*, 0)(z(\mu) - z_*) + o(z(\mu) - z_*).$$

Comme $F'_z(z_*, 0)$ est inversible et que $\tilde{e} \sim 1$, on en déduit que $(z(\mu) - z_*) \sim \mu$. □

Le but de l'analyse asymptotique que nous présentons ci-après est de déterminer les fonctions $\varepsilon^{\mathrm{D}}(\cdot)$ et $\varepsilon^{\mathrm{C}}(\cdot)$ et de préciser la règle de décroissance de μ en $\mu_+ = \phi(\mu)$ à l'étape 1 de l'algorithme : cette dernière doit faire tendre μ vers zéro le plus rapidement possible, tout en assurant que $z_+ \in V_{\mu_+}$ lorsque $z \in V_\mu$. Le résultat principal de cette analyse est de montrer que la décroissance de μ peut être sous-quadratique dans l'algorithme 16.35, alors qu'elle était seulement linéaire dans l'algorithme 16.31 des petits déplacements. Évidemment les hypothèses ne sont pas identiques dans les deux cas. L'analyse est asymptotique dans le sens où z est supposé proche de z_*. Nous l'avons répartie sur plusieurs lemmes, de manière à bien localiser les besoins sur $\varepsilon^{\mathrm{D}}(\cdot)$, $\varepsilon^{\mathrm{C}}(\cdot)$ et $\phi(\cdot)$. Ces lemmes cherchent à établir la dépendance en μ de différentes grandeurs afin d'établir que $z - z_*$, tout comme $z(\mu) - z_*$ (proposition 16.37), se comporte comme μ, avec une suite de μ convergeant vers zéro sous-quadratiquement.

Lemme 16.38 (estimation de $c(x)$ et λ dans V_μ) *Si* (H_1) *et* (H_3) *ont lieu et si* z *est voisin de* z_* *et vérifie* $\|C(x)\lambda - \mu e\|_2 \leqslant \varepsilon^{\mathrm{C}}(\mu)$, *avec* $\varepsilon^{\mathrm{C}}(\mu) \leqslant \kappa\mu$ *et* $0 < \kappa < 1$, *alors*

$$\begin{aligned} i \in I^0 \quad &\Longrightarrow \quad c_i(x) \sim \mu \quad \text{et} \quad \lambda_i \sim 1, \\ i \in I^- \quad &\Longrightarrow \quad c_i(x) \sim 1 \quad \text{et} \quad \lambda_i \sim \mu. \end{aligned}$$

Démonstration. Par hypothèse, on a pour tout i, $|\lambda_i c_i(x) - \mu| \leqslant \varepsilon^{\mathrm{C}}(\mu) \leqslant \kappa\mu$. Dès lors

$$\forall i : \quad (1 - \kappa)\mu \leqslant \lambda_i c_i(x) \leqslant (1 + \kappa)\mu. \tag{16.31}$$

Pour $i \in I^0$, $(\lambda_*)_i > 0$ par complémentarité stricte, si bien que $\lambda_i \sim 1$ lorsque $z = (x, \lambda)$ est voisin de z_*; de (16.31), on déduit que $c_i(x) \sim \mu$. Pour $i \in I^-$, $c_i(x_*) > 0$, si bien que $c_i(x) \sim 1$ lorsque $z = (x, \lambda)$ est voisin de z_*; de (16.31), on déduit que $\lambda_i \sim \mu$. □

Lemme 16.39 *Si les hypothèses 16.36 ont lieu, si* $\mu > 0$ *est petit, si* $z \in V_\mu$ *est voisin de* z_*, *alors*

1) $z - z(\mu) = O(\varepsilon^{\mathrm{D}}(\mu)) + O(\varepsilon^{\mathrm{C}}(\mu))$,
2) $z - z_* = O(\varepsilon^{\mathrm{D}}(\mu)) + O(\varepsilon^{\mathrm{C}}(\mu)) + O(\mu)$.

Démonstration. Par définition de $z(\mu)$ on a $F(z(\mu), \mu) = 0$ et un calcul simple montre que $F'_z(z(\mu), \mu) = F'_z(z(\mu), 0)$. Alors un développement de Taylor de $F(z, \mu)$ autour de $z(\mu)$ donne

$$F(z, \mu) = F'_z(z(\mu), 0)(z - z(\mu)) + o(\|z - z(\mu)\|).$$

Comme $F'_z(z(\mu), 0)$ est d'inverse borné pour μ proche de zéro (proposition 16.37), on en déduit le point 1 :

$$(z - z(\mu)) \sim F(z,\mu) = O(\varepsilon^{\mathrm{D}}(\mu)) + O(\varepsilon^{\mathrm{C}}(\mu)),$$

la dernière estimation provenant du fait que $z \in V_\mu$. Le point 2 se déduit alors de

$$z - z_* = (z - z(\mu)) + (z(\mu) - z_*)$$

du point 1 et du fait que $z(\mu) - z_* = O(\mu)$ (proposition 16.37). □

Lemme 16.40 (estimation de $d_{\mu_+}(z)$) *Si les hypothèses 16.36 ont lieu et si z est voisin de z_*, alors*

$$d_{\mu_+}(z) \sim F(z,\mu_+).$$

Si, de plus, $z \in V_\mu$ et $0 \leqslant \mu_+ \leqslant \mu$, on a

$$d_{\mu_+}(z) = O(\varepsilon^{\mathrm{D}}(\mu)) + O(\varepsilon^{\mathrm{C}}(\mu)) + O(\mu).$$

Démonstration. La première estimation vient de

$$d_{\mu_+}(z) = -F'_z(z,\mu_+)^{-1}F(z,\mu_+) = -F'_z(z,0)^{-1}F(z,\mu_+)$$

et du fait que $F'_z(z,0)$ est borné et d'inverse borné pour z proche de z_* (proposition 16.37). Par la seconde estimation, on utilise le majorant suivant de $F(x,\mu_+)$:

$$F(z,\mu_+) = \underbrace{F(z,\mu)}_{O(\varepsilon^{\mathrm{D}}(\mu))+O(\varepsilon^{\mathrm{C}}(\mu))} + \underbrace{(\mu-\mu_+)\tilde{e}}_{\|\cdot\|_2 \leqslant \mu m^{1/2}}.$$

□

Le lemme ci-dessous cherche à établir des conditions pour que l'itéré suivant $z_+ := z + d_{\mu_+}(z)$ soit strictement admissible. Une de ces conditions impose que μ_+ ne soit pas trop petit devant μ, ce qui est naturel : il faut que $\mu^2 = o(\mu_+)$, ce qui interdit une décroissance quadratique du paramètre de pénalisation ($\mu_+ = O(\mu^2)$). Elle est réalisée, par exemple, en prenant $\mu_+ = \mu^{2-\tau}$, où $0 < \tau < 1$. On parle de *vitesse de convergence sous-quadratique* de la suite des μ vers zéro.

Lemme 16.41 (admissibilité stricte de z_+) *Si les hypothèses 16.36 ont lieu, si $z \in V_\mu$ est voisin de z_*, si $0 < \mu_+ \leqslant \mu$ avec μ assez petit et si*

$$\varepsilon^{\mathrm{D}}(\mu) = O(\mu), \quad \varepsilon^{\mathrm{C}}(\mu) \leqslant \kappa\mu \text{ avec } 0 < \kappa < 1 \quad \text{et} \quad \mu^2 = o(\mu_+),$$

alors $c(x_+) > 0$ et $\lambda_+ > 0$.

Démonstration. Observons que

$$d_\mu(z) = -F'_z(z,0)^{-1}F(z,\mu) = O(F(z,\mu)) = O(\varepsilon^{\mathrm{D}}(\mu)) + O(\varepsilon^{\mathrm{C}}(\mu)) = O(\mu). \quad (16.32)$$

Considérons d'abord les indices $i \in I^0$. On utilise systématiquement les estimations $c_i(x) \sim \mu$ et $\lambda_i \sim 1$ obtenues au lemme 16.38, ainsi que (16.32). Pour λ proche de λ_*, on a

$$(\lambda_+)_i = \underbrace{\lambda_i}_{\sim 1} + \underbrace{d\lambda_i}_{O(\mu)} > 0.$$

Par ailleurs, la i-ième composante de la seconde équation définissant $d_\mu(z) = (dx, d\lambda)$ s'écrit

$$\lambda_i c_i'(x)dx + c_i(x)d\lambda_i = \mu_+ - \lambda_i c_i(x). \tag{16.33}$$

Cette relation permet d'évaluer $c_i(x) + c_i'(x)dx$ et donc d'estimer $c_i(x_+)$ à partir de son développement de Taylor autour de x ($\gamma > 0$ est une constante indépendante de μ) :

$$\begin{aligned} c_i(x_+) &= c_i(x) + c_i'(x)dx + O(\|dx\|^2) \\ &= \underbrace{\lambda_i^{-1}}_{\sim 1}\mu_+ - \underbrace{\lambda_i^{-1}}_{\sim 1}\underbrace{c_i(x)}_{\sim \mu}\underbrace{d\lambda_i}_{O(\mu)} + \underbrace{O(\|dx\|^2)}_{O(\mu^2)} \\ &= \gamma\mu_+ + O(\mu^2) > 0, \end{aligned}$$

grâce à l'hypothèse $\mu^2 = o(\mu_+)$.

On s'y prend de manière semblable pour les indices $i \in I^-$, en utilisant systématiquement les estimations $c_i(x) \sim 1$ et $\lambda_i \sim \mu$ obtenues au lemme 16.38, ainsi que (16.32). Pour x proche de x_*, on a

$$c_i(x_+) = \underbrace{c_i(x)}_{\sim 1} + \underbrace{O(dx)}_{O(\mu)} > 0.$$

En partant de (16.33) et en utilisant l'hypothèse $\mu^2 = o(\mu_+)$, on obtient ($\gamma > 0$ est une constante indépendante de μ) :

$$(\lambda_+)_i = \lambda_i + d\lambda_i = \underbrace{c_i(x)^{-1}}_{\sim 1}\mu_+ - \underbrace{c_i(x)^{-1}}_{\sim 1}\underbrace{\lambda_i}_{\sim \mu}\underbrace{c_i'(x)dx}_{O(\mu)} \geqslant \gamma\mu_+ + O(\mu^2) > 0.$$

□

Lemme 16.42 (l'algorithme est bien posé) *Si les hypothèses 16.36 ont lieu, si $z \in V_\mu$ est voisin de z_*, si $0 < \mu_+ \leqslant \mu$ et si*

$$\varepsilon^{\mathrm{D}}(\mu) \sim \mu, \quad \varepsilon^{\mathrm{C}}(\mu) \sim \mu \quad \text{et} \quad \mu^2 = o(\mu_+),$$

alors $F(x_+, \mu_+) = o(\mu_+)$ et $z_+ \in V_{\mu_+}$.

Démonstration. Un développement de Taylor, la définition de $d_{\mu_+}(z)$, le lemme 16.31, les hypothèses sur $\varepsilon^{\mathrm{D}}(\mu)$ et $\varepsilon^{\mathrm{C}}(\mu)$ et enfin $\mu^2 = o(\mu_+)$ donnent

$$F(z_+, \mu_+) = \underbrace{F(z, \mu_+) + F_z'(z, \mu_+)d_{\mu_+}(z)}_{0} + O(\|d_{\mu_+}(z)\|^2) = O(\mu^2) = o(\mu_+).$$

ce qui démontre la première estimation annoncée. Par ailleurs, $\nabla_x \ell(x_+, \lambda_+)$ et $C(x_+)\lambda_+ - \mu_+ e$, que l'on vient d'estimer à des $o(\mu_+)$, vérifient donc certainement

$$\|\nabla_x \ell(x_+, \lambda_+)\|_2 \leqslant \varepsilon^{\mathrm{D}}(\mu_+) \qquad \text{et} \qquad \|C(x_+)\lambda_+ - \mu_+ e\|_2 \leqslant \varepsilon^{\mathrm{C}}(\mu_+),$$

ce qui veut dire que $z_+ \in V_{\mu_+}$. □

Les lemmes précédents montrent que l'algorithme est bien posé si $\varepsilon^{\mathrm{D}}(\mu)$, $\varepsilon^{\mathrm{C}}(\mu)$ et μ_+ vérifient les règles suivantes :

$$\text{il existe des constantes } 0 < \kappa_1^{\mathrm{D}} \leqslant \kappa_2^{\mathrm{D}} \text{ telles que } \kappa_1^{\mathrm{D}}\mu \leqslant \varepsilon^{\mathrm{D}}(\mu) \leqslant \kappa_2^{\mathrm{D}}\mu,$$
$$\text{il existe des constantes } 0 < \kappa_1^{\mathrm{C}} \leqslant \kappa_2^{\mathrm{C}} < 1 \text{ telles que } \kappa_1^{\mathrm{C}}\mu \leqslant \varepsilon^{\mathrm{C}}(\mu) \leqslant \kappa_2^{\mathrm{C}}\mu,$$
$$\mu^2 = o(\mu_+).$$

Il reste à monter que $(z_+ - z) \sim \mu_+$, ce qui permettra de conclure que la convergence de $z \to z_*$ est sous-quadratique.

Lemme 16.43 (z converge tangentiellement au chemin central) *Si les hypothèses précédentes ont lieu, on a*

$$z_+ = z_* + \mu_+ z'(0) + o(\mu_+), \tag{16.34}$$

où $z'(0)$ est la dérivée de $z(\cdot)$ en zéro.

Démonstration. Comme $F'_z(z_*, 0)$ est inversible, en dérivant $F(z(\mu), \mu)$ en $\mu = 0$, on obtient

$$z'(0) = -F'(z_*, 0)^{-1} F'_\mu(z_*, 0) = -F'(z_*, 0)^{-1} \begin{pmatrix} 0 \\ -e \end{pmatrix}. \tag{16.35}$$

Par un développement de Taylor, $F(z_+, \lambda_+) = o(\mu_+)$ (lemme 16.42), on a

$$\underbrace{F(z_+, \lambda_+)}_{o(\mu_+)} = \underbrace{F(z_*, 0)}_{0} + F'_z(z_*, 0)(z_+ - z_*) + F'_\mu(z_*, 0)\mu_+ + o(\|z_+ - z_*\|) + o(\mu_+).$$

En multipliant à gauche par $F'(z_*, 0)^{-1}$, en tenant compte de (16.35) et en réarrangeant, on trouve

$$z_+ - z_* = \mu_+ z'(0) + o(\|z_+ - z_*\|) + o(\mu_+).$$

On en déduit que $z_+ - z_* = O(\mu_+)$ et donc le résultat. □

Proposition 16.44 (convergence sous-quadratique de PINL) *Si les hypothèses 16.36 ont lieu, si $z \in V_\mu$ est voisin de z_*, si $0 < \mu_+ \leqslant \mu$ et si*

$$\varepsilon^{\mathrm{D}}(\mu) \sim \mu, \quad \varepsilon^{\mathrm{C}}(\mu) \sim \mu \quad \text{et} \quad \mu_+ = O(\mu^{2-\tau}), \text{ avec } 0 < \tau < 1,$$

on a $(z_+ - z_*) \sim \mu_+$. En particulier, la convergence de z vers z_* est sous-quadratique.

Démonstration. D'après (16.34), $z_+ - z_* = O(\mu_+)$. Réciproquement, en multipliant cette équation (16.34) par $z'(0)$, on trouve

$$\|z'(0)\|_2 \, \|z_+ - z_*\|_2 \geqslant z'(0)^{\mathsf{T}}(z_+ - z_*) = \mu_+ \|z'(0)\|_2^2 + o(\mu_+).$$

D'après (16.35), $z'(0) \neq 0$, si bien que $\mu_+ = O(\|z_+ - z_*\|)$. □

Notes

La présentation de la section 16.1 suit d'assez près l'approche de Renegar [512 ; 2001] et peut être vue comme une introduction à ce très bel opuscule. L'ouvrage de référence est celui de Nesterov et Nemirovskii [455 ; 1993] et nous en avons parfois repris certains résultats non présents chez Renegar (voir aussi Nesterov [452 ; 2004]). La théorie des fonctions autoconcordantes s'enrichit considérablement lorsque leur domaine est l'intérieur d'un cône convexe fermé K. Elle est utile lorsque le problème d'optimisation à résoudre a un ensemble admissible qui est l'intersection de K et d'un sous-espace affine. On parle d'*optimisation conique*. On peut alors introduire des algorithmes de points intérieurs primaux-duaux (les variables duales sont associées aux contraintes affines). Les méthodes vues aux chapitres 18 et 20 en sont des cas particuliers. Les bases de l'optimisation conique sont présentées dans les deux ouvrages précités, ainsi que dans [44].

L'analyse asymptotique de la section 16.2.2 est reprise de [271]. Elle peut se mener sous des conditions plus faibles que les hypothèses 16.36, voir [628].

Exercices

16.1. *Autre expression de* lb. Un *ellipsoïde non dégénéré* centré en zéro est un ensemble de la forme $\mathscr{E}_A = \{v \in \mathbb{R}^n : v^\mathsf{T} A v \leqslant 1\}$, où $A \succ 0$. Si $x \in \mathbb{R}^n_{++}$, $\mathrm{lb}(x) = \log \sigma_n / V(x)$ où σ_n est le volume de la sphère unité dans $\mathbb{R}^n$ et $V(x)$ est le volume du plus volumineux *ellipsoïde inscrit dans* $\mathbb{R}^n_+$ et centré en x. Montrez que cet ellipsoïde est la boule unité $\bar{B}_x(0,1)$ pour la norme locale en x associée à lb.

16.2. *Exemples et contre-exemples de fonctions et de barrières autoconcordantes* [452].

1) Une fonction quadratique strictement convexe $f : x \in \mathbb{R}^n \mapsto \frac{1}{2} x^\mathsf{T} A x + b^\mathsf{T} x$, où $A \succ 0$ et $b \in \mathbb{R}^n$, est une fonction autoconcordante, mais n'est pas une barrière autoconcordante.

Partie IV

Étude de problèmes particuliers

17 Optimisation linéaire : théorie et algorithme du simplexe

Curiously, up to 1947 when I first proposed that a model based on linear inequalities be used for planning activities of large-scale enterprises, linear inequality theory had produced only 40 or so papers, in contrast to linear equation theory and the related subjects of linear algebra and approximation, which had produced a vast literature.

G.B. Dantzig (1914-2005). Origins of the simplex method [154 ; 1990].

Connaissances supposées. Une grande partie de la section 2.4 nous sera utile, en particulier pour les notions de polyèdre convexe et de sommet, ainsi que pour le résultat d'existence de solution de problème d'optimisation linéaire (théorème 2.20). Les conditions d'optimalité d'un problème d'optimalité linéaire seront directement déduites des conditions de Karush, Kuhn et Tucker (théoreme 4.32 et proposition 4.33). Le problème dual d'un problème d'optimisation linéaire s'obtiendra par la dualité min-max (section 14.1).

17.1 Introduction

17.1.1 Le problème à résoudre

Un *problème d'optimisation linéaire* (OL) est un problème d'optimisation dans lequel le critère et les fonctions définissant les contraintes sont linéaires (on devrait dire affines) ; il s'agit donc de minimiser une fonction linéaire sur un polyèdre convexe. Comme nous le verrons, la formulation suivante du problème est tout à fait générale. Il s'agit de trouver la solution $x \in \mathbb{R}^n$ du problème

$$(P_L) \quad \begin{cases} \inf_x \ c^\mathsf{T} x \\ Ax = b \\ x \geqslant 0. \end{cases} \tag{17.1}$$

Un problème d'optimisation linéaire écrit de cette manière est dit sous *forme standard*. On appelle *critère* du problème, la fonction linéaire $x \mapsto c^\mathsf{T} x$. Les données dans (P_L) sont deux vecteurs $c \in \mathbb{R}^n$ (appelé *coût* du problème) et $b \in \mathbb{R}^m$ (en général $m \leqslant n$), et une matrice A de dimension $m \times n$. La contrainte d'inégalité $x \geqslant 0$ veut dire

que toutes les composantes de x doivent être positives : $x_i \geqslant 0$ pour tout $i \in [1:n]$. On notera $x > 0$, lorsque toutes les composantes de x seront strictement positives. L'ensemble

$$\mathbb{R}^n_+ = \{x \in \mathbb{R}^n : x \geqslant 0\}$$

est appelé l'*orthant positif* de $\mathbb{R}^n$.

On rappelle que l'on dit que le problème (P_L) est *réalisable* si son *ensemble admissible*

$$\mathcal{F}_{\text{P}} := \{x \in \mathbb{R}^n : Ax = b,\ x \geqslant 0\}$$

est non vide. Un point de $\mathcal{F}_{\text{P}}$ est dit *admissible*. On dit que (P_L) est *borné* si la valeur optimale de (P_L) est finie.

Dans ce chapitre, nous allons étudier le problème (P_L) (structure et propriétés) et donner un algorithme pour le résoudre : l'algorithme du simplexe. Les algorithmes de points intérieurs seront décrits et étudiés au chapitre 18. Remarquons déjà que le problème (P_L) n'est « intéressant », c'est-à-dire qu'il mérite une étude approfondie, que par la présence des contraintes d'inégalité. En effet, sans ces contraintes le problème se réduit à deux cas complémentaires triviaux. Si $c \in \mathcal{N}(A)^\perp$, tout point x vérifiant $Ax = b$ est solution. Si $c \notin \mathcal{N}(A)^\perp$, il existe une direction d telle que $Ad = 0$ et $c^\mathsf{T} d < 0$ et donc $c^\mathsf{T}(x + td) \to -\infty$ lorsque $t \to +\infty$: si le problème est réalisable (en l'occurrence, il existe un x tel que $Ax = b$), il n'est pas borné.

En ce qui concerne les algorithmes de résolution de (P_L), on distingue deux classes principales de méthodes.

- Les *méthodes de points-frontière* génèrent des itérés sur la frontière de l'ensemble admissible $\mathcal{F}_{\text{P}}$. La méthode typique de cette classe est l'*algorithme du simplexe*, introduit par Dantzig en 1947 [153 ; 1951] ; il consiste à faire des déplacements le long des arêtes du polyèdre convexe $\mathcal{F}_{\text{P}}$; nous le présenterons et l'étudierons à la section 17.4. Cet algorithme n'est pas polynomial dans le pire des cas, bien qu'il soit polynomial en moyenne et en perturbations moyennées (section 17.4.5).
- Les *méthodes de points intérieurs*, développées à partir des travaux de Karmarkar [346 ; 1984], génèrent des itérés dans l'intérieur relatif de l'ensemble admissible $\mathcal{F}_{\text{P}}$. Ces méthodes peuvent être polynomiales et sont particulièrement intéressantes lorsqu'il y a beaucoup de contraintes d'inégalité, parce qu'elles ne ressentent pas l'irrégularité du bord de l'ensemble admissible. Nous présenterons et étudierons quelques algorithmes de points intérieurs au chapitre 18.

17.1.2 Formulations canoniques

Un problème d'optimisation linéaire se présente souvent sous l'une des deux formes suivantes : la forme standard (P_L) donnée ci-dessus et la *forme canonique*

$$(P'_L) \quad \begin{cases} \min (c')^\mathsf{T} x \\ A'x \leqslant b'. \end{cases}$$

Ces deux formulations sont équivalentes dans le sens où le problème (P_L) peut s'écrire sous la forme (P'_L) :

$$\left\{ \begin{array}{l} \min c^{\mathsf{T}} x \\ \begin{pmatrix} A \\ -A \\ -I \end{pmatrix} x \leqslant \begin{pmatrix} b \\ -b \\ 0 \end{pmatrix} \end{array} \right.$$

et inversement, le problème (P'_L) peut s'écrire sous la forme (P_L) en décomposant $x = u - v$, avec $u \geqslant 0$ et $v \geqslant 0$, et en introduisant des *variables d'écart* $z \geqslant 0$:

$$\left\{ \begin{array}{l} \min \begin{pmatrix} c^{\mathsf{T}} & -c^{\mathsf{T}} & 0 \end{pmatrix} \begin{pmatrix} u \\ v \\ z \end{pmatrix} \\ \\ A'u - A'v + z = b' \\ u \geqslant 0, \quad v \geqslant 0, \quad z \geqslant 0. \end{array} \right.$$

On a pris l'habitude de décrire et d'étudier les algorithmes de résolution des problèmes d'optimisation linéaire mis sous forme standard (P_L) et c'est ce que nous ferons dans ce chapitre et le suivant. Cependant, il peut être plus avantageux de considérer une autre formulation, en fonction du problème que l'on a à résoudre. D'autre part, les codes d'optimisation linéaire généralistes acceptent le plus souvent des contraintes écrites sous la forme :

$$Ax = b \quad \text{et} \quad l \leqslant Bx \leqslant u,$$

où les matrices A et B et les vecteurs b, l et u ont des dimensions appropriées. Il faut donc parfois adapter les algorithmes au cadre que l'on considère ; ceci se fait en général sans difficulté.

17.1.3 Exemples : problèmes d'optimisation dans des réseaux

Un *réseau* n'est pas autre chose qu'un *graphe* (orienté ou non), c'est-à-dire une collection de nœuds et d'arcs qui relient ces nœuds. Disons qu'il y a m nœuds, indicés par $i = 1, \ldots, m$, et n arcs, indicés par $j = 1, \ldots, n$. Ces derniers peuvent aussi être spécifiés par des couples (ordonnés ou non, suivant que le graphe est ou n'est pas orienté) formés des nœuds qu'ils relient : l'arc (i_1, i_2) relie les nœuds i_1 et i_2.

Dans la discussion qui suit, le réseau est supposé être utilisé comme support à l'acheminement d'un produit (eau, gaz, électricité, voitures, trains, avions, télécommunications, messages internet, quantité abstraite, *etc*). La quantité de produit qui transite par l'arc j est notée x_j. Il est normal de supposer que c'est une grandeur positive. L'équilibre du réseau s'exprime en écrivant que la quantité de produit qui sort du nœud i est égale à la quantité qui y entre plus la quantité b_i qui y est produite (« loi de Kirchhoff »). Pour tout i, on doit avoir

$$\sum_{\substack{j\,:\,\text{arc sortant} \\ \text{du nœud } i}} x_j = \sum_{\substack{j\,:\,\text{arc entrant} \\ \text{du nœud } i}} x_j + b_i. \tag{17.2}$$

La quantité b_i produite au nœud i peut représenter ce qui y est apporté (auquel cas $b_i > 0$) ou ce qui y est consommé ou demandé (auquel cas $b_i < 0$). Les équations précédentes s'écrivent sous forme matricielle :

$$Ax = b \quad \text{et} \quad x \geqslant 0, \tag{17.3}$$

où la matrice A est appelée la *matrice d'incidence du réseau*. Cette matrice peut avoir des dimensions considérables en pratique, mais elle est très creuse puisqu'elle n'a que deux éléments non nuls par colonne. En effet, au vu de (17.2), la colonne j associée à l'arc j contient $A_{ij} = +1$ si i est le nœud d'origine de l'arc j, $A_{ij} = -1$ si i est le nœud de destination de l'arc j et $A_{ij} = 0$ dans les autres cas. En particulier, la somme des lignes de A est nulle, ce qui implique que A n'est pas surjective. Ceci peut poser des difficultés algorithmiques et requiert au moins une condition de compatibilité sur b, à savoir $b \in \mathcal{R}(A)$, ce qui conduit à

$$\sum_{i=1}^{m} b_i = 0. \tag{17.4}$$

On montre cependant que A est de rang $m-1$ lorsque le graphe associé au réseau est *connexe*, ce qui veut dire que deux nœuds distincts i et i' peuvent être reliés par une chemin (une suite d'arcs (i_1, i_2), (i_2, i_3), $\ldots$, (i_{p-1}, i_p), avec $i_1 = i$ et $i_p = i'$) ; voir [57] par exemple.

Un exemple de problème d'optimisation linéaire classique dans les réseaux est le *problème du transport à coût minimal*. On cherche à acheminer des produits identiques depuis certains *nœuds d'origine* (disons pour $i \in O$) jusqu'à des *nœuds de destination* (disons pour $i \in D$, avec $O \cap D = \varnothing$). Ceci s'exprime par le positionnement des composantes de b aux valeurs désirées : $b_O > 0$, $b_D < 0$ et $b_i = 0$ si $i \notin O \cup D$, tout en respectant (17.4). D'autre part, en supposant que le coût de transport sur chaque arc est linéaire en x_j (coût c_j par unité transportée), le coût total du transport sera

$$\sum_{j=1}^{n} c_j x_j = c^{\mathsf{T}} x.$$

C'est la quantité que l'on cherche à minimiser. On reconnaît dans la minimisation de ce critère sous les contraintes de réseau (17.3), le problème d'optimisation linéaire sous forme standard (P_L).

Un autre exemple de problème d'optimisation linéaire classique et celui du *plus court chemin dans un graphe*. C'est un cas particulier du problème précédent : $b_{i_0} = +1$ au nœud de départ i_0, $b_{i_1} = -1$ au nœud d'arrivée i_1 et les autres b_i sont nuls, tandis que $c_j > 0$ donne la longueur du chemin représenté par l'arc j. Si $\bar{x}$ est la solution du problème, $\bar{x}_j = 0$ signifie que l'on ne passe pas par l'arc j et $\bar{x}_j = 1$ signifie que l'on y passe. On peut avoir des valeurs entre 0 et 1 si le problème a plusieurs solutions.

17.2 Étude du problème

17.2.1 Structure de l'ensemble admissible

L'ensemble admissible $\mathcal{F}_{\mathrm{P}}$ d'un problème d'optimisation linéaire est un polyèdre convexe (voir la section 2.4), exprimé dans une représentation duale. Les développements théoriques et algorithmiques en optimisation linéaire se font en général sur la forme suivante de l'ensemble admissible

$$\mathcal{F}_{\text{P}} = \{x \in \mathbb{R}^n : Ax = b,\ x \geqslant 0\},$$

où A est $m \times n$ et $b \in \mathbb{R}^m$. On supposera alors toujours que

$$A \text{ est surjective.}$$

Théoriquement, ce n'est pas une hypothèse restrictive (en pratique, c'est plus délicat). En effet si $b \notin \mathcal{R}(A)$, l'ensemble admissible est vide et le problème est mal posé. Sinon, on peut toujours éliminer les équations redondantes de $Ax = b$ pour se ramener à une matrice surjective (si des lignes de A sont linéairement dépendantes [$A^{\mathsf{T}}y = 0$ pour un $y \neq 0$], les éléments de b obéissent à la même dépendance [$b^{\mathsf{T}}y = 0$]).

On adopte une notation déjà utilisée. Si $x \in \mathbb{R}^n_+$, on écrit

$$I^+(x) := \{i : x_i > 0\} \quad \text{et} \quad I^0(x) := \{i : x_i = 0\}.$$

On rappelle que $x \in \mathcal{F}_{\text{P}}$ est un sommet de $\mathcal{F}_{\text{P}}$ si la sous-matrice $A_{:I^+(x)}$, formée des colonnes de A avec indices dans $I^+(x)$, est injective (proposition 2.21). On a donc nécessairement $|I^+(x)| \leqslant m$, et on peut bien sûr avoir $|I^+(x)| < m$.

Définition 17.1 On dit qu'un sommet $x \in \mathcal{F}_{\text{P}}$ est *dégénéré* si $|I^+(x)| < m$ et qu'il est *non dégénéré* si $|I^+(x)| = m$. □

Un concept jouant un rôle-clé dans l'algorithme du simplexe (section 17.4) est celui de base d'indices.

Définitions 17.2 On appelle *base d'indices* un ensemble B de m indices pris dans $[1:n]$ tels que la sous-matrice $A_{:B}$ formée des m colonnes correspondantes de A soit inversible.

Si $x \in \mathbb{R}^n$, les composantes x_i avec $i \in B$ sont alors dites *basiques* et celles avec $i \notin B$ sont dites *non basiques*.

On dit qu'une base d'indices B est *associée* à un sommet $x \in \mathcal{F}_{\text{P}}$ si $I^+(x) \subseteq B$. □

À une base d'indices B on associe l'ensemble d'indices complémentaire $N = [1:n] \setminus B$. Après permutations éventuelles des colonnes, on pourra donc écrire

$$A = \begin{pmatrix} A_{:B} & A_{:N} \end{pmatrix},$$

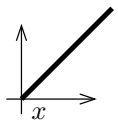
avec $A_{:B}$ inversible. On note (x_B, x_N) la partition correspondante d'un point $x \in \mathbb{R}^n$.

Un sommet non dégénéré n'a, bien sûr, qu'une seule base d'indices qui lui est associée, à savoir $I^+(x)$. Si x est un sommet dégénéré, le nombre de bases d'indices qui lui sont associées peut être très grand, bien que majoré par

$$\binom{n - |I^+(x)|}{m - |I^+(x)|} = \binom{n - |I^+(x)|}{n - m}.$$

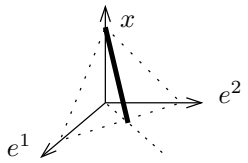
Exemples 17.3 Dans $\mathbb{R}^2$, $x = 0$ est un sommet dégénéré de l'ensemble défini par

$$\begin{cases} x_1 - x_2 = 0 \\ x \geqslant 0. \end{cases}$$



Dans $\mathbb{R}^3$, le point $x = (0, 0, 1)$ est un sommet dégénéré de l'ensemble défini par

$$\begin{cases} x_1 + x_2 + x_3 = 1 \\ x_1 - x_2 \qquad = 0 \\ x \geqslant 0. \end{cases}$$



□

On connait des problèmes d'optimisation linéaire pratiques dans lesquels l'ensemble admissible a un nombre factoriel de sommets et chaque sommet a un nombre exponentiel de bases d'indices [34 ; 1972]. On a donc imaginé des techniques capables de faire face à cette dégénérescence.

L'optimisation linéaire a son propre jargon, que l'on doit reconnaître si l'on veut comprendre les ouvrages et articles écrits par les spécialistes de la discipline, en particulier les contributions fondatrices. Certains termes, chargés d'histoire, apportent pourtant d'inutiles complications et confusions ; nous les éviterons. Il en va ainsi de *solution* pour désigner un point (pour nous, une *solution* sera une solution du problème d'optimisation linéaire), de *solution admissible* pour désigner un point admissible ou encore de *solution basique admissible* pour désigner un sommet de $\mathcal{F}_{\text{P}}$. Plus d'un demi-siècle après l'introduction de cette terminologie compliquée, nous nous sommes permis de la simplifier et de l'accorder avec celle utilisée en analyse convexe et en optimisation non linéaire.

Jargon de l'optimisation linéaire	Terminologie adoptée
base	base d'indices
solution	point
solution admissible	point admissible
solution basique	—
solution basique admissible	sommet
solution admissible optimale	solution
solution basique admissible optimale	solution-sommet

On trouvera dans le lexique en fin d'ouvrage la terminologie anglo-saxone dont sont issus les termes de la première colonne.

17.2.2 Existence de solution et conditions d'optimalité

On considère le problème (P_L) sous la forme standard

$$(P_L) \quad \begin{cases} \min c^{\mathsf{T}} x \\ Ax = b \\ x \geqslant 0. \end{cases}$$

On note

$$\inf(P_L) := \inf_{x \in \mathcal{F}_{\text{P}}} c^{\mathsf{T}} x$$

la valeur optimale de (P_L).

Pour le problème (P_L), l'existence de solution ne résulte pas directement du théorème sur la minimisation de fonctions continues sur un compact (théorème 1.2). En effet, l'ensemble admissible $\mathcal{F}_P$ n'est pas nécessairement compact. De plus, $x \mapsto c^T x$ ne tend pas nécessairement vers $+\infty$ lorsque $\|x\| \to \infty$ dans $\mathcal{F}_P$ (sauf si l'ensemble des solutions est borné et si $\mathcal{F}_P$ n'est pas borné, voir l'exercice 17.4). Il est donc nécessaire de donner une démonstration spécifique. Une possibilité est d'utiliser la remarque 2.19 comme dans la démonstration de la proposition 2.20. Nous rappelons le résultat ci-dessous.

Théorème 17.4 (existence de solution) *Le problème (P_L) a une solution si, et seulement si, il est réalisable ($\mathcal{F}_P \neq \varnothing$) et borné ($\inf f > -\infty$).*

L'algorithme du simplexe repose sur le résultat suivant. Cet algorithme va rechercher en effet une solution-sommet, dont l'existence est assurée par la proposition qui suit. Ce résultat dépend fortement de la représentation de l'ensemble admissible adoptée dans (P_L) et ne serait plus vrai pour un ensemble admissible de la forme $\{x \in \mathbb{R}^n : Ax \leqslant b\}$.

Proposition 17.5 (existence de solution-sommet) *Si le problème (P_L) a une solution, alors il a une solution en un sommet de $\mathcal{F}_P$.*

Démonstration. L'ensemble des solutions de (P_L),

$$\mathrm{Sol}(P_L) = \{x \in \mathbb{R}^n : c^T x = \mathrm{val}(P_L),\ Ax = b,\ x \geqslant 0\},$$

est un polyèdre convexe écrit sous forme standard. Il a donc un sommet (proposition 2.21), disons $\hat{x}$. D'autre part, $\mathrm{Sol}(P_L)$ est une face de $\mathcal{F}_P$ (exercice 2.11). Donc son sommet $\hat{x}$ est aussi un sommet de $\mathcal{F}_P$ (exercice 2.16, point 2). □

Voici le résultat donnant les conditions d'optimalité du problème (P_L).

Proposition 17.6 (conditions d'optimalité) *Le point x est solution de (P_L) si, et seulement si, il existe $y \in \mathbb{R}^m$ et $s \in \mathbb{R}^n$ tels que*

$$\begin{cases} (a) & A^T y + s = c, \quad s \geqslant 0 \\ (b) & Ax = b, \quad x \geqslant 0 \\ (c) & x^T s = 0. \end{cases} \tag{17.5}$$

Démonstration. Ce sont les conditions de KKT qui sont nécessaires (on utilise ici le fait que les contraintes de (P_L) sont qualifiées, car affines) et suffisantes (car le problème (P_L) est convexe). On les obtient en introduisant le *lagrangien*

$$\ell(x, y, s) = c^T x - y^T (Ax - b) - x^T s.$$

□

La nécessité des conditions (17.5) peut aussi se voir en observant que, par optimalité de x, le critère augmente le long des directions d tangentes en x à l'ensemble admissible (voir le théorème 4.6), ce qui s'exprime par

$$c \in \{d : Ad = 0,\ d_I \geqslant 0\}^+, \quad \text{avec } I = \{i : x_i = 0\},$$

et en utilisant le lemme de Farkas. D'autre part, le fait que les conditions (17.5) soient suffisantes pour impliquer l'optimalité de x peut aussi se montrer directement : pour tout $x' \in \mathcal{F}_{\text{P}}$, on a

$$\begin{aligned} c^{\mathsf{T}}x &= s^{\mathsf{T}}x + y^{\mathsf{T}}Ax \qquad [\text{par } (a)] \\ &= y^{\mathsf{T}}b \qquad [\text{par } (b) \text{ et } (c)] \\ &\leqslant s^{\mathsf{T}}x' + y^{\mathsf{T}}Ax' \qquad [\text{par } Ax' = b,\ x' \geqslant 0 \text{ et } s \geqslant 0] \\ &= c^{\mathsf{T}}x' \qquad [\text{par } (a)]. \end{aligned}$$

Le couple (y, s) donné par la proposition précédente est appelé *solution duale* du problème linéaire (P_L) et x est alors appelé *solution primale*. La variable s est aussi appelée *variable d'écart duale*, car elle joue le rôle de variable d'écart dans la relation $A^{\mathsf{T}}y \leqslant c$, qui est une autre façon d'écrire l'équation (17.5)(a). On note $\mathcal{S}_{\text{P}} \subseteq \mathbb{R}^n$ l'ensemble des solutions primales et $\mathcal{S}_{\text{D}} \subseteq \mathbb{R}^m \times \mathbb{R}^n$ l'ensemble des solutions duales. On note aussi $\mathcal{S}_{\text{PD}}$ l'ensemble des triplets (x, y, s) qui sont solutions du système d'optimalité (17.5). Malgré la non-linéarité de (17.5)(c), l'ensemble $\mathcal{S}_{\text{PD}}$ est un convexe, comme le montre la proposition suivante.

Proposition 17.7 (produit cartésien des solutions) *On a*

$$\mathcal{S}_{\text{PD}} = \mathcal{S}_{\text{P}} \times \mathcal{S}_{\text{D}},$$

en particulier, $\mathcal{S}_{\text{PD}}$ est un polyèdre convexe. Autrement dit, si (x^1, y^1, s^1) et (x^2, y^2, s^2) sont des solutions primales-duales alors (x^1, y^2, s^2) et (x^2, y^1, s^1) le sont aussi.

Démonstration. Par hypothèse, on a

$$\begin{cases} A^{\mathsf{T}}y^1 + s^1 = c, \quad s^1 \geqslant 0 \\ Ax^1 = b, \quad x^1 \geqslant 0 \\ (x^1)^{\mathsf{T}}s^1 = 0 \end{cases} \quad \text{et} \quad \begin{cases} A^{\mathsf{T}}y^2 + s^2 = c, \quad s^2 \geqslant 0 \\ Ax^2 = b, \quad x^2 \geqslant 0 \\ (x^2)^{\mathsf{T}}s^2 = 0. \end{cases}$$

On voit qu'il suffit de montrer que $(x^1)^{\mathsf{T}}s^2 = 0$ et $(x^2)^{\mathsf{T}}s^1 = 0$. Comme $c^{\mathsf{T}}x^1 = c^{\mathsf{T}}x^2$, on a

$$\begin{aligned} (x^1)^{\mathsf{T}}s^2 &= (x^1)^{\mathsf{T}}(c - A^{\mathsf{T}}y^2) \\ &= (x^1)^{\mathsf{T}}c - b^{\mathsf{T}}y^2 \\ &= (x^2)^{\mathsf{T}}c - (Ax^2)^{\mathsf{T}}y^2 \\ &= (x^2)^{\mathsf{T}}(c - A^{\mathsf{T}}y^2) \\ &= (x^2)^{\mathsf{T}}s^2 \\ &= 0. \end{aligned}$$

De même pour $(x^2)^\mathsf{T} s^1 = 0$.

On note finalement que, comme ensembles de solutions de problèmes d'optimisation linéaires, $\mathcal{S}_\text{P}$ et $\mathcal{S}_\text{D}$ sont des polyèdres convexes, si bien qu'il en est de même de $\mathcal{S}_\text{PD}$ (point 8 de l'exercice 2.20). □

On montrera plus loin (en section 17.3) que les couples formés des solutions primales et duales sont les points-selles du lagrangien ℓ sur $\mathbb{R}^n \times (\mathbb{R}^m \times \mathbb{R}^n_+)$. Le résultat précédent peut donc aussi se déduire de cette observation puisque l'ensemble des points-selles est le produit cartésien de l'ensemble des solutions primales par l'ensemble des solutions duales (corollaire 14.4).

L'identité (17.5)(c) forme ce que l'on appelle les *conditions de complémentarité*. Comme $x \geqslant 0$ et $s \geqslant 0$, elle s'écrit de manière équivalente

$$\forall i \in [1:n], \quad x_i > 0 \quad \Longrightarrow \quad s_i = 0,$$

qui exprime que si la i-ième contrainte d'inégalité n'est pas active, le multiplicateur associé est nul. On dit qu'une solution primale-duale (x, y, s) de (P_L) est *strictement complémentaire* si

$$\forall i \in [1:n], \quad x_i > 0 \quad \Longleftrightarrow \quad s_i = 0.$$

Toutes les solutions ne sont pas nécessairement strictement complémentaires. Le résultat suivant montre cependant qu'en optimisation linéaire on peut toujours trouver une solution strictement complémentaire.

Introduisons deux ensembles d'indices $\mathfrak{B}$ et $\mathfrak{N}$, qu'il faudra se garder de confondre avec la base d'indices B et son complémentaire N dans $[1:n]$, introduits au début de ce chapitre :

$$\begin{aligned} \mathfrak{B} &:= \{i \in [1:n] : \exists\, x \in \mathcal{S}_\text{P} \text{ vérifiant } x_i > 0\} \\ \mathfrak{N} &:= \{i \in [1:n] : \exists\, (y,s) \in \mathcal{S}_\text{D} \text{ vérifiant } s_i > 0\}. \end{aligned}$$

Par les conditions de complémentarité et du fait que l'ensemble des solutions primales-duales est un produit cartésien (proposition 17.7), ces deux ensembles n'ont pas d'indice commun : $\mathfrak{B} \cap \mathfrak{N} = \varnothing$. En effet, si $i \in \mathfrak{B}$, alors $s_i = 0$ pour toute solution duale (y, s), et donc $i \notin \mathfrak{N}$. Compte tenu de cette observation, montrer qu'il existe une solution primale-duale strictement complémentaire revient à montrer que $(\mathfrak{B}, \mathfrak{N})$ forme une partition de $[1:n]$.

Proposition 17.8 (existence d'une solution strictement complémentaire) *Si le problème (P_L) a une solution, alors il a une solution primale-duale strictement complémentaire, et donc*

$$\mathfrak{B} \cap \mathfrak{N} = \varnothing \quad \textit{et} \quad \mathfrak{B} \cup \mathfrak{N} = [1:n]. \tag{17.6}$$

Démonstration. Construisons d'abord la solution primale. Pour que celle-ci ait le plus grand nombre de composantes non nulles possible, on l'écrit comme combinaison convexe de solutions ayant chacune au moins une composante strictement positive :

par définition de $\mathfrak{B}$, on sait que pour tout $i \in \mathfrak{B}$, on peut choisir une solution $\bar{x}^i$ de (P_L) avec $\bar{x}^i_i > 0$. On prend

$$\bar{x} = \frac{1}{|\mathfrak{B}|} \sum_{i \in \mathfrak{B}} \bar{x}^i,$$

où $|\mathfrak{B}|$ désigne le nombre d'éléments de $\mathfrak{B}$. C'est aussi une solution (c'est une combinaison convexe de solutions d'un problème convexe, ou on vérifie directement que $c^\mathsf{T}\bar{x} = \inf f$, $A\bar{x} = b$, $\bar{x} \geqslant 0$) et elle vérifie $\bar{x}_\mathfrak{B} > 0$ (c'est une solution avec le moins de contraintes actives possible, située dans l'intérieur relatif de la face optimale, voir l'exercice 17.7). D'après les conditions d'optimalité de la proposition 17.6, il existe $(y_0, s_0) \in \mathbb{R}^m \times \mathbb{R}^n$ tel que

$$\begin{cases} A^\mathsf{T} y_0 + s_0 = c, \\ 0 \leqslant s_0 \perp \bar{x} \geqslant 0. \end{cases} \tag{17.7}$$

D'après la proposition 17.7, il reste à trouver une solution duale $(\bar{y}, \bar{s})$ avec $\bar{s}_i > 0$ pour $i \notin \mathfrak{B}$. Dans ce but, on considère le problème d'optimisation linéaire suivant

$$\begin{cases} \min_x -e_{\mathfrak{B}^c}^\mathsf{T} x \\ Ax = b \\ x \geqslant 0 \\ c^\mathsf{T} x \leqslant \mathrm{val}(P_L), \end{cases}$$

où l'on a noté $e_{\mathfrak{B}^c}$ le vecteur de $\mathbb{R}^n$ dont toutes composantes avec indices dans le complémentaire $\mathfrak{B}^c$ de $\mathfrak{B}$ valent 1 et les autres composantes sont nulles. L'ensemble admissible de ce problème est formé des solutions de (P_L) et par définition de $\mathfrak{B}$, sa valeur optimale est nulle. Alors $\bar{x}$ est solution de ce problème et l'on peut trouver des multiplicateurs optimaux $(y_1, s_1, t_1) \in \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}$ tels que

$$\begin{cases} A^\mathsf{T} y_1 + s_1 + e_{\mathfrak{B}^c} = t_1 c \\ 0 \leqslant s_1 \perp \bar{x} \geqslant 0 \\ 0 \leqslant t_1 \perp (\mathrm{val}(P_L) - c^\mathsf{T}\bar{x}) \geqslant 0. \end{cases}$$

Si t_1 n'est pas nul, on peut conclure avec la solution duale $(y_1, s_1 + e_{\mathfrak{B}^c})/t_1$, mais rien ne garantit que $t_1 \neq 0$. On combine alors cette solution duale avec celle donnée en (17.7) :

$$\bar{y} := \frac{1}{1+t_1}(y_0 + y_1) \qquad \text{et} \qquad \bar{s} := \frac{1}{1+t_1}(s_0 + s_1 + e_{\mathfrak{B}^c}),$$

qui est clairement une solution duale de (P_L). De plus, comme $s_0 \geqslant 0$ et $s_1 \geqslant 0$, on a $\bar{s}_{\mathfrak{B}^c} > 0$. □

En réalité, comme l'exercice 17.7 demande de le montrer, l'ensemble $\mathcal{S}^{\mathrm{sc}}_{\mathrm{PD}}$ des solutions primales-duales strictement complémentaires se confond avec l'intérieur relatif de l'ensemble des solutions primales-duales :

$$\mathcal{S}^{\mathrm{sc}}_{\mathrm{PD}} = \mathcal{S}^{\dot{\Theta}}_{\mathrm{PD}}.$$

Les ensembles d'indices $\mathfrak{B}$ et $\mathfrak{N}$ permettent de donner une description simple de l'ensemble $\mathcal{S}_\mathrm{P}$ des solutions primales et l'ensemble $\mathcal{S}_\mathrm{D}$ des solutions duales de (P_L). Pour une raison qui apparaîtra claire plus loin, on note

$$\mathcal{F}_{\text{D}} := \{(y,s) \in \mathbb{R}^m \times \mathbb{R}^n : A^{\mathsf{T}} y + s = c,\ s \geqslant 0\}.$$

Proposition 17.9 (faces optimales primale et duale) *Si (P_L) a une solution, alors l'ensemble de ses solutions primales est la face exposée de $\mathcal{F}_{\text{P}}$ définie par*

$$\begin{aligned} \mathcal{S}_{\text{P}} &= \{x \in \mathcal{F}_{\text{P}} : x_{\mathfrak{N}} = 0\} \\ &= \{x \in \mathbb{R}^n : Ax = b,\ x_{\mathfrak{B}} \geqslant 0,\ x_{\mathfrak{N}} = 0\} \end{aligned} \tag{17.8a}$$

et l'ensemble de ses solutions duales est la face exposée de $\mathcal{F}_{\text{D}}$ définie par

$$\begin{aligned} \mathcal{S}_{\text{D}} &= \{(y,s) \in \mathcal{F}_{\text{D}} : s_{\mathfrak{B}} = 0\} \\ &= \{(y,s) \in \mathbb{R}^m \times \mathbb{R}^n : A^{\mathsf{T}} y + s = c,\ s_{\mathfrak{B}} = 0,\ s_{\mathfrak{N}} \geqslant 0\}. \end{aligned} \tag{17.8b}$$

Démonstration. Le fait que $\mathcal{S}_{\text{P}}$ soit une face exposée de l'ensemble convexe $\mathcal{F}_{\text{P}}$ provient de l'expression $\mathcal{S}_{\text{P}} = \{x \in \mathcal{F}_{\text{P}} : c^{\mathsf{T}} x \leqslant \text{val}(P_L)\}$ de l'ensemble des solutions primales.

Démontrons à présent la première égalité dans (17.8a) (la seconde se déduit facilement de la première et de (17.6)). Si $x \in \mathcal{S}_{\text{P}}$, alors $x \in \mathcal{F}_{\text{P}}$ et $x_{\mathfrak{N}} = 0$ (par la proposition 17.8 et la définition de $\mathfrak{B}$). Inversement, si $x \in \mathcal{F}_{\text{P}}$ et $x_{\mathfrak{N}} = 0$, alors en prenant une solution duale (y,s) arbitraire (elle existe car (P_L) a une solution), le triplet (x,y,s) vérifie les conditions d'optimalité (17.5) (car $s_{\mathfrak{B}} = 0$ d'après la proposition 17.8 et la complémentarité) ; donc x est solution de (P_L).

Les affirmations concernant $\mathcal{S}_{\text{D}}$ se démontrent de la même manière. □

L'exercice 17.7 demande également de démontrer que l'intérieur relatif de $\mathcal{S}_{\text{P}}$ est donné par

$$\begin{aligned} \mathcal{S}_{\text{P}}^{\oplus} &= \{x \in \mathcal{S}_{\text{P}} : x_{\mathfrak{B}} > 0\} \\ &= \{x \in \mathbb{R}^n : Ax = b,\ x_{\mathfrak{B}} > 0,\ x_{\mathfrak{N}} = 0\} \end{aligned} \tag{17.9a}$$

et l'intérieur relatif de $\mathcal{S}_{\text{D}}$ est donné par

$$\begin{aligned} \mathcal{S}_{\text{D}}^{\oplus} &= \{x \in \mathcal{S}_{\text{D}} : s_{\mathfrak{N}} > 0\} \\ &= \{(y,s) \in \mathbb{R}^m \times \mathbb{R}^n : A^{\mathsf{T}} y + s = c,\ s_{\mathfrak{B}} = 0,\ s_{\mathfrak{N}} > 0\}. \end{aligned} \tag{17.9b}$$

17.3 Dualité

17.3.1 Dualité en optimisation linéaire

Nous ferons référence ici aux notions vues à la section 14.1. On part du problème d'optimisation linéaire

$$(P_L) \quad \begin{cases} \min c^{\mathsf{T}} x \\ Ax = b \\ x \geqslant 0 \end{cases}$$

et on introduit la fonction de couplage

$$\varphi(x,y) = c^{\mathsf{T}}x - y^{\mathsf{T}}(Ax-b),$$

par laquelle on n'a dualisé que les contraintes d'égalité. On a

$$\inf_{x\geqslant 0}\sup_{y}\varphi(x,y) \quad = \quad \inf_{x\geqslant 0}\begin{cases} c^{\mathsf{T}}x & \text{si } Ax=b \\ +\infty & \text{sinon.}\end{cases}$$

On retrouve donc le problème (P_L), si l'on adopte la convention que la valeur minimale de (P_L) vaut $+\infty$ lorsque l'ensemble admissible est vide.

Alors, le problème dual de (P_L) pour la fonction φ ci-dessus s'écrit

$$\sup_{y}\inf_{x\geqslant 0}\Big((c-A^{\mathsf{T}}y)^{\mathsf{T}}x + b^{\mathsf{T}}y\Big) = \sup_{y}\begin{cases} b^{\mathsf{T}}y & \text{si } A^{\mathsf{T}}y\leqslant c \\ -\infty & \text{sinon.}\end{cases}$$

On trouve finalement comme problème dual

$$(D_L) \quad \begin{cases} \max b^{\mathsf{T}}y \\ A^{\mathsf{T}}y \leqslant c, \end{cases}$$

si l'on prend la convention que la valeur maximale dans (D_L) vaut $-\infty$ lorsque l'ensemble admissible de (D_L) est vide. C'est aussi un problème d'optimisation linéaire.

On peut obtenir un problème dual de (P_L) en dualisant également la contrainte $x\geqslant 0$, c'est-à-dire en utilisant le lagrangien $\ell(x,y,s) = c^{\mathsf{T}}x - y^{\mathsf{T}}(Ax-b) - x^{\mathsf{T}}s$ comme fonction de couplage (il faut alors imposer $s\geqslant 0$ dans la dualité). On obtient alors comme problème dual

$$\begin{cases} \max b^{\mathsf{T}}y \\ A^{\mathsf{T}}y + s = c \\ s\geqslant 0, \end{cases}$$

qui est équivalent à (D_L).

Remarquons enfin que les conditions d'optimalité de (D_L) sont aussi celles de (P_L) et que le dual du dual est le primal.

On dit que (y,s) est admissible pour le problème dual (D_L), si $A^{\mathsf{T}}y + s = c$ et si $s\geqslant 0$. On note

$$\begin{aligned} \mathcal{F}_{\text{P}} &:= \{x\in\mathbb{R}^n : Ax=b,\ x\geqslant 0\}, \\ \mathcal{F}_{\text{D}} &:= \{(y,s)\in\mathbb{R}^m\times\mathbb{R}^n : A^{\mathsf{T}}y+s=c,\ s\geqslant 0\} \end{aligned}$$

les ensembles admissibles de (P_L) et de (D_L) respectivement.

Appliquons à présent les résultats obtenus dans le cadre général au problème d'optimisation linéaire (P_L). De la proposition 14.2, on obtient le résultat suivant.

Proposition 17.10 (dualité faible en optimisation linéaire) *On a*

$$\sup_{A^{\mathsf{T}}y\leqslant c} b^{\mathsf{T}}y \leqslant \inf_{\substack{Ax=b\\ x\geqslant 0}} c^{\mathsf{T}}x. \tag{17.10}$$

Ce résultat s'obtient également facilement par calcul. En effet, si $x \in \mathcal{F}_{\text{P}}$ et si $(y, s) \in \mathcal{F}_{\text{D}}$, la différence entre la valeur du critère primal et celle du critère dual est positive et donnée par la formule

$$\boxed{c^{\mathsf{T}}x - b^{\mathsf{T}}y = x^{\mathsf{T}}s \geqslant 0.} \tag{17.11}$$

On a en effet, $c^{\mathsf{T}}x = (A^{\mathsf{T}}y+s)^{\mathsf{T}}x = b^{\mathsf{T}}y+x^{\mathsf{T}}s$. La positivité de $x^{\mathsf{T}}s$ conduit à l'inégalité de dualité faible (17.10).

Théorème 17.11 (dualité forte en optimisation linéaire) *Les propriétés suivantes sont équivalentes :*
(i) (P_L) *et* (D_L) *sont réalisables,*
(ii) (P_L) *a une solution,*
(iii) (D_L) *a une solution.*
Lorsque ces propriétés ont lieu il n'y a pas de saut de dualité, c'est-à-dire que l'on a égalité en (17.10).

Démonstration. [$(i) \Rightarrow (ii)$] Si (i) a lieu, l'inégalité de dualité faible (17.10) montre que (P_L) est réalisable et borné. Ce problème a donc une solution (théorème 17.4).

[$(ii) \Rightarrow (iii)$] Les conditions d'optimalité (17.5), qui ont lieu si (P_L) a une solution, sont aussi celles de (D_L), qui a donc aussi une solution.

[$(iii) \Rightarrow (i)$] Les conditions d'optimalité (17.5) de (D_L) montrent que (P_L) est réalisable.

Supposons que $\bar{x}$ soit une solution de (P_L). Alors il existe un couple $(\bar{y}, \bar{s})$ tel que les conditions d'optimalité (17.5) soient vérifiées. Comme $\bar{x} \in \mathcal{F}_{\text{P}}$, $(\bar{y}, \bar{s}) \in \mathcal{F}_{\text{D}}$ et $\bar{x}^{\mathsf{T}}\bar{s} = 0$, (17.11) montre qu'il n'y a pas de saut de dualité. □

D'après ce théorème, si $\text{val}(P_L) \in \mathbb{R}$ ou si $\text{val}(D_L) \in \mathbb{R}$, il n'y a pas de saut de dualité. Mais on peut très bien avoir

- $\text{val}(D_L) = \text{val}(P_L) = -\infty$ (pas de saut de dualité) : $A = 0$, $b = 0$ et $c < 0$,
- $\text{val}(D_L) = \text{val}(P_L) = +\infty$ (pas de saut de dualité) : $A = 0$, $b \neq 0$ et $c = 0$,
- $\text{val}(D_L) = -\infty$ et $\text{val}(P_L) = +\infty$ (saut de dualité infini) : $A = 0$, $b \neq 0$ et $c < 0$.

Dès lors, pour qu'il n'y ait pas de saut de dualité il faut et il suffit que (P_L) ou (D_L) soit réalisable.

Corollaire 17.12 *Il n'y a pas de saut de dualité si, et seulement si,* (P_L) *ou* (D_L) *est réalisable.*

Démonstration. Supposons en effet que $\text{val}(P_L) = \text{val}(D_L)$. Si ces valeurs sont réelles, alors (P_L) et (D_L) sont réalisables. Si ces valeurs égalent $-\infty$ [resp. $+\infty$], seul (P_L) [resp. (D_L)] est réalisable.

Réciproquement, supposons que (P_L) soit réalisable (le raisonnement est le même si c'est (D_L) qui est réalisable). Si (P_L) est aussi borné, il a une solution et le résultat se déduit du théorème. Si (P_L) n'est pas borné, alors $\text{val}(P_L) = -\infty$, donc $\text{val}(D_L) = -\infty$ par dualité faible et il n'y a pas de saut de dualité. □

Le point (i) est souvent un moyen commode de montrer qu'un problème d'optimisation linéaire a une solution : on montre qu'il est réalisable et que son dual l'est également. Ce théorème permet de retrouver le lemme de Farkas (voir l'exercice 17.8). Quelques conséquences de ce théorème sont données à l'exercice 17.10.

17.3.2 Une relation entre le primal et le dual

Le théorème de dualité forte a montré que les propriétés des problèmes primal et dual sont très liées. Dans cette section nous donnons d'autres exemples de relations entre les deux problèmes.

On dit que $x \in \mathcal{F}_{\text{P}}$ est *strictement admissible* pour le problème primal (P_L) si $x > 0$ (toutes les composantes de x sont > 0). De même, on dit que $(y, s) \in \mathcal{F}_{\text{D}}$ est *strictement admissible* pour le problème dual (D_L) si $s > 0$. On définit les ensembles de points strictement admissibles

$$\begin{aligned} \mathcal{F}_{\text{P}}^{s} &:= \{x \in \mathbb{R}^n : Ax = b,\ x > 0\} \\ \mathcal{F}_{\text{D}}^{s} &:= \{(y, s) \in \mathbb{R}^m \times \mathbb{R}^n : A^{\mathsf{T}} y + s = c,\ s > 0\}. \end{aligned}$$

S'ils sont non vides, ce sont les intérieurs relatifs de $\mathcal{F}_{\text{P}}$ et $\mathcal{F}_{\text{D}}$ respectivement.

Pour les problèmes d'optimisation convexes, le rapprochement des propositions 4.43 [(QC-S) $\iff$ (QC-MF)] et 4.45 [(QC-MF) $\iff$ Λ_* borné] suggère que l'existence d'un point *strictement* admissible primal équivaut au caractère *borné* de l'ensemble des solutions duales. En optimisation linéaire, le dual du dual est le primal, si bien que l'existence d'une solution duale strictement admissible devrait aussi équivaloir au caractère borné de l'ensemble des solutions primales. C'est une manière de trouver naturelles les équivalences énoncées dans la proposition suivante. Nous en donnons une démonstration directe ci-dessous ; l'exercice 17.11 propose d'autres approches moins techniques et qui pourront être utiles pour étendre ce résultat à d'autres problèmes. On y a noté

$$\mathcal{S}_{\text{D},s} = \{s : \text{il existe } y \in \mathbb{R}^m \text{ tel que } (y, s) \in \mathcal{S}_{\text{D}}\}$$

la projection de $\mathcal{S}_{\text{D}}$ sur $\mathbb{R}^n$. On note aussi $e = (1\ \cdots\ 1)^{\mathsf{T}}$.

Proposition 17.13 *Supposons que les problèmes primal et dual soient réalisables* ($\mathcal{F}_{\text{P}} \neq \varnothing$ *et* $\mathcal{F}_{\text{D}} \neq \varnothing$). *Alors*

$$\mathcal{S}_{\text{P}} \text{ est borné} \iff \mathcal{F}_{\text{D}}^{s} \neq \varnothing, \tag{17.12}$$

$$\mathcal{S}_{\text{D},s} \text{ est borné} \iff \mathcal{F}_{\text{P}}^{s} \neq \varnothing. \tag{17.13}$$

Démonstration. Comme

$$\mathcal{S}_{\text{P}} = \{x \in \mathbb{R}^n : Ax = b,\ x \geqslant 0,\ c^{\mathsf{T}} x \leqslant \text{val}(P_L)\}$$

est un convexe fermé non vide, $\mathcal{S}_{\text{P}}$ est borné si, et seulement si, son cône asymptotique (exercice 2.20)

$$\mathcal{S}_{\text{P}}^{\infty} = \{d \in \mathbb{R}^n : Ad = 0,\ d \geqslant 0,\ c^{\mathsf{T}} d \leqslant 0\}$$

est réduit au singleton $\{0\}$ (proposition 2.8). Si $\mathcal{F}_{\text{D}}^s$ est non vide, il existe (y,s) : $c = A^{\mathsf{T}}y + s$, avec $s > 0$. Alors pour $d \in \mathcal{S}_{\text{P}}^{\infty}$, on a $0 \geqslant c^{\mathsf{T}}d = y^{\mathsf{T}}Ad + s^{\mathsf{T}}d = s^{\mathsf{T}}d$, ce qui implique que $d = 0$ (car $d \geqslant 0$ et $s > 0$). Inversement, si $\mathcal{S}_{\text{P}}^{\infty} = \{0\}$, son cône dual est $\mathbb{R}^n$. Celui-ci s'écrit (corollaire 2.46, de Farkas)

$$(\mathcal{S}_{\text{P}}^{\infty})^+ = \{A^{\mathsf{T}}y + s - \alpha c : s \in \mathbb{R}_+^n,\ \alpha \in \mathbb{R}_+\}.$$

Il existe donc $y \in \mathbb{R}^m$, $s \in \mathbb{R}_+^m$ et $\alpha \in \mathbb{R}_+$ tels que $A^{\mathsf{T}}y + s - \alpha c = -e + c$, ce qui s'écrit aussi $A^{\mathsf{T}}y + (s+e) = (\alpha+1)c$. Alors $(y, s+e)/(\alpha+1) \in \mathcal{F}_{\text{D}}^s$.

La seconde partie de la proposition peut se démontrer de la même manière. $\mathcal{S}_{\text{D}}$ étant un polyèdre convexe, il en est de même de $\mathcal{S}_{\text{D},s}$ (proposition 2.18), qui est donc aussi fermé. Alors

$$\mathcal{S}_{\text{D},s} = \{s \in \mathbb{R}_+^n : \text{il existe } y \in \mathbb{R}^m \text{ tel que } A^{\mathsf{T}}y + s = c,\ b^{\mathsf{T}}y \geqslant \text{val}(D_L)\}$$

est borné si, et seulement si, son cône asymptotique

$$\mathcal{S}_{\text{D},s}^{\infty} = \{r \in \mathbb{R}_+^n : \text{il existe } z \in \mathbb{R}^m \text{ tel que } A^{\mathsf{T}}z + r = 0 \text{ et } b^{\mathsf{T}}z \geqslant 0\} \tag{17.14}$$

est réduit à $\{0\}$. Si $\mathcal{F}_{\text{P}}^s \neq \varnothing$, il existe $x > 0$ tel que $Ax = b$. Quel que soit $r \in \mathcal{S}_{\text{D},s}^{\infty}$, on a $r^{\mathsf{T}}x = -z^{\mathsf{T}}Ax = -z^{\mathsf{T}}b \leqslant 0$. Comme $x > 0$ et $r \geqslant 0$, on en déduit que $r = 0$. Donc $\mathcal{S}_{\text{D},s}$ est borné. Inversement, si $\mathcal{S}_{\text{D},s}^{\infty}$ est réduit à $\{0\}$, son cône dual (on a noté x_0 une solution particulière de $Ax_0 = b$)

$$(\mathcal{S}_{\text{D},s}^{\infty})^+ = \mathbb{R}_- x_0 + \mathcal{N}(A) + \mathbb{R}_+^n, \tag{17.15}$$

est $\mathbb{R}^n$. Alors, on peut écrire $-e = -\alpha x_0 + h + p$, avec $\alpha \in \mathbb{R}_+$, $h \in \mathcal{N}(A)$ et $p \in \mathbb{R}_+^n$; donc $A(p+e) = \alpha b$ et $(x_0+p+e)/(1+\alpha) \in \mathcal{F}_{\text{P}}^s$. □

Évidemment, si $\mathcal{S}_{\text{D},s}$ n'est pas borné, $\mathcal{S}_{\text{D}}$ ne l'est pas non plus, ni $\mathcal{S}_{D,y} := \{y$: il existe $s \in \mathbb{R}^n$ tel que $(y,s) \in \mathcal{S}_{\text{D}}\}$ d'ailleurs. Par ailleurs, si A est surjective, il revient au même de dire que $\mathcal{S}_{\text{D},s}$ est borné ou que $\mathcal{S}_{\text{D}}$ est borné.

17.4 Algorithmes du simplexe

In the summer of 1947, when I began to work on the simplex method for solving linear programs, the first idea that occurred to me is one that would occur to any trained mathematician, namely the idea of step-by-step descent (with respect to the objective function) along edges of the convex polyhedral set from one vertex to an adjacent one. I rejected this algorithm outright on intuitive grounds—it had to be inefficient because it proposed to solve the problem by wandering along some path of outside edges until the optimal vertex was reached.

G.B. Dantzig (1914-2005). Origins of the simplex method [154 ; 1990].

On considère le problème d'optimisation linéaire sous forme standard

$$(P_L) \quad \begin{cases} \min c^{\mathsf{T}} x \\ Ax = b \\ x \geqslant 0, \end{cases}$$

dont les conditions d'optimalité s'écrivent (proposition 17.6) : $\exists\, (y, s) \in \mathbb{R}^m \times \mathbb{R}^n$, tels que

$$\begin{cases} A^{\mathsf{T}} y + s = c, \quad s \geqslant 0 \\ Ax = b, \quad x \geqslant 0 \\ x^{\mathsf{T}} s = 0. \end{cases}$$

On sait que si ce problème a une solution il a une solution sur un sommet du polyèdre (proposition 17.5)

$$\mathcal{F}_{\text{P}} = \{x : Ax = b,\ x \geqslant 0\}.$$

Si m et n sont grands, il peut y avoir beaucoup de sommets, ce qui exclut de les explorer tous. L'*algorithme du simplexe* de Dantzig [153 ; 1951] sélectionne les sommets à explorer. L'algorithme se déroule en général en deux étapes :

- Phase I : trouver un sommet de $\mathcal{F}_{\text{P}}$,
- Phase II : passer d'un sommet à l'autre de manière à faire décroître f, jusqu'à ce que l'on trouve une solution-sommet.

Dans la phase I, on résout un problème d'optimisation linéaire auxiliaire dont on connaît un sommet (voir la section 17.4.4). On peut donc utiliser l'algorithme du simplexe pour résoudre ce problème. Dans la phase II, on évite de parcourir tous les sommets en faisant décroître le critère à chaque itération, si possible strictement. On ne visite donc plus les sommets dont la valeur du critère est supérieure à celle au sommet courant.

17.4.1 Algorithme du simplexe primal

> *In my mind for example, solving a linear programming problem by the simplex method means moving from vertex to vertex of a vertical convex polyhedron until the bottom is reached. Therefore such terms as shadow prices, complementarity slackness, tableaux and nonbasic variables leave me cold.*
>
> M.J.D. Powell [495 ; 1990].

Hypothèse. On suppose que A a plein rang $m < n$.

La version de l'algorithme du simplexe que nous présentons dans cette section est celle connue sous le nom d'*algorithme du simplexe révisé*. Cet algorithme est géométriquement très simple : chaque itération consiste à passer d'un sommet du polyèdre convexe, qu'est l'ensemble admissible, à un sommet adjacent en suivant une arête particulière de ce polyèdre, de manière à faire décroître la fonction coût. Il est bien de garder cette idée générale à l'esprit car la description algébrique d'une itération est relativement longue et doit prendre en compte quelques cas particuliers (problème non borné et pas nul) qui distraient.

Une itération de l'algorithme démarre donc en un sommet $\hat{x}$. Nous l'avons déjà dit : le calcul d'un tel sommet n'est pas une opération triviale, mais nous verrons à la section 17.4.4 comment on peut la réaliser. On note B une base d'indices associée à ce sommet et N le complémentaire de B dans $[1:n]$. Au départ de l'itération, on a donc

$$\hat{x}_B \geqslant 0, \quad \hat{x}_N = 0 \quad \text{et} \quad A_{:B}\hat{x}_B = b \ \ (\text{donc } \hat{x}_B = A_{:B}^{-1}b),$$

où, comme précédemment, $A_{:B}$ (resp. $A_{:N}$) désigne la sous-matrice de A formée de ses colonnes avec indices dans B (resp. N)

L'algorithme du simplexe génère en réalité une suite de bases d'indices plutôt qu'une suite de sommets. Il y a une distinction entre les deux notions lorsque le sommet est dégénéré, auquel cas il peut y avoir plusieurs bases d'indices correspondant à un même sommet. Si l'algorithme du simplexe visite un sommet dégénéré, il est possible qu'il ne change pas de sommet à l'itération suivante, mais il changera alors de base d'indices. Cependant décrire l'algorithme en termes de bases d'indices fait perdre l'aspect géométrique de l'algorithme, qu'il nous semble précieux de conserver. Dès lors, nous considérerons que l'itéré de l'algorithme est un sommet, mais que certaines itérations font un déplacement nul.

1 *Reconnaître l'optimalité.*

Soit x un point de l'ensemble admissible $\mathcal{F}_{\text{P}}$, qui vérifie donc $Ax = b$. Comme B est une base d'indices, on peut exprimer les composantes basiques x_B de x en fonction de b et de ses composantes non basiques x_N :

$$x_B = A_{:B}^{-1}(b - A_{:N}x_N).$$

On peut également exprimer le coût $c^{\mathsf{T}}x$ en fonction de x_N :

$$c^{\mathsf{T}}x = c_B^{\mathsf{T}}A_{:B}^{-1}(b - A_{:N}x_N) + c_N^{\mathsf{T}}x_N.$$

Son gradient par rapport à x_N est appelé le *coût réduit*. Il s'écrit

$$\boxed{r = c_N - A_{:N}^{\mathsf{T}}A_{:B}^{-\mathsf{T}}c_B.} \tag{17.16}$$

Dans l'algorithme du simplexe, ce coût réduit sert, d'une part, à détecter l'optimalité éventuelle de l'itéré courant $\hat{x}$ (cette question est examinée dans la proposition 17.14 ci-dessous) et, d'autre part, à sélectionner une arête de $\mathcal{F}_{\text{P}}$ le long de laquelle la fonction coût décroît lorsque $\hat{x}$ n'est pas optimal (voir le point 2 ci-dessous).

Proposition 17.14 *Un sommet $\hat{x}$ de $\mathcal{F}_{\text{P}}$ est solution du problème (P_L) si, et seulement si, il existe une base d'indices $B \subseteq [1:n]$ associée à $\hat{x}$ telle que le coût réduit $r \geqslant 0$.*

Démonstration. [⇒] Supposons que $\hat{x}$ soit un sommet de $\mathcal{F}_{\text{P}}$ qui soit optimal pour le problème (P_L). Alors, par les conditions d'optimalité (17.5), il existe une solution duale $(y, s) \in \mathbb{R}^m \times \mathbb{R}^n$ telle que

$$A^{\mathsf{T}}y + s = c, \quad s \geqslant 0 \quad \text{et} \quad \hat{x}^{\mathsf{T}}s = 0.$$

Comme $\hat{x}$ est un sommet, $A_{:I^+(\hat{x})}$ est injective (en particulier $|I^+(\hat{x})| \leqslant m$).

Montrons que l'on peut trouver une base d'indices B associée à $\hat{x}$ telle que, après modification éventuelle de y et s dans les conditions d'optimalité ci-dessus, on ait $s_B = 0$. On s'y prend en un nombre fini d'étapes en commençant par prendre $B' := \{i : s_i = 0\}$. De $\hat{x}_{I(\hat{x})} > 0$ et $\hat{x}^\mathsf{T} s = 0$, il vient $s_{I(\hat{x})} = 0$ et donc $I^+(\hat{x}) \subseteq B'$.

1. Si $\operatorname{rg} A_{:B'} = m$, on n'a pas besoin de modifier y et s. En effet, l'affirmation est démontrée car, par le théorème de la base incomplète (théorème B.1), on peut trouver $B \subseteq B'$ tel que $I^+(\hat{x}) \subseteq B$, $|B| = m$ et $A_{:B}$ inversible (c'est-à-dire B est une base d'indices associée à $\hat{x}$). Cette base d'indices B est bien telle que $s_B = 0$.
2. Si $\operatorname{rg} A_{:B'} < m$, on peut trouver une direction $\mu \in \mathbb{R}^m$ telle que $A^\mathsf{T}_{:B'}\mu = 0$ et $A^\mathsf{T}_{:N'}\mu \neq 0$ (avec $N' := [1:n] \setminus B'$) ; c'est possible car autrement on aurait $\mathcal{N}(A^\mathsf{T}_{:B'}) \subseteq \mathcal{N}(A^\mathsf{T})$ ou $\mathcal{R}(A) \subseteq \mathcal{R}(A_{:B'})$, ce qui contredirait $\dim \mathcal{R}(A) = m$ et $\dim \mathcal{R}(A_{:B'}) < m$. On peut supposer que $A^\mathsf{T}_{:N'}\mu \not\leqslant 0$. On prend alors le plus grand $t \geqslant 0$ tel que $s'_{N'} = s_{N'} - tA^\mathsf{T}_{:N'}\mu \geqslant 0$. Ce t est strictement positif, car $s_{N'} > 0$ par définition de B' et N'. On note j un indice de N' tel que $s'_j = 0$. Alors avec $y' = y + t\mu$ et $s' = s - tA^\mathsf{T}\mu$. Alors, on a

$$A^\mathsf{T} y' + s' = c, \quad s' \geqslant 0 \quad \text{et} \quad \hat{x}^\mathsf{T} s' = 0.$$

À la fin du point 2, on se retrouve avec des conditions d'optimalité vérifiées par le triplet $(\hat{x}, y', s')$ et avec un ensemble d'indices $B'' := \{i : s'_i = 0\} \supseteq B' \cup \{j\}$. On peut alors revenir au point 1 avec B'' plutôt que B'. Comme chaque cycle augmente les ensembles d'indices B', B'', … d'au moins un indice et comme $\operatorname{rg} A = m$, le cycle est nécessairement interrompu au point 1 après un nombre fini d'étapes.

Il reste à montrer que pour la base d'indices B associée à $\hat{x}$ telle que $s_B = 0$, on a un coût réduit r positif. Comme $s_B = 0$, on a $y = A^{-\mathsf{T}}_{:B} c_B$, si bien $s_N = c_N - A^\mathsf{T}_{:N} A^{-\mathsf{T}}_{:B} c_B$ n'est autre que le coût réduit. Comme $s_N \geqslant 0$, on a ben montré que le coût réduit est positif.

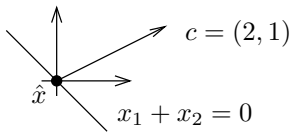
[⇐] Supposons que, pour une base d'indice B associée à $\hat{x}$, le coût réduit $r \geqslant 0$. Si l'on introduit le multiplicateur $y = A^{-\mathsf{T}}_{:B} c_B$, on trouve $r = c_N - A^\mathsf{T}_{:N} y$. En réarrangeant ces deux dernières équations et en posant $s = (0, r)$, on trouve

$$A^\mathsf{T} y + s = c, \quad s \geqslant 0 \quad \text{et} \quad \hat{x}^\mathsf{T} s = 0.$$

Comme $A\hat{x} = b$ et $\hat{x} \geqslant 0$, $(\hat{x}, y, s)$ vérifie les conditions d'optimalité (17.5), ce qui implique que $\hat{x}$ est optimal. □

Si le sommet optimal est dégénéré, il peut y avoir un coût réduit $r \ngeqslant 0$ pour une base d'indices arbitraire associée à ce sommet, comme le montre l'exemple suivant :

$$\begin{cases} \min 2x_1 + x_2 \\ x_1 + x_2 = 0 \\ x \geqslant 0. \end{cases}$$



L'ensemble admissible est réduit au singleton $\{(0, 0)\}$. La solution du problème est donc forcément $\hat{x} = (0, 0)$, qui est un sommet dégénéré. Si l'on prend pour base d'indices $B = \{1\}$, le coût réduit $r = -1$ strictement négatif. Ceci signifie que l'on

peut faire décroître le critère en augmentant $x_N = x_2$ tout en satisfaisant la contrainte d'égalité (c'est le sens du coût réduit). Mais ici, on ne peut pas augmenter x_2 sans sortir de l'ensemble admissible (ce ne serait pas le cas si le sommet était non dégénéré). À l'inverse, si l'on prend pour base d'indices $B = \{2\}$, le coût réduit $r = 1$ est positif, ce qui est annoncé par la proposition 17.14 (il n'y a pas d'autre base d'indices).

La technique utilisée par l'algorithme du simplexe pour détecter l'optimalité éventuelle du sommet courant $\hat{x}$ est la positivité du coût réduit, calculé en utilisant la base d'indices B courante. Il s'agit d'un critère essentiellement primal (il ne fait pas intervenir de multiplicateur ou variable duale). Lorsque l'itéré courant est un sommet-solution non dégénéré, il n'y a qu'une seule base d'indices associée à ce sommet, si bien que le coût réduit est positif et l'algorithme s'interrompt. À l'inverse, lorsque l'itéré courant est un sommet-solution *dégénéré*, il se peut que la base d'indices courante ne permette pas d'avoir un coût réduit positif. Il est donc important que l'algorithme dispose d'un mécanisme lui permettant de changer de base d'indices si cela est nécessaire jusqu'à en trouver une permettant d'avoir un coût réduit positif (comme dans l'exemple ci-dessus). Un mécanisme permettant d'obtenir la convergence de l'algorithme du simplexe (c'est-à-dire d'éviter son cyclage) est appelé *règle d'anti-cyclage*; les principales règles d'anti-cyclage seront vues à la section 17.4.2.

2 *Déplacement le long d'une arête.*

Si r a une composante strictement négative, disons $r_j < 0$, cela veut dire que l'on peut faire décroître le coût en augmentant la composante j de $\hat{x}_N$. On est donc tenté de chercher un nouveau point admissible en faisant un déplacement suivant une direction d, c'est-à-dire

$$x(\alpha) = \hat{x} + \alpha d,$$

telle que la composante non basique de d soit

$$d_N = e^j_N. \tag{17.17}$$

On a noté e^j le j-ième vecteur de base de $\mathbb{R}^n$ et $e^j_N = (e^j)_N$. Pour que ce déplacement $d = (d_B, d_N)$ soit acceptable, il faut d'abord que l'on ait $Ax(\alpha) = b$, donc $Ad = 0$, ce qui détermine sa composante basique :

$$d_B = -A^{-1}_{:B} A_{:N} e^j_N. \tag{17.18}$$

Sur le choix de l'indice j. Remarquons que le coût décroît bien le long de d puisque l'on a

$$c^\mathsf{T} d = r_j < 0.$$

Si r a plusieurs composantes strictement négatives, il semble donc raisonnable de choisir l'indice j parmi ceux donnant la composante de r la plus négative. C'est ce que l'on appelle la *règle du coût réduit minimal*. Cette règle ne garantit cependant pas l'efficacité globale de l'algorithme qui est principalement liée au nombre total d'itérations, c'est-à-dire au nombre de sommets visités (aspect global), ce qui ne peut se déduire d'un calcul de dérivée (aspect local). D'autres règles existent (comme celles introduites par les règles d'anti-cyclage décrites à la section 17.4.2) et les algorithmes du simplexe diffèrent en particulier par l'heuristique adoptée à cette étape.

Il est intéressant d'observer que le déplacement porté par la direction d se fait le long d'une arête de $\mathcal{F}_{\text{P}}$.

Proposition 17.15 *Soient d défini comme ci-dessus et $D := \{\hat{x}+\alpha d \in \mathcal{F}_{\text{P}} : \alpha \geqslant 0\}$. Alors soit D est réduit au sommet $\{\hat{x}\}$, soit c'est une arête de $\mathcal{F}_{\text{P}}$.*

Démonstration. Si $\hat{x} + \alpha d \notin \mathcal{F}_{\text{P}}$ pour tout $\alpha > 0$, on a $D = \{\hat{x}\}$.

Supposons maintenant que $\hat{x} + \alpha d \in \mathcal{F}_{\text{P}}$ pour tout $\alpha > 0$ petit. D'après la proposition 2.21, il suffit de montrer que pour $\alpha > 0$ petit :

$$\dim \mathcal{N}(A_{:\,I^+(\hat{x}+\alpha d)}) = 1.$$

Observons que, si $(\hat{x} + \alpha d)_i = 0$, on a nécessairement $d_i = 0$ (c'est clairement le cas si $i \in N \setminus \{j\}$; ce l'est aussi si $i \in B$ car alors $\hat{x}_i = 0$) ; alors, le fait que $Ad = 0$ montre que

$$A_{:\,I^+(\hat{x}+\alpha d)} d_{I^+(\hat{x}+\alpha d)} = 0, \qquad d_{I^+(\hat{x}+\alpha d)} \neq 0,$$

et donc que $\dim \mathcal{N}(A_{:\,I^+(\hat{x}+\alpha d)}) \geqslant 1$. Par ailleurs, l'inversibilité de $A_{:\,B}$ et $I^+(\hat{x}+\alpha d) \subseteq B \cup \{j\}$, montre que $\dim \mathcal{N}(A_{:\,I^+(\hat{x}+\alpha d)}) \leqslant 1$. □

3 *Détection d'un problème non borné ($d_B \geqslant 0$).*

Si $d_B \geqslant 0$, alors

$$\forall \alpha > 0 : \qquad x(\alpha) \geqslant 0,$$

et comme le coût $c^{\mathsf{T}} x(\alpha) = c^{\mathsf{T}} \hat{x} + \alpha\, c^{\mathsf{T}} d$ décroît strictement le long de la direction de descente d, le problème n'est pas borné.

4 *Pivotage ($d_B \not\geqslant 0$).*

Si $d_B \not\geqslant 0$, on ne peut plus prendre un pas arbitrairement grand. Pour que l'on ait $x(\alpha)_B \geqslant 0$, il faut que $\alpha \leqslant \hat{\alpha}$, où

$$\hat{\alpha} = \min \left\{ -\frac{\hat{x}_i}{d_i} : i \in B,\ d_i < 0 \right\}. \tag{17.19}$$

Lorsque le sommet $\hat{x}$ est dégénéré (il y a des $\hat{x}_i = 0$ pour $i \in B$), ce pas maximal $\hat{\alpha}$ peut être nul (voir ci-après). Soit k un indice donnant le min ci-dessus. Alors, $\hat{x}_k + \hat{\alpha} d_k = 0$ et on peut faire sortir l'indice k de la base d'indices B, et y faire entrer l'indice j. La nouvelle base d'indices s'écrit

$$B_+ = (B \cup \{j\}) \setminus \{k\}.$$

Proposition 17.16 *L'ensemble B_+ est une base d'indices.*

DÉMONSTRATION. Il s'agit de montrer que $A_{:B_+}$ est inversible. Si ce n'est pas le cas, l'inversibilité de $A_{:B}$ implique que A^j est combinaison linéaire des $\{A^i : i \in B \setminus \{k\}\}$, c'est-à-dire

$$A_{:N} e_N^j = A_{:B} u, \quad \text{avec } u_k = 0.$$

On en déduit que $u = -d_B$. On a alors une contradiction, car d'une part $d_k = 0$ (car $u_k = 0$) et $d_k < 0$ (car k donne le min en (17.19)). □

L'opération de mise à jour de la base d'indices B en B_+, qui consiste à lui adjoindre l'indice j et à lui ôter l'indice k, est parfois appelée *pivotage* et la règle déterminant le choix des indices j et k est alors appelée *règle de pivotage*.

5 *Progrès ou stagnation.*

Deux situations peuvent maintenant se présenter.

Si $\hat{\alpha} > 0$, le coût décroît strictement et on peut passer à l'itération suivante avec $\hat{x}_+ := \hat{x} + \hat{\alpha} d$ comme nouveau sommet et B_+ comme nouvelle base d'indices.

Si $\hat{\alpha} = 0$ (ceci ne peut se produire que si le sommet $\hat{x}$ est dégénéré), il y a un changement de base d'indices sans changer de sommet (le pas $\hat{\alpha}$ est nul). Si l'on ne prend pas quelques précautions, l'algorithme peut cycler (par exemple en faisant entrer k dans la base et en faisant sortir j à l'itération suivante). On a mis au point des règles d'anti-cyclage pour faire face à cette situation. Certaines d'entre elles sont présentées dans la section suivante.

17.4.2 Règles d'anti-cyclage ▲

Nous énonçons ci-dessous quelques règles d'anti-cyclage et renvoyons le lecteur aux articles qui les introduisent pour une démonstration de leur propriété d'anti-cyclage. Ces articles sont souvent difficiles à comprendre, si l'on n'est pas familier avec le jargon développé par les spécialistes de l'algorithme du simplexe, en particulier avec la description de l'algorithme sous forme de tableau.

Règle des petites perturbations ▲

Règle lexicographique ▲

Règle des plus petits indices de Bland

La règle consiste à faire entrer dans la base B_+ le plus petit indice $j \in N$ tel que le coût réduit $r_j < 0$ (voir le point 2 ci-dessus) et à en faire sortir le plus petit indice $k \in \arg\min\{-\hat{x}_i/d_i : i \in B,\, d_i < 0\}$ (voir la formule (17.19)). Cette règle a été proposée par Bland [64 ; 1977].

17.4.3 Énoncé de l'algorithme

On peut résumer l'algorithme décrit ci-dessus comme suit.

Algorithme 17.17 (simplexe révisé) On suppose au départ de l'itération que l'on dispose d'un sommet $\hat{x}$ de $\mathcal{F}_{\text{P}}$ et d'une base d'indices associée B. Une itération calcule un nouveau sommet $\hat{x}_+$ et une nouvelle base d'indices B_+, à moins qu'il ne soit observé que $\hat{x}$ est solution ou que le problème est non borné. Voici les différentes étapes de cette itération.

1. *Coût réduit.* Calculer le multiplicateur $y \in \mathbb{R}^m$, solution du système linéaire
$$A_{:B}^{\mathsf{T}} y = c_B$$
et en déduire le coût réduit
$$r = c_N - A_{:N}^{\mathsf{T}} y.$$
2. *Optimalité.* Si $r \geqslant 0$, on s'arrête : $\hat{x}$ est solution du problème (P_L).
3. *Direction de descente.* Soit j un indice tel que $r_j < 0$, respectant une règle d'anti-cyclage de la section 17.4.2. On définit la direction de descente $d \in \mathcal{N}(A)$ du critère $x \mapsto c^{\mathsf{T}} x$ par
$$d_B = -A_{:B}^{-1} A_{:N} e_N^j \quad \text{et} \quad d_N = e_N^j,$$
où e^j est le j-ième vecteur de base de $\mathbb{R}^{|N|}$.
4. *Problème non borné.* Si $d_B \geqslant 0$, on s'arrête car le problème (P_L) n'est pas borné : $c^{\mathsf{T}}(\hat{x} + \alpha d) \to -\infty$ lorsque $\alpha \to \infty$.
5. *Pas maximal.* On calcule le pas maximal $\hat{\alpha}$ jusqu'à la frontière de l'ensemble admissible $\mathcal{F}_{\text{P}}$:
$$\hat{\alpha} := \min \left\{ -\frac{\hat{x}_i}{d_i} : i \in B,\ d_i < 0 \right\}.$$
Ce pas peut être nul. On note k un des indices donnant le minimum ci-dessus et respectant une règle d'anti-cyclage de la section 17.4.2.
6. *Nouveau sommet.* $\hat{x}_+ = \hat{x} + \hat{\alpha} d$.
7. *Nouvelle base d'indices.* $B_+ = (B \cup \{j\}) \setminus \{k\}$.

Théorème 17.18 (convergence de l'algorithme du simplexe révisé) *Si le problème d'optimisation linéaire, écrit sous la forme standard (P_L), est réalisable (c'est-à-dire $\mathcal{F}_{\text{P}} \neq \varnothing$), l'algorithme du simplexe révisé décrit ci-dessus termine après un nombre fini d'étapes, soit en déterminant que le problème (P_L) est non borné, soit en en trouvant une solution-sommet.*

17.4.4 Démarrage de l'algorithme du simplexe

Pour utiliser l'algorithme du simplexe présenté ci-dessus, il faut disposer d'un itéré initial qui est un sommet de l'ensemble admissible $\mathcal{F}_{\text{P}} := \{x \in \mathbb{R}^n : Ax = b,\ x \geqslant 0\}$.

Nous présentons dans cette section plusieurs manières de faire face à cette exigence. Certaines méthodes de pivotage se déroulent en une seule phase, comme l'*algorithme en croix* [115, 578, 579, 615 ; 1979-1987].

Technique des deux phases

Comme son nom l'indique, la technique des deux phases décompose la résolution d'un problème d'optimisation linéaire en deux étapes. La phase/étape I consiste à résoudre un problème d'optimisation linéaire auxiliaire, dont on connait un sommet, par l'algorithme du simplexe. La résolution de ce problème auxiliaire fournit un sommet du polyèdre convexe $\mathcal{F}_{\text{P}}$ (pourvu que celui-ci soit non vide) ou indique que $\mathcal{F}_{\text{P}} = \varnothing$. Dans la phase II, on résout le problème (P_L) par l'algorithme du simplexe, à partir du sommet obtenu dans la première phase.

La phase I consiste donc à trouver un point du polyèdre convexe écrit sous forme standard, à trouver un $x \in \mathbb{R}^n$ tel que $Ax = b$ et $x \geqslant 0$. Elle réalise cette tâche en résolvant le problème d'optimisation linéaire suivant :

$$\left\{ \begin{array}{l} \min \sum_{i=1}^{m} z_i \\ Ax + Dz = b \\ x \geqslant 0, \quad z \geqslant 0, \end{array} \right. \tag{17.20a}$$

où D est la matrice diagonale définie par

$$D_{ii} = \left\{ \begin{array}{rl} 1 & \text{si } b_i \geqslant 0 \\ -1 & \text{si } b_i < 0. \end{array} \right. \tag{17.20b}$$

Proposition 17.19 *Le point $(0, Db)$ est un sommet de l'ensemble admissible du problème (17.20a), lequel a toujours une solution. Si ce problème est résolu par l'algorithme du simplexe en partant de ce point, il obtient pour solution un point $(\hat{x}, \hat{z})$. Si $\hat{z} \neq 0$, le problème (P_L) n'est pas réalisable. Si $\hat{z} = 0$, $\hat{x}$ est un sommet de l'ensemble admissible de (P_L).*

Démonstration. Il est clair que $(0, Db)$ est un sommet de l'ensemble admissible de (17.20a), car les colonnes de D sont linéairement indépendantes (proposition 2.21). Comme (17.20a) est borné ($\sum_{i=1}^{m} z_i \geqslant 0$), il a une solution (proposition 17.4). Si l'algorithme du simplexe est démarré en $(0, Db)$, il trouve une solution $(\hat{x}, \hat{z})$.

Si $\hat{z} \neq 0$, la valeur optimale de (17.20a) est non nulle et le problème (P_L) n'est pas réalisable (car si $x \in \mathcal{F}_{\text{P}}$, $(x, 0)$ est admissible pour (17.20a) et donne une valeur nulle au critère).

Si $\hat{z} = 0$, on a $A\hat{x} = b$ et $\hat{x} \geqslant 0$, donc $\hat{x} \in \mathcal{F}_{\text{P}}$. D'autre part, $(\hat{x}, 0)$ étant un sommet de l'ensemble admissible de (17.20a), les colonnes $\{A^j : \hat{x}_j > 0\}$ sont linéairement indépendantes ; donc $\hat{x}$ est un sommet de $\mathcal{F}_{\text{P}}$. □

Technique du grand M

Cette technique combine les phases I et II de manière à ne devoir résoudre qu'un seul problème d'optimisation linéaire, à savoir le problème

$$\begin{cases} \min\ c^{\mathsf{T}}x + M\ (\sum_{i=1}^{m} z_i) \\ Ax + Dz = b \\ x \geqslant 0, \quad z \geqslant 0, \end{cases} \tag{17.21}$$

où D est pris comme en (17.20b) et $M > 0$ est une constante choisie « suffisamment grande ». On ne connait pas a priori la valeur qu'il faut donner à M pour que les problèmes (17.21) et (P_L) soient équivalents, c'est-à-dire pour qu'en une solution $(\hat{x}, \hat{z})$ de (17.21) on ait $\hat{z} = 0$. Le raisonnement suivant montre pourquoi ils le seront si M est « suffisamment grand ».

On peut en effet voir (17.21) comme la pénalisation exacte, au moyen de la norme ℓ_1, de la contrainte $z = 0$ dans

$$\begin{cases} \min\ c^{\mathsf{T}}x \\ Ax + Dz = b \\ z = 0 \\ x \geqslant 0, \quad z \geqslant 0. \end{cases}$$

Ce dernier problème est équivalent à (P_L). Dès lors, selon la théorie développée à la section 13.5, si M est supérieur à la norme ℓ_∞ (norme duale de la norme ℓ_1 pour le produit scalaire euclidien) de tout multiplicateur optimal associé à la contrainte $z = 0$ dans ce dernier problème, et si $\mathcal{F}_{\mathrm{P}} \neq \varnothing$, alors toute solution $(\hat{x}, \hat{z})$ de (17.21) sera telle que $\hat{z} = 0$ et donc $\hat{x}$ sera solution de (P_L).

17.4.5 Complexité

L'algorithme du simplexe n'est pas polynomial. Klee et Minty [354 ; 1972] ont en effet construit un problème, dans lequel l'ensemble admissible est un cube de $\mathbb{R}^n$ légèrement déformé, pour lequel l'algorithme du simplexe avec la règle du coût réduit minimal visite les 2^n sommets de l'ensemble admissible. Des variantes de l'exemple de Klee et Minty existent pour la plupart des règles de pivotage autres que celle du coût réduit minimal, bien que l'on ne sache pas s'il n'existe pas de règle permettant d'éviter ce comportement exponentiel (voir [581] et ses références). Le nombre d'opérations pour résoudre le problème par l'algorithme du simplexe dans le pire des cas n'est donc pas nécessairement borné par un polynôme dépendant des dimensions de celui-ci ou de la longueur des données le définissant. On parle de *complexité pire-cas*, celle que l'on peut observer dans le cas le moins favorable.

Cependant, en pratique, ces algorithmes sont souvent très efficaces. La raison tient peut-être au fait que, bien que non polynomial dans le pire des cas, il est polynomial *en moyenne* pour des données distribuées aléatoirement suivant diverses lois de probabilité (voir Borgwardt [75, 76], Smale [556], Megiddo [418] et Sharmir [544]) et en *perturbations moyennées* (technique qualifiée d'*analyse lisse* par ses auteurs [566] et qui consiste à étudier la complexité moyenne de l'algorithme pour des données variant aléatoirement autour de données fixées arbitraires).

La non-polynomialité pire-cas de l'algorithme du simplexe a stimulé la recherche d'algorithmes pouvant être polynomiaux en optimisation linéaire, un problème jugé suffisamment « simple » pour admettre un tel algorithme. Ceci a conduit aux algorithmes de points intérieurs, que nous verrons au chapitre 18.

Notes

L'énoncé du problème dual d'un problème d'optimisation linéaire est attribué à J. Von Neumann. L'existence de solutions strictement complémentaires (proposition 17.8) a été montré par Goldman et Tucker (1956).

On pourra trouver d'autres résultats sur l'optimisation linéaire dans les références suivantes : Minoux [422 ; 1983] ; Chvátal [125 ; 1983] a une approche très opérationnelle de l'optimisation linéaire et de l'algorithme du simplexe et s'intéresse à la modélisation de problèmes concrets sous forme de problèmes d'optimisation linéaire, notamment des problèmes de transport dans les réseaux ; Schrijver [540 ; 1986] ; Fletcher [212 ; 1987] ; Ciarlet [127 ; 1988] ; Goldfarb et Todd [260 ; 1989] ; Nering et Tucker [451 ; 1993] ; Saigal [534 ; 1995], Helgason et Kennington [308 ; 1995] discutent de l'utilisation de l'algorithme du simplexe dans divers problèmes d'optimisation linéaire rencontrés dans les réseaux ; le même commentaire s'applique à l'ouvrage de Bertsimas et Tsitsiklis [57 ; 1997], qui contient des chapitres sur les problèmes de grande taille avec leurs méthodes de décomposition, sur les problèmes d'OL dans les réseaux et sur les problèmes en nombres entiers ; Vanderbei [604 ; 1997] ; Martin [408 ; 1999] aborde, dans la partie IV de son ouvrage, les méthodes de décomposition pour résoudre les grands problèmes linéaires (décomposition de Benders, de Dantzig-Wolfe et lagrangienne) et traite, dans le chapitre 14, de l'optimisation dans les réseaux ; Padberg [468 ; 1999].

Exercices

17.1. *Optimisation linéaire et admissiblité polyédrique.* Montrez que résoudre un problème d'optimisation linéaire revient à trouver un point admissible d'un polyèdre convexe, dont les données se déduisent aisément de celles du problème d'optimisation linéaire.

17.2. *Effet d'une perturbation du coût.* On considère le problème d'optimisation linéaire, sous sa forme standard (P_L). On s'intéresse à la valeur optimale $\varphi : \mathbb{R}^n \to \varphi(c) := \inf\{c^\mathsf{T} x : Ax = b,\ x \geqslant 0\} \in \mathbb{R}$, comme fonction du coût c.

- (i) La fonction φ est concave.
- (ii) Le point $\bar{x}$ est l'*unique* solution de (P_L) si, et seulement si, il existe $\epsilon > 0$ tel que si l'on remplace c par $\tilde{c}$ dans (P_L), avec $\|\tilde{c} - c\| \leqslant \epsilon$, $\bar{x}$ est encore (l'unique) solution du problème.
- (iii) Si pour $c = c_0$, (P_L) a une unique solution $\bar{x}_0$, alors φ est *linéaire* dans un voisinage de c_0 et $\nabla\varphi(c_0) = \bar{x}_0$.

17.3. *Hypothèse de non dégénérescence et qualification des contraintes.* Soient $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ et x un sommet du polyèdre convexe $\{x : Ax = b,\ x \geqslant 0\}$. Montrez que x est non dégénéré si, et seulement si, $(A^\mathsf{T}\ I_{I^0(x)})$ est injective, où $I_{I^0(x)}$ est la matrice formée des colonnes d'indices $i \in I^0(x)$ de la matrice identité (ce qui n'est autre que la condition de qualification des contraintes (QC-IL)).

17.4. *Ensemble de solutions borné.* Supposons que l'ensemble admissible $\mathcal{F}_{\text{P}}$ du problème d'optimisation linéaire (P_L) ne soit pas borné et que son ensemble de solution $\mathcal{S}_{\text{P}}$ le soit. Montrez qu'alors le critère $c^{\mathsf{T}}x$ tend vers l'infini à l'infini dans $\mathcal{F}_{\text{P}}$.

17.5. *Polyèdres convexes imbriqués.* Soient P_A et P_B deux polyèdres convexes de $\mathbb{R}^n$ définis par

$$P_A := \{x \in \mathbb{R}^n : Ax \leqslant a\} \qquad \text{et} \qquad P_B := \{x \in \mathbb{R}^n : Bx \leqslant b\},$$

où $A \in \mathbb{R}^{m_A \times n}$, $a \in \mathbb{R}^{m_A}$, $B \in \mathbb{R}^{m_B \times n}$ et $b \in \mathbb{R}^{m_B}$. On suppose que $P_A \neq \varnothing$. Montrez que les propriétés suivantes sont équivalentes :

(*i*) $P_A \subseteq P_B$,
(*ii*) $\exists\, Y \in \mathbb{R}_+^{m_B \times m_A}$ tel que $YA = B$ et $Ya \leqslant b$.

17.6. *Inclusions polaires.*

1) Soient $\mathbb{E}$ un espace euclidien et A et B deux ensembles convexes fermés non vides de $\mathbb{E}$. On note σ_A et σ_B les fonctions d'appui de ces ensembles. Alors, les propriétés suivantes sont équivalentes :
   (*i*) $A \subseteq B$,
   (*ii*) $\sigma_A \leqslant \sigma_B$,
   (*iii*) $\forall\, \alpha \in \mathbb{R}$, $\{d \in \mathbb{E} : \sigma_B(d) \leqslant \alpha\} \subseteq \{d \in \mathbb{E} : \sigma_A(d) \leqslant \alpha\}$,
   (*iv*) $\forall\, \alpha \in \{-1, 1\}$, $\{d \in \mathbb{E} : \sigma_B(d) \leqslant \alpha\} \subseteq \{d \in \mathbb{E} : \sigma_A(d) \leqslant \alpha\}$.

   Les équivalences (*i*) $\Leftrightarrow$ (*iii*) $\Leftrightarrow$ (*iv*) permettent d'associer des inclusions, *polaires* l'une de l'autre, dans certains cas.

2) Si $P := \{x \in \mathbb{R}^n : Ax \leqslant b\}$ est le polyedre convexe non vide, formé à partir de $A \in \mathbb{R}^{m \times n}$ et $b \in \mathbb{R}^m$, et si $\alpha \in \mathbb{R}$, alors

$$\{d \in \mathbb{R}^n : \sigma_P(d) \leqslant \alpha\} = A^{\mathsf{T}}\{y \in \mathbb{R}_+^m : b^{\mathsf{T}}y \leqslant \alpha\}. \tag{17.22}$$

3) On note $\mathscr{E}(M) := \{x \in \mathbb{R}^n : x^{\mathsf{T}}Mx \leqslant 1\}$ l'ellipsoïde de $\mathbb{R}^n$ centré en 0, façonné par une matrice $M \succ 0$ d'ordre n. Alors, pour $H \succ 0$, on a

$$\{d \in \mathbb{R}^n : \sigma_{\mathscr{E}(H)}(d) \leqslant 1\} = \mathscr{E}(H^{-1}). \tag{17.23}$$

4) *Polyèdres convexes imbriqués.* Avec les notations de l'exercice 17.5, on a

$$P_A \subseteq P_B \iff \begin{cases} \forall\, \alpha \in \mathbb{R}, \\ B^{\mathsf{T}}\{z \in \mathbb{R}_+^{m_B} : b^{\mathsf{T}}z \leqslant \alpha\} \subseteq \\ A^{\mathsf{T}}\{y \in \mathbb{R}_+^{m_A} : a^{\mathsf{T}}y \leqslant \alpha\}. \end{cases} \tag{17.24a}$$

$$\iff \begin{cases} \forall\, \alpha \in \{-1, 1\}, \\ B^{\mathsf{T}}\{z \in \mathbb{R}_+^{m_B} : b^{\mathsf{T}}z \leqslant \alpha\} \subseteq \\ A^{\mathsf{T}}\{y \in \mathbb{R}_+^{m_A} : a^{\mathsf{T}}y \leqslant \alpha\}. \end{cases} \tag{17.24b}$$

   On peut retrouver l'équivalence de l'exercice 17.5 à partir de (17.24).

5) *Circonscription ellipsoïdique d'un polyèdre convexe.* Soient $H \succ 0$ et $P := \{x \in \mathbb{R}^n : Ax \leqslant b\} \neq \varnothing$ le polyèdre convexe de $\mathbb{R}^n$ défini au moyen de $A \in \mathbb{R}^{m \times n}$ et $b \in \mathbb{R}^m$. Alors

$$P \subseteq \mathscr{E}(H) \iff \mathscr{E}(H^{-1}) \subseteq A^{\mathsf{T}}\{y \in \mathbb{R}_+^m : b^{\mathsf{T}}y \leqslant 1\}. \tag{17.25}$$

17.7. *Solutions strictement complementaires.* Soit $\mathcal{S}_{\text{PD}}^{\text{sc}}$ l'ensemble des solution primales-duales strictement complémentaires (proposition 17.8). Montrez

1) les formules de $\mathcal{S}_{\text{P}}^{\oplus}$ et $\mathcal{S}_{\text{D}}^{\oplus}$ données par (17.9),
2) $\mathcal{S}_{\text{PD}}^{\text{sc}} = \mathcal{S}_{\text{PD}}^{\oplus} = \mathcal{S}_{\text{P}}^{\oplus} \times \mathcal{S}_{\text{D}}^{\oplus}$,
3) $x \in \mathcal{S}_{\text{P}}^{\oplus} \Leftrightarrow \exists\, (y, s) \in \mathcal{S}_{\text{D}}$ tel que $(x, y, s) \in \mathcal{S}_{\text{PD}}^{\text{sc}}$ (nécessairement $(y, s) \in \mathcal{S}_{\text{D}}^{\oplus}$),
4) $(y, s) \in \mathcal{S}_{\text{D}}^{\oplus} \Leftrightarrow \exists\, x \in \mathcal{S}_{\text{P}}$ tel que $(x, y, s) \in \mathcal{S}_{\text{PD}}^{\text{sc}}$ (nécessairement $x \in \mathcal{S}_{\text{P}}^{\oplus}$).

17.8. *Dualité linéaire et lemme de Farkas.*

1) Retrouvez le lemme de Farkas, sous la forme $\{Ax : x \geqslant 0\} = \{y : A^{\mathsf{T}}y \geqslant 0\}^{+}$ (voir le corollaire 2.46), à partir des résultats de dualité en optimisation linéaire (théorèmes 17.10 et 17.11).
2) Soient A une matrice $m \times n$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$ et $\alpha \in \mathbb{R}$, tels que $\{x : Ax \geqslant b\} \neq \varnothing$. Montrez que les deux propriétés suivantes sont équivalentes :
   (*i*) tout x vérifiant $Ax \geqslant b$ vérifie aussi $c^{\mathsf{T}}x \geqslant \alpha$.
   (*ii*) il existe $y \in \mathbb{R}^m_+$ tel que $A^{\mathsf{T}}y = c$ et $b^{\mathsf{T}}y \geqslant \alpha$.

17.9. *Théorèmes de l'alternative par l'optimisation linéaire*, que l'on écrit ici sous la forme d'équivalences. Voir aussi l'exercice 2.39.

1) *Théorème de l'alternative de Motzkin non-homogène* [295 ; théorème 3.17]. Soient $A \in \mathbb{R}^{m_A \times n}$, $B \in \mathbb{R}^{m_B \times n}$ et $C \in \mathbb{R}^{m_C \times n}$ des matrices ayant un même nombre de colonnes et $a \in \mathbb{R}^{m_A}$, $b \in \mathbb{R}^{m_B}$ et $c \in \mathbb{R}^{m_C}$ des vecteurs. Alors, les affirmations suivantes sont équivalentes :
   (*i*) $\exists\, x \in \mathbb{R}^n : Ax = a$, $Bx \leqslant b$ et $Cx < c$,
   (*ii*) $\forall\, (\alpha, \beta, \gamma) \in \mathbb{R}^{m_A} \times \mathbb{R}^{m_B}_+ \times \mathbb{R}^{m_C}_+$ vérifiant $A^{\mathsf{T}}\alpha + B^{\mathsf{T}}\beta + C^{\mathsf{T}}\gamma = 0$ et $a^{\mathsf{T}}\alpha + b^{\mathsf{T}}\beta + c^{\mathsf{T}}\gamma \leqslant 0$, on a $a^{\mathsf{T}}\alpha + b^{\mathsf{T}}\beta = 0$ et $\gamma = 0$.

17.10. *Dualité.* On considère le problème d'optimisation linéaire (P_L) et son dual (D_L). Démontrez les affirmations suivantes.

1) Si (P_L) est réalisable et non borné, alors (D_L) n'est pas réalisable,
2) Si (P_L) n'est pas réalisable et (D_L) est réalisable, alors (D_L) est non borné.

17.11. *Contributions à la proposition 17.13.*

1) Vérifiez l'identité (17.14).
2) Vérifiez l'identité (17.15).
3) Autre démonstration de (17.12).
   (*i*) Montrez que $\mathcal{F}^s_{\mathrm{D}}$ est vide si, et seulement si, il existe un indice $i \in [1:n]$ tel que le problème $\inf\{-s_i : A^{\mathsf{T}}y + s = c,\ s \geqslant 0\}$ est réalisable et a une valeur optimale nulle.
   (*ii*) Conclure en utilisant un argument de dualité.
4) Autre démonstration de (17.13).
   (*i*) Montrez que $\mathcal{F}^s_{\mathrm{P}}$ est vide si, et seulement si, il existe un indice $i \in [1:n]$ tel que le problème $\min\{-x_i : Ax = b,\ x \geqslant 0\}$ est réalisable et a une valeur optimale nulle.
   (*ii*) Conclure en utilisant un argument de dualité.

17.12. *Identité du minimax de von Neumann* [612 ; 1928]. Soient A une matrice de type $m \times n$ et Δ_p le simplexe unitaire de $\mathbb{R}^p$. Montrez que

$$\max_{y \in \Delta_m} \min_{x \in \Delta_n} y^{\mathsf{T}}Ax = \min_{x \in \Delta_n} \max_{y \in \Delta_m} y^{\mathsf{T}}Ax. \tag{17.26}$$

17.13. *Détection d'une solution unique par l'algorithme du simplexe* [260 ; corollaire 3.1]. Montrez que si le coût réduit r défini par (17.16) a ses composantes *strictement* positives, alors le sommet courant est l'unique solution de (P_L).

17.14. *Lemme de Hoffman* [319 ; 1952]. Le lemme de Hoffman est une *borne d'erreur*, c'est-à-dire une estimation (ici, une majoration) de la distance à un ensemble par des quantités plus facilement calculables que la distance elle-même. On s'intéresse ici à la distance au polyèdre convexe, défini par $P_{A,b} = \{x \in \mathbb{R}^n : Ax \leqslant b\}$, où $A \in \mathbb{R}^{m \times n}$ et $b \in \mathbb{R}^m$. On note $\mathcal{B}_A$ le cône convexe des vecteurs $b \in \mathbb{R}^m$ tels que $P_{A,b} \neq \varnothing$. On se donne également une norme $\|\cdot\|$ sur $\mathbb{R}^n$. Enfin, on note

$$\mathrm{d}_{P_{A,b}}(x) := \min_{z \in P_{A,b}} \|z - x\|$$

la distance de $x \in \mathbb{R}^n$ à $P_{A,b}$, mesurée au moyen de la norme $\|\cdot\|$. Le lemme de Hoffman affirme que

$$\forall\, A \in \mathbb{R}^{m\times n},\ \exists\, h > 0,\ \forall\, b \in \mathcal{B}_A,\ \forall\, x \in \mathbb{R}^n : \quad \mathrm{d}_P(x) \leqslant h\, \|(Ax-b)^+\|_2, \tag{17.27}$$

où $(\cdot)^+ = \max(0, \cdot)$, composante par composante, et $\|\cdot\|_2$ est la norme euclidienne sur $\mathbb{R}^m$. La constante h ne dépend donc que de A. Autrement dit, on peut estimer la distance de x à $P_{A,b}$ par la norme du résidu $(Ax-b)^+$.

(i) Montrez que l'on peut trouver des $A \in \mathbb{R}^{m\times n}$ tels que h est arbitrairement grand.

(ii) Démonstration de (17.27) [295 ; 2010, section 11.8].

(a) On note $\|\cdot\|_{\mathrm{D}}$ la norme duale de la norme $\|\cdot\|$. Montrez que

$$\exists\, u \in \mathbb{R}^n : \quad \|u\|_{\mathrm{D}} = 1 \quad \text{et} \quad \mathrm{d}_P(x) = \sup_{\substack{y \geqslant 0 \\ A^{\mathsf{T}} y = u}} (Ax-b)^{\mathsf{T}} y. \tag{17.28}$$

(b) Montrez que l'on peut *choisir* une solution du problème d'optimisation linéaire dans (17.28), qui soit majorée en norme par une constante ne dépendant que de A.

(c) Démontrez (17.27).

18 Optimisation linéaire : algorithmes de points intérieurs

My view of interior point methods for optimization calculations with linear constraints is that it seems silly to introduce nonlinearities and iterative procedures for following central paths, because these complications are not present in the original problem. On the other hand, when the number of constraints is huge, then algorithms that treat constraints individually are also unattractive, especially if the attention to detail causes the number of iterations to be about the number of constraints. It is possible, however, to retain linear constraints explicitly, and to take advantage of the situation where the boundary of the feasible region has so many linear facets that it seems to be smooth. This is done by the TOLMIN software that I developed in 1989, for example, but the number of variables is restricted to a few hundred, because quadratic models with full second derivative matrices are employed. Therefore eventually I expect interior point methods to be best only if the number of variables is large. Another reservation about this field is that it seems to be taking far more than its share of research activity.

M.J.D. Powell [497 ; 2003].

Ce chapitre se limite à l'étude de quelques aspects des algorithmes de suivi de chemin primaux-duaux en optimisation linéaire. Ce sont souvent les algorithmes de points intérieurs les plus efficaces en pratique. D'autres approches de points intérieurs ont été proposées et peuvent être utiles dans des contextes particuliers : algorithme de l'*ellipsoïde de Dikin*, algorithmes à réduction de potentiel, *etc.* Nous renvoyons le lecteur aux notes de fin de chapitre pour des références sur ceux-ci.

Comme leur nom l'indique les algorithmes de suivi de chemin génèrent des itérés qui suivent ce que l'on appelle aujourd'hui le *chemin central*. Celui-ci sert de guide, conduisant les itérés vers une solution particulière du problème (ou vers la solution s'il n'y en a qu'une). Nous l'étudions en détail à la section 18.1.

La partie algorithmique de ce chapitre est importante et débute par des remarques générales sur les approches primales-duales (section 18.2). Vient ensuite l'étude de trois algorithmes supposant que l'on dispose d'un premier itéré admissible (section 18.3) : les algorithmes des petits déplacements, des grands déplacements et prédicteur-correcteur. Les algorithmes sont motivés et décrits ; leur convergence et leur complexité itérative sont étudiées. Le premier algorithme a la complexité itérative théorique

la meilleure, mais ce sont les deux derniers qui sont les plus efficaces en pratique, le dernier étant le plus souvent implémenté (dans une version plus élaborée, il est vrai). Enfin, la section 18.4 décrit et analyse un algorithme n'imposant pas aux itérés d'être admissibles.

Connaissances supposées. Les bases de l'optimisation linéaire (chapitre 17) ; la notion de fonction asymptotique (section 3.3.4), qui est utilisée pour montrer l'existence du chemin central ; l'algorithme de Newton (section 10.1.1).

Rappel et notations

On considère le problème d'optimisation linéaire dans $\mathbb{R}^n$ suivant, dit sous forme standard

$$(P) \quad \begin{cases} \inf_x \ c^\mathsf{T} x \\ Ax = b \\ x \geqslant 0, \end{cases}$$

où, rappelons-le, $c \in \mathbb{R}^n$, A est une matrice $m \times n$ et $b \in \mathbb{R}^m$. Son ensemble admissible est noté

$$\mathcal{F}_\text{P} := \{x \in \mathbb{R}^n : Ax = b,\ x \geqslant 0\}.$$

L'ensemble des itérés strictement admissibles est noté

$$\mathcal{F}_\text{P}^s := \{x \in \mathbb{R}^n : Ax = b,\ x > 0\}.$$

Lorsqu'il est non vide, c'est l'intérieur relatif de $\mathcal{F}_\text{P}$ (exercice 2.20). L'ensemble des solutions de (P) est noté $\mathcal{S}_\text{P}$.

La dualisation lagrangienne de (P) conduit au problème dual suivant (voir la section 17.3.1) :

$$(D) \quad \begin{cases} \sup \ b^\mathsf{T} y \\ A^\mathsf{T} y + s = c \\ s \geqslant 0. \end{cases}$$

Son ensemble admissible est noté

$$\mathcal{F}_\text{D} := \{(y, s) \in \mathbb{R}^m \times \mathbb{R}^n : A^\mathsf{T} y + s = c,\ s \geqslant 0\}.$$

On note aussi

$$\mathcal{F}_\text{D}^s := \{(y, s) \in \mathbb{R}^m \times \mathbb{R}^n : A^\mathsf{T} y + s = c,\ s > 0\},$$

qui, lorsqu'il est non vide, est l'intérieur relatif de $\mathcal{F}_\text{D}$. L'ensemble des solutions de (D) est noté $\mathcal{S}_\text{D}$.

Les conditions d'optimalité de (P) affirment que x est solution de ce problème si, et seulement si, il existe un couple $(y, s) \in \mathbb{R}^m \times \mathbb{R}^n$ tel que l'on ait

$$\begin{cases} A^\mathsf{T} y + s = c, \quad s \geqslant 0 \\ Ax = b, \quad x \geqslant 0 \\ x^\mathsf{T} s = 0. \end{cases} \tag{18.1}$$

Ce sont aussi les conditions d'optimalité du problème dual : (y, s) est solution de (D) si, et seulement si, il existe $x \in \mathbb{R}^n$ tel que l'on ait (18.1).

Nous désignerons par $X = \mathrm{Diag}(x_1, \ldots, x_n)$ la matrice diagonale portant les composantes du vecteur x sur sa diagonale ; de même $S = \mathrm{Diag}(s_1, \ldots, s_n)$, *etc.* On utilisera constamment le vecteur

$$e := \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

dont la dimension se déduira clairement du contexte. Observons que $Xe = x$.

18.1 Le chemin central primal-dual

Le *chemin central* est une courbe située dans

$$\mathcal{F}^s := \mathcal{F}^s_{\mathrm{P}} \times \mathcal{F}^s_{\mathrm{D}} = \{z = (x, y, s) : Ax = b,\ A^{\mathsf{T}} y + s = c,\ x > 0,\ s > 0\}$$

paramétrée par un scalaire $\mu > 0$. C'est donc l'image d'une application

$$\mu \in \,]0, +\infty[\mapsto z(\mu) := (x(\mu), y(\mu), s(\mu)) \in \mathcal{F}^s.$$

Celui-ci sert de guide aux itérés, les conduisant vers une solution primale-duale particulière du problème (P). Pour qu'il soit défini, il faut évidemment que $\mathcal{F}^s_{\mathrm{P}}$ et $\mathcal{F}^s_{\mathrm{D}}$ soient non vides, ce qui, d'après les propositions 17.11 et 17.13, peut s'exprimer en termes primaux seulement :

$$\boxed{\mathcal{F}^s_{\mathrm{P}} \neq \varnothing \text{ et } \mathcal{S}_{\mathrm{P}} \text{ est non vide et borné.}} \tag{18.2}$$

On fera cette hypothèse dans toute cette section. La variable duale $y(\mu)$ ne sera définie de façon univoque que si A est surjective.

Comme son nom l'indique, les points $z(\mu)$ sont « bien centrés » dans l'ensemble admissible. On montrera que, sous certaines conditions, $z(\mu)$ émane (pour $\mu \to +\infty$) du *centre analytique* de l'ensemble admissible $\mathcal{F} := \mathcal{F}_{\mathrm{P}} \times \mathcal{F}_{\mathrm{D}}$ pour aboutir (lorsque $\mu \downarrow 0$) au *centre analytique de la face optimale primale-duale* $\mathcal{S} := \mathcal{S}_{\mathrm{P}} \times \mathcal{S}_{\mathrm{D}}$.

Définition et existence

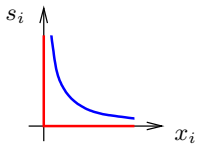
Il y a plusieurs manières d'introduire le chemin central primal-dual (voir aussi l'exercice 18.5). Nous commencerons par utiliser la technique de perturbation des conditions d'optimalité, qui est celle qui se généralise le plus facilement à des problèmes ne faisant pas partie de l'optimisation (comme les problèmes de complémentarité). L'existence du chemin central est cependant plus aisément démontrable si l'on revient dans le domaine de l'optimisation en interprétant le système perturbé comme les conditions d'optimalité d'un problème pénalisé. Nous considérerons les cas de la pénalisation des problèmes dual et primal.

Les équations d'optimalité (18.1) sont formées de deux équations linéaires $A^{\mathsf{T}} y + s = c$ et $Ax = b$, qui à elles seules ne présentent pas trop de difficulté, ainsi que des conditions

$$x^{\mathsf{T}} s = 0, \quad x \geqslant 0, \quad s \geqslant 0.$$

Ces dernières renferment toute la *combinatoire* du problème. Il y a en effet 2^n façons de les réaliser en choisissant, pour tout i, si c'est x_i ou s_i qui est nul. On peut faire disparaître cette combinatoire en perturbant chaque condition $x_i s_i = 0$ en $x_i s_i = \mu$ ($\mu > 0$ est un paramètre) et en imposant la stricte positivité de x et s :

$$\begin{cases} A^{\mathsf{T}} y + s = c, \quad s > 0 \\ Ax = b, \quad x > 0 \\ Xs = \mu e. \end{cases} \tag{18.3}$$



On a noté

$$X := \mathrm{Diag}(x_1, \ldots, x_n) \qquad \text{et} \qquad e := (1 \ \cdots \ 1)^{\mathsf{T}}.$$

Le graphique à droite dans (18.3) montre comment, dans l'espace (x_i, s_i), la courbe *non différentiable* définie par $x_i s_i = 0$ est transformée en la courbe *différentiable* définie par $x_i s_i = \mu$. Au passage, cette observation montre le lien qu'il peut y avoir entre combinatoire et non différentiabilité. Lorsqu'elle existe, la solution de ce système perturbé (18.3) est un point central $z(\mu) = (x(\mu), y(\mu), s(\mu))$.

Nous venons d'exposer l'approche par perturbation. On peut aussi obtenir le système (18.3) par une technique de pénalisation. En effet, (18.3) sont les conditions d'optimalité du problème obtenu à partir du problème (P) par pénalisation logarithmique de sa contrainte de positivité :

$$(P_\mu) \quad \begin{cases} \inf \ c^{\mathsf{T}} x + \mu \, \mathrm{lb}(x) \\ Ax = b, \end{cases} \tag{18.4}$$

où $\mathrm{lb} : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ est la *fonction log-barrière* définie en $x \in \mathbb{R}^n$ par (c'est une fonction autoconcordante, voir l'exemple 16.2)

$$\mathrm{lb}(x) = \begin{cases} -\sum_{i=1}^n \log x_i & \text{si } x > 0 \\ +\infty & \text{sinon.} \end{cases}$$

En effet, comme les contraintes de (P_μ) sont qualifiées (elles sont affines), il existe un multiplicateur de Lagrange $y \in \mathbb{R}^m$ tel que l'on ait les conditions d'optimalité suivantes :

$$\begin{cases} A^{\mathsf{T}} y + \mu X^{-1} e = c \\ Ax = b. \end{cases}$$

En posant $s = \mu X^{-1} e$, on retrouve (18.3). On appelle (P_μ) le *problème barrière primal* (les logarithmes dans lb forment une barrière empêchant x de quitter l'orthant positif).

Une troisième manière d'introduire le chemin central est de pénaliser la contrainte de positivité $s \geqslant 0$ du problème dual. Lorsque $\mathcal{F}_{\mathrm{D}}^s \neq \varnothing$, on peut considérer le problème :

$$(D_\mu) \quad \begin{cases} \sup \ b^{\mathsf{T}} y - \mu \, \mathrm{lb}(s) \\ A^{\mathsf{T}} y + s = c. \end{cases} \tag{18.5}$$

Comme il s'agit d'un problème de maximisation, la pénalisation est obtenue en retranchant la barrière logarithmique (comparez avec le problème (P_μ)).

Intéressons-nous à présent à des conditions assurant l'existence du chemin central. Le fait que le système d'optimalité (18.1), qui est non linéaire, ait une solution n'assure

en rien qu'il en est de même pour le système perturbé (18.3). Il faut certainement que $\mathcal{F}^s \neq \varnothing$ (équivalente à (18.2)) et il se fait que cette condition est aussi suffisante. Elle joue ici le rôle de la *condition de qualification de Slater* du système (18.1).

Proposition 18.1 (existence du chemin central) *Si (18.2) a lieu et $\mu > 0$, alors le système (18.3) a une solution. Celle-ci est unique en x et s. Elle est également unique en y si A est surjective.*

Démonstration. D'après ce qui précède (x, y, s) est solution de (18.3) si, et seulement si, x est solution de (P_μ) de multiplicateur y et si $s = \mu X^{-1}e$ (la condition est suffisante du fait de la convexité du problème (P_μ)). Il reste à montrer que (P_μ) a une solution, car du fait de la stricte convexité de $x \mapsto c^\mathsf{T}x - \mu \sum_{i=1}^n \log x_i$, celle-ci est nécessairement unique en x et en $s = \mu X^{-1}e$.

On considère la fonction $\psi_\mu \in \overline{\mathrm{Conv}}(\mathbb{R}^n)$ définie par

$$\psi_\mu(x) = c^\mathsf{T}x + \mu \,\mathrm{lb}(x) + \mathcal{I}_{\{x\in\mathbb{R}^n : Ax=b\}}(x),$$

où $\mathcal{I}_P$ désigne la fonction indicatrice de l'ensemble P. Ses minimiseurs sont ceux de (P_μ). Il suffit donc que sa fonction asymptotique ψ_μ^∞ vérifie $\psi_\mu^\infty(d) > 0$ pour tout $d \neq 0$, pour que (P_μ) ait un ensemble non vide (et borné) de solutions (proposition 3.29). On observe que

$$\mathrm{lb}^\infty(d) = \begin{cases} 0 & \text{si } d \geqslant 0 \\ +\infty & \text{sinon.} \end{cases}$$

Dès lors

$$\psi_\mu^\infty(d) = \begin{cases} c^\mathsf{T}d & \text{si } Ad = 0 \text{ et } d \geqslant 0 \\ +\infty & \text{sinon.} \end{cases}$$

Si $\psi_\mu^\infty(d) \leqslant 0$, on a

$$d \neq 0, \quad c^\mathsf{T}d \leqslant 0, \quad Ad = 0 \quad \text{et} \quad d \geqslant 0.$$

La bornitude de $\mathcal{S}_\mathrm{P}$ entraîne alors que $d = 0$, ce qui démontre l'existence d'un minimiseur de ψ_μ, c'est-à-dire d'une solution de (P_μ). □

On notera

$$z(\mu) := (x(\mu), y(\mu), s(\mu))$$

un triplet solution de (18.3) (unique en x et s). Le chemin central est donc l'image de l'application

$$\mu \in \,]0, +\infty[\mapsto z(\mu).$$

Propriétés

Nous allons montrer que, sous certaines conditions, le chemin central émane du centre analytique de l'ensemble admissible $\mathcal{F}_\mathrm{P}$ (lorsque $\mu \to +\infty$ et si $\mathcal{F}_\mathrm{P}^s$ est non vide et borné) pour aboutir au centre analytique de la face optimale (lorsque $\mu \downarrow 0$ et si $\mathcal{S}_\mathrm{P}^o$ est non vide et borné). Il nous faut d'abord définir ce que sont ces centres analytiques.

Proposition 18.2 (centre analytique de l'ensemble admissible) *Si $\mathcal{F}_{\mathrm{P}}^s$ est non vide et borné, le problème*

$$\begin{cases} \inf\ \mathrm{lb}(x) \\ Ax = b \end{cases} \tag{18.6}$$

a une solution unique $\check{x} \in \mathbb{R}^n$ caractérisée par les conditions suivantes :

$$A\check{x} = b, \quad \check{x} > 0, \quad \check{X}^{-1}e \in \mathcal{R}(A^{\mathsf{T}}). \tag{18.7}$$

De même, si $\mathcal{F}_{\mathrm{D}}^s$ est non vide et borné et si A est surjective, le problème

$$\begin{cases} \inf\ \mathrm{lb}(s) \\ A^{\mathsf{T}}y + s = c \quad (s > 0) \end{cases} \tag{18.8}$$

a une solution unique $(\check{y}, \check{s}) \in \mathbb{R}^m \times \mathbb{R}^n$ caractérisée par les conditions suivantes :

$$A^{\mathsf{T}}\check{y} + \check{s} = c, \quad \check{s} > 0, \quad \check{S}^{-1}e \in \mathcal{N}(A). \tag{18.9}$$

Démonstration. Le problème (18.6) est celui obtenu par pénalisation logarithmique du problème

$$\begin{cases} \inf\ 0 \\ Ax = b \\ x \geqslant 0. \end{cases}$$

Son ensemble de solution étant $\mathcal{F}_{\mathrm{P}}$, qui par hypothèse est non vide et borné, la proposition 18.1 implique que (18.6) a une solution unique, soit $\check{x} \in \mathbb{R}^n$. Celle-ci est caractérisée par les conditions d'optimalité de (18.6) qui peuvent s'exprimer par (18.7).

La seconde partie se démontre de la même manière. □

Lorsqu'elles existent, les solutions uniques de (18.6) et (18.8) sont appelées les *centres analytiques* des ensembles admissibles $\mathcal{F}_{\mathrm{P}}$ et $\mathcal{F}_{\mathrm{D}}$, respectivement. On utilise le qualificatif « analytique », car ce n'est pas une notion géométrique dans le sens où elle ne dépend pas que de la forme de $\mathcal{F}_{\mathrm{P}}$ et $\mathcal{F}_{\mathrm{D}}$, mais également des équations que définissent ces ensembles. Ainsi, si l'on ajoute une contrainte superflue, c'est-à-dire une contrainte ne modifiant pas l'ensemble admissible, son centre analytique est déplacé (voir l'exercice 18.4). On pourrait penser que la notion géométrique de *centre de gravité* serait un concept plus attrayant, mais son calcul n'est pas aisé, même si l'ensemble est comme ici un polyèdre convexe [633 ; section 2.1.1].

Venons-en maintenant à la notion de centre analytique de la face optimale. On rappelle les notations de la section 17.2.2 :

$$\begin{aligned} \mathfrak{B} &:= \{i \in [1:n] : \exists\, x \in \mathcal{S}_{\mathrm{P}} \text{ vérifiant } x_i > 0\} \\ \mathfrak{N} &:= \{i \in [1:n] : \exists\, (y,s) \in \mathcal{S}_{\mathrm{D}} \text{ vérifiant } s_i > 0\}. \end{aligned}$$

On sait que $\mathfrak{B}$ et $\mathfrak{N}$ forment une partition de $[1:n]$ (propositions 17.7 et 17.8) et que l'ensemble des solutions de (P) est la face de $\mathcal{F}_{\mathrm{P}}$, appelée aussi *face optimale* de $\mathcal{F}_{\mathrm{P}}$,

définie par

$$\mathcal{S}_{\text{P}} = \{x \in \mathcal{F}_{\text{P}} : x_{\mathfrak{N}} = 0\}.$$

On note son intérieur relatif

$$\mathcal{S}_{\text{P}}^{o} = \{x \in \mathcal{F}_{\text{P}} : x_{\mathfrak{B}} > 0,\ x_{\mathfrak{N}} = 0\}. \tag{18.10}$$

Proposition 18.3 (centre analytique de la face optimale) 1) *Si $\mathcal{S}_{\text{P}}$ est non vide et borné et si $\mathfrak{B} \neq \varnothing$, le problème*

$$\begin{cases} \inf\ \text{lb}(x_{\mathfrak{B}}) \\ Ax = b, \quad x_{\mathfrak{N}} = 0 \quad (x_{\mathfrak{B}} > 0) \end{cases} \tag{18.11}$$

a une solution unique $\hat{x} \in \mathbb{R}^n$, caractérisée par les conditions

$$A\hat{x} = b, \quad \hat{x}_{\mathfrak{B}} > 0, \quad \hat{x}_{\mathfrak{N}} = 0, \quad \hat{X}_{\mathfrak{B}}^{-1} e \in \mathcal{R}(A_{:\mathfrak{B}}^{\mathsf{T}}). \tag{18.12}$$

2) *De même si $\mathcal{S}_{\text{D}}$ est non vide et borné (donc A est surjective) et si $\mathfrak{N} \neq \varnothing$, le problème*

$$\begin{cases} \inf\ \text{lb}(s_{\mathfrak{N}}) \\ A^{\mathsf{T}} y + s = c, \quad s_{\mathfrak{B}} = 0 \quad (s_{\mathfrak{N}} > 0) \end{cases} \tag{18.13}$$

a une solution unique $(\hat{y}, \hat{s}) \in \mathbb{R}^m \times \mathbb{R}^n$, caractérisée par les conditions

$$A^{\mathsf{T}}\hat{y} + \hat{s} = c, \quad \hat{s}_{\mathfrak{B}} = 0, \quad \hat{s}_{\mathfrak{N}} > 0, \quad A_{:\mathfrak{N}} \hat{S}_{\mathfrak{N}}^{-1} e \in \mathcal{R}(A_{:\mathfrak{B}}). \tag{18.14}$$

Démonstration. En ne considérant que les variables d'indices dans $\mathfrak{B}$ ($x_{\mathfrak{N}}$ étant fixé à 0), le problème (18.11) est obtenu par pénalisation logarithmique du problème

$$\begin{cases} \inf\ 0^{\mathsf{T}} x_{\mathfrak{B}} \\ A_{:\mathfrak{B}} x_{\mathfrak{B}} = b \\ x_{\mathfrak{B}} \geqslant 0. \end{cases}$$

Par l'hypothèse $\mathfrak{B} \neq \varnothing$, $\{x_{\mathfrak{B}} : Ax_{\mathfrak{B}} = b, x_{\mathfrak{B}} > 0\}$ est non vide et l'ensemble des solutions de ce problème (qui se confond avec son ensemble admissible, c'est-à-dire S_P) est non vide et borné. Par la proposition 18.1, (18.11) a une solution unique. Celle-ci est caractérisée par les conditions d'optimalité de (18.11) qui peuvent s'exprimer par (18.12).

La seconde partie se démontre de la même manière. □

Lorsqu'ils existent, les points $\hat{x}$ et $(\hat{y}, \hat{s})$ définis par (18.12) et (18.14) sont appelés les *centres analytiques des faces optimales*.

On peut à présent décrire le chemin central. On sera amené à comparer le comportement de quantités (éventuellement vectorielles) qui dépendent du paramètre μ. Ainsi si $\mu \mapsto v(\mu)$ et $\mu \mapsto w(\mu)$ sont deux fonctions de μ, on notera :

- $v = O(w)$ s'il existe une constante positive C (indépendante de μ) telle que lorsque $\mu \downarrow 0$ on a $v(\mu) \leqslant Cw(\mu)$ (inégalité vectorielle),

- $v \sim w$ si $v = O(w)$ et $w = O(v)$.

Avec ces notations, $v = O(1)$ signifie que $\{v(\mu)\}_{\mu\downarrow 0}$ est bornée et $v \sim 1$ signifie que $\{v(\mu)\}_{\mu\downarrow 0}$ et $\{v(\mu)^{-1}\}_{\mu\to 0^+}$ sont bornées.

On trouvera à l'exercice 18.6 un résultat précisant le point (ii).

Proposition 18.4 (description du chemin central) *Supposons que* (18.2) *ait lieu. Alors*

(i) *si A est surjective, $x > 0$ et $s > 0$, alors la matrice*

$$\begin{pmatrix} 0 & A^{\mathsf{T}} & I \\ A & 0 & 0 \\ S & 0 & X \end{pmatrix}$$

est inversible et l'application $\mu \mapsto z(\mu)$ définissant le chemin central est de classe C^∞ ;

(ii) *si μ croît, $c^{\mathsf{T}}x(\mu)$ et $x(\mu)^{\mathsf{T}}s(\mu)$ croissent et $b^{\mathsf{T}}y(\mu)$ décroît ;*

(iii) *pour $\bar\mu > 0$ fixé, les ensembles $\{x(\mu) : 0 < \mu \leqslant \bar\mu\}$ et $\{s(\mu) : 0 < \mu \leqslant \bar\mu\}$ sont bornés ; si A est surjective, l'ensemble $\{y(\mu) : 0 < \mu \leqslant \bar\mu\}$ est aussi borné ;*

(iv) *si $\mathcal{F}_{\text{P}}$ est borné, $x(\mu)$ converge vers le centre analytique de $\mathcal{F}_{\text{P}}$, lorsque $\mu \to +\infty$; de même si $\mathcal{F}_{\text{D}}$ est borné (donc A surjective), $(y(\mu), s(\mu))$ converge vers le centre analytique de $\mathcal{F}_{\text{D}}$, lorsque $\mu \to +\infty$;*

(v) *si $\mu \downarrow 0$, on a $x_{\mathfrak{B}}(\mu) \sim 1$, $x_{\mathfrak{N}}(\mu) \sim \mu$, $s_{\mathfrak{B}}(\mu) \sim \mu$, $s_{\mathfrak{N}}(\mu) \sim 1$ et $(x(\mu), s(\mu)) \to (\hat x, \hat s)$ le centre analytique des faces optimales.*

Démonstration. (i) L'application $\mu \mapsto z(\mu)$ est fonction implicite de $F(z, \mu) = 0$, où

$$F(z,\mu) = \begin{pmatrix} A^{\mathsf{T}}y + s - c \\ Ax - b \\ Xs - \mu e \end{pmatrix}.$$

Comme F est C^∞, la fonction implicite sera C^∞ si la matrice $F'_z(z, \mu)$ (celle donnée au point (ii)) est inversible pour $z = z(\mu)$. C'est bien le cas, car si (dx, dy, ds) est dans son noyau, on trouve successivement

$$ds = -X^{-1}Sdx, \quad dx = S^{-1}XA^{\mathsf{T}}dy \quad \text{et} \quad AS^{-1}XA^{\mathsf{T}}dy = 0.$$

Comme A est surjective et $(x, s) > 0$, la matrice $AS^{-1}XA^{\mathsf{T}}$ est inversible, ce qui implique que $(dx, dy, ds) = 0$.

(ii) Ce résultat se démontre comme dans la théorie de la pénalisation (proposition 13.2), en considérant les problèmes pénalisés (P_μ) et (D_μ). Pour le saut de dualité, on utilise $x(\mu)^{\mathsf{T}}s(\mu) = c^{\mathsf{T}}x(\mu) - b^{\mathsf{T}}y(\mu)$. (Voir aussi l'exercice 18.6 pour un résultat plus précis et une démonstration directe lorsque A est surjective.)

(iii) Comme $\mathcal{S}_{\text{P}}$ est borné, l'ensemble $\{x \in \mathbb{R}^n : Ax = b,\ x \geqslant 0,\ c^{\mathsf{T}}x \leqslant c^{\mathsf{T}}x(\bar\mu)\}$ est aussi borné (ils ont le même cône asymptotique). Or $x(\mu)$ est dans ce dernier ensemble lorsque $\mu \in]0, \bar\mu]$ (par le point (ii)), donc $\{x(\mu)\}_{0<\mu\leqslant\bar\mu}$ est borné.

De même, comme $\mathcal{F}_{\text{P}}^s \neq \varnothing$, l'ensemble $\{s \in \mathbb{R}_+ :$ il existe $y \in \mathbb{R}^m$ tel que $A^{\mathsf{T}}y + s = c$ et $b^{\mathsf{T}}y \geqslant b^{\mathsf{T}}y(\bar{\mu})\}$ est borné (les deux ensembles ont le même cône asymptotique $\{r \in \mathbb{R}_+ :$ il existe $z \in \mathbb{R}^m$ tel que $A^{\mathsf{T}}z + r = 0$, et $b^{\mathsf{T}}z \geqslant 0\}$). Comme $s(\mu)$ appartient à cet ensemble lorsque $\mu \in]0, \bar{\mu}]$ (par le point (ii)), $\{s(\mu)\}_{0<\mu\leqslant\bar{\mu}}$ est bornée.

Si A est surjective, l'équation $A^{\mathsf{T}}y(\mu) + s(\mu) = c$ montre que $\{y(\mu) : 0 < \mu \leqslant \bar{\mu}\}$ est aussi borné.

(iv) Si $\mathcal{F}_{\text{P}}$ est borné, $\{x(\mu)\} \subseteq \mathcal{F}_{\text{P}}$ est bornée. Il suffit donc de montrer que cette suite a un unique point d'adhérence quand $\mu \to \infty$, qui est le centre analytique $\check{x}$. Soit $\bar{x}$ un point adhérent à $\{x(\mu)\}$: pour une sous-suite $x(\mu) \to \bar{x}$. Comme $x(\mu)$ est la solution de (P_μ), on a quel que soit $x \in \mathcal{F}_{\text{P}}^s$:

$$c^{\mathsf{T}}x(\mu) + \mu \operatorname{lb}(x(\mu)) \leqslant c^{\mathsf{T}}x + \mu \operatorname{lb}(x).$$

En divisant par $\mu > 0$ et en passant à la limite quand $\mu \to \infty$, on trouve pour tout $x \in \mathcal{F}_{\text{P}}^s$:

$$\operatorname{lb}(\bar{x}) \leqslant \operatorname{lb}(x).$$

Ceci montre que $\bar{x} = \check{x}$.

On s'y prend de la même manière pour $(y(\mu), s(\mu))$. Par optimalité, pour tout couple $(y, s) \in \mathcal{F}_{\text{D}}^s$, on a

$$b^{\mathsf{T}}y(\mu) - \mu \operatorname{lb}(s(\mu)) \geqslant b^{\mathsf{T}}y - \mu \operatorname{lb}(s).$$

(v) Soit $(\bar{x}, \bar{y}, \bar{s})$ une solution strictement complémentaire (proposition 17.8). On réduit le saut de dualité $x^{\mathsf{T}}s$ en remplaçant x par $\bar{x}$:

$$\bar{x}^{\mathsf{T}}s(\mu) \leqslant x(\mu)^{\mathsf{T}}s(\mu) = n\mu.$$

Comme $\bar{x}_{\mathfrak{N}} = 0$, $\bar{x}_{\mathfrak{B}}^{\mathsf{T}}s_{\mathfrak{B}}(\mu) \leqslant n\mu$. Mais $\bar{x}_{\mathfrak{B}} > 0$, donc $s_{\mathfrak{B}}(\mu) = O(\mu)$. Alors, de $x_i(\mu)s_i(\mu) = \mu$, pour tout i, on déduit que $1 = O(x_{\mathfrak{B}}(\mu))$. Mais $\{x_{\mathfrak{B}}(\mu)\}$ est bornée, donc $x_{\mathfrak{B}}(\mu) \sim 1$. En utilisant à nouveau $x_i(\mu)s_i(\mu) = \mu$, pour tout i, on déduit que $s_{\mathfrak{B}}(\mu) \sim \mu$.

De la même manière,

$$x(\mu)^{\mathsf{T}}\bar{s} \leqslant x(\mu)^{\mathsf{T}}s(\mu) = n\mu.$$

Comme $\bar{s}_{\mathfrak{B}} = 0$ et $\bar{s}_{\mathfrak{N}} > 0$, on a $x_{\mathfrak{N}}(\mu) = O(\mu)$. En utilisant alors $x_i(\mu)s_i(\mu) = \mu$, on a $s_{\mathfrak{N}}(\mu) \sim 1$ ($\{s(\mu)\}$ bornée) et donc $x_{\mathfrak{N}}(\mu) \sim \mu$.

Il reste à montrer que $(x(\mu), y(\mu), s(\mu))$ converge vers le centre analytique de la face optimale lorsque $\mu \downarrow 0$. Comme $\{(x(\mu), s(\mu))\}$ est bornée, il existe des sous-suites telles que

$$x(\mu) \to \tilde{x} \quad \text{et} \quad s(\mu) \to \tilde{s}.$$

Passons à la limite dans (18.3). Comme $A^{\mathsf{T}}y(\mu) = c - s(\mu)$ converge et que $\mathcal{R}(A^{\mathsf{T}})$ est fermé, il existe $\tilde{y} \in \mathbb{R}^m$ tel que l'on ait

$$\begin{cases} A^{\mathsf{T}}\tilde{y} + \tilde{s} = c \\ A\tilde{x} = b \\ \tilde{X}\tilde{s} = 0. \end{cases}$$

Donc $\tilde{z} = (\tilde{x}, \tilde{y}, \tilde{s})$ est solution primale-duale. Dès lors $\tilde{x}_{\mathfrak{N}} = 0$ et $\tilde{s}_{\mathfrak{B}} = 0$. D'autre part, $(x_{\mathfrak{B}}(\mu), s_{\mathfrak{N}}(\mu)) \sim 1$ implique que $(\tilde{x}_{\mathfrak{B}}, \tilde{s}_{\mathfrak{N}}) > 0$. Il reste à montrer que $\tilde{X}_{\mathfrak{B}}^{-1} e \in \mathcal{R}(A_{:\mathfrak{B}}^{\mathsf{T}})$ et $A_{:\mathfrak{N}} \tilde{S}_{\mathfrak{N}}^{-1} e \in \mathcal{R}(A_{:\mathfrak{B}})$ pour conclure.

Comme $(x_{\mathfrak{N}}(\mu), s_{\mathfrak{B}}(\mu)) \sim \mu$, il existe une sous-suite telle que $(x_{\mathfrak{N}}(\mu), s_{\mathfrak{B}}(\mu))/\mu \to (x'_{\mathfrak{N}}, s'_{\mathfrak{B}})$. À la limite dans $X(\mu)s(\mu)/\mu = e$, on trouve

$$\tilde{X}_{\mathfrak{B}}^{-1} e = s'_{\mathfrak{B}} \quad \text{et} \quad \tilde{S}_{\mathfrak{N}}^{-1} e = x'_{\mathfrak{N}}.$$

D'autre part $Ax(\mu) = b = A_{:\mathfrak{B}} \bar{x}_{\mathfrak{B}}$ (car $\bar{x}_{\mathfrak{N}} = 0$), donc

$$-A_{:\mathfrak{B}} \frac{x_{\mathfrak{B}}(\mu) - \bar{x}_{\mathfrak{B}}}{\mu} = A_{:\mathfrak{N}} \frac{x_{\mathfrak{N}}(\mu)}{\mu} \to A_{:\mathfrak{N}} x'_{\mathfrak{N}}.$$

Comme $\mathcal{R}(A_{:\mathfrak{B}})$ est fermé, $A_{:\mathfrak{N}} x'_{\mathfrak{N}} \in \mathcal{R}(A_{:\mathfrak{B}})$. De même $A_{:\mathfrak{B}}^{\mathsf{T}} y(\mu) + s_{\mathfrak{B}}(\mu) = c_{\mathfrak{B}} = A_{:\mathfrak{B}}^{\mathsf{T}} \bar{y}$ (car $\bar{s}_{\mathfrak{B}} = 0$), donc

$$A_{:\mathfrak{B}}^{\mathsf{T}} \frac{\bar{y} - y(\mu)}{\mu} = \frac{s_{\mathfrak{B}}(\mu)}{\mu} \to s'_{\mathfrak{B}}.$$

Comme $\mathcal{R}(A_{:\mathfrak{B}}^{\mathsf{T}})$ est fermé, $s'_{\mathfrak{B}} \in \mathcal{R}(A_{:\mathfrak{B}}^{\mathsf{T}})$. □

18.2 Éléments constitutifs des algorithmes

Cheminement

Résoudre le problème d'optimisation (P) revient à résoudre ses conditions d'optimalité (18.1), lesquelles sont nécessaires et suffisantes. En apparence simple, ce système d'équations et d'inéquations présente plusieurs difficultés, toutes liées aux conditions de complémentarité $0 \leqslant s \perp x \geqslant 0$. D'une part, l'équation $s^{\mathsf{T}} x = 0$ qui exprime la perpendicularité de s et x est non linéaire. D'autre part, elle présente une « combinatoire » importante. Elle s'écrit en effet, du fait de la positivité de s et x : $x_i s_i = 0$, pour tout $i \in [1:n]$; il faut donc décider pour tout indice i si $x_i = 0$ ou $s_i = 0$, et il y a 2^n possibilités.

Si l'on a un premier itéré primal-dual $z := (x, y, s)$ avec $x > 0$ et $s > 0$, on pourrait songer à résoudre le système d'optimalité (18.1) directement par des itérations de Newton amorties : à chaque itération, on détermine un pas $\alpha > 0$ le long de la direction de Newton $d := (dx, dy, ds)$ de telle sorte que l'itéré suivant $z_+ := (x_+, y_+, s_+) = z + \alpha d$ vérifie encore $x_+ > 0$ et $s_+ > 0$. On sait en effet que la stricte positivité des variables x et s assure que l'équation de Newton

$$\begin{pmatrix} 0 & A^{\mathsf{T}} & I \\ A & 0 & 0 \\ S & 0 & X \end{pmatrix} \begin{pmatrix} dx \\ dy \\ ds \end{pmatrix} = - \begin{pmatrix} A^{\mathsf{T}} y + s - c \\ Ax - b \\ Xs \end{pmatrix},$$

associée au système d'optimalité est bien définie (la matrice est inversible par la proposition 18.4). L'expérience à montré que cette stratégie, qui est suivie par l'*algorithme affine* (« *affine scaling algorithm* »), ne conduit pas à des algorithmes polynomiaux. La

raison provient probablement du fait que, lorsque z est proche du bord de l'ensemble admissible primal-dual, le pas $\alpha > 0$ le long de d, assurant l'admissibilité des itérés, peut devenir très petit, empêchant tout progrès significatif vers la solution. Une des techniques mises au point pour obtenir la polynomialité consiste à forcer les itérés de rester proche du chemin central, d'une part, et à être moins gourmand, d'autre part, en ne cherchant pas à résoudre le système non linéaire (18.1) directement.

Il est difficile de renoncer à la direction de Newton, dont on connaît les qualités, si bien que le fait de faire des déplacements le long de telles directions est conservé dans les algorithmes de points intérieurs considérés dans ce chapitre. De manière à prévenir le phénomène des petits pas décrit ci-dessus, les algorithmes vont maintenir les itérés suffisamment proches du chemin central étudié à la section 18.1. Nous montrerons en effet dans chaque cas que, dans ces conditions, le pas α pris le long de la direction de Newton est borné inférieurement par une constante strictement positive en $O(n^{-\omega})$, où $\omega > 0$ dépend de l'approche algorithmique, en particulier du type de voisinage du chemin central où sont maintenus les itérés. Même si l'on peut regretter que la borne inférieure sur le pas dépende de n (il ne semble pas possible d'éviter cela), on est au moins assuré que les pas ne deviendront pas arbitrairement petits. Cette dépendance en n aura une incidence directe sur la *complexité itérative* des algorithmes, c'est-à-dire sur le nombre d'itérations qu'ils requièrent pour s'approcher d'une solution à $\varepsilon > 0$ près.

Pour des raisons évidentes, on dit que les algorithmes qui viennent d'être brièvement décrits sont des *méthodes de suivi de chemin*. La solution que l'on recherche par ces algorithmes, le centre analytique de la face optimale situé au bout de chemin central, est un point singulier parce que la jacobienne du système de Newton n'y est en général pas inversible. On ne peut donc pas appliquer directement les techniques de suivi de chemin que l'on utilise dans les méthodes d'homotopie par exemple. Celles mises en œuvre pour suivre le chemin central dans les méthodes de points intérieurs sont originales. Elles dépendent des algorithmes, mais elles relèvent presque toujours d'un principe que l'on peut voir comme la *poursuite d'un objectif fuyant*, ce que l'on a schématisé à la figure 18.1 : en l'itéré $z_k \in \mathcal{F}^s$ de l'itération k, on se fixe pour objectif

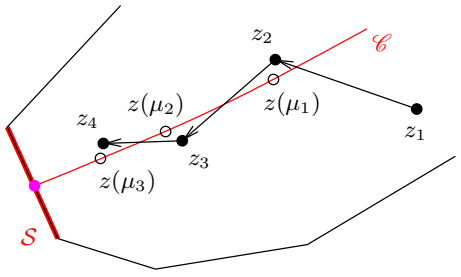


Fig. 18.1. Poursuite d'un objectif fuyant

un point $z(\mu_k)$ sur le chemin central $\mathscr{C}$, mais après avoir fait un pas de Newton dans sa direction (parfois plusieurs pas) conduisant à z_{k+1}, on change d'objectif en visant un autre point $z(\mu_{k+1})$ sur le chemin central, plus près de la solution ($\mu_{k+1} < \mu_k$).

On se rapproche ainsi petit à petit de la solution cherchée, laquelle est en quelque sorte inaccessible directement. Ce principe s'est avéré fécond.

Certains algorithmes imposent aux itérés d'être strictement admissibles (section 18.3). Leur complexité itérative est en $O(n^\omega \log \varepsilon^{-1})$, avec $\omega = \frac{1}{2}$ ou 1, ce qui veut dire que le nombre d'itérations pour atteindre une solution à $\varepsilon > 0$ près est majoré par une constante (indépendante de n et de ε) fois $n^\omega \log \varepsilon^{-1}$ (une définition précise de cette complexité itérative sera donnée plus loin). La complexité itérative en $O(n^{1/2} \log \varepsilon^{-1})$ est la meilleure que l'on ait obtenue ; mais les algorithmes qui la réalisent demandent que l'on dispose d'un premier itéré strictement admissible. D'autres algorithmes autorisent les itérés à ne pas satisfaire les équations linéaires de (18.1), ce qui peut être utile s'il n'y a pas de point primal-dual strictement admissible (c'est-à-dire si $\mathcal{F}^s = \varnothing$) ou si l'on ne dispose pas initialement d'un tel point. Leur complexité itérative est moins bonne ; elle est en $O(n^2 \log \varepsilon^{-1})$ pour l'algorithme étudié à la section 18.4.

Voici à présent quelques concepts qui jouent un rôle-clé dans l'étude des algorithmes de points intérieurs.

Mesures du saut de dualité et du centrage

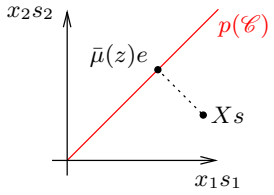
Le contrôle des itérés dans les algorithmes de points intérieurs se fait par plusieurs « mesures » : mesure du saut de dualité, mesure du centrage et mesure de l'admissibilité.

Si l'on veut se donner une cible sur le chemin central primal-dual $\mathscr{C}$, il est nécessaire de savoir près de quel point central l'itéré courant z se trouve. Il n'y aurait en effet pas de sens à se donner une cible qui soit plus éloignée de la solution que ne l'est l'itéré courant. Trouver le point central le plus proche de z n'est cependant pas un problème simple, ni d'ailleurs bien posé car, le chemin central n'étant pas un convexe fermé, la projection de z sur $\mathscr{C}$ n'est en général pas bien définie (proposition 2.25). Par contre, l'image de $\mathscr{C}$ par l'application surjective (et bijective si A est surjective, voir l'exercice 18.2)

$$p : z = (x, y, s) \in \mathcal{F}^s \mapsto (x_1 s_1, \ldots, x_n s_n) \in \mathbb{R}^n_{++} \tag{18.15}$$

est la demi-droite $\{\mu e : \mu > 0\}$, si bien que la projection dans l'espace d'arrivée de cette application se fait trivialement en résolvant le problème

$$\min_{\mu \in \mathbb{R}} \|Xs - \mu e\|_2.$$



Sa solution est la moyenne arithmétique des produits $x_i s_i$:

$$\bar{\mu} \equiv \bar{\mu}(z) := \frac{x^\mathsf{T} s}{n}.$$

On l'appelle la *mesure du saut de dualité* (ou plus simplement le *saut de dualité* ; il s'agit alors d'un abus de langage car, d'après le théorème 17.11, il n'y a pas de

saut de dualité en la solution d'un problème d'optimisation linéaire). Les algorithmes chercheront à faire tendre $\bar{\mu}(z)$ vers 0.

Pour $z = (x, y, s) \in \mathcal{F}^s$:

$$z \in \mathscr{C} \quad \Longleftrightarrow \quad Xs = \bar{\mu}(z)e.$$

Un point $z \in \mathcal{F}^s$ est donc proche du chemin central si $\|(Xs)/\bar{\mu}(z) - e\|$ est petit devant 1 ou encore si $\|Xs - \bar{\mu}(z)e\|$ est petit devant $\bar{\mu}(z)$. On appelle *mesure du centrage* pour la norme ℓ_p, $p \in [1, \infty]$, la quantité

$$\gamma_p(z) := \|Xs - \bar{\mu}(z)e\|_p.$$

Divers voisinages du chemin central sont associés à ce concept de centrage.

On définit une première famille de voisinages, paramétrés par $p \in [1, \infty]$ et $\theta \in [0, 1[$, par

$$V_p(\theta) := \{z \in \mathcal{F}^s : \|Xs - \bar{\mu}(z)e\|_p \leqslant \theta\bar{\mu}(z)\}. \tag{18.16}$$

Dans cette famille de voisinages, on utilisera essentiellement $V_2(\theta)$, qui donne les meilleurs résultats de complexité. Cependant, il est parfois trop petit, ne laissant pas assez de liberté aux itérés. On montre en effet que pour $n \geqslant 2$, $\bigcup_{0 \leqslant \theta < 1} V_2(\theta) \neq \mathcal{F}^s$. Cette affirmation est examinée à l'exercice 18.7.

On fera aussi usage de

$$V_\infty^-(\theta) := \{z \in \mathcal{F}^s : Xs \geqslant (1-\theta)\bar{\mu}(z)e\}, \tag{18.17}$$

où $\theta \in [0, 1[$. Ce voisinage contient $V_\infty(\theta)$ et est assez grand puisque l'on peut montrer que $\bigcup_{0 \leqslant \theta < 1} V_\infty^-(\theta) = \mathcal{F}^s$. On se référera à nouveau à l'exercice 18.7 pour un examen de ces affirmations.

Sur la complexité itérative des algorithmes

Contrairement à l'algorithme du simplexe, les méthodes de points intérieurs ne sont pas des algorithmes à *terminaison finie* : ils ne trouvent pas la solution en un nombre fini d'étapes. En effet, à chaque itération, on a $x > 0$ et $s > 0$, ce qui n'est jamais le cas en une solution de (P) (voir la troisième condition dans (18.1)). On ne peut donc estimer que le nombre d'opérations pour trouver une solution à $\varepsilon > 0$ près. La proximité de la solution se mesurera ici par la petitesse du saut de dualité $\bar{\mu}(z)$. En réalité, il existe des procédures, dites de *purification* (section **??**), permettant de déterminer une solution exacte strictement complémentaire par quelques opérations d'algèbre linéaire à partir d'un itéré généré par un algorithme de points intérieurs, suffisamment proche de la face optimale.

On ne s'intéressera ici qu'à la *complexité itérative* des algorithmes. On veut dire par là que l'on cherche à estimer le nombre d'*itérations* nécessaires pour obtenir une solution à $\varepsilon > 0$ près, dans le pire des cas. Les résultats que nous donnerons sur cette question feront usage du lemme suivant. On note $\{z_k\}$ la suite des itérés générés par l'algorithme considéré et $\bar{\mu}_k = \bar{\mu}(z_k)$.

Lemme 18.5 (de complexité) *Supposons qu'il existe des constantes $\delta > 0$ et $\omega > 0$, telles que pour tout $k \geqslant 0$, les itérés $\{z_k\}$ vérifient*

$$\bar{\mu}_{k+1} \leqslant \left(1 - \frac{\delta}{n^\omega}\right) \bar{\mu}_k.$$

Alors, pour tout $\varepsilon \in]0,1]$ et tout $k \geqslant (n^\omega \log \varepsilon^{-1})/\delta$, on a

$$\frac{\bar{\mu}_k}{\bar{\mu}_0} \leqslant \varepsilon.$$

DÉMONSTRATION. En prenant le logarithme de l'inégalité donnée dans l'énoncé et en utilisant le fait que $\log(1+t) \leqslant t$, on a

$$\log \bar{\mu}_{k+1} \leqslant \log\left(1 - \frac{\delta}{n^\omega}\right) + \log \bar{\mu}_k \leqslant -\frac{\delta}{n^\omega} + \log \bar{\mu}_k.$$

Par récurrence

$$\log \frac{\bar{\mu}_k}{\bar{\mu}_0} \leqslant -\frac{k\delta}{n^\omega}.$$

Dès lors $\bar{\mu}_k/\bar{\mu}_0 \leqslant \varepsilon$ si $-k\delta/n^\omega \leqslant \log \varepsilon$, ce qui s'écrit aussi $k \geqslant (n^\omega \log \varepsilon^{-1})/\delta$. □

En pratique, on cherche à avoir un saut de dualité inférieur à un seuil donné $\varepsilon > 0$:

$$\bar{\mu}_k \leqslant \varepsilon,$$

plutôt que l'inégalité relative $\bar{\mu}_k \leqslant \varepsilon\, \bar{\mu}_0$ fournie par le lemme. Le résultat reste le même, mais avec K qui dépend de $\bar{\mu}_0 > 0$. En effet, on peut écrire $\bar{\mu}_0 = \varepsilon^{1-\kappa}$ pour un certain $\kappa > 0$ (éventuellement grand) dépendant donc de l'itéré initial. On utilise alors le lemme avec $\varepsilon/\bar{\mu}_0 = \varepsilon^\kappa$ au lieu de ε : *pour tout $\varepsilon > 0$ et tout $k \geqslant K := O(n^\omega \log \varepsilon^{-1})$, on a $\bar{\mu}_k \leqslant \varepsilon$.*

18.3 Algorithmes avec itérés admissibles

18.3.1 Préliminaires

Les algorithmes étudiés dans cette section génèrent des itérés strictement admissibles, c'est-à-dire dans $\mathcal{F}^s$. Dès lors, les résidus sont toujours nuls : $r_c = 0$ et $r_b = 0$. Chaque nouvel itéré z_+ est obtenu à partir du précédent z en se déplaçant le long de la direction de Newton d, solution de (18.34), qui s'écrit ici

$$\begin{pmatrix} 0 & A^\mathsf{T} & I \\ A & 0 & 0 \\ S & 0 & X \end{pmatrix} \begin{pmatrix} dx \\ dy \\ ds \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \mu e - Xs \end{pmatrix}. \tag{18.18}$$

Du fait de la linéarité des résidus et de l'utilisation de la direction de Newton pour mettre à jour les itérés, on voit qu'il suffit que les premiers résidus r_c et r_b soient nuls pour que les suivants le soient aussi. Par exemple, si $r_b(z) = 0$, on a en $z_+ = z + \alpha d$: $r_b(z_+) = r_b(z) + \alpha A dx = 0$. De même pour r_c.

Pour que les algorithmes soient bien posés et convergent, il faudra que les itérés conservent la stricte positivité de x et s et fassent décroître le saut de dualité $\bar{\mu}(z)$ vers 0. On comprend que l'analyse des algorithmes passe par l'examen de l'évolution du saut de dualité $\bar{\mu}(z)$ et de la mesure du centrage $\gamma_2(z) := \|Xs - \bar{\mu}(z)e\|_2$ le long de la direction de Newton. C'est ce qu'étudie le lemme suivant. On suppose que le paramètre μ est de la forme $\sigma\bar{\mu}$, avec $\sigma \geqslant 0$.

Lemme 18.6 (évolution du saut de dualité et du centrage) *Soit $d = (dx, dy, ds)$ la direction de Newton en $z = (x, y, s) \in \mathcal{F}^s$, avec $\mu := \sigma\bar{\mu}$, $\sigma \geqslant 0$. On note $z(\alpha) := z + \alpha d$. Alors les relations suivantes sont vérifiées :*
(i) $\bar{\mu}(z(\alpha)) = [1 - \alpha(1 - \sigma)]\bar{\mu}$,
(ii) $\gamma_2(z(\alpha)) \leqslant (1 - \alpha)\gamma_2(z) + \alpha^2\|dXds\|_2$.

Démonstration. On a

$$x_i(\alpha)s_i(\alpha) = (x_i + \alpha dx_i)(s_i + \alpha ds_i) = x_i s_i + \alpha(s_i dx_i + x_i ds_i) + \alpha^2 dx_i ds_i. \quad (18.19)$$

En sommant et en utilisant le fait que $dx^{\mathsf{T}} ds = 0$ (car $Adx = 0$ et $A^{\mathsf{T}} dy + ds = 0$) :

$$\bar{\mu}(z(\alpha)) = \bar{\mu} + \frac{\alpha}{n}(s^{\mathsf{T}} dx + x^{\mathsf{T}} ds).$$

On utilise ensuite le fait que $s^{\mathsf{T}} dx + x^{\mathsf{T}} ds = e^{\mathsf{T}}(Sdx + Xds) = e^{\mathsf{T}}(\sigma\bar{\mu}e - Xs) = n(\sigma - 1)\bar{\mu}$ pour trouver (i).

En reprenant (18.19) et en utilisant $s_i dx_i + x_i ds_i = \sigma\bar{\mu} - x_i s_i$ ainsi que la valeur trouvée pour $\bar{\mu}(z(\alpha))$, on a

$$\begin{aligned} x_i(\alpha)s_i(\alpha) - \bar{\mu}(z(\alpha)) &= (1 - \alpha)x_i s_i + (\alpha - 1)\bar{\mu} + \alpha^2 dx_i ds_i \\ &= (1 - \alpha)(x_i s_i - \bar{\mu}) + \alpha^2 dx_i ds_i. \end{aligned}$$

On en déduit (ii). □

Le point (i) du lemme précédent montre qu'en prenant $\sigma = 0$ et en faisant un pas unité ($\alpha = 1$), on annule le saut de dualité. Il est peu probable cependant que cela conduise en une solution car il y a de fortes chances pour qu'en chemin on ait perdu la positivité des variables x et s (même si l'on a conservé la *somme* des $x_i s_i$ positive).

Le point (ii) montre que le contrôle du centrage $\gamma_2(z(\alpha))$ passe par celui de $\|dXds\|_2$. L'estimation de $dXds$ sera toujours un passage-clé des démonstrations de convergence des algorithmes. On utilisera à plusieurs reprises la technique suivante. On observe que la dernière équation du système de Newton (18.18) s'écrit $Sdx + Xds = \mu - Xs$. En la multipliant par $(XS)^{-\frac{1}{2}}$, elle devient

$$D^{-1}dx + Dds = (XS)^{-\frac{1}{2}}(\mu e - Xs), \quad (18.20)$$

où

$$D = X^{\frac{1}{2}} S^{-\frac{1}{2}}.$$

Le lemme de Mizuno ci-dessous donne alors un moyen de relier la quantité $dXds = (D^{-1}dX)(Dds)$ qui nous intéresse à la somme des deux vecteurs $D^{-1}dx + Dds$, dont on a une autre expression en (18.20).

Lemme 18.7 (Mizuno [424]) *Si u, $v \in \mathbb{R}^n$ vérifient $u^\mathsf{T} v \geqslant 0$ et si $U := \mathrm{Diag}(u_1, \ldots, u_n)$, alors*

$$\|Uv\|_2 \leqslant \frac{1}{\sqrt{8}} \|u + v\|_2^2.$$

Démonstration. On a

$$\|Uv\|_2^2 = \sum_{i=1}^n (u_i v_i)^2 \leqslant \left(\sum_{u_i v_i < 0} u_i v_i \right)^2 + \left(\sum_{u_i v_i > 0} u_i v_i \right)^2,$$

parce que les doubles produits à droite sont tous positifs. La relation $u^\mathsf{T} v \geqslant 0$ s'écrit aussi $\sum_{u_i v_i < 0} |u_i v_i| \leqslant \sum_{u_i v_i > 0} u_i v_i$ et comme $\alpha\beta \leqslant (\alpha + \beta)^2/4$, on obtient

$$\|Uv\|_2^2 \leqslant 2 \left(\sum_{u_i v_i > 0} u_i v_i \right)^2 \leqslant \frac{1}{8} \left(\sum_{u_i v_i > 0} (u_i + v_i)^2 \right)^2 = \frac{1}{8} \|u + v\|_2^4.$$

□

18.3.2 Algorithme des petits déplacements

L'algorithme des petits pas est conceptuellement l'un des plus simples. Sa très faible complexité itérative est obtenue de la manière suivante (voir la figure 18.2). L'algorithme fait des déplacements de Newton complets, avec pas $\alpha = 1$, si bien

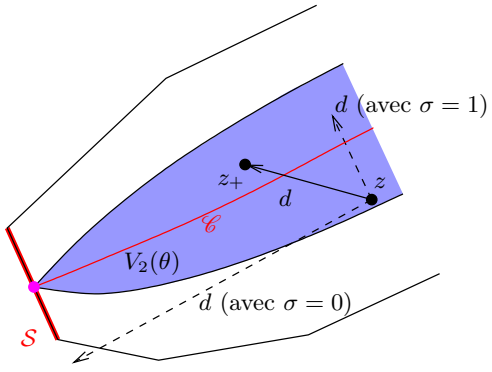
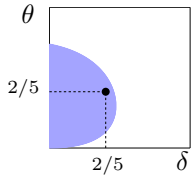


Fig. 18.2. Algorithme des petits déplacements

que $\bar\mu(z_+) = \sigma\bar\mu(z)$ (lemme 18.6). On ne peut pas alors prendre σ trop petit, sous peine de sortir de $\mathcal{F}^s$, comme lorsque $\sigma = 0$. En prenant $\sigma = 1 - \delta/n^{1/2}$ comme facteur de réduction de $\bar\mu$ ($\delta > 0$ étant une constante), on s'assure d'une part de rester dans un voisinage $V_2(\theta)$ (avec $\theta \in]0,1[$ fixé) et d'autre part d'une complexité en $O(n^{1/2}\log\varepsilon^{-1})$ (voir le lemme de complexité 18.5). Pour que les itérés restent dans le voisinage $V_2(\theta)$, une relation doit relier θ et δ, à savoir :

$$\frac{\delta^2+\theta^2}{(1-\delta)(1-\theta)\theta} \leqslant \sqrt{8}. \tag{18.21}$$



Elle est vérifiée pour (δ,θ) dans la zone représentée dans la figure ci-dessus, par exemple pour $\delta = \theta = 2/5$.

Algorithme 18.8 (PIL petits pas)

On se donne θ et δ vérifiant (18.21) et :
– un voisinage du chemin central $V_2(\theta)$,
– un facteur de réduction de $\bar\mu$: $\sigma = 1 - \delta/n^{1/2}$.
L'itéré courant z est supposé dans $V_2(\theta)$.
L'itéré suivant $z_+ \in V_2(\theta)$ s'obtient par les étapes suivantes.
1. Calcul de la direction de Newton d, solution de (18.18) avec $\mu = \sigma\bar\mu(z)$.
2. Nouvel itéré : $z_+ = z + d$.

L'analyse de cet algorithme commence par l'estimation de $\|dXds\|_2$ (lemme 18.9), suivie d'une estimation de $\bar\mu(z_+)$ (lemme 18.10). On peut alors conclure grâce au lemme de complexité.

Lemme 18.9 *Dans l'algorithme 18.8, on a*

$$\|dXds\|_2 \leqslant \frac{\theta^2 + n(1-\sigma)^2}{\sqrt{8}(1-\theta)}\,\bar\mu.$$

Démonstration. L'identité (18.20) et le lemme de Mizuno (qui s'applique car $(D^{-1}dx)^{\mathsf T}(Dds) = 0$) permettent d'écrire

$$\|dXds\|_2 = \|(D^{-1}dX)(Dds)\|_2 \leqslant \frac{1}{\sqrt{8}}\|(XS)^{-\frac12}(\sigma\bar\mu e - Xs)\|_2^2.$$

D'une part, lorsque $z \in V_2(\theta)$, on a $x_i s_i \geqslant (1-\theta)\bar\mu$. Dès lors

$$\|(XS)^{-\frac12}\|_2^2 = \frac{1}{\min_i x_i s_i} \leqslant \frac{1}{(1-\theta)\bar\mu}.$$

D'autre part, on écrit $\sigma\bar\mu e - Xs = (\bar\mu e - Xs) - (1-\sigma)\bar\mu e$ et on observe que $e^{\mathsf T}(\bar\mu e - Xs) = 0$. Dès lors, pour des $z \in V_2(\theta)$, on a

$$\|\sigma\bar{\mu}e - Xs\|_2^2 = \|\bar{\mu}e - Xs\|_2^2 + n(1-\sigma)^2\bar{\mu}^2 \leqslant \theta^2\bar{\mu}^2 + n(1-\sigma)^2\bar{\mu}^2.$$

En utilisant ces deux dernières majorations dans la première, on obtient le résultat. □

Lemme 18.10 *Si $\theta \in]0,1[$ et $\delta \in]0,1[$ vérifient (18.21), alors $z_+ \in V_2(\theta)$ et*

$$\bar{\mu}(z_+) = \left(1 - \frac{\delta}{n^{1/2}}\right)\bar{\mu}(z).$$

Démonstration. D'après le point (i) du lemme 18.6 et $\alpha = 1$, on a

$$\bar{\mu}(z_+) = \sigma\bar{\mu}.$$

En utilisant l'expression choisie pour σ, on en déduit la formule de $\bar{\mu}(z_+)$ de l'énoncé.

Il reste à montrer que $z_+ \in V_2(\theta)$. D'après le point (ii) du même lemme, le lemme précédent, on a pour tout $\alpha \in [0,1]$:

$$\begin{aligned}\gamma_2(z(\alpha)) &\leqslant (1-\alpha)\gamma_2(z) + \alpha^2\frac{\theta^2 + n(1-\sigma)^2}{\sqrt{8}(1-\theta)}\,\bar{\mu} \\ &\leqslant \theta\left(1 - \alpha + \alpha\frac{\theta^2 + n(1-\sigma)^2}{\sqrt{8}(1-\theta)\theta}\right)\bar{\mu}, \qquad (18.22)\end{aligned}$$

où on a utilisé le fait que $z \in V_2(\theta)$ et $\alpha \leqslant 1$. Dès lors, compte tenu du fait que $\bar{\mu}(z(\alpha)) = (1 - \alpha + \alpha\sigma)\bar{\mu}$ (voir le point (i) du lemme 18.6), on a

$$\gamma_2(z(\alpha)) \leqslant \theta\bar{\mu}(z(\alpha)), \quad \forall\alpha \in [0,1], \qquad (18.23)$$

si la fraction dans (18.22) est inférieure à σ. C'est ce qui impose à σ d'être une fonction de n. En prenant $\sigma = 1 - \delta/n^{1/2}$ comme dans l'algorithme, on voit qu'il faut que l'on ait

$$\frac{\theta^2 + \delta^2}{\sqrt{8}(1-\theta)\theta} \leqslant 1 - \frac{\delta}{n^{1/2}}.$$

Cette inégalité est la plus restrictive lorsque $n = 1$: c'est l'inégalité (18.21) qui est vérifiée. L'estimation (18.23) que nous venons de démontrer implique que $\gamma_2(z_+) \leqslant \theta\bar{\mu}(z_+)$ et que $x_i(\alpha)s_i(\alpha) > 0$ pour tout $\alpha \in [0,1]$. Donc les composantes de $x(\alpha)$ et $s(\alpha)$ ne peuvent s'annuler. Comme elles sont strictement positives en $\alpha = 0$, elles le reste en $\alpha = 1$. Dès lors $z_+ \in V_2(\theta)$. □

L'estimation obtenue dans le lemme précédent et le lemme de complexité 18.5 permettent d'obtenir facilement le résultat de convergence et de complexité itérative suivant.

Théorème 18.11 (convergence et complexité de l'algorithme des petits déplacements) *L'algorithme 18.8 converge ($\bar{\mu}_k \to 0$ q-linéairement) et pour tout $\varepsilon > 0$, il existe un indice $K = O(n^{1/2} \log \varepsilon^{-1})$ tel que $\bar{\mu}_k \leqslant \varepsilon \bar{\mu}_0$, dès que $k \geqslant K$.*

Même si la réduction du saut de dualité $\bar{\mu}_k$ par le facteur $1 - \delta/n^{1/2}$ assure à l'algorithme des petits déplacements la meilleure complexité itérative obtenue à ce jour, $\bar{\mu}_k$ tend vers zéro assez lentement lorsque n est grand. Par exemple, si on prend $\delta = 2/5$ et si $n = 10^4$, il faudra 575 itérations pour réduire le saut de dualité d'un facteur 10. C'est beaucoup. Cet algorithme n'est donc guère utilisé en pratique. Il nous a toutefois permis de présenter de manière concise un des meilleurs résultats de complexité que peuvent apporter les méthodes de points intérieurs en optimisation linéaire.

18.3.3 Algorithme des grands déplacements

On obtient de plus grands déplacements, et donc une efficacité numérique meilleure en pratique, en utilisant $V_\infty^-(\theta)$ comme voisinage du chemin central devant contrôler les itérés. La complexité itérative théorique est cependant moins bonne ; elle est en $O(n \log \varepsilon^{-1})$. Dans l'algorithme proposé ci-dessous, on prend un pas α le long de la direction de Newton le plus grand possible.

Algorithme 18.12 (PIL grands pas)

On se donne :
– un voisinage du chemin central $V_\infty^-(\theta)$, avec $\theta \in]0,1[$,
– des bornes $0 < \sigma_{\min} < \sigma_{\max} < 1$ pour le facteur de réduction de $\bar{\mu}$.
L'itéré courant z est supposé dans $V_\infty^-(\theta)$.
L'itéré suivant $z_+ \in V_\infty^-(\theta)$ s'obtient par les étapes suivantes.

1. Choix de $\sigma \in [\sigma_{\min}, \sigma_{\max}]$.
2. Calcul de la direction de Newton d, solution de (18.18) avec $\mu = \sigma \bar{\mu}(z)$.
3. Calcul du pas α le plus grand possible dans $]0,1]$ tel que $z + \alpha d \in V_\infty^-(\theta)$.
4. Nouvel itéré : $z_+ = z + \alpha d$.

Typiquement, on prend $\theta = 0.99$, ce qui permet à $V_\infty^-(\theta)$ d'occuper une grande partie de $\mathcal{F}^s$. L'étape 3 demande de calculer le plus grand pas α permettant à $z(\alpha) = z + \alpha d$ de rester dans $V_\infty^-(\theta)$. Il faut donc vérifier que $x_i(\alpha) s_i(\alpha) \geqslant (1-\theta)\bar{\mu}(z(\alpha))$ pour tout i. Comme $x_i(\alpha)$, $s_i(\alpha)$ et $\bar{\mu}(z(\alpha))$ dépendent linéairement de α (lemme 18.6), on voit que cela revient à trouver les racines de n fonctions quadratiques.

L'analyse de l'algorithme 18.12 commence par l'estimation de $\|dXds\|_2$.

Lemme 18.13 *Dans l'algorithme 18.12, on a*

$$\|dXds\|_2 \leqslant \frac{n(2-\theta)}{\sqrt{8}(1-\theta)}\,\bar{\mu}.$$

DÉMONSTRATION. Comme $(D^{-1}dx)^{\mathsf{T}}(Dds) = 0$, on peut appliquer le lemme de Mizuno sur l'identité (18.20). Cela donne

$$\|dXds\|_2 \leqslant \frac{1}{\sqrt{8}}\|(XS)^{-\frac{1}{2}}(\sigma\bar{\mu}e - Xs)\|_2^2.$$

Lorsque $z \in V_\infty^-(\theta)$, on a $x_i s_i \geqslant (1-\theta)\bar{\mu}$. Dès lors, en développant le carré

$$\begin{aligned}\|dXds\|_2 &\leqslant \frac{1}{\sqrt{8}}\|\sigma\bar{\mu}(XS)^{-\frac{1}{2}}e - (XS)^{\frac{1}{2}}e\|_2^2 \\ &\leqslant \frac{1}{\sqrt{8}}\left(\sigma^2\bar{\mu}^2\frac{n}{(1-\theta)\bar{\mu}} - 2n\sigma\bar{\mu} + n\bar{\mu}\right) \\ &\leqslant \frac{1}{\sqrt{8}}\left(\frac{1}{(1-\theta)} + 1\right)n\bar{\mu},\end{aligned}$$

où on a majoré $\sigma^2 \leqslant 1$ et négligé $-2n\sigma\bar{\mu} \leqslant 0$. □

Lemme 18.14 *Dans l'algorithme 18.12, il existe une constante δ indépendante de n telle que*

$$\bar{\mu}(z_+) \leqslant \left(1 - \frac{\delta}{n}\right)\bar{\mu}(z).$$

DÉMONSTRATION. Il s'agit d'obtenir une borne inférieure sur le pas maximal α permettant de rester dans $V_\infty^-(\theta)$, qui soit en $O(n^{-1})$. Le résultat se trouve alors en appliquant le point (i) du lemme 18.6.

D'après le lemme précédent, il existe une constante $C_1 > 0$ indépendante de n telle $dx_i ds_i \geqslant -C_1 n\bar{\mu}$. En utilisant (18.19) comme dans le lemme 18.6, l'équation de Newton et $x_i s_i \geqslant (1-\theta)\bar{\mu}$ (car $z \in V_\infty^-(\theta)$), on trouve pour $t > 0$:

$$\begin{aligned}x_i(t)s_i(t) &= (1-t)x_i s_i + t\sigma\bar{\mu} + t^2 dx_i ds_i \\ &\geqslant (1-t)(1-\theta)\bar{\mu} + t\sigma\bar{\mu} - C_1 t^2 n\bar{\mu}.\end{aligned}$$

D'après le point (i) du lemme 18.6, $\bar{\mu}(z(t)) = (1-t+t\sigma)\bar{\mu}$. Alors l'inégalité ci-dessus montre que l'on aura $x_i(t)s_i(t) \geqslant (1-\theta)\bar{\mu}(z(t)) \equiv (1-\theta)(1-t+t\sigma)\bar{\mu}$ et $x_i(t)s_i(t) > 0$ si

$$t \leqslant \frac{\theta\sigma}{C_1 n} \quad \text{et} \quad t < \frac{1}{1-\sigma}.$$

Comme α est le plus grand $t > 0$ tel que $x_i(t)s_i(t) \geqslant (1-\theta)\bar{\mu}(z(t))$, $x_i(t) > 0$ et $s_i(t) > 0$, pour tout i, on a

$$\alpha \geqslant \min\left(\frac{\theta\sigma}{C_1 n}, \frac{1}{2(1-\sigma)}\right).$$

On note à présent qu'avec les bornes encadrant $\sigma \in [\sigma_{\min}, \sigma_{\max}]$, on a $\sigma(1-\sigma) \geqslant C_2$, une constante strictement positive indépendante de n. Dès lors, la formule de $\bar{\mu}(z(\alpha))$ du lemme 18.6, conduit à l'estimation cherchée :

$$\bar{\mu}(z(\alpha)) = [1 - \alpha(1-\sigma)]\bar{\mu} \leqslant \left(1 - \frac{\delta}{n}\right)\bar{\mu},$$

où $\delta = \min(\theta C_2/C_1, 1/2)$. □

On peut maintenant conclure aisément en appliquant le lemme de complexité 18.5.

Théorème 18.15 (convergence et complexité de l'algorithme des grands déplacements) *L'algorithme 18.12 converge ($\bar{\mu}_k \to 0$ q-linéairement) et pour tout $\varepsilon > 0$, il existe un indice $K = O(n \log \varepsilon^{-1})$ tel que $\bar{\mu}_k \leqslant \varepsilon \bar{\mu}_0$, dès que $k \geqslant K$.*

18.3.4 Un algorithme prédicteur-correcteur

Les algorithmes précédents souffrent du fait que le facteur de réduction σ du saut de dualité est imposé (il vaut $1 - \delta/n^{1/2}$ dans l'algorithme des petits déplacements) ou minoré (par $\sigma_{\min}$ dans l'algorithme des grands déplacements) a priori. Il semble en effet préférable de laisser l'algorithme choisir lui-même une réduction maximale de $\bar{\mu}(z)$. C'est ce qui motive l'algorithme prédicteur-correcteur, que nous présentons dans cette section.

Cet algorithme doit son nom au fait qu'il alterne des *phases de prédiction* et des *phases de correction* (voir la figure 18.3). La phase de prédiction en $z \in \mathcal{F}^s$ a pour

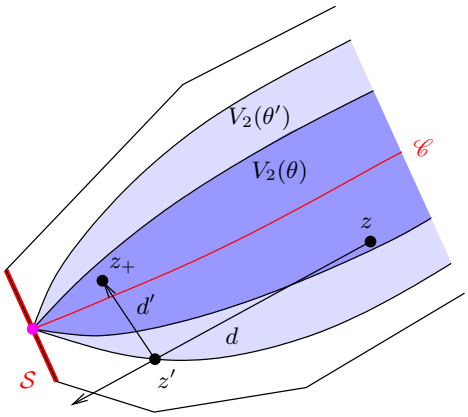


Fig. 18.3. Un algorithme prédicteur-correcteur

but de faire décroître le saut de dualité $\bar{\mu}$ le plus possible, ce qui s'obtient en suivant la direction de Newton d avec $\sigma = 0$ (voir la proposition 18.6 (i)) et en prenant le plus grand pas $\alpha > 0$ le long de cette direction, de manière toutefois à ce que $z' = z + \alpha d$ reste dans un voisinage $V_2(\theta')$ (avec un pas unité, z' serait vraisemblablement en dehors de l'ensemble admissible). Le point z' ne peut être le nouvel itéré car une direction de prédiction en ce point aurait de grande chance de sortir immédiatement de $V_2(\theta')$. L'étape de correction a donc pour but de revenir dans un voisinage du chemin central $V_2(\theta)$ plus petit (avec $0 < \theta < \theta' < 1$). Ce recentrage se fait par un

pas unité le long de la direction de Newton d' avec $\sigma = 1$, donc sans modification du saut de dualité. Il faut que θ et θ' vérifient les inégalités

$$0 < \theta < \theta' < 1 \quad \text{et} \quad \frac{(\theta')^2}{1-\theta'} \leqslant \sqrt{8}\theta, \tag{18.24}$$

pour que le pas unité le long de d' conduise en un point z_+ dans $V_2(\theta)$. Les valeurs $\theta = \frac{1}{4}$ et $\theta' = \frac{1}{2}$ conviennent. On considère qu'une itération est formée d'une phase de prédiction suivie d'une phase de correction.

Algorithme 18.16 (PIL prédicteur-correcteur)

On considère deux voisinages $V_2(\theta)$ et $V_2(\theta')$, avec θ et θ' vérifiant (18.24).
On suppose que l'itéré courant $z := (x, y, s) \in V_2(\theta)$.
L'itéré suivant $z_+ \in V_2(\theta)$ s'obtient par les étapes suivantes.

1. *Phase de prédiction* :
   - 1.1. Calcul du déplacement de Newton d, solution de (18.18) avec $\mu = 0$.
   - 1.2. Calcul d'un pas α le plus grand possible dans $]0, 1]$ tel que $z + \alpha d \in V_2(\theta')$.
   - 1.3. Itéré intermédiaire : $z' = z + \alpha d$.
2. *Phase de correction* :
   - 2.1. Calcul du déplacement de Newton d', solution de (18.18) en $z = z'$, avec $\mu = \bar{\mu}' := \bar{\mu}(z')$.
   - 2.2. Nouvel itéré : $z_+ = z' + d'$ (pas unité).

L'étape 1.2 demande de calculer le plus grand pas α permettant à $z(\alpha) = z + \alpha d$ de rester dans $V_2(\theta')$. Il faut donc vérifier que $\|X(\alpha)s(\alpha) - \bar{\mu}(z(\alpha))e\|_2^2 = (\theta')^2\bar{\mu}(z(\alpha))^2$. Comme $x(\alpha)$, $s(\alpha)$ et $\bar{\mu}(z(\alpha))$ dépendent linéairement de α (lemme 18.6), on voit que cela revient à trouver la racine maximale d'un polynôme quartique (d'ordre 4).

Nous analysons dans les deux lemmes suivants les phases de prédiction et de correction, avant de conclure par un résultat de convergence et de complexité. Dans la phase de prédiction la décroissance du saut de dualité se fait par un facteur semblable à celui de l'algorithme des petits déplacements, mais cette fois l'algorithme a la possibilité de faire décroître $\bar{\mu}$ plus rapidement.

Lemme 18.17 (phase de prédiction) *Si $0 < \theta < \theta' < 1$, il existe une constante $\delta > 0$ indépendante de n telle que, lorsque $z \in V_2(\theta)$, la phase de prédiction conduit à un point $z' \in V_2(\theta')$ avec*

$$\bar{\mu}' \leqslant \left(1 - \frac{\delta}{n^{1/2}}\right)\bar{\mu}.$$

Démonstration. On cherche à estimer le pas maximal α tel que $z + \alpha d$ reste dans $V_2(\theta')$. Ceci se fait en contrôlant l'évolution du centrage le long de d.

On applique le lemme de Mizuno en considérant (18.20) avec $\sigma = 0$, ce qui est licite car $(D^{-1}dx)^{\mathsf{T}}(Dds) = 0$. On a

$$\|dXds\|_2 = \|(D^{-1}dX)(Dds)\|_2 \leqslant \frac{1}{\sqrt{8}}\|(XS)^{-\frac{1}{2}}Xs\|_2^2 = \frac{n}{\sqrt{8}}\bar{\mu}.$$

Alors le point (ii) du lemme 18.6 donne pour $z \in V_2(\theta)$ et $z(t) := z + td$:

$$\gamma_2(z(t)) \leqslant \left[(1-t)\theta + t^2\frac{n}{\sqrt{8}}\right]\bar{\mu}.$$

D'après le point (i) du lemme 18.6, $\bar{\mu}(z(t)) = (1-t)\bar{\mu} > 0$, si $t < 1$. Dès lors $z(t)$ sera certainement dans $V_2(\theta')$ si $t < 1$ et si le membre de droite ci-dessus est inférieur à $\theta'\bar{\mu}(z(t)) = \theta'(1-t)\bar{\mu}$, ce qui s'écrit encore

$$\frac{n}{\sqrt{8}}t^2 \leqslant (\theta' - \theta)(1-t).$$

Cette inégalité est vérifiée pour tout $t \in [0, \delta/\sqrt{n}]$, où $\delta > 0$ est fixé indépendamment de n par :

$$\frac{\delta^2}{\sqrt{8}} \leqslant (\theta' - \theta)(1-\delta).$$

Il suffit en effet d'observer que pour $t \in [0, \delta/\sqrt{n}]$, on a $nt^2 \leqslant \delta^2$ et $1 - \delta \leqslant 1 - \delta/\sqrt{n} \leqslant 1 - t$. Comme la phase de prédiction détermine le plus grand pas α tel que $z(\alpha) \in V_2(\theta')$, ce pas vérifie $\alpha \geqslant \delta/\sqrt{n}$. Dès lors

$$\bar{\mu}(z(\alpha)) = (1-\alpha)\bar{\mu} \leqslant \left(1 - \frac{\delta}{n^{1/2}}\right)\bar{\mu}.$$

□

Lemme 18.18 (phase de correction) *Si θ et θ' vérifient (18.24) et si $z' \in V_2(\theta')$, alors la phase de correction donne un point $z_+ \in V_2(\theta)$ tel que $\bar{\mu}_+ = \bar{\mu}'$.*

Démonstration. Le point (i) du lemme 18.6 et $\sigma = 1$ montrent immédiatement que $\bar{\mu}_+ = \bar{\mu}'$. Il reste à montrer que $z_+ \in V_2(\theta)$.

D'après le lemme de Mizuno et (18.20) avec $\mu = \bar{\mu}$, la direction de Newton $d' = (dx', dy', ds')$ calculée en z' vérifie

$$\|dX'ds'\|_2 \leqslant \frac{1}{\sqrt{8}}\|(X'S')^{-\frac{1}{2}}(\bar{\mu}'e - X's')\|_2^2 = \frac{1}{\sqrt{8}}\sum_{i=1}^{n}\frac{(x_i's_i' - \bar{\mu}')^2}{x_i's_i'}.$$

Puisque $z' \in V_2(\theta')$, on a pour tout i : $\|X's' - \bar{\mu}'e\|_2 \leqslant \theta'\bar{\mu}'$, donc $x_i's_i' \geqslant (1-\theta')\bar{\mu}'$, si bien que

$$\|dX'ds'\|_2 \leqslant \frac{1}{\sqrt{8}(1-\theta')\bar{\mu}'}\|X's' - \bar{\mu}'e\|_2^2 \leqslant \frac{(\theta')^2}{\sqrt{8}(1-\theta')}\bar{\mu}'.$$

Le point (ii) du lemme 18.6 avec $\alpha = 1$ et (18.24) donnent

$$\|X_+s_+ - \bar{\mu}_+e\|_2 \leqslant \frac{(\theta')^2}{\sqrt{8}(1-\theta')}\bar{\mu}_+ \leqslant \theta\bar{\mu}_+.$$

Pour conclure que $z_+ \in V_2(\theta)$, il reste à montrer que $x_+ > 0$, $s_+ > 0$, mais ceci se déduit du fait aisément démontrable que $z' + td' \in V_2(\theta')$ pour tout $t \in [0,1]$. □

Théorème 18.19 (convergence et complexité de l'algorithme prédicteur-correcteur) *L'algorithme prédicteur-correcteur 18.16 converge* ($\bar\mu_k \to 0$ *q-linéairement*) *et pour tout* $\varepsilon > 0$, *il existe un indice* $K = O(n^{1/2}\log\varepsilon^{-1})$ *tel que* $\bar\mu_k \leqslant \varepsilon\bar\mu_0$, *dès que* $k \geqslant K$.

Démonstration. D'après les lemmes 18.17 et 18.18, on a

$$\bar\mu_{k+1} \leqslant \left(1 - \frac{\delta}{n^{1/2}}\right)\bar\mu_k.$$

On utilise alors le lemme de complexité 18.5. □

En pratique, le pas $\alpha > 0$ le long de la direction de prédiction est souvent proche de 1, si bien que l'algorithme converge souvent très rapidement. On peut d'ailleurs montrer que la convergence de $\bar\mu_k \to 0$ est superlinéaire. Malgré cela, l'algorithme peut être gêné par le petitesse du voisinage $V_2(\theta)$, surtout dans les premières itérations.

18.4 Un algorithme sans admissibilité forcée

On ne dispose pas toujours d'un itéré primal-dual strictement admissible. Parfois même il n'y en a pas, alors que le problème est parfaitement bien posé. Cette situation se rencontre, par exemple, si l'on transforme un problème d'optimisation linéaire de la forme $\inf\{c^{\mathsf T}x : Ax \leqslant b\}$ sous la forme standard : $\inf\{c^{\mathsf T}u - c^{\mathsf T}v : Au - Av + z = b, (u,v,z) \geqslant 0\}$, où z est un vecteur de variables d'écart et $u - v$ joue le rôle de x. Dans ce cas, les contraintes du problème dual s'écrivent : $A^{\mathsf T}y + s^1 = c$, $-A^{\mathsf T}y + s^2 = -c$, $y + s^3 = 0$ et $(s^1, s^2, s^3) \geqslant 0$. En sommant les deux premières contraintes, on voit que s^1 et s^2 sont nuls : il n'y a pas de point strictement admissible pour le dual. Il est donc intéressant de développer des algorithmes dans lesquels les itérés ne vérifient pas les contraintes linéaires de (18.1).

Dans les algorithmes n'obligeant pas les itérés à rester dans $\mathcal{F}^s$, quoique maintenant $x > 0$ et $s > 0$ (c'est leur aspect « points intérieurs »), il y a deux objectifs à réaliser : annuler les résidus $(r_b, r_c) = (Ax - b, A^{\mathsf T}y + s - c)$ et le saut de dualité $\bar\mu = x^{\mathsf T}s/n$. Ces algorithmes opèrent comme précédemment, en faisant des déplacements le long de la direction de Newton d, solution de (18.34).

On s'est longtemps demandé si l'on pourrait un jour trouver des algorithmes polynomiaux dans lesquels les itérés ne sont pas admissibles. Une façon d'y arriver est de ne pas faire décroître le saut de dualité plus vite que l'admissibilité. La possibilité de réaliser ce principe est fondé sur les observations de l'évolution des résidus et du saut de dualité le long de la direction de Newton, disons en $z + \alpha d$, $\alpha \geqslant 0$, alors que le point courant z n'annule pas les résidus. On note

$$r(z) := (r_b(z), r_c(z)) = (Ax - b, A^{\mathsf T}y + s - c).$$

On vérifie facilement que les résidus varient de façon affine :

$$r(z + \alpha d) = (1 - \alpha)r(z).$$

De même pour les normes, pour autant que $\alpha \leqslant 1$:

$$\|r(z+\alpha d)\|_2 = (1-\alpha)\|r(z)\|_2. \tag{18.25}$$

Le saut de dualité évolue quant à lui de façon non linéaire. En utilisant la troisième équation de (18.34), on a $s^{\mathsf{T}}dx + x^{\mathsf{T}}ds = n\mu - x^{\mathsf{T}}s = n(\mu - \bar{\mu})$ et donc

$$\bar{\mu}(z+\alpha d) = \frac{1}{n}(x+\alpha dx)^{\mathsf{T}}(s+\alpha ds) = \bar{\mu} + \alpha(\mu - \bar{\mu}) + \frac{\alpha^2}{n}dx^{\mathsf{T}}ds.$$

Sans admissibilité, on n'a plus $dx^{\mathsf{T}}ds = 0$, si bien que la décroissance de $\bar{\mu}(z(\alpha))$ pour de petit pas $\alpha > 0$ n'est assurée que si on prend $\mu < \bar{\mu}$ (on n'est pas maître du signe de $dx^{\mathsf{T}}ds$) et on se fera à l'idée de prendre comme précédemment

$$\mu = \sigma\bar{\mu}, \quad \text{avec un } \sigma \in]0,1[.$$

On verra par la suite qu'il faut être plus contraignant sur la valeur de σ. On a alors

$$\bar{\mu}(z+\alpha d) = (1-\alpha)\bar{\mu} + \alpha\sigma\bar{\mu} + \frac{\alpha^2}{n}dx^{\mathsf{T}}ds. \tag{18.26}$$

On en déduit que

$$\bar{\mu}(z+\alpha d) \geqslant (1-\alpha)\bar{\mu}, \quad \text{pour } \alpha > 0 \text{ petit}. \tag{18.27}$$

En comparant (18.25) et (18.27), on voit que si $\|r\|_2 \leqslant \rho\bar{\mu}$ à l'itéré courant (pour une constante $\rho > 0$), on a pour $\alpha > 0$ petit :

$$\|r(z+\alpha d)\|_2 \leqslant (1-\alpha)\rho\bar{\mu} \leqslant \rho\bar{\mu}(z+\alpha d).$$

L'inégalité $\|r\|_2 \leqslant \rho\bar{\mu}$ est donc conservée à l'itération suivante. Autrement dit, de petits pas le long de la direction de Newton maintiennent les itérés dans le voisinage du chemin central suivant :

$$V_\infty^- \equiv V_\infty^-(\theta,\rho) := \{z := (x,y,s) : (x,s) > 0,\ \|r(z)\|_2 \leqslant \rho\bar{\mu}(z),\ Xs \geqslant (1-\theta)\bar{\mu}(z)e\}.$$

On prend $\theta \in]0,1[$ et de manière à ce que le premier itéré soit dans ce voisinage, on prend

$$\rho = \rho' \frac{\|r^0\|_2}{\bar{\mu}_0}, \quad \text{avec } \rho' \geqslant 1. \tag{18.28}$$

Ce voisinage diffère de $V_\infty^-(\theta)$ par le fait que le résidu $r = (r_b, r_c)$ ne doit pas nécessairement être nul.

Nous donnons ci-dessous une itération d'un algorithme de suivi de chemin fondé sur les observations précédentes. Il ne demande donc pas d'avoir un premier itéré admissible et permet de faire de grands déplacements via l'utilisation du voisinage V_∞^-. L'algorithme est très simple. À chaque itération, on fait un pas $\alpha > 0$ le long de la direction de Newton d, de manière à rester dans V_∞^-, tout en faisant décroître le saut de dualité $\bar{\mu}$. Une recherche linéaire sur $\alpha \mapsto \bar{\mu}(z+\alpha d)$ assure la décroissance du saut de dualité et du même coup celle des résidus (puisque les itérés restent dans V_∞^-). Cette stratégie est possible car, comme nous l'avons montré ci-dessus, la direction de Newton rentre dans V_∞^- et est une direction de descente de $\bar{\mu}$.

Algorithme 18.20 (PIL sans admissibilité)

On se donne des constantes :
– $\omega \in]0,1[$ est utilisé dans la recherche linéaire sur $\bar{\mu}(\cdot)$,
– $\sigma \in]0, 1-\omega[$ est le facteur de réduction de $\bar{\mu}$ dans le système de newton,
– $\theta \in]0,1[$ et ρ vérifiant (18.28) définissent le voisinage $V_\infty^-(\theta,\rho)$.
On suppose que l'itéré courant $z \in V_\infty^-(\theta,\rho)$.
L'itéré suivant $z_+ \in V_\infty^-(\theta,\rho)$ s'obtient par les étapes suivantes.

1. *Direction de Newton.* Calcul de la direction de Newton $d = (dx, dy, ds)$, solution de (18.34), avec $\mu = \sigma\bar{\mu}(z)$.
2. *Recherche linéaire.* Prendre un pas α le plus grand possible dans $]0,1]$ tel que $z + \alpha d \in V_\infty^-(\theta,\rho)$ et
$$\bar{\mu}(z+\alpha d) \leqslant (1-\omega\alpha)\bar{\mu}(z). \tag{18.29}$$
3. *Nouvel itéré.* $z^+ := z + \alpha d$.

Les valeurs typiques des constantes sont

$$\omega = 10^{-2}, \quad \sigma = 0.25 \quad \text{et} \quad \theta = 0.99.$$

On note z^k les itérés, α_k le pas pris à l'étape 2, $r_b^k := Ax^k - b$, $r_c^k := A^\mathsf{T} y^k + s^k - c$, $r^k := (r_b^k, r_c^k)$, $\bar{\mu}_k := \bar{\mu}(z^k)$ et

$$D_k := X_k^{1/2} S_k^{-1/2}.$$

On simplifiera $\bar{\mu}(z^k + \alpha d^k)$ en $\bar{\mu}_k(\alpha)$ sans que cela ne porte à confusion (la dimension de l'argument change).

La preuve de convergence vise essentiellement à montrer que le pas α_k pris à l'étape 2 de l'algorithme est borné inférieurement par une constante $\bar{\alpha} > 0$. Dans ce cas, grâce à (18.29), $\bar{\mu}_{k+1} \leqslant (1-\omega\bar{\alpha})\bar{\mu}_k$, ce qui montre la convergence q-linéaire de $\bar{\mu}_k$. De même, en utilisant (18.25), on a $\|r^{k+1}\|_2 \leqslant (1-\bar{\alpha})\|r^k\|_2$, si bien que la convergence q-linéaire des résidus sera aussi une conséquence immédiate de l'estimation $\alpha_k \geqslant \bar{\alpha}$.

La convergence de cet algorithme s'obtient sans hypothèse supplémentaire, mais on n'a pu démontrer sa polynomialité qu'en supposant z^0 de la forme

$$z^0 = (\zeta_0 e, 0, \zeta_0 e), \quad \text{avec un scalaire } \zeta_0 \geqslant \|(x^*, s^*)\|_\infty, \tag{18.30}$$

où z^* est une solution primale-duale arbitraire [629]. Évidemment, en pratique, on ne peut pas réaliser (18.30) car on ne connaît pas de solution primale-duale, mais on a observé que l'algorithme est plus rapide si l'itéré initial est bien centré ($X^0 s^0 = \zeta_0^2 e$ sous l'hypothèse (18.30)) avec un rapport $\|r^0\|_2/\bar{\mu}_0$ petit (il est de l'ordre $1/\zeta_0$ sous l'hypothèse (18.30)). En pratique, on prend donc $z^0 = (\zeta_0 e, 0, \zeta_0 e)$, avec ζ_0 « grand ».

La formule (18.26) montre que nous avons besoin d'une estimation de $|(dx^k)^\mathsf{T}(ds^k)|$. Celle-ci découlera des estimations de $D_k^{-1} dx^k$ et $D_k ds^k$ données dans le lemme très technique suivant.

Lemme 18.21 *Il existe une constante $C_1 > 0$ telle que pour tout $k \geqslant 0$:*

$$\|D_k^{-1} dx^k\|_2 \leqslant C_1 \bar{\mu}_k^{1/2} \quad et \quad \|D_k ds^k\|_2 \leqslant C_1 \bar{\mu}_k^{1/2}.$$

Si z^0 vérifie (18.30), alors $C_1 = \frac{9\rho'}{(1-\theta)^{1/2}} n$ est indépendante de ζ_0.

Démonstration. D'après (18.25) et par récurrence, on a

$$r^k = \beta_k r^0, \quad \text{où} \quad \beta_k := \prod_{i=0}^{k-1} (1 - \alpha_k). \tag{18.31}$$

Première étape : *il existe une constante C_1' $(= 4\rho' n/\zeta_0$ si (18.30) a lieu) telle que*

$$\beta_k \|(x^k, s^k)\|_1 \leqslant C_1' \bar{\mu}_k. \tag{18.32}$$

Soit z^* une solution du problème (P). On introduit

$$\bar{z} := \beta_k z^0 + (1 - \beta_k) z^* - z^k.$$

On voit facilement que $A\bar{x} = 0$ et $A^T \bar{y} + \bar{s} = 0$, si bien que $\bar{x}^T \bar{s} = 0$, ce que l'on exploite :

$$\begin{aligned} 0 &= \left(\beta_k x^0 + (1-\beta_k) x^* - x^k\right)^T \left(\beta_k s^0 + (1-\beta_k) s^* - s^k\right) \\ &= \beta_k^2 (x^0)^T s^0 + (1-\beta_k)^2 (x^*)^T s^* + (x^k)^T s^k \\ &\quad + \beta_k (1-\beta_k) \left[(x^0)^T s^* + (x^*)^T s^0\right] - \beta_k \left[(x^0)^T s^k + (x^k)^T s^0\right] \\ &\quad - (1-\beta_k) \left[(x^*)^T s^k + (x^k)^T s^*\right]. \end{aligned}$$

La majoration que l'on cherche vient de l'avant dernier terme que l'on fait passer dans le membre de gauche. D'autre part, le dernier terme est négatif ($(x^*, s^*, x^k, s^k) \geqslant 0$ et $\beta_k \leqslant 1$) et on le néglige, $(x^0)^T s^0 = n\bar{\mu}_0$, $(x^*)^T s^* = 0$ par optimalité, $(x^k)^T s^k = n\bar{\mu}_k$, $(x^0)^T s^* \leqslant \|x^0\|_\infty \|s^*\|_1$ et $(x^*)^T s^0 \leqslant \|s^0\|_\infty \|x^*\|_1$. Cela donne

$$\beta_k \left[(x^0)^T s^k + (x^k)^T s^0\right] \leqslant n\beta_k^2 \bar{\mu}_0 + n\bar{\mu}_k + \beta_k (1-\beta_k) \|(x^0, s^0)\|_\infty \|(x^*, s^*)\|_1.$$

Ensuite, le membre de gauche est minoré par $\beta_k \xi_0 \|(x^k, s^k)\|_1$, où

$$\xi_0 := \min_{1 \leqslant i \leqslant n} \min\{x_i^0, s_i^0\}.$$

On obtient aussi un facteur commun $\bar{\mu}_k$ à tous les termes à droite, car, d'après (18.31) et l'appartenance de $z^k \in V_\infty^-$, on a $\beta_k = \|r_k\|_2 / \|r^0\|_2 \leqslant \rho\bar{\mu}_k / \|r^0\|_2$. En utilisant également $0 \leqslant \beta_k \leqslant 1$, on obtient finalement l'inégalité

$$\beta_k \xi_0 \|(x^k, s^k)\|_1 \leqslant n \frac{\rho\bar{\mu}_0}{\|r^0\|_2} \bar{\mu}_k + n\bar{\mu}_k + \frac{\rho}{\|r^0\|_2} \bar{\mu}_k \|(x^0, s^0)\|_\infty \|(x^*, s^*)\|_1,$$

qui est bien de la forme (18.32).

Supposons à présent que z^0 vérifie (18.30). On note

$$\rho' := \frac{\rho\bar{\mu}_0}{\|r^0\|_2},$$

qui est $\geqslant 1$ et peut être considéré comme une constante ne dépendant pas de l'itéré initial. D'autre part, en utilisant (18.30) : $\xi_0 = \zeta_0$, $\bar\mu_0 = \zeta_0^2$, $\|(x^0, s^0)\|_\infty = \zeta_0$ et $\|(x^*, s^*)\|_1 \leqslant 2n\|(x^*, s^*)\|_\infty \leqslant 2n\zeta_0$. Dès lors, l'inégalité précédente devient

$$\beta_k\zeta_0\|(x^k, s^k)\|_1 \leqslant n\rho'\bar\mu_k + n\bar\mu_k + 2n\rho'\bar\mu_k \leqslant 4n\rho'\bar\mu_k.$$

Ce qui montre que l'on peut prendre $C_1 = 4\rho' n/\zeta_0$.

Deuxième étape : *démonstration du lemme*. Comme dans la première partie on construit deux vecteurs orthogonaux, ici formés à partir de dx^k et ds^k. Dans ce but, on introduit $\bar d = (\bar dx, \bar dy, \bar ds)$ défini par

$$\bar d := d^k + \beta_k(z^0 - z^*),$$

où z^* est une solution du problème (P). On vérifie facilement que $A\bar dx = 0$ et $A^{\mathsf T}\bar dy + \bar ds = 0$, si bien que $\bar dx^{\mathsf T}\bar ds = 0$, ce que l'on va exploiter.

Auparavant, observons que par la troisième équation du système de (18.34)

$$\begin{aligned} &S_k\left(dx^k + \beta_k(x^0 - x^*)\right) + X_k\left(ds^k + \beta_k(s^0 - s^*)\right)\\ &\quad = \quad -(X_kS_ke - \sigma\bar\mu_k e) + \beta_kS_k(x^0 - x^*) + \beta_kX_k(s^0 - s^*). \end{aligned}$$

Après multiplication par $(X_kS_k)^{-1/2}$, on a

$$\begin{aligned} &D_k^{-1}\left(dx^k + \beta_k(x^0 - x^*)\right) + D_k\left(ds^k + \beta_k(s^0 - s^*)\right)\\ &\quad = \quad -(X_kS_k)^{-1/2}(X_kS_ke - \sigma\bar\mu_k e) + \beta_kD_k^{-1}(x^0 - x^*) + \beta_kD_k(s^0 - s^*). \end{aligned}$$

On utilise maintenant le fait démontré ci-dessus, que les deux premiers termes sont orthogonaux. En prenant la norme au carré, on obtient

$$\begin{aligned} &\|D_k^{-1}(dx^k + \beta_k(x^0 - x^*))\|_2^2 + \|D_k(ds^k + \beta_k(s^0 - s^*))\|_2^2\\ &\quad \leqslant \quad \Big(\|(X_kS_k)^{-1/2}(X_kS_ke - \sigma\bar\mu_k e)\|_2 + \beta_k\|D_k^{-1}(x^0 - x^*)\|_2 + \beta_k\|D_k(s^0 - s^*)\|_2\Big)^2. \end{aligned}$$

On traite maintenant séparément les deux termes du membre de gauche en utilisant l'inégalité triangulaire :

$$\begin{aligned} &\|D_k^{-1}dx^k\|_2 \text{ et } \|D_kds^k\|_2\\ &\quad \leqslant \quad \|(X_kS_k)^{-1/2}(X_kS_ke - \sigma\bar\mu_k e)\|_2 + 2\beta_k\|D_k^{-1}(x^0 - x^*)\|_2 + 2\beta_k\|D_k(s^0 - s^*)\|_2. \end{aligned}$$

Il reste à montrer que chaque terme du membre de droite de cette inégalité est en $O(\bar\mu_k^{1/2})$.

Pour le premier terme, on a en utilisant l'appartenance de z^k à $V_\infty^-(\theta, \rho)$:

$$\|(X_kS_k)^{-1/2}\|_2 = \max_{1\leqslant i\leqslant n}\frac{1}{(x_i^ks_i^k)^{1/2}} \leqslant \frac{1}{(1-\theta)^{1/2}\bar\mu_k^{1/2}}$$

D'autre part, comme $\sigma < 1$ et $(x^k)^{\mathsf T}s^k = n\bar\mu_k$:

$$\|X_kS_ke - \sigma\bar\mu_k e\|_2^2 = \|X_kS_ke\|_2^2 - 2\sigma\bar\mu_k(x^k)^{\mathsf T}s^k + n\sigma^2\bar\mu_k^2 \leqslant \|X_kS_ke\|_1^2 = n^2\bar\mu_k^2.$$

Le premier terme est donc majoré par $(1-\theta)^{-1/2}n\bar\mu_k^{1/2}$.

Pour le second terme, on observe d'abord que

$$\|D_k^{-1}\|_2 = \max_{1\leqslant i\leqslant n} \frac{s_i^k}{(x_i^k s_i^k)^{1/2}} \leqslant \frac{\|s^k\|_\infty}{(1-\theta)^{1/2}\bar\mu_k^{1/2}} \leqslant \frac{\|s^k\|_1}{(1-\theta)^{1/2}\bar\mu_k^{1/2}}.$$

En utilisant alors le résultat (18.32) de la première partie de la démonstration, on a

$$\beta_k\|D_k^{-1}(x^0-x^*)\|_2 \leqslant \frac{\beta_k\|s^k\|_1}{(1-\theta)^{1/2}\bar\mu_k^{1/2}}\,\|x^0-x^*\|_2 \leqslant \frac{C_1'}{(1-\theta)^{1/2}}\,\bar\mu_k^{1/2}\|x^0-x^*\|_2,$$

qui est bien un majorant en $O(\bar\mu_k^{1/2})$. On s'y prend de la même manière pour le troisième terme, avec la majoration $\|D_k\|_2 \leqslant \|x^k\|_1\bar\mu_k^{1/2}/(1-\theta)^{1/2}$.

Enfin, si z^0 vérifie (18.30), on s'y prend différemment pour majorer les deux derniers termes. Dans ce cas, on a

$$0 \leqslant x^0 - x^* \leqslant \zeta_0 \quad \text{et} \quad 0 \leqslant s^0 - s^* \leqslant \zeta_0.$$

On en déduit, avec (18.32), que

$$\begin{aligned}
&\beta_k\|D_k^{-1}(x^0-x^*)\|_2 + \beta_k\|D_k(s^0-s^*)\|_2\\
&\quad\leqslant\ \beta_k\zeta_0\left(\|D_k^{-1}e\|_2 + \|D_k e\|_2\right)\\
&\quad=\ \beta_k\zeta_0\left(\|(X_kS_k)^{-1/2}s^k\|_2 + \|(X_kS_k)^{-1/2}x^k\|_2\right)\\
&\quad\leqslant\ \zeta_0\|(X_kS_k)^{-1/2}\|_2(\beta_k\|(x^k,s^k)\|_1)\\
&\quad\leqslant\ \frac{\zeta_0}{(1-\theta)^{1/2}\bar\mu_k^{1/2}}\,\frac{4\rho' n}{\zeta_0}\,\bar\mu_k\\
&\quad=\ \frac{4\rho'}{(1-\theta)^{1/2}}\,n\bar\mu_k^{1/2}.
\end{aligned}$$

Dès lors la constante C_1 peut être prise comme suit

$$\frac{1}{(1-\theta)^{1/2}}\,n + \frac{8\rho'}{(1-\theta)^{1/2}}\,n \leqslant \frac{9\rho'}{(1-\theta)^{1/2}}\,n =: C_1. \qquad \square$$

Lemme 18.22 *Il existe une constante $\bar\alpha \in\ ]0,1]$, telle que pour tout $\alpha \in [0,\bar\alpha]$, on a*

$$\begin{aligned}
(1-\alpha)\bar\mu_k \leqslant \bar\mu(z^k+\alpha d^k) \leqslant (1-\omega\alpha)\bar\mu_k\\
(x_i^k+\alpha dx_i^k)(s_i^k+\alpha ds_i^k) \geqslant (1-\theta)\bar\mu(z^k+\alpha d^k).
\end{aligned}$$

Si z^0 vérifie (18.30), alors il existe une constante $C_2 > 0$ indépendante de n telle que $\bar\alpha \geqslant C_2 n^{-2}$.

Démonstration. On laisse tomber l'indice k. Avec la constante C_1 donnée par le lemme 18.21 et l'inégalité de Cauchy-Schwarz, on a $dx^{\mathsf T}ds \geqslant -C_1^2\bar\mu$. Alors (18.26) conduit à

$$\bar\mu(z+\alpha d) \geqslant (1-\alpha)\bar\mu + \alpha\bar\mu\left(\sigma - \frac{\alpha C_1^2}{n}\right) \geqslant (1-\alpha)\bar\mu,$$

dès que

$$\alpha \leqslant \bar{\alpha}_a := \frac{n\sigma}{C_1^2}.$$

En majorant cette fois $dx^T dx \leqslant C_1^2 \bar{\mu}$, on obtient de (18.26)

$$\begin{aligned} \bar{\mu}(z(\alpha)) &\leqslant \bar{\mu} - \alpha\bar{\mu} + \alpha\sigma\bar{\mu} + \frac{\alpha^2 C_1^2}{n}\bar{\mu} \\ &\leqslant \bar{\mu} - \omega\alpha\bar{\mu}, \end{aligned} \tag{18.33}$$

dès que

$$\alpha \leqslant \bar{\alpha}_b := \frac{n(1-\omega-\sigma)}{C_1^2}.$$

En utilisant le fait que $x_i s_i \geqslant (1-\theta)\bar{\mu}$ lorsque $z \in V_\infty^-(\theta,\rho)$, on trouve

$$\begin{aligned} (x_i + \alpha dx_i)(s_i + \alpha ds_i) &= x_i s_i + \alpha(\underbrace{x_i ds_i + s_i dx_i}_{\sigma\bar{\mu} - x_i s_i}) + \alpha^2 dx_i ds_i \\ &\geqslant (1-\alpha)(1-\theta)\bar{\mu} + \alpha\sigma\bar{\mu} + \alpha^2 dx_i ds_i, \quad [\text{si } \alpha \leqslant 1] \\ &\geqslant ((1-\alpha)(1-\theta) + \alpha\sigma - \alpha^2 C_1^2)\bar{\mu} \\ &\geqslant (1-\theta)\left(1 - \alpha + \alpha\sigma + \frac{1}{n}\alpha^2 C_1^2\right)\bar{\mu} \\ &\geqslant (1-\theta)\bar{\mu}(z(\alpha)), \quad [\text{par (18.33)}], \end{aligned}$$

où l'avant-dernière inégalité est vraie dès que

$$\alpha \leqslant \bar{\alpha}_c := \min\left(1, \frac{\theta\sigma}{C_1^2(1 + \frac{1-\theta}{n})}\right).$$

Il suffit alors de prendre $\bar{\alpha} := \min(\bar{\alpha}_a, \bar{\alpha}_b, \bar{\alpha}_c) \leqslant 1$. On observe aussi que si z^0 vérifie (18.30), $C_1 = O(n)$ par le lemme 18.21. On en déduit alors que $\bar{\alpha} \geqslant C_2/n^2$. □

Théorème 18.23 (convergence de l'algorithme 18.20) *L'algorithme 18.20 converge : $\bar{\mu} \to 0$ et $(r_b^k, r_c^k) \to 0$ q-linéairement. De plus, si z^0 vérifie (18.30), alors, pour tout $\varepsilon > 0$, il existe un indice $K = O(n^2 \log \varepsilon^{-1})$ tel que $\bar{\mu}_k \leqslant \varepsilon\bar{\mu}_0$, dès que $k \geqslant K$.*

Démonstration. Comme l'algorithme choisit le plus grand pas tel que l'on ait (18.29), α_k est certainement supérieur à $\bar{\alpha} > 0$ donné par le lemme 18.22. Alors l'inégalité (18.29) montre que l'on a

$$\bar{\mu}_{k+1} \leqslant (1 - \omega\bar{\alpha})\bar{\mu}_k.$$

Ce qui implique la convergence q-linéaire de $\bar{\mu}_k$ vers zéro. Le même raisonnement et

$$\|(r_b^{k+1}, r_c^{k+1})\|_2 = (1-\alpha_k)\|(r_b^k, r_c^k)\|_2 \leqslant (1-\bar{\alpha})\|(r_b^k, r_c^k)\|_2$$

montre que (r_b^k, r_c^k) converge vers zéro q-linéairement.

Enfin, si z^0 vérifie (18.30), $\bar{\alpha} \geqslant C_2/n^2$ pour une constante $C_2 > 0$ indépendante de n, si bien que

$$\bar{\mu}_{k+1} \leqslant \left(1 - \frac{\omega C_2}{n^2}\right) \bar{\mu}_k.$$

On conclut en utilisant le lemme de complexité 18.5. □

18.5 Mise en œuvre

18.5.1 Calcul du déplacement de Newton

Les algorithmes primaux-duaux calculent l'itéré suivant en se déplaçant le long de la direction de Newton obtenue par linéarisation du système d'optimalité perturbé (18.3). Cette direction est solution de l'*équation de Newton*

$$\begin{pmatrix} 0 & A^T & I \\ A & 0 & 0 \\ S & 0 & X \end{pmatrix} \begin{pmatrix} dx \\ dy \\ ds \end{pmatrix} = \begin{pmatrix} -r_c \\ -r_b \\ \mu e - Xs \end{pmatrix}. \tag{18.34}$$

On a noté les *résidus*

$$r_c \equiv r_c(z) := A^T y + s - c \quad \text{et} \quad r_b \equiv r_b(z) := Ax - b.$$

Intéressons-nous aux méthodes de calcul d'une solution du système de Newton (18.34). On peut éliminer ds grâce à la dernière équation, qui donne

$$ds = -X^{-1}Sdx + \mu X^{-1}e - s.$$

On obtient alors ce que l'on appelle le *système linéaire augmenté*

$$\begin{pmatrix} -X^{-1}S & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} dx \\ dy \end{pmatrix} = \begin{pmatrix} -r_c - \mu X^{-1}e + s \\ -r_b \end{pmatrix}.$$

La matrice de ce système est symétrique, mais indéfinie, et garde le caractère creux éventuel de A. On peut le résoudre par la *factorisation de Bunch et Parlett* ou de *Bunch et Kaufman*.

On peut poursuivre la réduction de la dimension du système linéaire à résoudre en éliminant dx grâce à la première équation

$$dx = XS^{-1}A^T dy + \mu S^{-1}e - x + XS^{-1}r_c.$$

On obtient ce que l'on appelle le *système normal* ou l'*équation normale*

$$(AXS^{-1}A^T)dy = -r_b - A(\mu S^{-1}e - x + XS^{-1}r_c).$$

Si A est surjective, la matrice de ce système est symétrique définie positive. En pratique, c'est souvent ce système qui est résolu, par *factorisation de Cholesky creuse*.

Le mauvais conditionnement dû au facteur XS^{-1} requiert toutefois une factorisation adaptée [630]. Par ailleurs, si A a une colonne dense, la matrice $AXS^{-1}A^{\mathsf{T}}$ perd son caractère creux. Si celles-ci ne sont pas trop nombreuses, on peut faire face à la difficulté de la manière suivante. Si C (resp. D) désigne l'ensemble des indices des colonnes creuses (resp. denses) de A, on écrit

$$AXS^{-1}A^{\mathsf{T}} = A_C X_C S_C^{-1} A_C^{\mathsf{T}} + A_D X_D S_D^{-1} A_C^{\mathsf{T}}.$$

On calcule les facteurs de Cholesky creux de $A_C X_C S_C^{-1} A_C^{\mathsf{T}}$ et on prend en compte le second terme $A_D X_D S_D^{-1} A_C^{\mathsf{T}}$ par la technique de Sherman-Morrison-Woodbury.

18.5.2 Logiciels

La difficulté principale des méthodes de points intérieurs est due à la résolution des systèmes linéaires mal conditionnés que l'on y rencontre (section 18.5.1). Si l'on dispose de solveurs linéaires adéquats, les algorithmes peuvent s'écrire facilement et être ainsi adaptés à des situations particulières.

Logiciels généralistes : CPLEX [536], HOPDM [268], LIPSOL (en Matlab avec solveurs linéaires en Fortran) [642], LOQO, MOSEK [12, 537], OSX, PCx (en C) [150] et SEDUMI [575, 538].

Notes

L'algorithme des petits déplacements de la section 18.3.2 a été mis au point par Kojima, Mizuno et Yoshise [359 ; 1989] et par Monteiro et Adler [427 ; 1989] ; celui des grands déplacements de la section 18.3.3 est dû à Kojima, Mizuno et Yoshise [358 ; 1989]. L'algorithme prédicteur-correcteur de la section 18.3.4 fut énoncé et analysé par Mizuno, Todd et Ye [425 ; 1993], mais son principe est fondé sur les travaux antérieurs de Monteiro et Adler [427 ; 1989] et de Sonnevend, Stoer et Zhao [560, 561 ; 1989-1991]. L'algorithme prédicteur-correcteur de Mehrotra [419 ; 1992], que nous n'avons pas présenté, en est une version plus élaborée. Il suit le chemin central avec plus de précision en en construisant une approximation d'ordre plus élevé ; c'est important pour ne pas perdre du temps dans le suivi des coudes que peut former le chemin central. Il permet aussi d'adapter à l'itération courante le facteur σ de réduction de $\bar{\mu}$, alors qu'il est donné *a priori* dans l'algorithme 18.16. C'est cette méthode qui est la plus souvent implémentée dans les codes de points intérieurs pour l'optimisation linéaire.

Le lecteur trouvera des compléments à cette brève introduction dans la monographie de S. Wright [629], qui est une très bonne référence sur les algorithmes de points intérieurs primaux-duaux en optimisation linéaire ; nous nous en sommes souvent inspiré dans la partie algorithmique de ce chapitre. D'autres aspects de l'approche par points intérieurs en optimisation linéaire sont exposés par den Hertog [165], Saigal [534] (algorithmes affines), Terlaky [580], Jansen [334], Roos, Terlaky et Vial [529], Vanderbei [604] et Ye [633].

Les méthodes de points intérieurs sont utilisées pour résoudre d'autres classes de problèmes convexes, ayant un côté combinatoire provenant de relations de complémentarité. On pourra consulter [73] pour les problèmes de complémentarité linéaire

monotone (incluant l'optimisation quadratique convexe), [626] pour l'optimisation semi-définie positive (voir aussi le chapitre 20), [512] pour l'optimisation conique (qui permet une extension à la dimension infinie), [456] pour l'optimisation convexe (voir aussi la section 16.1). Ces sujets sont aussi abordés par Ben-Tal et Nemirovski [44], qui présentent de nombreux exemples et applications.

L'utilisation des points intérieurs pour résoudre les problèmes d'optimisation non linéaire remonte à l'ouvrage pionnier de Fiacco et McCormick [207 ; 1968]. Il faut également citer l'article parfois oublié de McLinden [416 ; 1980], qui contient des résultats qualitatifs sur les problèmes convexes. En optimisation non linéaire, les techniques sont moins bien maîtrisées, les algorithmes ne sont pas encore stabilisés et l'intérêt de l'approche est débattu [241]. On trouvera dans [101] une étude de convergence d'un algorithme se ramenant à une suite de problèmes-barrières sous contraintes d'égalité non linéaires globalisés par régions de confiance (voir aussi la section 16.2) ; les codes KNITRO [459] et OPINL [225] s'en inspirent chacun à leur manière. Un état de l'art est présenté dans [217].

Exercices

18.1. On suppose que c, A et b sont donnés comme en optimisation linéaire (avec A surjective) et que (18.2) a lieu. Soit $\mu > 0$. Démontrez les affirmations suivantes.

(*i*) Le problème $(D_\mu) \equiv$ (18.5) a une solution et une seule.

(*ii*) Si l'on note (x_μ, y_μ, s_μ) le point central correspondant à μ, c'est-à-dire la solution unique de (18.3), la solution de (D_μ) n'est autre que (y_μ, s_μ) et x_μ est le multiplicateur optimal associé à la contrainte de (D_μ).

18.2. On suppose que c, A et b sont donnés comme en optimisation linéaire et que (18.2) a lieu. Soit $w \in \mathbb{R}^n_{++}$. Montrez qu'il existe un unique triplet (x, y, s), unique si A est surjective, vérifiant

$$\begin{cases} A^\mathsf{T} y + s = c, \quad s > 0 \\ Ax = b, \quad x > 0 \\ Xs = w. \end{cases}$$

En conséquence, l'application p définie en (18.15) est surjective (et bijective si A est surjective).

18.3. *Problèmes singuliers.* Supposons que (P) ait une solution $\bar{x}$ telle que $\bar{x} > 0$. Alors

(*i*) $c \perp \mathcal{N}(A)$,

(*ii*) tout point admissible de (P) est solution primale,

(*iii*) si (18.2) a lieu et $\mu \mapsto (x_\mu, s_\mu, s_\mu)$ est le chemin central, alors $x_\mu = \check{x}$ pour tout $\mu > 0$, où $\check{x}$ est le centre analytique de $\mathcal{F}_\text{P}$.

Supposons maintenant que (D) ait une solution $(\bar{y}, \bar{s})$ telle que $\bar{s} > 0$. Alors

(*iv*) $b = 0$,

(*v*) tout point admissible de (D) est solution duale,

(*vi*) si (18.2) a lieu, si A est surjective et si $\mu \mapsto (x_\mu, y_\mu, s_\mu)$ est le chemin central, alors $(y_\mu, s_\mu) = (\check{y}, \check{s})$ pour tout $\mu > 0$, où $(\check{y}, \check{s})$ est le centre analytique de $\mathcal{F}_\text{D}$.

18.4. *Le centre analytique n'est pas géométrique.* Montrez que le centre analytique de l'ensemble $\{x \in \mathbb{R}^2 : x \geqslant 0,\ x_1 + x_2 \leqslant 1\}$ est $(\frac{1}{3}, \frac{1}{3})$, alors que si le même ensemble est décrit par $\{x \in \mathbb{R}^2 : x \geqslant 0,\ x_1 + x_2 \leqslant 1,\ x_1 \leqslant 1\}$, on trouve $(\frac{1}{4}, \frac{3}{8})$.

18.5. *Autre définition du chemin central.* On considère le problème d'optimisation linéaire (P) et on suppose que (18.2) a lieu. Montrez que pour $\alpha > \text{val}(P)$, $X_\alpha := \{x \in \mathcal{F}_\text{P} :$

$c^\mathsf{T} x \leqslant \alpha\}$ est non vide, qu'il a un centre analytique et que celui-ci est situé sur le chemin central de (P).

18.6. *Tangente au chemin central.* Soit $z(\mu) = (x(\mu), y(\mu), s(\mu)) = (x, y, s)$ un point central, vérifiant donc $A^\mathsf{T} y + s = c$, $Ax = b$ et $Xs = \mu e$ pour un certain $\mu > 0$. On suppose que A est surjective. Soit $d = (dx, dy, ds)$ la dérivée de $\mu \mapsto (x(\mu), y(\mu), s(\mu))$ en μ. Montrez que l'on a

$$\begin{aligned} c^\mathsf{T} dx &= \frac{1}{\mu} c^\mathsf{T} Z^- (Z^{-\mathsf{T}} X^{-1} S Z^-)^{-1} Z^{-\mathsf{T}} c, \\ b^\mathsf{T} dy &= -\frac{1}{\mu} b^\mathsf{T} (AXS^{-1} A^\mathsf{T})^{-1} b, \\ s^\mathsf{T} dx + x^\mathsf{T} ds &= n, \end{aligned}$$

où Z^- est une matrice dont les colonnes forment une base de $\mathcal{N}(A)$.

Remarque. On voit que le critère primal $\mu \mapsto c^\mathsf{T} x(\mu)$ croît strictement (si $Z^{-\mathsf{T}} c \neq 0$ c'est-à-dire si $c \notin \mathcal{R}(A^\mathsf{T})$), que le critère dual $\mu \mapsto b^\mathsf{T} y(\mu)$ décroît strictement (si $b \neq 0$) et que le saut de dualité $\mu \mapsto x(\mu)^\mathsf{T} s(\mu)$ croît strictement. On pourra rapprocher cela du résultat de la proposition 18.4 (ii).

18.7. *Voisinages du chemin central.* On suppose que c, A et b sont donnés comme en optimisation linéaire (avec A surjective) et que (18.2) a lieu. On considère les voisinages $V_p(\theta)$ et $V_\infty^-(\theta)$ définis par (18.16) et (18.17), avec $p \in [1, \infty]$ et $\theta \in [0, 1[$.

1) On dit qu'un espace topologique E est *connexe par arcs* si deux quelconques de ses points peuvent être joints par un *chemin continu*, c'est-à-dire une application continue $\gamma : [0, 1] \mapsto E$ telle que $\gamma(0)$ est le premier point et $\gamma(1)$ le second. Montrez que $V_p(\theta)$ et $V_\infty^-(\theta)$ sont connexes par arcs si ces ensembles sont munis de la topologie induite de celle de $\mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n$.
2) Montrez que tout point (x, y, s) de $V_p(\theta)$ et de $V_\infty^-(\theta)$ vérifie $x > 0$ et $s > 0$, ce qui peut s'écrire $V_p(\theta) \subseteq \mathcal{F}^s$ et $V_\infty^-(\theta) \subseteq \mathcal{F}^s$.
3) Démontrez les affirmations suivantes.
   (a) Si $1 \leqslant p < +\infty$ et $n \geqslant 2$, alors $\bigcup_{0 \leqslant \theta < 1} V_p(\theta) \neq \mathcal{F}^s$.
   (b) Si $n \geqslant 3$, alors $\bigcup_{0 \leqslant \theta < 1} V_\infty(\theta) \neq \mathcal{F}^s$.
   (c) Par contre, $\bigcup_{0 \leqslant \theta < 1} V_\infty^-(\theta) = \mathcal{F}^s$.

19 Systèmes non déterminés

On rencontre souvent des problèmes de moindres-carrés. Mathématiquement, ces problèmes consistent à minimiser le carré de la norme euclidienne d'une fonction à valeurs vectorielles, qui peut être linéaire (on parle alors de moindres-carrés linéaire, voir la section 19.1) ou non linéaire (on parle alors de moindres-carrés non linéaire, voir la section 19.3). Ce chapitre leur est consacré. Étant donné la grande variété de problèmes qui peuvent être modélisés comme problème de moindres-carrés, on les rencontre sous des appellations différentes : *estimation de paramètres*, problème *d'identification*, de *calibration* ou de *régression* en statistiques [199, 333].

19.1 Moindres-carrés linéaire

> *Presque chaque soir, je fais une nouvelle édition du tableau, qu'il est facile d'améliorer n'importe où. Contre la monotonie du travail d'arpentage, c'est toujours une plaisante distraction ; on peut aussi voir immédiatement si quelque chose de douteux s'est glissé, ce qui reste à obtenir, etc. Je vous recommande cette méthode comme modèle. Il ne vous arrivera presque plus jamais de pratiquer une élimination directe, du moins quand vous avez plus de deux inconnues. La procédure indirecte peut se faire à moitié endormi, ou en pensant à autre chose.*
>
> C.F. Gauss, extrait d'une lettre à G.L. Gerling, datée du 26 décembre 1823, dans laquelle il loue les mérites de sa méthode de relaxation, par rapport aux méthodes directes, pour résoudre l'équation normale d'un problème de moindres-carrés linéaire. Traduction de A. Michel-Pajus [112].

19.1.1 Définition du problème

Un *problème de moindres-carrés linéaire* (MCL) est un problème d'optimisation qui s'écrit de la manière suivante :

$$\min_{x\in\mathbb{R}^n} \left(f(x) = \frac{1}{2}\|Ax-b\|_2^2 \right), \tag{19.1}$$

où $\|\cdot\|_2$ est la norme ℓ_2, A est une matrice de type $m \times n$ et $b \in \mathbb{R}^m$. Lorsque $m = n$, on parle de problème *d'interpolation linéaire*. Ce problème peut se voir de différentes manières.

Le premier point de vue est « concret ». On cherche à déterminer des *paramètres* $x \in \mathbb{R}^n$ d'un « système » au moyen de *mesures* $b \in \mathbb{R}^m$ réalisées sur celui-ci. La loi qui relie les paramètres x aux *quantités mesurées* Ax est supposée linéaire. Le problème de moindres-carrés linéaire permet de déterminer les paramètres x qui donnent des quantités mesurées Ax au plus proche des mesures b. De ce point de vue, le problème consiste à projeter b sur l'image de A au moyen du produit scalaire euclidien (voir la section 2.5.1, on pourrait prendre d'autres produits scalaires d'ailleurs).

Le second point de vue est « abstrait ». On s'intéresse à la résolution du système linéaire $Ax = b$. On n'impose pas que $b \in \mathcal{R}(A)$, si bien que ce système n'a peut-être pas de solution. Le problème de moindres-carrés linéaire cherche alors à résoudre ce système linéaire « au mieux », en minimisant le résidu $Ax - b$.

Il s'agit donc d'un problème « fondamental », auquel sont rattachés divers concepts bien connus en algèbre linéaire. On note

$$r = \operatorname{rg} A$$

le rang de A.

19.1.2 Exemple : régression linéaire

En statistiques, un problème de *régression linéaire* consiste à déterminer au mieux la dépendance *affine* supposée entre m mesures et $n-1$ variables. En général $m \gg n$. Pour prendre les notations de la section 19.1.1, la i-ième mesure b_i résulte des valeurs $(a_{i,1}, \dots, a_{i,n-1})$ données au vecteur dont elle dépend. Si la mesure dépend affinement du vecteur, il doit y avoir un vecteur de paramètres $x \in \mathbb{R}^n$, inconnu mais indépendant des mesures prises, tel que le i-ième résidu

$$r_i = b_i - \left(\sum_{j=1}^{n-1} a_{i,j} x_j + x_n \right)$$

soit nul (x_n est le coefficient indépendant des mesures rendant la dépendance *affine* plutôt que *linéaire*). Si l'on note

$$b = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} \in \mathbb{R}^m \qquad \text{et} \qquad A = \begin{pmatrix} a_{1,1} & \cdots & a_{1,n-1} & 1 \\ \vdots & & \vdots & \vdots \\ a_{m,1} & \cdots & a_{1,n-1} & 1 \end{pmatrix} \in \mathbb{R}^{m \times n},$$

le résidu

$$r = b - Ax \in \mathbb{R}^m$$

doit être nul. En présence d'erreurs de mesure ou d'une relation mesures-données non linéaire, on peut chercher à annuler r au mieux, ce qui conduit à résoudre le problème de moindres-carrés linéaire (19.1).

Dans le cas où $n = 2$, les couples $\{(b_i, a_{i,1})\}_{i=1}^m$ représentent un nuage de m points dans le plan. Pour la solution $x \in \mathbb{R}^2$ du problème de moindres-carrés, l'application affine $a \in \mathbb{R} \mapsto ax_1 + x_2$ donne la relation affine « la plus proche » des couples $\{(b_i, a_{i,1})\}_{i=1}^m$. La figure 19.1 reprise de [619 ; 2013, figure 2.2] donne ainsi la droite passant au plus près, au sens des moindres-carrés, des mesures de l'*indice des problèmes sanitaires et sociaux* en fonction des *inégalités de revenus* ; il y a une mesure (un point) pour chacun des 21 pays considérés dans l'étude.

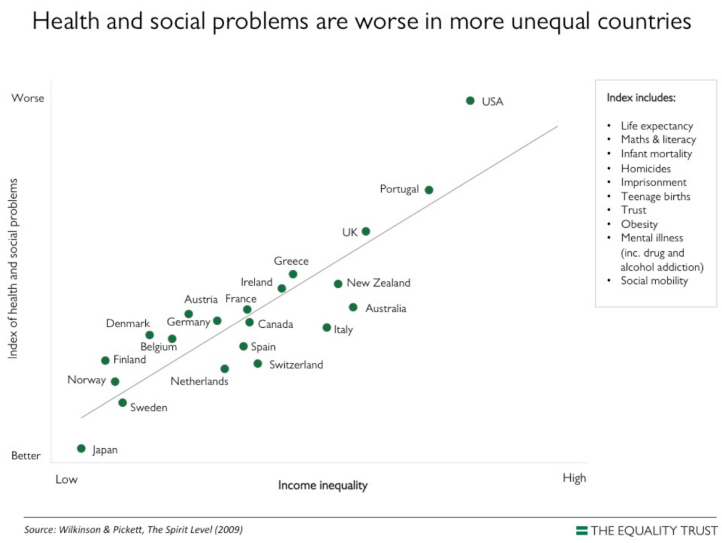


Fig. 19.1. Dépendance affine « la plus proche » de 21 couples (indice des problèmes sanitaires et sociaux, inégalités de revenus), avec un couple par pays considéré.

19.1.3 L'ensemble des solutions

La condition d'optimalité du premier ordre de ce problème s'écrit $\nabla f(x) = 0$ ou encore

$$A^\mathsf{T} A x = A^\mathsf{T} b. \tag{19.2}$$

Celle-ci porte le nom d'*équation normale* de (19.1).

La proposition suivante règle la question de l'existence et de l'unicité des solutions de (19.1).

Proposition 19.1 (ensemble des solutions) *Le problème* (19.1) *est convexe et admet toujours une solution. Celle-ci est unique si, et seulement si, A est injective. L'ensemble des solutions de* (19.1) *s'écrit $x_p + \mathcal{N}(A)$, où x_p est une solution particulière de* (19.1) *et la valeur optimale est $\frac{1}{2}\|Pb\|_2^2$, où P est le projecteur orthogonal (pour le produit scalaire euclidien) sur $\mathcal{R}(A)^\perp$.*

Démonstration. Le problème (19.1) consiste à projeter b sur l'image de A (voir la section 2.5.1), qui est un convexe fermé non vide. Il existe donc un unique élément $y \in \mathcal{R}(A)$ qui est le plus proche de b (proposition 2.25). Cet élément est de la forme $y = Ax$, où x est une solution de (19.1).

Si A est injective f est strictement convexe ($A^\mathsf{T}A$ est définie positive) et donc (19.1) a une solution unique. Si A n'est pas injective et x est solution, tous les points de $x + \mathcal{N}(A)$ ($\neq \{x\}$) sont aussi solutions ; par ailleurs, si x et x' sont deux solutions de (19.1), elles vérifient l'équation normale et donc $x - x' \in \mathcal{N}(A^\mathsf{T}A) = \mathcal{N}(A)$.

Enfin, $Q = I - P$ étant le projecteur orthogonal sur $\mathcal{R}(A)$, la valeur optimale s'écrit $\frac{1}{2}\|Qb - b\|_2^2 = \frac{1}{2}\|Pb\|_2^2$. □

Il existe de nombreuses démonstrations de l'existence d'une solution de (19.1). Celle donnée ci-dessus considère directement le problème d'optimisation. On peut aussi s'intéresser à son système d'optimalité (19.2) (du fait de la convexité du critère — sa hessienne $A^\mathsf{T}A$ est semi-défini positif — il y a équivalence entre les solutions (19.1) et celles de son système d'optimalité ; théorème 4.9). Pour montrer que ce dernier a toujours une solution, il suffit d'observer que $A^\mathsf{T}b \in \mathcal{R}(A^\mathsf{T}A)$.

L'ensemble des solutions de (19.1) sont donc les solutions de l'équation normale (19.2). On peut s'intéresser à la *solution de norme minimale*, qui est donc définie par

$$\begin{cases} \min \frac{1}{2}\|x\|_2^2 \\ A^\mathsf{T}Ax = A^\mathsf{T}b. \end{cases} \tag{19.3}$$

On peut bien parler de « la » solution de norme minimale, car le critère de ce problème étant strictement convexe, il y a exactement une unique solution de norme minimale. On peut caractériser la solution de ce problème. Comme celui-ci est convexe, ses conditions d'optimalité du premier ordre sont nécessaires et suffisantes (on notera en effet que les contraintes sont qualifiées, car affines). Donc $\hat{x}$ est la solution de norme minimale si, et seulement si, il existe un multiplicateur $\lambda \in \mathbb{R}^n$ (non nécessairement unique) tel que

$$\begin{cases} \hat{x} + A^\mathsf{T}A\lambda = 0 \\ A^\mathsf{T}A\hat{x} = A^\mathsf{T}b. \end{cases} \tag{19.4}$$

La première condition s'écrit aussi $\hat{x} \in \mathcal{R}(A^\mathsf{T}A) = \mathcal{R}(A^\mathsf{T})$. Dès lors, $\hat{x}$ est solution de norme minimale de (19.1), c'est-à-dire solution de (19.3), si, et seulement si,

$$\begin{cases} \hat{x} \in \mathcal{R}(A^\mathsf{T}) \\ A^\mathsf{T}A\hat{x} = A^\mathsf{T}b. \end{cases} \tag{19.5}$$

Comme $\hat{x}$ est univoquement déterminé par le système linéaire d'optimalité (19.4) (il peut cependant avoir de multiples solutions en λ si A n'est pas injective), qui se récrit

$$\begin{pmatrix} I & A^\mathsf{T}A \\ A^\mathsf{T}A & 0 \end{pmatrix} \begin{pmatrix} \hat{x} \\ \lambda \end{pmatrix} = \begin{pmatrix} 0 \\ A^\mathsf{T}b \end{pmatrix},$$

l'application qui à $b \in \mathbb{R}^m$ fait correspondre $\hat{x}$ est linéaire et la matrice qui la représente est appelée le pseudo-inverse (de Moore-Penrose) de A. On note cette matrice $A^\dagger$ et la solution de norme minimale s'écrit

$$\hat{x} = A^\dagger b.$$

On peut voir $\hat{x}$ comme le résultat d'une tentative de trouver une solution au système $Ax = b$, qui n'en n'a pas nécessairement (on se rappelle, en particulier, que A n'est pas carrée), au moyen de deux opérations :

- la première force l'*existence* d'une solution en relaxant « $Ax = b$ » en un problème de minimisation, celui que l'on trouve dans (19.1),
- la seconde force l'*unicité* de la solution par le problème d'optimisation (19.3).

19.1.4 Résolution numérique

On distingue les algorithmes qui utilisent l'équation normale (19.2) (résolution par factorisation de Cholesky ou par gradient conjugué) et ceux qui s'attaquent directement à la formulation originale (19.1) (résolution par factorisation QR ou par factorisation en valeurs singulières).

Résolution de l'équation normale par factorisation

L'approche la plus simple est de faire la *factorisation de Cholesky* de $A^\mathsf{T}A$ et de calculer la solution de l'équation normale en résolvant les deux systèmes triangulaires qui en découlent. Cette approche n'est pas sans inconvénients. D'une part elle oblige de former la matrice $A^\mathsf{T}A$, ce qui demande $O(mn^2)$ opérations (ce n'est souvent pas négligeable). Ensuite, on perd aussi en précision, du fait de l'annulation de l'influence des petits éléments de A (en particulier dans les termes diagonaux $(A^\mathsf{T}A)_{ii} = \sum_j A_{ji}^2$). Enfin, cette approche peut éventuellement détruire la creusité éventuelle de A (si A a une ligne pleine, $A^\mathsf{T}A$ sera en général une matrice pleine).

Une autre possibilité, bien adaptée aux matrices A creuses est de résoudre ce que l'on appelle le *système linéaire augmenté*, équivalent à l'équation normale, que l'on obtient à partir de celle-ci en posant $y = -Ax$:

$$\begin{pmatrix} I & A \\ A^\mathsf{T} & 0 \end{pmatrix} \begin{pmatrix} y \\ x \end{pmatrix} = \begin{pmatrix} 0 \\ -A^\mathsf{T}b \end{pmatrix}.$$

La matrice K de ce système linéaire est symétrique, mais n'est pas définie positive. D'autre part, l'ordre $n+m$ de K peut être beaucoup plus important que l'ordre n de l'équation normale ; mais si A est creuse, K l'est aussi. On peut alors la factoriser par des méthodes pouvant prendre en compte son caractère creux (comme les solveurs MA27/MA47 de Duff et Reid [181, 182, 183]). On se rappellera que K n'est en général pas définie positive (si A est injective, K a n valeurs propres strictement négatives et m valeurs propres strictement positives), si bien qu'une factorisation de Cholesky ne convient pas.

Résolution de l'équation normale par gradient conjugué

Dans cette approche, on minimise f par l'algorithme du gradient conjugué (GC), ce qui revient à résoudre l'équation normale par le même algorithme. Théoriquement, cet algorithme n'est bien défini que lorsque la matrice du système à résoudre, ici $A^\mathsf{T}A$, est définie positive. Toutefois, grâce à la structure de l'équation normale, le GC peut être utilisé pour résoudre cette équation. C'est ce qu'affirme la proposition suivante. On rappelle que $r = \operatorname{rg} A$.

Proposition 19.2 *L'algorithme du gradient conjugué pour minimiser* (19.1) *est bien défini et converge en au plus r itérations. De plus, si l'itéré initial est pris dans $\mathcal{R}(A^\mathsf{T})$ (par exemple $x_1 = 0$), les itérés convergent vers la solution de norme minimale de* (19.1).

Démonstration. C'est une application directe du point 2 de la proposition 8.11. Le système linéaire à résoudre est l'équation normale (19.2) : on a bien $A^{\mathsf{T}}A \succcurlyeq 0$, $A^{\mathsf{T}}b \in \mathcal{R}(A^{\mathsf{T}}) = \mathcal{R}(A^{\mathsf{T}}A)$ et $\dim \mathcal{R}(A^{\mathsf{T}}A) = \dim \mathcal{R}(A^{\mathsf{T}}) = \dim \mathcal{R}(A) = r$. □

Le principal avantage du GC est d'être utilisable pour les grands problèmes. Cependant, cette approche est sensible aux erreurs d'arrondi et présente des problèmes de stabilité numérique (difficultés si A a des valeurs singulières nulles ou presque nulles). En particulier, *il ne faut pas implémenter l'algorithme du GC standard directement sur l'équation normale*. L'algorithme recommandé dans [63], appelé CGLS ci-dessous, est celui déjà proposé par Hestenes et Steifel [312, 571]. Ces deux algorithmes sont équivalents en arithmétique exacte, mais c'est ce dernier qui a les meilleures propriétés de stabilité numérique.

L'algorithme CGLS génère des itérés x_k, en faisant un pas $\alpha_k > 0$ le long d'une direction d_k : $x_{k+1} = x_k + \alpha_k d_k$. C'est ici le résidu $r_k := b - Ax_k$, dont on cherche à minimiser la norme, qui est mis à jour récursivement (par $r_{k+1} = x_k - \alpha_k p_k$, où $p_k := Ad_k$) et non pas le résidu de l'équation normale, lequel est calculé par $s_k := A^{\mathsf{T}}r_k$. C'est essentiellement par cette mise-à-jour du résidu que CGLS diffère d'une application directe du GC standard.

Algorithme 19.3 (CGLS)

1. On se donne $x_1 \in \mathbb{R}^n$;
2. On calcule le résidu $r_1 = b - Ax_1$, l'opposé du gradient $s_1 = A^{\mathsf{T}}r_1$ et sa norme au carré $\gamma_1 = \|s_1\|_2^2$;
3. Pour $k = 1, 2, \ldots$ faire :
   - 3.1. Si $\gamma_k \simeq 0$, on s'arrête ;
   - 3.2. *Paramètre de conjugaison* : si $k \geqslant 2$, $\beta_k = \gamma_k/\gamma_{k-1}$;
   - 3.3. *Déplacement en x* :

$$d_k = \begin{cases} s_k & \text{si } k = 1 \\ s_k + \beta_k d_{k-1} & \text{si } k \geqslant 2 \,; \end{cases}$$

   - 3.4. *Déplacement en r* : $p_k = Ad_k$;
   - 3.5. *Calcul du pas* : $\alpha_k = \gamma_k/\|p_k\|_2^2$;
   - 3.6. *Nouveau point* : $x_{k+1} = x_k + \alpha_k d_k$;
   - 3.7. *Nouveau résidu* : $r_{k+1} = r_k - \alpha_k p_k$;
   - 3.8. *Nouveau gradient* : $s_{k+1} = A^{\mathsf{T}}r_{k+1}$; $\gamma_{k+1} = \|s_{k+1}\|_2^2$;

On peut encore noter que l'algorithme CGLS évite de faire les produits $A^{\mathsf{T}}(Ad)$ qui, en calcul flottant, peuvent détériorer la performance lorsque le système est mal conditionné ; voir [469 ; section 7.1] et [63 ; section 4.2]. Un autre algorithme à l'efficacité semblable est LSQR de Paige et Saunders [469]. C'est une autre version stable du GC, fondée sur la bidiagonalisation de Lanczos et la factorisation QR.

Résolution par factorisation QR de A

Soit

$$A = QR$$

une factorisation QR de A, où Q est une matrice orthogonale et R est de la forme

$$R = \begin{pmatrix} R_1 \\ 0_{(m-r)\times n} \end{pmatrix}.$$

On a noté $0_{(m-r)\times n}$ la matrice nulle de type $(m-r)\times n$. On sait que R_1 est triangulaire supérieure ($(R_1)_{ij} = 0$ si $i > j$) et que ses éléments diagonaux $(R_1)_{ii}$ $(1 \leqslant i \leqslant r)$ sont non nuls.

On a $\|Q^\mathsf{T} y\|_2 = \|y\|_2$ par orthogonalité de Q, si bien que

$$\begin{aligned} \|Ax - b\|_2^2 &= \|Q^\mathsf{T}(Ax - b)\|_2^2 \\ &= \|Rx - Q^\mathsf{T} b\|_2^2 \\ &= \sum_{i=1}^{r} (Rx - Q^\mathsf{T} b)_i^2 + \sum_{i=r+1}^{m} (Q^\mathsf{T} b)_i^2. \end{aligned}$$

La somme du second terme ne dépend pas de x et la somme du premier terme peut être annulée en résolvant

$$R_1 x = \tilde{b}, \tag{19.6}$$

où $\tilde{b} \in \mathbb{R}^n$ est défini par $\tilde{b}_i = (Q^\mathsf{T} b)_i$ pour $1 \leqslant i \leqslant r$. Du fait de la structure de R_1, cette équation a toujours une solution. Celle-ci est aussi une solution du problème (19.1) puisqu'elle donne au critère f sa valeur minimale $\frac{1}{2}\sum_{i=r+1}^{m} (Q^\mathsf{T} b)_i^2$.

La solution de (19.6) est unique si A est injective, car alors $r = n$ et le système (19.6) est carré. Si A n'est pas injective ($r < n$), ce système permet d'écrire les r premières composantes de x comme fonction de ses $n - r$ dernières composantes, celles-ci pouvant être choisies arbitrairement. On obtiendra la solution de (19.1) avec le plus de zéros en prenant $x_i = 0$ pour $r + 1 \leqslant i \leqslant n$.

Numériquement, la résolution de (19.1) par la factorisation QR de A est très stable. Mais la solution peut être très différente de celle de norme minimale. Mieux vaut utiliser la factorisation SVD (plus coûteuse).

Résolution par factorisation en valeurs singulières (SVD) de A ▲

La décomposition en valeurs singulières de A, de type $m \times n$ et de rang r, s'écrit

$$A = U \Sigma V^\mathsf{T},$$

où U est une matrice de type $m \times r$ dont les colonnes sont orthogonales ($U^\mathsf{T} U = I_r$), V est une matrice de type $n \times r$ dont les colonnes sont orthogonales ($V^\mathsf{T} V = I_r$) et Σ est une matrice d'ordre r diagonale définie positive ($\Sigma_{ij} = 0$ si $i \neq j$ et $\Sigma_{ii} = \sigma_i > 0$). Le pseudo-inverse de A s'écrit

$$A^\dagger = V \Sigma^{-1} U^\mathsf{T}.$$

19.1.5 Estimation numérique de la qualité de la solution ▲

La *matrice de résolution* d'un problème de moindres-carrés linéaire est définie par

$$R = A^\dagger A.$$

Cette formule montre que R est le projecteur orthogonal sur $\mathcal{R}(A^\dagger)$. En particulier, celui-ci projette toute solution x du problème de moindres-carrés sur la solution de norme minimale $\hat{x} = Rx$. On aura donc $R = I_n$ si, et seulement si, le problème de moindres-carrés a une unique solution. D'autre part, plus la colonne j de R est proche de e^j (le j-ième vecteur de base de $\mathbb{R}^m$), plus la j-ième composante des solutions est bien déterminée.

La *matrice de covariance des erreurs a posteriori sur la solution de norme minimale* est définie par

$$C_x = A^\dagger C_b (A^\dagger)^\mathsf{T},$$

où C_b est la *matrice de covariance des erreurs a priori sur les données* b. La matrice C_x spécifie comment des erreurs sur b se propagent à la solution de norme minimale.

Pour les problèmes dont la taille permet de calculer la factorisation SVD de A, disons n et m ne dépassant pas quelques milliers, on peut calculer R et C_x par

$$R = VV^\mathsf{T} \qquad \text{et} \qquad C_x = V\Sigma^{-1}U^\mathsf{T} C_b U \Sigma^{-1} V^\mathsf{T}.$$

Pour les problèmes de très grande taille, le calcul explicite de $A^\dagger$ n'est pas possible. Dans [461], la colonne i de $A^\dagger$, solution de norme minimale de

$$\inf_x \ \|Ax - e^i\|_2$$

est approchée par le point obtenu après une seule itération de gradient démarrant en zéro avec pas optimal. Ceci revient à approcher $A^\dagger$ par $A^\mathsf{T} D$, où D est la matrice diagonale des pas optimaux.

Voir [331, 618] pour des références en géophysique.

19.2 Moindres-carrés polyédrique

Dans les problèmes de moindres-carrés linéaires, on cherche à réaliser au mieux une équation linéaire, éventuellement sans solution. Comment s'y prendre si l'on cherche à réaliser au mieux un système d'équations et d'inéquations linéaires ? C'est à cette question que répond la modélisation sous forme de problème de moindres-carrés polyédrique (MCP).

19.2.1 Définition du problème

Un système d'équations et d'inéquations linéaires peut se représenter de manière compacte par un *problème d'inclusion linéaire dans un intervalle*, lequel s'écrit de la manière suivante

$$Ax \in [l, u], \tag{19.7}$$

où $A \in \mathbb{R}^{m\times n}$, $x \in \mathbb{R}^n$ est l'inconnue, $l \in (\mathbb{R} \cup \{-\infty\})^m$ et $u \in (\mathbb{R} \cup \{+\infty\})^m$. On supposera toujours que $l \leqslant u$. Avec les bornes l et u pouvant avoir des composantes infinies, il est nécessaire de préciser ce que l'on entend par l'intervalle $[l, u]$; celui-ci est défini par

$$[l, u] := \{y \in \mathbb{R}^m : l \leqslant y \leqslant u\}.$$

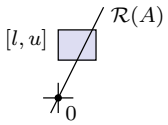
Il ne contient donc pas de valeurs infinies. Le problème d'inclusion (19.7) permet de représenter une égalité linéaire en prenant des bornes égales et une simple inégalité linéaire en prenant une borne infinie :

$$\begin{aligned} A_{i:}x = b_i \quad &\iff \quad A_{i:}x \in [l_i, u_i], \text{ avec } l_i = u_i = b_i, \\ A_{i:}x \leqslant b_i \quad &\iff \quad A_{i:}x \in [l_i, u_i], \text{ avec } l_i = -\infty \text{ et } u_i = b_i, \end{aligned}$$

où $A_{i:}$ désigne la ligne i de A.

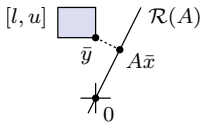
Clairement, on pourra trouver une solution de l'inclusion (19.7), c'est-à-dire un point x tel que Ax soit dans $[l, u]$, si, et seulement si, la condition géométrique suivante a lieu :

$$\mathcal{R}(A) \cap [l, u] \neq \varnothing. \tag{19.8}$$



Si cette condition n'est pas remplie, on pourra essayer de réaliser (19.7) « au mieux » en cherchant x tel que Ax soit le plus proche possible de $[l, u]$, au sens de la norme ℓ_2 (d'où vient le vocable « moindres-carrés »), ce qui s'écrit

$$\min_{x\in\mathbb{R}^n} \mathrm{d}_{[l,u]}(Ax), \tag{19.9}$$



où $\mathrm{d}_{[l,u]} : y \in \mathbb{R}^m \mapsto \mathrm{d}_{[l,u]}(y) := \min\{\|y - y'\|_2 : y' \in [l, u]\}$. On a utilisé l'opérateur « min » dans (19.9) plutôt que « inf », car on verra à la proposition 19.4 que ce problème a toujours une solution (comme les problèmes de moindres-carrés linéaires, voir la proposition 19.1). La valeur optimale de (19.9) est donc la distance euclidienne entre $\mathcal{R}(A)$ et $[l, u]$.

Nous donnerons à (19.9) le nom de *problème de moindres-carrés polyédrique* (MCP) ; le qualificatif *polyédrique* venant du fait que l'ensemble des x vérifiant (19.7) est un polyèdre convexe.

Le problème de moindres-carrés polyédrique (19.9) peut se récrire de différentes manières, qui aident à sa compréhension et à son analyse. En utilisant la définition de la distance $\mathrm{d}_{[l,u]}$, on voit que (19.9) se récrit comme suit

$$(19.9) \quad \iff \quad \min_{(x,y)\in\mathbb{R}^n\times[l,u]} \|y - Ax\|_2. \tag{19.10}$$

Ici aussi l'opérateur « min » est justifié par la proposition 19.4. En introduisant $s := y - Ax$ et en éliminant y, on obtient

$$(19.9) \quad \iff \quad \begin{cases} \min_{(x,s)\in\mathbb{R}^n\times\mathbb{R}^m} \|s\|_2 \\ l \leqslant Ax + s \leqslant u. \end{cases} \tag{19.11}$$

Les contraintes de ce problème sont toujours réalisables (par exemple en prenant $x = 0$ et $s \in [l, u]$ supposé non vide). Un vecteur s qui rend les contraintes de (19.11) réalisables est appelé une *translation réalisable*. L'ensemble des translations réalisables est clairement l'ensemble noté et défini par

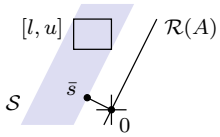
$$\mathcal{S} := [l, u] + \mathcal{R}(A).$$

On peut donc interpréter le problème (19.11) comme celui qui recherche à déterminer la plus petite translation s qui rende l'inclusion $Ax + s \in [l, u]$ réalisable ; on l'appelle la *plus petite translation réalisatible* et on la note $\bar{s}$. En éliminant x du problème (19.11), on obtient

$$(19.9) \quad \Longleftrightarrow \quad \min_{s \in \mathcal{S}} \|s\|_2, \tag{19.12}$$

qui est une expression du problème (19.9) ayant l'avantage d'être coercive. L'utilisation de l'opérateur « min » dans (19.12) est aussi justifié à la proposition 19.4. Ce problème montre aussi que $\bar{s}$ est la projection de zéro sur $\mathcal{S}$:

$$\bar{s} := \arg\min_{s \in \mathcal{S}} \|s\|_2. \tag{19.13}$$



Cette translation $\bar{s}$ est nulle si, et seulement si, (19.7) est réalisable.

19.2.2 Existence et unicité de solution

On note $\mathrm{P}_{[l,u]}$ le projecteur orthogonal sur $[l, u]$, pour le produit scalaire euclidien de $\mathbb{R}^m$.

Proposition 19.4 (existence de solution) *L'ensemble $\mathcal{S}$ est un polyèdre convexe non vide. Les problèmes (19.9), (19.10), (19.11) et (19.12) ont tous une solution. Leurs solutions $(\bar{x}, \bar{y}, \bar{s})$ se correspondent et on a*

$$\bar{s} = \bar{y} - A\bar{x} \qquad et \qquad \bar{y} = \mathrm{P}_{[l,u]}(A\bar{x}). \tag{19.14}$$

La plus petite translation réalisable $\bar{s}$ est déterminée de manière unique.

Démonstration. L'ensemble $\mathcal{S}$ est un polyèdre convexe non vide, comme somme de deux polyèdres convexes non vides.

L'existence et l'unicité de la plus petite translation réalisable $\bar{s}$, solution de (19.13), vient du fait qu'elle est la projection de zéro sur le convexe fermé non vide $\mathcal{S}$ (proposition 2.25).

L'existence de solution pour les problèmes (19.9), (19.10) et (19.11) s'en déduit, de même que la correspondance entre les solutions, $\bar{s} = \bar{y} - A\bar{x}$ et $\bar{y} = \mathrm{P}_{[l,u]}(A\bar{x})$. □

L'unicité du problème de moindres-carrés polyédrique, caractérisée dans la proposition 19.5 ci-dessous a un lien étroit avec l'unicité des solutions d'un système

d'inégalités linéaires : on trouve une caractérisation de l'unicité semblable à celle de la proposition 2.51 et même identique si l'inclusion (19.7) est réalisable. La caractérisation donnée dans la proposition ci-dessous généralise aussi celle de l'unicité de solution du problème de moindres-carrés linéaire, exprimée par l'injectivité de la matrice. Nous reviendrons sur ces deux points après la démonstration de la proposition. Dans cette proposition, on a noté $\mathrm{T}_{[l,u]}(y)$ le cône tangent à $[l,u]$ en $y \in [l,u]$ et $\mathrm{N}_{[l,u]}(y)$ le cône normal à $[l,u]$ en $y \in [l,u]$.

Proposition 19.5 (unicité de solution) *Soit $\bar{x}$ une solution du problème (19.9). Alors les propriétés suivantes sont équivalentes :*
(i) $\bar{x}$ est solution unique de (19.9),
(ii) tout d tel que $Ad \in \mathrm{T}_{[l,u]}(\mathrm{P}_{[l,u]}(A\bar{x}))$ est nul,
(iii) $A^{\mathsf{T}}(\mathrm{N}_{[l,u]}(\mathrm{P}_{[l,u]}(A\bar{x}))) = \mathbb{R}^n$.

Démonstration. [$(i) \Rightarrow (ii)$] Soit $d \in \mathbb{R}^n$ tel que $Ad \in \mathrm{T}_{[l,u]}(\mathrm{P}_{[l,u]}(A\bar{x}))$. Comme l'intervalle $[l,u]$ est polyédrique son cône tangent se confond avec son cône des directions admissibles (exercice 2.43), si bien que $\mathrm{P}_{[l,u]}(A\bar{x}) + tAd \in [l,u]$ pour $t > 0$ assez petit. Ceci a pour conséquence que $\bar{x}_t := \bar{x} + td$, avec $t > 0$ assez petit, vérifie

$$\mathrm{d}_{[l,u]}(A\bar{x}_t) \leqslant \|[\mathrm{P}_{[l,u]}(A\bar{x}) + tAd] - A\bar{x}_t\|_2 = \|\,\mathrm{P}_{[l,u]}(A\bar{x}) - A\bar{x}\|_2 = \mathrm{d}_{[l,u]}(A\bar{x}).$$

Comme $\bar{x}$ est solution de (19.9), $\bar{x}_t$ l'est aussi. L'unicité de la solution affirmée par (i) implique alors que $\bar{x}_t = \bar{x}$, c'est-à-dire que $d = 0$.

[$(ii) \Rightarrow (i)$] Soit $\bar{x}'$ une solution de (19.9) ; on cherche à montrer que $\bar{x}' = \bar{x}$. Par l'unicité de $\bar{s}$ (proposition 19.4), il existe des $\bar{y}$ et $\bar{y}' \in [l,u]$ tels que $\bar{s} = \bar{y} - A\bar{x} = \bar{y}' - A\bar{x}'$. Dès lors, $A(\bar{x}' - \bar{x}) = \bar{y}' - \bar{y}$. Comme $\bar{y} = \mathrm{P}_{[l,u]}(A\bar{x})$ (proposition 19.4) et $\bar{y}' \in [l,u]$, $\bar{y}' - \bar{y} \in \mathrm{T}_{[l,u]}(\mathrm{P}_{[l,u]}(A\bar{x}))$. Alors $A(\bar{x}' - \bar{x}) \in \mathrm{T}_{[l,u]}(\mathrm{P}_{[l,u]}(A\bar{x}))$ et (ii) impliquent que $\bar{x}' = \bar{x}$.

[$(ii) \Leftrightarrow (iii)$] On a

$$\begin{aligned}
(ii) &\iff \{0\} = \{d : Ad \in \mathrm{T}_{[l,u]}(\mathrm{P}_{[l,u]}(A\bar{x}))\} \\
&\iff \mathbb{R}^n = \{d : Ad \in \mathrm{T}_{[l,u]}(\mathrm{P}_{[l,u]}(A\bar{x}))\}^+ \qquad (19.15)\\
&\iff \mathbb{R}^n = -A^{\mathsf{T}}(\mathrm{N}_{[l,u]}(\mathrm{P}_{[l,u]}(A\bar{x}))) \qquad (19.16)\\
&\iff (iii),
\end{aligned}$$

où en (19.15) on a pris le dual de chaque membre et en (19.16) on a utilisé le lemme de Farkas (proposition 2.45), la relation de dualité reliant le cône tangent et le cône normal (point 4 de la proposition 2.55) et le caractère fermé de $A^{\mathsf{T}}(\mathrm{N}_{[l,u]}(\mathrm{P}_{[l,u]}(A\bar{x})))$ (image linéaire d'un polyèdre convexe ; proposition 2.18 et exercice 2.43). □

Voyons quelles formes prennent les conditions nécessaires et suffisantes d'unicité de la proposition 19.5 dans deux cas particuliers.

- Dans le problème de moindres-carrés linéaire (19.1), on cherche à réaliser au mieux le système linéaire $Ax = b$. On retrouve ce problème avec (19.9) en prenant pour intervalle $[l,u]$ le singleton $\{b\}$. Dans ce cas, le cône tangent $\mathrm{T}_{\{b\}}(\mathrm{P}_{\{b\}}(A\bar{x})) =$

$\mathrm{T}_{\{b\}}(b)$ est réduit à $\{0\}$ et le cône normal $\mathrm{N}_{\{b\}}(\mathrm{P}_{\{b\}}(A\bar{x})) = \mathrm{N}_{\{b\}}(b)$ est $\mathbb{R}^m$. Alors la condition (ii) exprime l'injectivité de A et (iii) la surjectivité de A^T, ce qui est bien la condition nécessaire et suffisante d'unicité révélée par la proposition 19.1.

On notera que la condition nécessaire et suffisante d'unicité de la solution du problème de moindres-carrés linéaire ne dépend que de A, alors que celle de la proposition 19.5 fait intervenir une solution $\bar{x}$.

- Si l'on cherche à réaliser au mieux un système d'inéquations $Ax \leqslant b$ par l'approche décrite à la section 19.2.1, on prendra $[l, u] =]-\infty, b]$ dans (19.7). Alors $\mathrm{d}_{]-\infty,b]}(Ax) = \mathrm{d}_{\mathbb{R}^m_-}(Ax-b) = \|\mathrm{P}_{\mathbb{R}^m_-}(Ax-b) - (Ax-b)\|_2 = \|(Ax-b)^- + (Ax-b)\|_2 = \|(Ax-b)^+\|_2$, si bien que le problème (19.9) s'écrit dans ce cas

$$\min_{x\in\mathbb{R}^n} \|(Ax-b)^+\|_2. \tag{19.17}$$

Ce problème a donc toujours une solution (proposition 19.4).

Pour déterminer les conditions d'unicité de la solution de ce problème au moyen de la proposition 19.5, on observe que $\mathrm{P}_{]-\infty,b]}(A\bar{x}) = b + \mathrm{P}_{\mathbb{R}^m_-}(A\bar{x}-b)$ ou

$$\mathrm{P}_{]-\infty,b]}(A\bar{x}) = b - (A\bar{x}-b)^-. \tag{19.18}$$

Alors $\mathrm{T}_{]-\infty,b]}(\mathrm{P}_{]-\infty,b]}(A\bar{x})) = \mathrm{T}_{\mathbb{R}^m_-}(-(A\bar{x}-b)^-) = \{d \in \mathbb{R}^n : A_I d \leqslant 0\}$, où I est l'ensemble des indices $i \in [1:m]$ tels que $(A\bar{x}-b)_i^- = 0$, c'est-à-dire

$$I = \{i \in [1:m] : (A\bar{x}-b)_i \geqslant 0\}.$$

Dès lors, d'après la proposition 19.5, le problème (19.17) a une solution unique si, et seulement si, l'une des conditions équivalentes suivantes a lieu :

$$\begin{gathered}\text{tout } d \in \mathbb{R}^n \text{ tel que } A_I d \leqslant 0 \text{ est nul,} \\ A_I^\mathsf{T}(\mathbb{R}_+^{|I|}) = \mathbb{R}^n.\end{gathered} \tag{19.19}$$

Si le coût optimal de (19.17) est nul, l'ensemble des solutions du problème est le polyèdre convexe $\{x \in \mathbb{R}^n : Ax \leqslant b\}$. On sait par la proposition 2.51 que celui-ci est le singleton $\{\bar{x}\}$ si, et seulement si, tout vecteur d vérifiant $A_I d \leqslant 0$ est nul, condition dans laquelle $I := \{i \in [1:m] : (A\bar{x}-b)_i = 0\}$. C'est bien ce qu'affirme (19.19), car les deux ensembles d'indices I sont identiques dans ce cas.

Évidemment la condition (ii) de la proposition est plus forte que l'injectivité de A et la condition (iii) de la proposition est plus forte que la surjectivité de A ; ainsi (ii) ou $(iii) \Rightarrow A$ injective ou A^T surjective, mais la réciproque n'est pas garantie. Comme contre-exemple, considérons le problème (19.17) avec $m = n = 1$, $A = 1$ et $b = 0$, qui consiste à minimiser x^+ sur $\mathbb{R}$: A est injective (ou $A^\mathsf{T} = A$ est surjective), mais l'ensemble des solutions $\mathbb{R}_-$ n'est pas un singleton, donc (ii) n'a pas lieu [en effet $\mathbb{R}_- \neq \{0\}$] (ou (iii) n'a pas lieu [en effet $\mathbb{R}_+ \neq \mathbb{R}$]).

19.2.3 Condition d'optimalité

Les problèmes (19.9), (19.10), (19.11) et (19.12) sont convexes. On peut donc en donner des conditions nécessaires et suffisantes d'optimalité. On note $\mathrm{Sol}(P)$ l'ensemble des solutions du problème P, tout en évitant les doubles parenthèses.

Proposition 19.6 (conditions d'optimalité) *Les conditions nécessaires et suffisantes d'optimalité suivantes ont lieu :*

$$\bar{x} \in \mathrm{Sol}\,(19.9) \iff A^{\mathsf{T}}\big(\mathrm{P}_{[l,u]}(A\bar{x}) - A\bar{x}\big) = 0, \tag{19.20}$$

$$(\bar{x},\bar{y}) \in \mathrm{Sol}\,(19.10) \iff A^{\mathsf{T}}(\bar{y} - A\bar{x}) = 0 \text{ et } \bar{y} = \mathrm{P}_{[l,u]}(A\bar{x}), \tag{19.21}$$

$$(\bar{x},\bar{s}) \in \mathrm{Sol}\,(19.11) \iff A^{\mathsf{T}}\bar{s} = 0 \text{ et } \bar{s} = \mathrm{P}_{[l,u]}(A\bar{x}) - A\bar{x}, \tag{19.22}$$

$$\bar{s} \in \mathrm{Sol}\,(19.12) \iff \bar{s} = \mathrm{P}_{[l,u]+\mathcal{R}(A)}(0). \tag{19.23}$$

Démonstration. On ne change pas le problème (19.9) en élevant son critère au carré. Ce nouveau critère est alors convexe différentiable avec un gradient pour le produit scalaire euclidien qui, en $\bar{x}$, vaut $2A^{\mathsf{T}}(\mathrm{P}_{[l,u]}(A\bar{x}) - A\bar{x})$. La condition nécessaire et suffisante d'optimalité (19.20) découle alors de la condition de Fermat (proposition 4.9).

Les autres conditions d'optimalité (19.21) et (19.22) se déduisent de (19.20) et du fait que les solutions des problèmes (19.9), (19.10) et (19.11) se correspondent avec les relations (19.14). La condition d'optimalité (19.23) est immédiate. □

Appliquons ce résultat au problème (19.17). D'après (19.18), $A\bar{x} - \mathrm{P}_{]-\infty,b]}(A\bar{x}) = (A\bar{x} - b)^+$, si bien que

$$\bar{x} \text{ est solution de (19.17)} \iff A^{\mathsf{T}}(A\bar{x} - b)^+ = 0,$$

ce que l'on aurait pu obtenir directement en annulant la dérivée de la fonction différentiable $x \in \mathbb{R}^n \mapsto \|(Ax - b)^+\|_2^2$, comme dans la démonstration de la proposition.

19.3 Moindres-carrés non linéaire

19.3.1 Définition du problème

Soit $r : \mathbb{R}^n \to \mathbb{R}^m$ une fonction régulière, que l'on appelle ci-dessous le *résidu*. On cherche à annuler celui-ci au sens des moindres carrés, c'est-à-dire :

$$\min_{x\in\mathbb{R}^n} \Big(f(x) = \frac{1}{2}\|r(x)\|^2\Big), \tag{19.24}$$

où $\|\cdot\|$ désigne la norme ℓ_2. Un problème de ce type porte le nom de *problème de moindres-carrés non linéaire*, parce qu'on y minimise la somme des carrés des résidus non linéaires r_i. C'est une version non linéaire du problème de moindres-carrés linéaire (19.1), puisque l'on retrouve ce dernier en définissant r en x par $r(x) = Ax - b$. En pratique, on a $m \gg n$, mais, sauf mention contraire, nous ne faisons pas d'hypothèse systématique sur la grandeur relative de m et n ci-dessous.

On note

$$J(x) = r'(x)$$

la jacobienne de r en x, qui est une matrice de type $m \times n$. On calcule aisément le gradient et la hessienne du critère de (19.24) pour le produit scalaire euclidien :

$$\nabla f(x) = J(x)^{\mathsf{T}} r(x) \quad \text{et} \quad \nabla^2 f(x) = J(x)^{\mathsf{T}} J(x) + \sum_{i=1}^{p} r_i(x) \nabla^2 r_i(x). \tag{19.25}$$

Dans un contexte où une suite $\{x_k\} \subseteq \mathbb{R}^n$ est définie, on note encore

$$r_k := r(x_k), \qquad g_k := \nabla f(x_k) \qquad \text{et} \qquad J_k := J(x_k). \tag{19.26}$$

19.3.2 Algorithme de Gauss-Newton

L'algorithme et sa cohérence

L'application de la méthode de Newton pour minimiser la fonction f définie en (19.24) requiert le calcul de la hessienne $\nabla^2 f(x)$ et donc, comme le montre la formule (19.25), le calcul des dérivées secondes du résidu. Pour certains problèmes de moindres-carrés, ce calcul peut être très coûteux, si bien que l'on s'intéresse à l'algorithme qui néglige ces termes de la hessienne et calcule donc en x_k une direction d_k en résolvant le système linéaire suivant :

$$\boxed{\left(J_k^{\mathsf{T}} J_k\right) d_k = -J_k^{\mathsf{T}} r_k.} \tag{19.27}$$

On a noté $J_k = J(x_k)$ et $r_k = r(x_k)$ (il y a une petite ambiguïté de notation — r_k n'est pas la k-ième composant du résidu — mais celle-ci sera toujours facilement levée par le contexte). La perte d'efficacité locale par rapport à l'algorithme de Newton sera donc faible si le terme négligé $\sum_i r_i(x) \nabla^2 r_i(x)$ de la hessienne de f est relativement petit, ce qui sera le cas si r est (presque) linéaire ou si $r(x)$ est petit (cas des problèmes d'identification).

On peut obtenir la même direction en raisonnant comme suit. On considère le modèle quadratique de f obtenu, non pas par son développement au deuxième ordre (ce serait l'algorithme de Newton et l'on cherche à simplifier celui-ci ici), mais en linéarisant le résidu à l'intérieur de la norme dans (19.24). Ceci donne :

$$\min_{d \in \mathbb{R}^n} \frac{1}{2} \|r_k + J_k d\|^2. \tag{19.28}$$

On observe que ce problème définit les mêmes directions d_k que précédemment, car l'équation d'optimalité du premier ordre de (19.28) n'est autre que (19.27). De plus celle-ci est nécessaire et suffisante car le problème (19.28) est convexe.

Le résultat suivant montre qu'il est raisonnable de construire une méthode à direction de descente en prenant la direction d_k définie par (19.27) ou (19.28) comme direction le long desquelles on se déplace.

Proposition 19.7 (direction de descente de Gauss-Newton) *Le système linéaire (19.27) a une solution d_k. Si x_k n'est pas un point stationnaire du problème de moindres-carrés non linéaire (19.24), alors d_k est une direction de descente de f en x_k.*

Démonstration. Le problème (19.28) a toujours au moins une solution car c'est un problème de moindres-carrés linéaire (proposition 19.1). D'autre part, en prenant le produit scalaire des deux membres de (19.27) avec d_k, on trouve

$$g_k^T d_k = -\|J_k d_k\|^2,$$

qui est strictement négatif lorsque x_k n'est pas stationnaire (car alors $J_k^T r_k \neq 0$ et donc $J_k d_k \neq 0$ par (19.27)). □

Ceci nous conduit à l'*algorithme de Gauss-Newton*, dont nous décrivons une itération ci-dessous.

Algorithme 19.8 (Gauss-Newton) L'algorithme calcule l'itéré x_{k+1} à partir de l'étéré x_k de la manière suivante.

1. *Test d'arrêt.* Si $J_k^T r_k \simeq 0$, arrêt de l'algorithme.
2. *Direction de descente.* Calculer une solution d_k de (19.27).
3. *Recherche linéaire.* Trouver un pas $\alpha_k > 0$ en x_k le long de d_k par une règle de recherche linéaire satisfaisant la condition de Zoutendijk (6.19) ou (6.26a).
4. *Nouvel itéré.* $x_{k+1} := x_k + \alpha_k d_k$.

À l'étape 3, une règle de recherche linéaire appropriée est celle d'Armijo ou de Goldstein (voir la section 6.3.3).

Convergence globale et complexité itérative

Des conditions de convergence globale et de complexité itérative globale de cet algorithme sont données dans le théorème qui suit. On y requiert une hypothèse assez forte, qui est celle de l'*injectivité uniforme* de la suite $\{J_k\}$, ce qui veut dire qu'il existe une constante $\alpha_J > 0$ telle que pour tout indice $k \geqslant 1$ et tout vecteur $v \in \mathbb{R}^n$, on ait

$$\|J_k v\| \geqslant \alpha_J \|v\|. \tag{19.29}$$

Dans ce cas, la direction de Gauss-Newton est déterminée de manière unique par (19.27) et la convergence globale *vers un point stationnaire* de f est assurée (point 1) avec une complexité itérative en $O(\varepsilon^{-2})$ (point 2) ; cette complexité ne dépend pas des dimensions n et m du problème.

Les résultats s'améliorent beaucoup s'il s'avère que $\{J_k^T\}$ est uniformément injective, puisqu'alors, l'algorithme assure la convergence de $\{r_k\}$ vers zéro (point 3), avec une complexité itérative en $O(\log \varepsilon^{-1})$ (point 4), ce qui est spectaculairement mieux ; cette complexité ne dépend pas de n et m. D'ailleurs, dans ce cas, on doit avoir $m = n$ (car J_k est déclarée bijective) et l'on cherche donc un zéro du système de n équations à n inconnues $r(x) = 0$.

Théorème 19.9 (convergence et complexité itérative de l'algorithme de Gauss-Newton) *Supposons que f donnée par (19.24) soit $\mathcal{C}^{1,1}$ dans un voisinage de $\mathcal{N}_1 := \{x \in \mathbb{R}^n : f(x) \leqslant f(x_1)\}$. Soit $\{x_k\}$ une suite générée par l'algorithme 19.8 de Gauss-Newton, telle que $\{J_k\}$ soit bornée et uniformément injective. Alors,*

1) $J_k^T r_k \to 0$,
2) *il existe une constante $C > 0$, indépendante de n et m, telle que, pour tout $\varepsilon > 0$, $\|J_k^T r_k\| \leqslant \varepsilon$ pour un indice k inférieur à $\lceil C\varepsilon^{-2} \rceil$,*
3) *si, de plus, $\{J_k^T\}$ est uniformément injective, alors*
   a) *$f(x_k) \to 0$ linéairement,*
   b) *il existe une constante $C' > 0$, indépendante de $n = m$, telle que, pour tout $\varepsilon > 0$, $\|r_k\| \leqslant \varepsilon \|r_0\|$ pour tout indice $k \geqslant \lceil C' \log \varepsilon^{-1} \rceil$.*

Démonstration. 1) Si la condition de Zoutendijk (6.19) a lieu, on utilise la proposition 6.8 et sa conclusion (6.21). Pour cela, on observe que $f_* := \inf_k f(x_k) \in [0, f(x_1)]$ est bien fini et que le cosinus de l'angle θ_k entre d_k et $-g_k$ est minoré par une constante strictement positive :

$$\cos\theta_k := \frac{-g_k^T d_k}{\|g_k\|\,\|d_k\|} = \frac{\|J_k d_k\|^2}{\|J_k^T r_k\|\,\|d_k\|} \geqslant \frac{\alpha_J \|J_k d_k\|}{\|J_k^T J_k d_k\|} \geqslant \frac{\alpha_J}{\beta_J}, \tag{19.30}$$

où l'on a utilisé (19.29) (injectivité uniforme de $\{J_k\}$) et la borne β_J sur $\|J_k^T\| = \|J_k\|$. Alors, (6.21) implique que $g_k = J_k^T r_k \to 0$.

Si la condition (6.26a) a lieu, $|g_k^T d_k| = \|J_k d_k\|^2$ doit tendre vers zéro (car $f(x_k)$ et $f(x_{k+1})$ tendent vers la même valeur) et comme $\|J_k d_k\| \geqslant \alpha_J \|d_k\|$ par (19.29) (injectivité uniforme de $\{J_k\}$), on a $d_k \to 0$. Par (19.27) et la bornitude de $\{J_k\}$, ceci implique aussi $g_k = J_k^T r_k \to 0$.

2) Si la condition de Zoutendijk (6.19) a lieu, le résultat se déduit de la seconde partie de la proposition 6.8 qui affirme que, pour tout $\varepsilon > 0$, $\|J_k^T r_k\| \leqslant \varepsilon$ pour au moins un indice k inférieur au K_ε donné par (6.22), avec $\gamma = \alpha_J/\beta_J$.

Considérons maintenant le cas où la condition (6.26a) a lieu. Par (19.27) et l'existence d'une borne β_J sur $\|J_k^T\| = \|J_k\|$, on a

$$\|g_k\|^2 = \|J_k^T J_k d_k\|^2 \leqslant \beta_J^2 \|J_k d_k\|^2 = \beta_J^2 |g_k^T d_k|.$$

Alors la condition (6.26a) se récrit

$$f(x_{k+1}) \leqslant f(x_k) - C\beta_J^{-2}\|g_k\|^2 \qquad \text{ou} \qquad C\beta_J^{-2}\|g_k\|^2 \leqslant f(x_k) - f(x_{k+1}). \tag{19.31}$$

En sommant de $k = 1$ à K, on trouve

$$\min_{k \in [1:K]} \|g_k\|^2 \leqslant \frac{1}{K} \sum_{k=1}^{K} \|g_k\|^2 \leqslant \frac{f(x_1) - f_*}{C\beta_J^{-2} K}.$$

Donc un des $\|g_k\|$, avec $k \in [1:K]$, est inférieur à $\varepsilon > 0$ si le membre de droite est inférieur à ε^2, c'est-à-dire si $K \geqslant C^{-1}\beta_J^2 (f(x_1) - f_*)\varepsilon^{-2}$.

3) Pour une constante strictement positive C_1, on a

$$f(x_{k+1}) \leqslant f(x_k) - C_1\|g_k\|^2,$$

soit par la condition de Zoutendijk (6.19) et (19.30), soit par la condition (6.26a) et (19.31). Alors, en utilisant $g_k = J_k^{\mathsf{T}} r_k$ et $\|J_k^{\mathsf{T}} r_k\|^2 \geqslant (\alpha'_J)^2\|r_k\|^2 = 2(\alpha'_J)^2 f(x_k)$ par l'injectivité uniforme de J_k^{T}, on trouve

$$f(x_{k+1}) \leqslant (1 - C_2) f(x_k),$$

pour la constante strictement positive $C_2 = 2C_1(\alpha'_J)^2$. Ceci montre la convergence linéaire de $f(x_k)$ vers zéro (point 3.a). Puis par récurrence,

$$\frac{f(x_k)}{f(x_0)} \leqslant (1 - C_2)^k.$$

Dès lors, pour un $\varepsilon > 0$ donné, $\|r_k\| \leqslant \varepsilon\|r_0\|$ si $f(x_k)/f(x_0) \leqslant \varepsilon^2$ et donc certainement si le membre de droite ci-dessus est inférieur à ε^2 :

$$(1 - C_2)^k \leqslant \varepsilon^2.$$

En prenant les logarithmes et en tenant compte du fait que $\log(1-C_2) < 0$, on obtient comme condition sur k :

$$k \geqslant \frac{2\log \varepsilon^{-1}}{\log(1 - C_2)^{-1}}.$$

On obtient donc le résultat du point 3.b avec la constante $C' := 2/\log(1-C_2)^{-1}$. □

Cette courte analyse montre que, contrairement à la méthode de Newton, l'algorithme de Gauss-Newton est bien défini à tout itéré x_k (la direction d_k existe toujours) et, pourvu que x_k ne soit pas un point stationnaire de la fonction de moindres carrés f définie en (19.24), la direction calculée d_k est de descente pour f. Il demande toutefois de résoudre le système linéaire (19.27), qui est moins bien conditionné que (10.3). Sa convergence est assurée sous des conditions souvent vérifiées, mais pas toujours. L'algorithme de Gauss-Newton peut en effet très bien générer une suite convergeant vers un point auquel f n'est pas stationnaire, a fortiori n'annulant pas le résidu. De plus, la vitesse de convergence peut être assez lente si le résidu n'est pas nul en la solution ou si r y est fortement non linéaire (r'' non nul), car l'algorithme s'écarte alors de la méthode de Newton par un terme significativement non nul $\nabla^2 f(x) - J(x)^{\mathsf{T}} J(x) = \sum_i r_i(x)\nabla^2 r_i(x)$ (voir (19.25)).

L'algorithme de Levenberg-Morrison-Marquardt de la section suivante apporte quelques remèdes à ces défauts ; en particulier, il permet de se défaire de la condition d'injectivité uniforme de $\{J_k\}$.

19.3.3 Algorithme de Levenberg-Morrison-Marquardt

L'algorithme et sa cohérence

L'algorithme de Levenberg-Morrison-Marquardt (LMM) cherche à résoudre le problème de moindres-carrés non linéaire (19.1) en générant une suite $\{x_k\} \subseteq \mathbb{R}^n$

de la manière suivante. Il définit le déplacement $s_k \in \mathbb{R}^n$ de l'itéré courant x_k à l'itéré suivant $x_{k+1} := x_k + s_k$ en résolvant le système linéaire

$$\boxed{\left(J_k^T J_k + \lambda_k M_k\right) s_k = -J_k^T r_k,} \tag{19.32}$$

où l'on a utilisé les notations (19.26). Dans ce système, $M_k \in \mathcal{S}^n_{++}$ est une matrice symétrique définie positive, dont le rôle et les règles qu'elle doit vérifier seront précisés plus loin, et $\lambda_k > 0$ est un paramètre, appelé tantôt *facteur de pénalisation*, tantôt *multiplicateur*, selon l'interprétation que l'on en fait. Ce multiplicateur joue un rôle important dans l'algorithme, qui le détermine lui-même à chaque itération, alors que M_k peut être laissé au choix de l'utilisateur. Ce multiplicateur est l'élément nouveau par rapport à la direction de l'algorithme de Gauss-Newton, calculée comme solution de (19.27), qui n'est autre que (19.32) avec $\lambda_k = 0$. Remarquons qu'ici, grâce à la stricte positivité de λ_k et la définie positivité de M_k, la matrice du système linéaire (19.32) est symétrique définie positive, si bien que s_k y est déterminé de manière unique. Lorsqu'il sera important de mentionner la dépendance en λ_k de la solution de (19.32), on notera celle-ci

$$s_k(\lambda_k).$$

Insistons sur le fait que s_k est le déplacement de x_k à x_{k+1} et pas une direction de déplacement le long de laquelle on ferait un pas pouvant être différent de un ; pour le dire autrement, l'algorithme présenté ci-dessous n'utilise pas de recherche linéaire pour déterminer l'itéré suivant, comme c'est le cas pour l'algorithme de Gauss-Newton de la section 19.3.2. Sur ce plan, l'algorithme s'apparente davantage aux méthodes à régions de confiance. En particulier, comme nous allons le voir, c'est en augmentant λ_k que l'on diminue la grandeur du déplacement $s_k(\lambda_k)$, ce qui explique pourquoi c'est l'algorithme lui-même qui doit prendre en charge la détermination de λ_k. C'est aussi ce réglage du multiplicateur λ_k qui permettra de faire décroître f suffisamment à chaque itération.

On voit que si $g_k = 0$, le déplacement donné par (19.32) est nul, si bien que l'algorithme ne peut pas progresser. Par conséquent, l'algorithme de LMM ne peut pas trouver mieux qu'un point stationnaire de la fonction de moindres-carrés f, mais un tel point peut être un zéro de r si sa jacobienne y est injective. C'est donc en ces termes que se formuleront le test d'arrêt de l'algorithme 19.10 et le résultat de convergence globale du théorème 19.12.

Il est utile de constater que le déplacement s_k de l'algorithme de LMM, solution de l'équation (19.32), est aussi solution du problème de moindres-carrés linéaire *pénalisé* suivant

$$\min_{s \in \mathbb{R}^n} \frac{1}{2} \|r_k + J_k s\|^2 + \frac{\lambda_k}{2} s^T M_k s, \tag{19.33}$$

où $\|\cdot\|$ désigne la norme ℓ_2 dans $\mathbb{R}^m$ (on pourra comparer ce problème d'optimisation quadratique associé à celui (19.28) qui était associé au système de Gauss-Newton (19.27)). En effet, la condition nécessaire et suffisante d'optimalité de ce problème quadratique convexe, c'est-à-dire la nullité du gradient de son critère par rapport à s, n'est autre que (19.32). De ce point de vue, λ est un *facteur de pénalisation*. Il sera

utile d'introduire une notation pour la fonction qui intervient dans le critère de ce problème, qui est la fonction $\varphi_k : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ définie en $(x, \lambda) \in \mathbb{R}^n \times \mathbb{R}$ par

$$\varphi_k(s, \lambda) := \frac{1}{2} \|J_k s + r_k\|^2 + \frac{\lambda}{2} s^T M_k s.$$

Au terme $-\lambda\Delta^2/2$ indépendant de s près, cette fonction est le lagrangien du problème

$$\begin{cases} \min_{s \in \mathbb{R}^n} \frac{1}{2} \|J_k s + r_k\|^2 \\ s^T M_k s \leqslant \Delta^2. \end{cases} \tag{19.34}$$

où $\Delta > 0$ est le *rayon de confiance* du problème (voir le chapitre 9). De ce point de vue, λ est un *multiplicateur* associé à la contrainte de région de confiance $\{s \in \mathbb{R}^n : s^T M_k s \leqslant \Delta^2\}$ de ce problème.

On déduit des points de vue présentés dans le paragraphe précédent que, $\|s_k(\lambda)\|$ décroît lorsque λ croît (proposition 13.2). De plus (proposition 19.11 et exercice 19.1), lorsque $\lambda \to \infty$ et $g_k \neq 0$, on a $s_k(\lambda) \to 0$ et $\lambda M_k s_k(\lambda) \to -g_k$, si bien que, dans ces circonstances, $s_k(\lambda)$ s'aligne sur $-M_k^{-1} g_k$, dans le sens où

$$s_k(\lambda) = -\frac{\|s_k(\lambda)\|}{\|M_k^{-1} g_k\|} M_k^{-1} g_k + o(\|s_k(\lambda)\|), \qquad \text{lorsque } \lambda \to \infty.$$

Par la différentiabilité supposée de la fonction f définie en (19.24), on en déduit que

$$f(x_k + s_k(\lambda)) = f(x_k) - \frac{g_k^T M_k^{-1} g_k}{\|M_k^{-1} g_k\|} \|s_k(\lambda)\| + o(\|s_k(\lambda)\|), \quad \text{lorsque } \lambda \to \infty.$$

Dès lors, toujours si $g_k \neq 0$, $f(x_k + s_k(\lambda)) < f(x_k)$ si λ est pris assez grand. C'est en ajustant λ à chaque itération que l'algorithme de LMM assure la décroissance suffisante de f et la convergence de g_k vers zéro.

Il y a plusieurs manières de déterminer λ_k à chaque itération. Nous présentons ci-dessous celle proposée par Osborne [465] qui prend racine sur la méthode des régions de confiance. Son intérêt est de ne pas devoir résoudre le problème de région de confiance (19.34) à chaque itération, lequel peut être coûteux pour certains problèmes. L'idée est similaire à celle utilisée par les régions de confiance et fait entrer en jeu le rapport ρ_k entre la décroissance réelle de f apportée par le déplacement $s_k(\lambda)$, pour un certain $\lambda > 0$, et la décroissance prédite par le modèle φ_k :

$$\rho_k(\lambda) := \frac{1}{2} \frac{f(x_k) - f(x_k + s_k(\lambda))}{f(x_k) - \varphi_k(s_k(\lambda), \lambda)}.$$

Le facteur $1/2$ est placé pour que la variation au premier ordre du numérateur, qui vaut $-g_k^T s_k$, soit identique à celle du dénominateur, si bien que fonction et modèle sont alors « tangents ». Le dénominateur vaut en effet

$$\begin{aligned} 2[f(x_k) - \varphi_k(s_k(\lambda), \lambda)] &= -2\, r_k^T J_k s_k(\lambda) - \|J_k s_k(\lambda)\|^2 - \lambda\, s_k(\lambda)^T M_k s_k(\lambda) \\ &= -g_k^T s_k(\lambda), \end{aligned} \tag{19.35}$$

où l'on a utilisé $g_k = J_k^T r_k$ et (19.32) pour obtenir la dernière égalité. Comme g_k est supposé non nul au cours de l'itération (l'algorithme s'arrête dès qu'il trouve

un point stationnaire de f) et que $s_k(\lambda)$ minimise $\varphi_k(\cdot,\lambda)$, le dénominateur de ρ_k est strictement positif. En effet, par (19.32), $s_k(\lambda)$ est non nul lorsque $g_k \neq 0$ et comme $s_k(\lambda)$ est l'unique point qui minimise le critère de (19.33), on a nécessairement $\varphi_k(s_k(\lambda),\lambda) < \varphi_k(0,\lambda) = f(x_k)$. Au passage, nous avons montré que, quel que soit $\lambda > 0$,

$$g_k^T s_k(\lambda) < 0.$$

Alors, si le rapport $\rho_k(\lambda)$ est supérieur à un seuil préfixé $\eta_1 \in \mathbb{R}_{++}$, l'on a grâce à (19.35)

$$f(x_k + s_k(\lambda)) \leqslant f(x_k) + \eta_1\, g_k^T s_k(\lambda), \tag{19.36}$$

qui rappelle la condition de décroissance suffisante (6.9) utilisée en recherche linéaire, sauf qu'ici le pas α_k est remplacé par le paramètre λ qui ajuste la grandeur de $s_k(\lambda)$. On considère qu'une décroissance suffisante de f est obtenue lorsque cette inégalité est vérifiée pour un certain $\lambda > 0$ (cet considération sera validée par le résultat de convergence ci-dessous) et on accepte $s_k(\lambda)$ comme déplacement de l'itération k. Dans le cas contraire, tant que l'inégalité ci-dessus n'est pas vérifiée, on augmente λ et on résout à nouveau (19.32) ; on montrera par la proposition 19.11 que cette procédure de réglage du λ se conclut en un nombre fini d'étapes.

Nous avons à présent sous la main et à l'esprit, tous les éléments permettant d'énoncer et de comprendre *une itération* de l'algorithme de LMM, dans la version d'Osborne [465]. On suppose qu'au début de l'itération k, l'on dispose d'un itéré $x_k \in \mathbb{R}^n$, d'un multiplicateur $\lambda_{k-1} > 0$ et d'une matrice $M_k \in \mathcal{S}^n_{++}$. L'itéré x_k et le multiplicateur λ_{k-1} sont mis à jour par l'itération et une nouvelle matrice M_k pourra être choisie en fin d'itération. Par constante, la description ci-dessous entend une valeur qui ne dépend pas de l'itération.

Algorithme 19.10 (Levenberg-Morrison-Marquardt revisité)
L'algorithme utilise les constantes suivantes : $0 < \tau_1 < 1 < \tau_2$ pour la mise à jour de λ_{k-1} et $0 < \eta_1 < \eta_2 < 1$ comme seuils de satisfaction de la décroissance de f. On suppose que $M_k \in \mathcal{S}^n_{++}$. Une itération de l'algorithme se déroule comme suit.

1. *Test d'arrêt.* Si $J_k^T r_k \simeq 0$, arrêt de l'algorithme.
2. *Déplacement.* Prendre $\lambda_{k,0} := \lambda_{k-1}$ et répéter les opérations suivantes pour $i \in \mathbb{N}$ jusqu'à satisfaction du test de sortie (19.38).
   2.1. Calculer la solution $s_{k,i}$ du système linéaire
   $$\left(J_k^T J_k + \lambda_{k,i} M_k\right) s_{k,i} = -J_k^T r_k, \tag{19.37}$$
   2.2. Si
   $$f(x_k + s_{k,i}) \leqslant f(x_k) + \eta_1 g_k^T s_{k,i}, \tag{19.38}$$
   sortir de la boucle courante avec $s_k := s_{k,i}$ (on va à l'étape 3), sinon $\lambda_{k,i+1} = \tau_2 \lambda_{k,i}$.
3. *Nouveau facteur de pénalisation.* Si

$$f(x_k + s_k) \leqslant f(x_k) + \eta_2 g_k^\mathsf{T} s_k, \tag{19.39}$$

$\lambda_k := \tau_1 \lambda_{k,i}$, sinon $\lambda_k := \lambda_{k,i}$.

4. *Nouvel itéré.* $x_{k+1} := x_k + s_k$.
5. *Nouvelle matrice.* Choix de $M_{k+1} \in \mathcal{S}^n_{++}$.

Voici quelques remarques sur cet algorithme.

- Le coût de cet algorithme est principalement lié au nombre de systèmes linéaires (19.37) qu'il faut résoudre.
- Typiquement, on prendra $\eta_1 \simeq 10^{-4}$, comme pour la constante ω_1 de (6.9) en recherche linéaire, de manière à rendre le test de deecroissance suffisante (19.38) aussi peu contraignant que possible.
- L'étape 3 et la constante η_2 ne jouent pas de rôle dans la convergence mais sont introduites pour ne pas imposer la croissance de la suite $\{\lambda_k\}$, ce qui ne serait pas adéquat (le premier multiplicateur λ_1 choisi peut être trop grand et donc conduire à des déplacements s_k trop petits, ce qui ralentirait la convergence). Si la décroissance de f est forte au sens de l'inégalité (19.39), alors l'algorithme s'autorise prendre λ_k plus petit que $\lambda_{k,i}$ à l'itération suivante.
- Au lieu de régler λ_k, certains auteurs [597] préfèrent prendre le multiplicateur de la forme $\lambda_k = \mu_k \|r_k\|^\delta$, avec $\delta \geqslant 0$ ($\delta = 0$ dans notre cas), et ajuster μ_k au lieu de λ_k à chaque itération.

Le caractère *bien défini* de l'algorithme de LMM, c'est-à-dire le fait qu'il sorte de la boucle de l'étape 2 en un nombre fini de cycles, est suggéré par la discussion qui précède son énoncé. Avec la proposition suivante, nous le montrons de manière rigoureuse.

Proposition 19.11 (descente suffisante) *Si r est différentiable en x_k, si le gradient $g_k := J_k^\mathsf{T} r_k$ est non nul, si $M_k \in \mathcal{S}^n_{++}$ et si $\eta_1 \in]0,1[$, alors*

1) *$s_k(\lambda) \neq 0$ pour tout $\lambda \geqslant 0$,*
2) *$s_k(\lambda) \to 0$ lorsque $\lambda \to \infty$,*
3) *$s_k(\lambda)/\|s_k(\lambda)\| \to -M_k^{-1} g_k / \|M_k^{-1} g_k\|$ lorsque $\lambda \to \infty$,*
4) *(19.36) est vérifiée pour tout λ suffisamment grand.*

Démonstration. 1) Selon (19.32), $s_k(\lambda)$ est non nul parce que $g_k \neq 0$.

2) La convergence de $s_k(\lambda)$ vers zéro est due au fait que $s_k(\lambda)$ minimise $\varphi_k(\cdot, \lambda)$ et donc

$$0 \leqslant \frac{\lambda}{2} s_k(\lambda)^\mathsf{T} M_k s_k(\lambda) \leqslant \varphi_k(s_k(\lambda), \lambda) \leqslant \varphi_k(0, \lambda) = f(x_k).$$

En divisant certains membres par $\lambda > 0$ et en faisant tendre $\lambda \to \infty$, on voit que $s_k(\lambda)^\mathsf{T} M_k s_k(\lambda) \to 0$ et donc $s_k(\lambda) \to 0$, car $M_k \in \mathcal{S}^n_{++}$.

3) D'après (19.32) et $s_k(\lambda) \to 0$, on voit que $\lambda M_k s_k(\lambda) \to -g_k$, ce qui implique que $\lambda s_k(\lambda) \to -M_k^{-1} g_k$ et donc $s_k(\lambda)/\|s_k(\lambda)\| \to -M_k^{-1} g_k / \|M_k^{-1} g_k\|$.

4) On raisonne par l'absurde en supposant que (19.36) n'est pas vérifiée pour une suite de $\lambda \to \infty$. Alors, pour ces $\lambda \to \infty$, on a

$$\frac{f(x_k + s_k(\lambda)) - f(x_k) - g_k^T s_k(\lambda)}{\|s_k(\lambda)\|} > (1 - \eta_1) \frac{-g_k^T s_k(\lambda)}{\|s_k(\lambda)\|}.$$

Par la différentiabilité de f en x_k, le membre de gauche tend vers zéro. Par le point 3, le membre de droite tend vers $(1-\eta_1) g_k^T M_k^{-1} g_k / \|M_k^{-1} g_k\|$, qui est strictement positif. On a obtenu la contradiction souhaitée. □

Convergence globale et complexité itérative

On pourrait penser que le surcoût de l'algorithme 19.10, dû au besoin de devoir résoudre un nouveau système linéaire (19.37) chaque fois que l'inégalité de décroissance suffisante (19.38) n'est pas vérifiée, est trop important et qu'il serait préférable de faire de la recherche linéaire bien moins coûteuse le long de la première direction $s_{k,0}$ calculée si celle-ci n'est pas acceptée par (19.38). C'est ce que propose de faire certains auteurs. Cependant, l'algorithme décrit a l'avantage d'avoir un résultat de convergence globale, sans requérir l'injectivité uniforme des J_k, alors que celle-ci est requise par l'algorithme de Gauss-Newton (théorème 19.9). Seule intervient la bornitude de la suite $\{(J_k, \lambda_k M_k)\}$ (la bornitude de $\{\lambda_k M_k\}$ peut-être contrôlée par un choix d'implémentation de l'algorithme, mais pas celle de $\{J_k\}$ qui dépend du problème considéré). Comme annoncé dans la description de l'algorithme, sans hypothèse supplémentaire, sa convergence s'exprime en termes du gradient $g_k = J_k^T r_k$.

Le résultat qui demande le moins d'hypothèse est le suivant. Il suppose qu'une suite est générée par l'algorithme de LMM et donc que celui-ci ne trouve pas un point stationnaire de f en un nombre fini d'itérations.

Théorème 19.12 (convergence de l'algorithme LMM) *Supposons que la fonction de moindres-carrés f soit différentiable. Soit $\{(x_k, \lambda_k)\}$ une suite générée par l'algorithme 19.10. Alors,*

1) *$\{f(x_k)\}$ converge,*
2) *pour toute partie infinie $\mathcal{K}$ de $\mathbb{N}$ telle que $\{(J_k, \lambda_k M_k)\}_{k\in\mathcal{K}}$ est bornée, on a $\{J_k^T r_k\}_{k\in\mathcal{K}} \to 0$ lorsque $k \to \infty$ dans $\mathcal{K}$.*

Démonstration. 1) La convergence de la suite $\{f(x_k)\}$ découle de sa décroissance et du fait qu'elle est bornée inférieurement (par zéro).

2) D'après l'inégalité (19.38) et la convergence de $f(x_k)$, on a

$$g_k^T s_k \to 0, \qquad \text{lorsque } k \to \infty. \tag{19.40}$$

Il s'agit à présent de montrer que la convergence dans (19.40) est due à un gradient g_k qui tend vers zéro et pas à la convergence de s_k vers zéro (qui a probablement aussi lieu). D'après (19.32) et la semi-définie positivité de M_k,

$$-g_k^T s_k = \|J_k s_k\|^2 + \lambda_k s_k^T M_k s_k = \|J_k s_k\|^2 + \lambda_k \|M_k^{1/2} s_k\|^2.$$

On déduit alors de (19.40) et de la positivité de λ_k que

$$J_k s_k \to 0 \quad \text{et} \quad \lambda_k^{1/2} M_k^{1/2} s_k \to 0, \qquad \text{lorsque } k \to \infty.$$

Dès lors, si $\{(J_k, \lambda_k M_k)\}_{k\in\mathcal{K}}$ est bornée, on a

$$J_k^T J_k s_k \to 0 \quad \text{et} \quad \lambda_k M_k s_k \to 0, \qquad \text{lorsque } k \to \infty \text{ dans } \mathcal{K}.$$

La définition (19.32) de l'itération implique maintenant que $g_k \to 0$ lorsque $k \to \infty$ dans $\mathcal{K}$. □

Pour avoir des résultats plus forts, on a besoin d'un peu plus de régularité sur f, à savoir le caractère lipschitzien de sa dérivée, et sur les matrices M_k, à savoir leur uniforme définie positivité. On peut alors montrer, dans un premier temps, que la suite $\{\lambda_k\}$ des multiplicateurs est bornée.

Lemme 19.13 (CS pour avoir des multiplicateurs bornés) *Supposons que la fonction de moindres-carrés f soit $\mathcal{C}^{1,1}_L$ et qu'il existe une constante $\beta_{M^{-1}}$ telle que pour tout k on ait $\|M_k^{-1}\| \leqslant \beta_{M^{-1}}$. Alors, pour tout $k \geqslant 1$, on a*

$$\lambda_k \leqslant \beta_\lambda := \max\left(\lambda_0, \frac{\tau_2 \beta_{M^{-1}} L}{1-\eta_1}\right). \tag{19.41}$$

Démonstration. Il suffit de trouver une borne supérieure pour les multiplicateurs λ_k tels que $\hat{\lambda}_k = \lambda_k/\tau_2$ n'a pas été accepté par l'inégalité de décroissance suffisante (19.38). Les autres multiplicateurs sont en effet plus petits qu'un de ces derniers ou plus petit que λ_0. Si l'on note $\hat{s}_k := s_k(\hat{\lambda}_k)$, on a $f(x_k+\hat{s}_k) > f(x_k) + \eta_1 g_k^T \hat{s}_k$ et donc

$$\frac{f(x_k+\hat{s}_k) - f(x_k) - g_k^T\hat{s}_k}{\|\hat{s}_k\|^2} > (1-\eta_1)\frac{-g_k^T\hat{s}_k}{\|\hat{s}_k\|^2}. \tag{19.42}$$

Le théorème des accroissements finis (corollaire C.13) et la L-lipschitzianité de f conduisent à la majoration suivante

$$|f(x_k+\hat{s}_k) - f(x_k) - g_k^T\hat{s}_k| \leqslant \left(\sup_{t\in]0,1[} \|\nabla f(x_k+t\hat{s}_k) - \nabla f(x_k)\|\right)\|\hat{s}_k\| \leqslant L\|\hat{s}_k\|^2.$$

Dès lors, le membre de gauche de l'inégalité (19.42) est majoré par L. Le membre de droite de cette même inégalité (19.42) peut être minoré en observant que $(J_k^T J_k + \hat{\lambda}_k M_k)\hat{s}_k = -g_k$ et donc que

$$-g_k\hat{s}_k = \hat{s}_k^T(J_k^T J_k + \hat{\lambda}_k M_k)\hat{s}_k \geqslant \hat{\lambda}_k \beta_{M^{-1}}^{-1}\|\hat{s}_k\|^2.$$

L'inégalité (19.42) permet donc d'écrire

$$L > (1-\eta_1)\beta_{M^{-1}}^{-1}\hat{\lambda}_k,$$

ce qui donne la majoration (19.41), puisque $\hat{\lambda}_k = \lambda_k/\tau_2$. □

On aurait pu affaiblir une hypothèse du résultat précédent en n'exigeant le caractère $\mathcal{C}^{1,1}_L$ de f que sur un voisinage suffisamment étendu de l'ensemble de sous-niveau $\mathcal{N}_1 := \{x \in \mathbb{R}^n : f(x) \leqslant f(x_1)\}$, auquel appartiennent les itérés x_k, pour que tous les points rejetés $x_k + \hat{s}_k$ (voir la démonstration) y soient contenus. Ce soin apporté à la démonstration est présent dans [596].

En bref, le résultat suivant nous apprend que, sans avoir besoin de l'injectivité uniforme de $\{J_k\}$, ce qui est à comparer avec le théorème 19.9 sur la convergence de l'algorithme de Gauss-Newton, la complexité itérative de l'algorithme 19.10 de LMM est au pire en $O(\varepsilon^{-2})$, qui est la complexité itérative des algorithmes du gradient (proposition 6.8) et de Gauss-Newton (théorème 19.9), ce qui veut dire que l'on peut obtenir un gradient g_k de norme inférieure à un seuil $\varepsilon > 0$ arbitraire en moins de $O(\varepsilon^{-2})$ itérations. C'est une borne très grande, mais elle a l'intérêt de ne pas dépendre de la dimension du problème. Lorsque la suite générée $\{J_k^T\}$ est uniformément injective, la situation devient beaucoup plus favorable, puisqu'alors la suite $\{f(x_k)\}$ converge linéairement vers zéro et la complexité itérative est en $O(\log \varepsilon^{-1})$.

Théorème 19.14 (complexité itérative de l'algorithme LMM) *Supposons que la fonction de moindres-carrés f soit $\mathcal{C}^{1,1}$. Soit $\{(x_k, \lambda_k)\}$ une suite générée par l'algorithme 19.10. On suppose en outre que la suite $\{(J_k, M_k, M_k^{-1})\}$ est bornée. Alors,*

1) *il existe une constante C, indépendante de n et m, telle que, pour tout $\varepsilon > 0$, $\|J_k^T r_k\| \leqslant \varepsilon$ pour un indice k inférieur à $\lceil C\varepsilon^{-2} \rceil$,*
2) *si, de plus, $\{J_k^T\}$ est uniformément injective, alors*
   a) *$f(x_k) \to 0$ linéairement,*
   b) *il existe une constante $C' > 0$, indépendante de n et m, telle que, pour tout $\varepsilon > 0$, $\|r_k\| \leqslant \varepsilon \|r_0\|$ pour tout indice $k \geqslant \lceil C' \log \varepsilon^{-1} \rceil$.*

Démonstration. 1) En sommant les K premières inégalités (19.38), on obtient

$$-\eta_1 \sum_{k=1}^{K} g_k^T s_k \leqslant f(x_1) - f(x_{K+1}) \leqslant f(x_1) - f_*.$$

où $f_* := \min_k f(x_k)$ est positif. Comme dans la démonstration du théorème 19.12, $-g_k^T s_k = \|J_k s_k\|^2 + \lambda_k \|M_k^{1/2} s_k\|^2$ et donc, par la positivité de λ_k, l'inégalité précédente apporte les majorations suivantes

$$\sum_{k=1}^{K} \|J_k s_k\|^2 \leqslant \frac{f(x_1) - f_*}{\eta_1} \qquad \text{et} \qquad \sum_{k=1}^{K} \lambda_k \|M_k^{1/2} s_k\|^2 \leqslant \frac{f(x_1) - f_*}{\eta_1}.$$

Soient β_J un majorant de $\{\|J_k\|\}$, β_M un majorant de $\{\|M_k\|\}$ et β_λ un majorant de $\{\|\lambda_k\|\}$, que existent par hypothèse et par le lemme 19.13 (ces majorants ne dépendent pas n et m). Il vient

$$\sum_{k=1}^{K} \|J_k^T J_k s_k\|^2 \leqslant \frac{\beta_J^2[f(x_1)-f_*]}{\eta_1} \qquad \text{et} \qquad \sum_{k=1}^{K} \lambda_k^2 \|M_k s_k\|^2 \leqslant \frac{\beta_\lambda \beta_M [f(x_1)-f_*]}{\eta_1}.$$

Par ailleurs,

$$\begin{aligned} \|g_k\|^2 &= \|J_k^T J_k s_k + \lambda_k M_k s_k\|^2 \qquad [(19.32)] \\ &\leqslant \left(\|J_k^T J_k s_k\| + \lambda_k \|M_k s_k\|\right)^2 \qquad [\text{inégalité triangulaire}] \\ &\leqslant 2\left(\|J_k^T J_k s_k\|^2 + \lambda_k^2 \|M_k s_k\|^2\right) \qquad [(a+b)^2 \leqslant 2(a^2+b^2)]. \end{aligned} \tag{19.43}$$

Avec les majorations précédentes, on trouve que

$$K\left(\min_{k\in[1:K]} \|g_k\|^2\right) \leqslant \sum_{k=1}^{K} \|g_k\|^2 \leqslant 2(\beta_J^2 + \beta_\lambda \beta_M)\frac{f(x_1)-f_*}{\eta_1}.$$

Dès lors, $\min\{\|g_k\| : k \in [1:K]\}$ est plus petit qu'un $\varepsilon > 0$ arbitraire donné, dès que le membre de droite ci-dessus est inférieur à $K\varepsilon^2$, ce qui s'écrit

$$K \geqslant K_\varepsilon := \left\lceil 2\varepsilon^{-2}(\beta_J^2 + \beta_\lambda \beta_M)\frac{f(x_1)-f_*}{\eta_1} \right\rceil.$$

On a donc montré qu'un des $\|g_k\|$, avec $k \leqslant K_\varepsilon$, est inférieur à ε.

2) On part de (19.38) :

$$f(x_{k+1}) \leqslant f(x_k) + \eta_1 g_k^T s_k \leqslant f(x_k) - C_1 \|g_k\|^2, \tag{19.44}$$

où la dernière inégalité provient de

$$\begin{aligned} &\|g_k\|^2 \\ &\quad\leqslant 2\left(\|J_k^T J_k s_k\|^2 + \lambda_k^2 \|M_k s_k\|^2\right) \qquad [(19.43)] \\ &\quad\leqslant 2\beta_J^2 \|J_k s_k\|^2 + 2\beta_\lambda \beta_M \lambda_k s_k^T M_k s_k \qquad [\|J_k^T\| \leqslant \beta_J,\ \|\lambda_k\| \leqslant \beta_\lambda,\ \|M_k\| \leqslant \beta_M] \\ &\quad\leqslant (\eta_1/C_1)(\|J_k s_k\|^2 + \lambda_k s_k^T M_k s_k) \qquad [C_1 := \eta_1 \max(2\beta_J^2, 2\beta_\lambda\beta_M)^{-1}] \\ &\quad= -(\eta_1/C_1) g_k^T s_k \qquad [(19.32)]. \end{aligned}$$

On poursuit comme dans la démonstration du point 3 du théorème 19.9 : $\|g_k\|^2 = \|J_k^T r_k\|^2 \geqslant (\alpha'_J)^2 \|r_k\|^2 = 2(\alpha'_J)^2 f(x_k)$ par l'injectivité uniforme de J_k^T, si bien que (19.44) devient

$$f(x_{k+1}) \leqslant (1 - C_2) f(x_k),$$

pour la constante strictement positive $C_2 = 2C_1(\alpha'_J)^2$. Ceci montre la convergence linéaire de $f(x_k)$ vers zéro (point 2.a). Puis par récurrence,

$$\frac{f(x_k)}{f(x_0)} \leqslant (1 - C_2)^k.$$

Dès lors, $\|r_k\| \leqslant \varepsilon \|r_0\|$ si $f(x_k)/f(x_0) \leqslant \varepsilon^2$ et donc certainement si le membre de droite ci-dessus est inférieur à ε^2 :

$$(1 - C_2)^k \leqslant \varepsilon^2.$$

En prenant le logarithme des deux membres et en utilisant $\log(1-C_2)<0$, on obtient comme condition sur k :

$$k \geqslant \frac{2\log \varepsilon^{-1}}{\log(1-C_2)^{-1}}.$$

On obtient donc le résultat du point 2.b avec la constante $C' := 2/\log(1-C_2)^{-1}$, qui est indépendante de n et m. □

19.4 Recherche de solution parcimonieuse ▲

19.4.1 Motivation

Supposons que l'on dispose d'une mesure $y := Ms + x$ (M est une matrice $n \times p$) d'un signal $s \in \mathbb{R}^p$, qui est perturbée sur *quelques* composantes par le vecteur $x \in \mathbb{R}^n$, lequel a donc beaucoup de composantes nulles. Dans ce cadre, M est injective : pour que l'opération de récupération de s à partir de y ait quelques chances de succès, il faut qu'il y ait plus de mesures que de composantes au signal recherché. Dans ce cadre, l'on dispose aussi d'un annihilant à gauche A de M : c'est une matrice $m \times n$ telle que $AM = 0$ (par exemple, A peut être le projecteur orthogonal sur $\mathcal{R}(A)^\perp$). Alors on connaît $b := Ay = Ax$ et on cherche à écarter les mesures i corrompues par les perturbations $x_i \neq 0$ en résolvant

$$\begin{cases} \inf \ \|x\|_0 \\ Ax = b, \end{cases} \tag{19.45}$$

où $\|x\|_0 := |\{i \in [1:n] : x_i \neq 0\}|$ est le *compteur de composante non nulle* de x, c'est-à-dire le nombre de ses composantes non nulles (ce n'est pas une norme, par manque de positivité homogène). Ce problème est NP-ardu [448 ; 1995].

Voir la suite, par exemple, chez Candès et Tao [105 ; 2005].

Problème d'*acquisition comprimée* (*compressed sensing*).

19.4.2 Poursuite de base

La *poursuite de base* (de l'anglais *basis pursuit*) est une technique d'optimisation utilisée initialement en *traitement du signal* qui revient à résoudre un problème de la forme

$$(P_1) \quad \begin{cases} \min \ \|x\|_1 \\ Ax = b, \end{cases} \tag{19.46}$$

où l'inconnue est un vecteur $x \in \mathbb{R}^n$, $\|\cdot\|_1$ est la norme ℓ_1, A est une matrice réelle $m \times n$ et $b \in \mathbb{R}^m$. Il s'agit donc de trouver le plus petit vecteur $x \in \mathbb{R}^n$, au sens de la norme ℓ_1, qui vérifie les contraintes affines $Ax = b$. Ce problème est convexe, mais non lisse.

Comme nous le verrons, l'intérêt du problème est de sélectionner une solution du système linéaire $Ax = b$, supposé sous-déterminé (donc avec plus d'une solution), ayant peu d'éléments non nuls. La non-différentiabilité de la norme ℓ_1 joue un rôle-clé dans l'obtention de cette propriété.

L'appellation *poursuite de base* vient de l'algorithme du simplexe qui était proposé dans l'article original [117 ; 1998] pour résoudre le problème ci-dessus, lequel détermine une *base optimale*. Dans la terminologie de cet algorithme, il s'agit d'une sélection de m colonnes de A telle que la sous-matrice B correspondante soit inversible et détermine la solution par $B^{-1}b$.

Motivation

Voici comment on peut être amené à résoudre un problème d'optimisation de la forme (P_1) ci-dessus. Un problème classique en *traitement du signal* consiste à trouver une décomposition *parcimonieuse* (c'est-à-dire formée de peu d'éléments) d'un signal donné dans un *dictionnaire surabondant* de signaux, contenant par exemple des sinusoïdes (décomposition de Fourier), des ondelettes, et bien d'autres signaux. Dans l'écriture ci-dessus, b est le signal à décomposer, les colonnes de A sont les éléments du dictionnaire de signaux et les composantes de x sont les coefficients recherchés pour représenter le signal au moyen des signaux du dictionnaire. On peut donc écrire

$$b = \sum_{j=1}^{m} x_j A^j,$$

où A^j est la colonne j de A. Lorsque le dictionnaire de signaux A^j est *surabondant*, $m > n$ et la décomposition de x comme ci-dessus n'est pas unique. Lorsqu'on cherche une décomposition parcimonieuse, l'on cherche à avoir le moins de coefficients x_j non nuls. C'est ce qui permet d'avoir une représentation compacte du signal (compression de celui-ci).

Annuler le plus de coefficients x_j revient à résoudre le problème

$$(P_0) \quad \begin{cases} \min \ \|x\|_0 \\ Ax = b, \end{cases}$$

où $\|x\|_0 := |\{i \in [1\,{:}\,n] : x_i \neq 0\}|$ est le nombre d'éléments non nuls de x (ce n'est pas une norme, car l'homogénéité n'a pas lieu, mais la limite de $\|x\|_p$ lorsque $p \to 0$, d'où la notation). Ce dernier problème est malheureusement NP-ardu [448], ce qui est aujourd'hui un handicap rédhibitoire lorsqu'on veut résoudre des problèmes de grande taille. Le problème (P_1) peut être vu comme une *approximation traitable* du problème (P_0), pour les raisons suivantes.

- Le problème (P_1) consiste à trouver un point du sous-espace affine $\mathcal{A} := \{x \in \mathbb{R}^n : Ax = b\}$ le plus proche de zéro au sens de la norme ℓ_1. Comme la boule unité B_1 de cette dernière est polyédrique, elle a un nombre fini de sommets et le problème (P_1) a tendance à trouver une solution en un sommet de $\mathrm{val}(P_1)B_1$ (on a noté $\mathrm{val}(P_1)$ la valeur optimale du problème (P_1)) ou sur une face contenant peu de sommets de cette boule. Or les sommets de $\mathrm{val}(P_1)B_1$ sont des multiples des vecteurs de base de $\mathbb{R}^n$, qui ont toutes leurs composantes nulles sauf une ! La solution de (P_1) aura donc tendance à avoir beaucoup d'éléments nuls.
- Par ailleurs, le problème (P_1) est un problème *convexe*, qui peut être récrit comme un problème d'optimisation linéaire (voir ci-dessous) et donc peut être résolu en temps polynomial.

Cette approche de résolution de (P_0) par (P_1) a été proposée par [117 ; 1998].

Analyse du problème

DUALITÉ

Le problème (P_1) s'écrit aussi

$$\inf_{x\in\mathbb{R}^n}\sup_{y\in\mathbb{R}^m}\Big(\|x\|_1 - y^{\mathsf{T}}(Ax-b)\Big).$$

Le dual lagrangien de (P_1) est donc le problème

$$\sup_{y\in\mathbb{R}^m}\inf_{x\in\mathbb{R}^n}\Big(\|x\|_1 - y^{\mathsf{T}}(Ax-b)\Big) = \sup_{y\in\mathbb{R}^m}\left(b^{\mathsf{T}}y - \sup_{x\in\mathbb{R}^n}\big((A^{\mathsf{T}}y)^{\mathsf{T}}x - \|x\|_1\big)\right).$$

En se rappelant l'expression de la conjuguée de la norme (exercice 3.30), le dernier supremum est nul si $\|A^{\mathsf{T}}y\|_\infty \leqslant 1$ et vaut $+\infty$ sinon, si bien que l'on obtient comme problème dual lagrangien

$$(D_1)\quad \begin{cases} \sup\ b^{\mathsf{T}}y \\ \|A^{\mathsf{T}}y\|_\infty \leqslant 1. \end{cases}$$

On a bien sûr le résultat de dualité faible

$$\operatorname{val}(D_1) \leqslant \operatorname{val}(P_1). \tag{19.47}$$

Nous allons montrer qu'il y a dualité forte, dans un sens à préciser, entre le primal (P_1) et le dual (D_1). Une manière de montrer cette dualité forte est de récrire ces problèmes sous la forme des problèmes d'optimisation linéaire (P_1') et (D_1') ci-dessous et d'utiliser la dualité forte entre (P_1') et (D_1') (théorème 17.11). On peut reformuler le problème (P_1) comme un problème d'optimisation linéaire en $(u,v)\in\mathbb{R}^n\times\mathbb{R}^n$ sous forme standard :

$$(P_1')\quad \begin{cases} \min\ e^{\mathsf{T}}u + e^{\mathsf{T}}v \\ \begin{pmatrix} A & -A \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix} = b, \\ u \geqslant 0, \quad v \geqslant 0. \end{cases}$$

où $e\in\mathbb{R}^n$ est le vecteur dont toutes les composantes valent 1. Le lien entre les variables (u,v) de (P_1') et la variable x de (P_1) est $x = u - v$. Quant au problème dual (D_1), on peut l'écrire comme le problème d'optimisation linéaire suivant

$$(D_1')\quad \begin{cases} \sup\ b^{\mathsf{T}}y \\ -e \leqslant A^{\mathsf{T}}y \leqslant e. \end{cases}$$

D'après la section 17.3.1, (D_1') est le dual lagrangien de (P_1'). On déduit alors du théorème 17.11 une partie du résultat de dualité forte suivant. Il y a une différence toutefois : il n'y a ici jamais de saut de dualité (essentiellement parce que le dual est toujours réalisable et sa valeur optimale ne peut donc pas être $-\infty$).

Proposition 19.15 (dualité forte en poursuite de base)

1) *Les propriétés suivantes sont équivalentes :*
   (*i*) (P_1) *est réalisable,*
   (*ii*) (P_1) *a une solution,*
   (*iii*) (D_1) *a une solution.*
2) *Il n'y a pas de saut de dualité : si les propriétés* (*i*)-(*iii*) *ont lieu* $\mathrm{val}(D_1) = \mathrm{val}(P_1) \in \mathbb{R}$, *sinon* $\mathrm{val}(D_1) = \mathrm{val}(P_1) = +\infty$.
3) *Un point* $\bar{x} \in \mathbb{R}^n$ *est solution de* (P_1) *et un point* $\bar{y} \in \mathbb{R}^m$ *est solution de* (D_1) *si, et seulement si,* $(\bar{x}, \bar{y})$ *est un point-selle du lagrangien* $(x, y) \in \mathbb{R}^n \times \mathbb{R}^m \mapsto \|x\|_1 - y^\mathsf{T}(Ax - b)$.

Démonstration. 1) Les équivalences entre (*i*)-(*iii*) se déduisent directement du théorème 17.11 et de l'équivalence entre (P_1) et (P_1'), d'une part, et (D_1) et (D_1'), d'autre part.

2) D'après le théorème 17.11, il n'y a pas de saut de dualité si les propriétés (*i*)-(*iii*) ont lieu et dans ce cas les valeurs optimales sont finies.

Supposons maintenant que les propriétés (*i*)-(*iii*) n'ont pas lieu. Alors (P_1) n'a pas de solution et donc $b \notin \mathcal{R}(A)$ (on le verra au point 1 de la proposition 19.16 ci-dessous, mais c'est évident dès à présent), ce qui revient à dire que $\mathrm{val}(P_1) = +\infty$. Mais si $b \notin \mathcal{R}(A)$, on peut écrire $b = Ax_0 + b_1$ avec b_1 non nul dans $\mathcal{R}(A)^\perp = \mathcal{N}(A^\mathsf{T})$. Alors, quel que soit $t \in \mathbb{R}$, $y_t := tb_1 \in \mathcal{N}(A^\mathsf{T})$ est admissible pour (D_1) et on a $b^\mathsf{T} y_t = t\|b_1\|_2^2 \to +\infty$ lorsque $t \to +\infty$. Donc $\mathrm{val}(D_1) = +\infty$.

3) C'est une conséquence des points 1 et 2 et du théorème 14.3. □

Existence de solution, unicité, Conditions d'optimalité

La condition d'existence de solution de (P_1) découlant du théorème de Weierstrass, de nature topologique, est triviale (point 2 de la proposition 19.16 ci-dessous). Mais le problème (P_1) est convexe, si bien que ses conditions d'existence de solution sont équivalentes à ses conditions d'optimalité ; cette remarque conduit aux conditions d'existence de solution (19.48) qui s'expriment, comme toutes conditions d'optimalité, en termes d'un candidat-solution $\bar{x}$.

On notera que la condition (19.49) prend en compte le cas où $b = 0$. En effet dans ce cas $\bar{x} = 0$ est l'unique solution de (P_1). Les conditions (19.49) n'accepte pas d'utres points, car si $\bar{x} \neq 0$, alors $I \neq \varnothing$ et l'injectivité de A_I et $A_I \bar{x}_I = 0$ impliquerait que $\bar{x}_I = 0$; une contradiction.

Proposition 19.16 (existence, unicité, conditions d'optimalité)
1) *L'ensemble des solution de* (P_1) *est un polyèdre convexe, éventuellement vide.*
2) *Le problème* (P_1) *a une solution si, et seulement si,* $b \in \mathcal{R}(A)$.
3) *Un point* $\bar{x} \in \mathbb{R}^n$ *est solution de* (P_1) *si, et seulement si,*

$$\begin{cases} A\bar{x} = b \\ \text{il existe } y \in \mathbb{R}^m \text{ tel que } A_I^\mathsf{T} y = s \text{ et } \|A_{I^c}^\mathsf{T} y\|_\infty \leqslant 1, \end{cases} \tag{19.48}$$

où $I := \{i \in [1:n] : \bar{x}_i \neq 0\}$ *et* $s := \operatorname{sgn}(\bar{x}_I)$.

4) *Un point* $\bar{x} \in \mathbb{R}^n$ *est solution unique de* (P_1) *si, et seulement si,*

$$\begin{cases} A\bar{x} = b \\ A_I \text{ est injective} \\ \text{il existe } y \in \mathbb{R}^m \text{ tel que } A_I^\mathsf{T} y = s \text{ et } \|A_{I^c}^\mathsf{T} y\|_\infty < 1, \end{cases} \tag{19.49}$$

où $I := \{i \in [1:n] : \bar{x}_i \neq 0\}$ *et* $s := \operatorname{sgn}(\bar{x}_I)$.

Démonstration. 1) L'ensemble des solutions de (P_1) est l'image par l'application linéaire $T : (u, v) \in \mathbb{R}^n \times \mathbb{R}^n \mapsto u - v \in \mathbb{R}^n$ de l'ensemble des solutions du problème d'optimisation linéaire (P_1'), qui est un polyèdre convexe. C'est donc un polyèdre convexe (proposition 2.18).

2) Évidemment, si (P_1) a une solution $\bar{x}$, $b = A\bar{x} \in \mathcal{R}(A)$. Inversement, c'est aussi évident : l'ensemble admissible est non vide (par la condition $b \in \mathcal{R}(A)$) et fermé (c'est un sous-espace affine en dimension finie) ; puis la coercivité du critère assure l'existence d'une solution (proposition 1.4).

3) Le problème (P_1) étant convexe, $\bar{x} \in \mathbb{R}^n$ en est solution si, et seulement si, zéro est dans le sous-différentiel de la fonction

$$x \in \mathbb{R}^n \mapsto \|x\|_1 + \mathcal{I}_{\mathcal{A}}(x), \tag{19.50}$$

où $\mathcal{A} := \{x \in \mathbb{R}^n : Ax = b\}$ (corollaire 3.61). Par la proposition 3.71, ceci peut s'écrire

$$0 \in \partial(\|\cdot\|_1)(\bar{x}) + \partial\mathcal{I}_{\mathcal{A}}(\bar{x}).$$

Forcément, $\bar{x} \in \mathcal{A}$, sinon $\mathrm{N}_{\mathcal{A}}(\bar{x}) = \varnothing$. Par ailleurs, $\partial(\|\cdot\|_1)(\bar{x}) = \{z \in \mathbb{R}^n : \|z\|_\infty \leqslant 1, z^\mathsf{T}\bar{x} = \|\bar{x}\|_1\} = \{z \in \mathbb{R}^n : z_I = s \text{ et } \|z_{I^c}\|_\infty \leqslant 1\} =: D$ (exercice 3.30) et $\partial\mathcal{I}_{\mathcal{A}}(\bar{x}) = \mathrm{N}_{\mathcal{A}}(\bar{x})$ (exercice 3.28). On voit facilement que $\mathrm{N}_{\mathcal{A}}(\bar{x}) = \mathcal{N}(A)^\perp = \mathcal{R}(A^\mathsf{T})$, si bien que l'élément de $\partial(\|\cdot\|_1)(\bar{x})$ dans la condition d'optimalité ci-dessus est de la forme $A^\mathsf{T}\bar{y}$ avec $\bar{y} \in \mathbb{R}^m$. Dès lors (19.48) a lieu.

4) La fonction (19.50) étant polyédrique, elle a un minimiseur unique si, et seulement si,

$$0 \in (\partial(\|\cdot\|_1)(\bar{x}) + \partial\mathcal{I}_{\mathcal{A}}(\bar{x}))^\circ = \big(D + \mathcal{R}(A^\mathsf{T})\big)^\circ,$$

où on a utilisé les expressions de $\partial(\|\cdot\|_1)(\bar{x})$ et $\partial\mathcal{I}_{\mathcal{A}}(\bar{x})$ déterminées dans la démonstration du point 3. Ceci peut s'exprimer par les deux conditions suivantes :

$$\operatorname{aff}\big(D + \mathcal{R}(A^\mathsf{T})\big) = \mathbb{R}^n, \tag{19.51}$$

$$0 \in D^\ominus + \mathcal{R}(A^\mathsf{T}). \tag{19.52}$$

La condition (19.51) s'exprime aussi par $\operatorname{aff} D + \mathcal{R}(A^\mathsf{T}) = \mathbb{R}^n$ ou $\{z \in \mathbb{R}^n : z_I = s\} + \mathcal{R}(A^\mathsf{T}) = \mathbb{R}^n$. Il n'est pas difficile de montrer par l'algèbre linéaire que cela équivaut à l'injectivité de A_I. La condition (19.52) est équivalente à l'existence d'un $y \in \mathbb{R}^m$ tel que $A^\mathsf{T} y \in D^\ominus = \{z \in \mathbb{R}^n : z_I = s \text{ et } \|z_{I^c}\|_\infty < 1\}$. On a donc montré que (19.49) équivaut au fait que $\bar{x}$ est la solution unique de (P_1). □

19.4.3 Problème LASSO ▲

Le *problème* LASSO (pour *Least Absolute Shrinkage and Selection Operator*) consiste à trouver un vecteur $x \in \mathbb{R}^n$ solution de

$$\begin{cases} \min_x \frac{1}{2}\|Ax-b\|_2^2 \\ \|x\|_1 \leqslant r, \end{cases} \tag{19.53}$$

où $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $r > 0$, et $\|\cdot\|_p$ désigne la norme ℓ_p. Ce problème est convexe (critère et contrainte convexe), mais l'on prendra garde à la non-différentiabilité de sa contrainte.

Le problème LASSO, dont le nom imagé a fait florès et a sans doute contribué à la notoriété de son initiateur [583], est utilisé (en statistiques, en fouille de données,...) pour résoudre approximativement des problèmes de moindres-carrés linéaires par des vecteurs ayant beaucoup de composantes nulles (c'est la raison d'être de la norme ℓ_1 en contrainte) [466, 467, 584, 585]. L'approche est surtout utile lorsque $m \ll n$ (il y a alors beaucoup de x vérifiant $Ax = b$ approximativement et la norme ℓ_1 est utilisée pour sélectionner ceux ayant beaucoup de composantes nulles), alors que le problème de moindres-carrés linéaire classique (19.1) (sans la contrainte en norme ℓ_1 dans (19.53)) est surtout utilisé lorsque $m \gg n$ (le plus souvent, il n'y a alors pas de x vérifiant $Ax = b$).

Commençons par des propriétés de la norme ℓ_1.

Lemme 19.17 (norme ℓ_1) *Le sous-différentiel $S(x) := \partial(\|\cdot\|_1)(x)$ de la norme ℓ_1 en $x \in \mathbb{R}^n$ s'écrit*

$$\begin{aligned} S(x) = \{x^* \in \mathbb{R}^n : {} & x_i^* = -1 \text{ si } x_i < 0, \\ & x_i^* \in [-1,+1] \text{ si } x_i = 0, \\ & x_i^* = +1 \text{ si } x_i > 0\}. \end{aligned} \tag{19.54}$$

Le cône tangent $\mathrm{T}_{rB_1}(x)$ à la boule rB_1 de rayon $r > 0$ pour la norme ℓ_1 en un point x de norme $\|x\|_1 = r$ ne dépend pas de r et est le cône dual négatif de $S(x)$:

$$\mathrm{T}_{rB_1}(x) = S(x)^-. \tag{19.55}$$

DÉMONSTRATION. Le sous-différentiel $S(x)$ est calculé à l'exercice 3.30.

Pour démontrer (19.55), il est utile d'introduire

$$\begin{aligned} S_0(x) = \{x^* \in \mathbb{R}^n : {} & x_i^* = -1 \text{ si } x_i < 0, \\ & x_i^* = \pm 1 \text{ si } x_i = 0, \\ & x_i^* = +1 \text{ si } x_i > 0\}, \end{aligned}$$

Comme la boule unité pour la norme ℓ_∞ (dont les éléments ont leurs composantes dans $[-1,1]$) est l'enveloppe convexe de ses sommets (dont les composantes valent ± 1), on a $S(x) = \operatorname{co} S_0(x)$. Par ailleurs, la boule rB_1 est un polyèdre convexe qui s'écrit $rB_1 = \{x \in \mathbb{R}^n : Ex \leqslant r\}$, où la matrice E a ses 2^n lignes formées de tous les vecteurs

différents de $\mathbb{R}^n$ dont les composantes valent ± 1. D'après l'exercice 2.43, si l'on note $I := \{i : (Ex)_i = r\}$, son cône tangent s'écrit

$$\begin{aligned} \mathrm{T}_{rB_1}(x) &= \{d \in \mathbb{R}^n : (Ed)_I \leqslant 0\} \\ &= \{d \in \mathbb{R}^n : s^{\mathsf{T}} d \leqslant 0,\ \forall\, s \in S_0(x)\} \\ &= S(x)^-. \end{aligned}$$

Pour la dernière égalité, on a utilisé le fait que le dual négatif de $S_0(x)$ est égal à celui de son enveloppe convexe $S(x)$ (point 3 de la proposition 2.43). □

Exhibons maintenant des conditions d'existence et d'unicité de solution du problème LASSO. Dans leur expression, on utilise le rayon r_0 défini comme suit :

$$X_0 := \arg\min\{\|Ax - b\|_2 : x \in \mathbb{R}^n\} \qquad \text{et} \qquad r_0 := \inf\{\|x\|_1 : x \in X_0\}.$$

Donc X_0 est l'ensemble des minimiseurs sans contrainte (il est non vide par la proposition 19.1) et r_0 est la distance de zéro à X_0 pour la norme ℓ_1 (elle est donc finie).

Proposition 19.18 (existence et unicité de solution) *Le problème* LASSO (19.53) *admet une solution. Une telle solution $\bar{x}$ est unique si, et seulement si, l'une des deux conditions suivantes est vérifiée*
(*i*) *A est injective,*
(*ii*) *$r \leqslant r_0$ et $\mathcal{N}(A) \cap S(\bar{x})^+ = \{0\}$.*

DÉMONSTRATION. Le problème a une solution par le théorème de Weierstrass (théorème 1.2), parce que le critère est continu et l'ensemble admissible est non vide et compact. Soit $\bar{x}$ une solution du problème.

[⇒] Supposons $\bar{x}$ soit l'unique solution du problème. Montrons dans un premier temps que l'on a nécessairement

$$\mathcal{N}(A) \cap S(\bar{x})^+ = \{0\}. \tag{19.56}$$

Soit $h \in \mathcal{N}(A) \cap S(\bar{x})^+$. D'après (19.55), $-h \in \mathcal{N}(A) \cap \mathrm{T}_{\|x\|_1 B_1}(x)$ et donc pour $\alpha > 0$ petit, $\bar{x}_\alpha := \bar{x} - \alpha h$ vérifie $\|A\bar{x}_\alpha - b\|_2 = \|A\bar{x} - b\|_2$ et, par la polyédricité de $\|x\|_1 B_1$, $\|\bar{x}_\alpha\|_1 \leqslant \|\bar{x}\|_1 \leqslant r$. Donc $\bar{x}_\alpha$ est aussi solution. L'unicité de celle-ci implique que $h = 0$.

On peut maintenant établir qu'une des conditions (*i*) ou (*ii*) a lieu. Si (*i*) n'a pas lieu, alors nécessairement $r \leqslant r_0$ (sinon il existe des solutions $\bar{x}'$ de norme $r_0 < r$ et $\bar{x}' + \alpha h$ est encore solution si $h \in \mathcal{N}(A) \setminus \{0\}$ et α est proche de zéro) et donc (*ii*) a lieu. Si (*ii*) n'a pas lieu, alors nécessairement $r > r_0$ (puisque (19.56) est vérifiée) et donc A doit être injective pour que l'unicité ait lieu (même raisonnement que précédemment).

[⇐] Réciproquement, si (*i*) a lieu, le critère est strictement convexe et l'unicité de solution s'en déduit.

Supposons maintenant que (*ii*) ait lieu et que $\bar{x}'$ soit une solution du problème. Il suffit de montrer que $h := \bar{x} - \bar{x}'$ est dans $\mathcal{N}(A) \cap S(\bar{x})^+$, ce qui impliquera que $h = 0$ et donc que $\bar{x}$ est la seule solution.

- Montrons que $h \in \mathcal{N}(A)$. Par convexité de l'ensemble des solutions, $\bar{x}_\alpha := \bar{x} - \alpha h$ est aussi solution pour tout $\alpha \in [0,1]$. De l'égalité $\|A\bar{x} - b\|_2 = \|A\bar{x}_\alpha - b\|_2$, on obtient $\|A\bar{x} - b\|_2^2 = \|(A\bar{x} - b) - \alpha Ah\|_2^2$, qui après développement donne

$$-2\alpha h^\mathsf{T} A^\mathsf{T}(A\bar{x} - b) + \alpha^2 \|Ah\|_2^2 = 0.$$

  Comme cette identité est vraie pour tout $\alpha \in [0,1]$, on en déduit que $h \in \mathcal{N}(A)$.
- Montrons que $h \in S(\bar{x})^+$. Dans ce but, on observe que

$$\|\bar{x}\|_1 = \|\bar{x}'\|_1 = r. \tag{19.57}$$

  En effet, si $\|\bar{x}\|_1 < r$, $\bar{x}$ est solution sans contrainte de (19.53) et donc $\bar{x} \in X_0$, ce qui implique que $\|\bar{x}\|_1 \geqslant r_0$, en contradiction avec $\|\bar{x}\| \leqslant r < r_0$; pour la même raison $\|\bar{x}'\| = r$. Alors $\bar{x}$ et $\bar{x}' \in rB_1$, si bien que $-h \in \mathrm{T}_{rB_1}(\bar{x}) = -S(\bar{x})^+$, par (19.55). □

Proposition 19.19 (conditions d'optimalité) *Soit $\bar{x} \in \mathbb{R}^n$. Les conditions suivantes sont équivalentes*

(i) $\bar{x}$ est solution du problème LASSO (19.53),
(ii) $\exists\, s \in S(\bar{x})$ tel que $-A^\mathsf{T}(A\bar{x} - b) = \|A^\mathsf{T}(A\bar{x} - b)\|_\infty s$,
(iii) $-(A\bar{x} - b)^\mathsf{T} A\bar{x} = \|A^\mathsf{T}(A\bar{x} - b)\|_\infty \Delta$.

Démonstration. □

Montrons que l'on peut prendre comme dual lagrangien du problème LASSO, le problème

$$\sup_{z \in \mathbb{R}^n} \; -\left(\frac{1}{2}\|Az - b\|_2^2 + \Delta \, \|A^\mathsf{T} Az\|_\infty\right). \tag{19.58}$$

Notes

La méthode des moindres-carrés pour l'identification de paramètres a été proposée par Gauss. Cette approche lui permit de déterminer l'orbite de la planète Cérès en 1801 [62], découverte fortuitement la même année par Piazzi. Cette méthode fut introduite indépendamment par Legendre dans ses « *Nouvelles méthodes pour la détermination des orbites des comètes* » en 1806 et celui-ci a un temps revendiqué cette découverte. Il s'est finalement avéré que Gauss fut le premier à en avoir fait l'exposition, mais dans des notes non publiées datant de 1799. La première référence de Gauss traitant de ce sujet est sa « *Theoria motus corporum coelestium* » de 1809 [277]. D'autres éléments historiques sont relatés par Goldstine [265 ; 1977].

Les problèmes de moindres-carrés linéaires et leurs techniques de résolution sont passés en revue dans les livres et monographies de Hanson et Lawson [304] et de Björck [61, 62]. Autres ouvrages sur l'estimation de paramètres par moindres-carrés :

Bard [37 ; 1974] ; Arnold et Beck [22 ; 1977]. Mentionnons aussi la technique de régularisation de ces problèmes due à Tikhonov [586 ; 1963]. Les *problèmes de moindres-carrés linéaires totaux* sont passés en revue par Van Huffel et Vandewalle [322].

Le résultat de complexité en $O(\log \varepsilon^{-1})$ des théorèmes 19.9 et 19.14 a été obtenu en reprenant un argument de [597] pour l'algorithme de Levenberg-Morrison-Marquardt, mais qui peut aussi s'utiliser pour l'analyse de l'algorithme de Gauss-Newton.

L'algorithme de Levenberg-Morrison-Marquardt (LMM) présenté et étudié à la section 19.3.3 a été proposé par Levenberg [389 ; 1944], avec des notations qui rendent son texte difficile à suivre aujourd'hui ; l'article ne donne pas de résultat de convergence. Sans référence à Levenberg, ni à quiconque d'ailleurs, Morrison [443 ; 1960] introduit un algorithme similaire en raisonnant à partir de ce que l'on appelle aujourd'hui la notion de *région de confiance* (chapitre 9) ; il en donne quelques propriétés mais pas de résultat de convergence. L'approche fut ensuite redécouverte et analysée par Marquardt [407 ; 1963]. La détermination du paramètre de pénalisation présentée ici est due à Osborne [465 ; 1976]. Le résultat de convergence globale (théorème 19.12) que nous donnons est plus fort que celui travaillé dans [465], car il ne requiert pas la bornitude de $\{\lambda_k\}$, ni d'hypothèse entraînant celle-ci. La complexité itérative de l'algorithme de LMM a été étudiée dans [596, 597] ; l'analyse que nous en donnons dans la démonstration du théorème 19.14 nous semble plus rapide et est directement inspirée de celle sur la complexité itérative de l'algorithme du gradient (proposition 6.8). Pour d'autres approches et contributions, voir [437, 643] et leurs références.

Exercices

19.1. *Problème de moindres-carrés linéaire régularisé.* Soient $\mathbb{E}$ et $\mathbb{F}$ deux espaces vectoriels normés de dimension finie et de normes notées $\|\cdot\|_{\mathbb{E}}$ et $\|\cdot\|_{\mathbb{F}}$ respectivement, $A : \mathbb{E} \to \mathbb{F}$ une application linéaire et $b \in \mathbb{F}$. On considère le problème

$$(P) \quad \inf_{x \in \mathbb{E}} \|Ax - b\|_{\mathbb{F}}$$

et sa version régularisée

$$(P_\varepsilon) \quad \inf_{x \in \mathbb{E}} \|Ax - b\|_{\mathbb{F}}^{\alpha} + \varepsilon \|x\|_{\mathbb{E}}^{\beta},$$

où $\alpha > 0$, $\beta > 0$ et $\varepsilon > 0$.

1) (P) a une solution.
2) (P_ε) a une solution. On en sélectionne une, notée $\bar{x}_\varepsilon$.
3) Lorsque $\varepsilon \to \infty$, $\bar{x}_\varepsilon \to 0$.
4) Lorsque $\varepsilon \downarrow 0$, $\|\bar{x}_\varepsilon\|_{\mathbb{E}}$ croît ; la suite $\{\bar{x}_\varepsilon\}_{\varepsilon \downarrow 0}$ est bornée et ses points d'adhérence sont solutions de $\min\{\|x\|_{\mathbb{E}} : x \in S\}$, où S est l'ensemble des solutions de (P).
5) Si $\mathbb{E}$ et $\mathbb{F}$ sont euclidiens, si $\|\cdot\|_{\mathbb{E}}$ et $\|\cdot\|_{\mathbb{F}}$ sont associées à un produit scalaire, si $\alpha = \beta = 2$, alors
   - (P_ε) a une solution unique,
   - $\varepsilon \bar{x}_\varepsilon$ converge vers $A^* b$ si $\varepsilon \to \infty$ (A^* est l'opérateur adjoint de $A : \mathbb{E} \to \mathbb{F}$) et si, de plus, $A^* b \neq 0$, alors $\bar{x}_\varepsilon / \|\bar{x}_\varepsilon\| \to A^* b / \|A^* b\|$ (où $\|\cdot\|$ est une norme quelconque),
   - $\bar{x}_\varepsilon$ converge vers la solution de norme minimale de (P) si $\varepsilon \downarrow 0$.

19.2. *Problème de moindres-carrés linéaire sous région de confiance.* Soient $A \in \mathbb{R}^{m\times n}$ une matrice, $b \in \mathbb{R}^m$ et $\Delta > 0$. On considère le problème suivant

$$\begin{cases} \min \|Ax - b\| \\ \|x\| \leqslant \Delta. \end{cases} \tag{19.59}$$

dans lequel $\|\cdot\|$ désigne la norme euclidienne sur $\mathbb{R}^m$ ou $\mathbb{R}^n$.

1) Montrez que le problème (19.59) a une solution.
2) *Justifiez* et *écrivez* les conditions d'optimalité de Karush, Kuhn et Tucker de (19.59) (de manière à rendre le calcul plus aisé, on pourra élever certaines quantités au carré). Montrez que ces conditions sont nécessaires et suffisantes pour l'optimalité.
3) Montrez que l'on peut trouver une solution de (19.59) qui soit dans $\mathcal{R}(A^\mathsf{T})$.
4) Montrez que (19.59) a une unique solution si, et seulement si,

$$\bar{B}(0,\Delta) \cap \{x : A^\mathsf{T}(Ax - b) = 0\} \text{ a au plus un élément.}$$

On a noté $\bar{B}(0,\Delta)$ la boule fermée de centre 0 et de rayon Δ pour la norme euclidienne.

19.3. *Approximation de Tchebychev d'un système linéaire surdéterminé* [86 ; ex. 5.6]. Soient A une matrice de type $m \times n$ et $b \in \mathbb{R}^m$. On considère le problème d'*approximation au sens de Tchebychev* ou en norme ℓ_∞ suivant :

$$v_{\text{tch}} := \min_{x\in\mathbb{R}^n} \|Ax - b\|_\infty. \tag{19.60}$$

Le problème (19.60) n'a pas nécessairement une unique solution, mais il en a au moins une, d'où l'utilisation du « min ».

1) Montrez que le problème (19.60) a une solution.

On ne connait pas d'expression analytique de v_{tch}, ni a fortiori l'expression analytique des solutions de (19.60), alors que les solutions du problème de moindres-carrés sont connues. En particulier, on connait la forme de ses solutions x_{mc} et son résidu optimal

$$r_{\text{mc}} := Ax_{\text{mc}} - b.$$

Il est donc naturel de chercher à savoir si ce résidu permet d'estimer v_{tch} et plus précisément d'en donner une borne inférieure (positive bien sûr).

2) Montrez que

$$\frac{1}{\sqrt{m}} \|r_{\text{mc}}\|_\infty \leqslant v_{\text{tch}}. \tag{19.61}$$

On montre maintenant que la dualité permet de resserrer cette estimation de v_{tch}.

3) Montrez dans quel sens on peut considérer que le problème

$$\max_{\substack{y\in\mathbb{R}^m \\ \|y\|_1\leqslant 1 \\ A^\mathsf{T}y=0}} b^\mathsf{T}y \tag{19.62}$$

est dual du problème (19.60). Montrez que (19.62) a une solution.
4) On suppose ici que $r_{\text{mc}} \neq 0$ (dans le cas contraire, $v_{\text{tch}} = 0$ et il n'y a donc pas lieu de trouver un minorant strictement positif de v_{tch}). En choisissant bien un point admissible du problème dual (19.62), montrez que l'on a

$$\frac{\|r_{\text{mc}}\|_2^2}{\|r_{\text{mc}}\|_1} \leqslant v_{\text{tch}}.$$

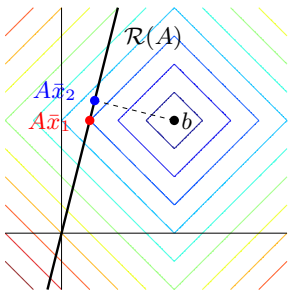
Montrez que ce minorant est meilleur que celui donné en (19.61).

19.4. *Meilleure résolution d'un système linéaire non réalisable.* On considère le problème d'optimisation en $x \in \mathbb{R}^n$ suivant

$$(P) \qquad \inf_{x\in\mathbb{R}^n} \|Ax-b\|,$$

où $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ et $\|\cdot\|$ est une norme *quelconque* sur $\mathbb{R}^m$. On note $f : \mathbb{R}^n \to \mathbb{R}$ la fonction *convexe* définie en $x \in \mathbb{R}^n$ par $f(x) = \|Ax-b\|$.

1) Montrez que le sous-différentiel de f en $x \in \mathbb{R}^n$, pour le produit scalaire euclidien, s'écrit
$$\partial f(x) = \{A^{\mathsf{T}}y : \|y\|_{\mathrm{D}} \leqslant 1,\ y^{\mathsf{T}}(Ax-b) = \|Ax-b\|\}, \tag{19.63}$$
où $\|\cdot\|_{\mathrm{D}}$ désigne la norme duale de $\|\cdot\|$ pour le produit scalaire euclidien.
2) Montrez dans quel sens on peut considérer le problème (D) ci-dessous comme un problème dual min-max (section 14.1.1) de (P) :
$$(D) \qquad \begin{cases} \sup_{y\in\mathbb{R}^m} b^{\mathsf{T}}y \\ A^{\mathsf{T}}y = 0 \\ \|y\|_{\mathrm{D}} \leqslant 1. \end{cases}$$
3) Montrez que (P) et (D) ont une solution et qu'il n'y a pas de saut de dualité.
4) Montrez que l'ensemble des solutions de (P) s'écrit
$$\mathrm{Sol}(P) = C + \mathcal{N}(A),$$
où C est le convexe $\{x \in \mathcal{R}(A^{\mathsf{T}}) : \|Ax-b\| = \mathrm{val}(P)\}$ et que C est réduit à un point lorsque $\|\cdot\|$ est la norme ℓ_2 de $\mathbb{R}^m$.
5) Montrez que, lorsque la norme utilisée dans (P) est la norme ℓ_1, le problème (P) peut s'écrire comme un problème d'optimisation linéaire.



Remarque. Si $\|\cdot\|$ est la norme ℓ_1, en résolvant (P), on cherche à satisfaire exactement le plus grand nombre possible d'équations scalaires du système $Ax = b$ (qui est éventuellement non réalisable), tout en minimisant l'erreur $\|Ax-b\|_1$, sans savoir a priori quelles sont ces équations scalaires qui peuvent être résolues exactement. Ceci est illustré en dimension $m = 2$ à la figure ci-jointe. On voit que la solution $\bar{x}_1$, obtenue avec la norme ℓ_1 rend la seconde composante de $A\bar{x}_1$ égale à celle de b, donc la seconde équation du système, c'est-à-dire $(A\bar{x}_1 - b)_2 = 0$, est vérifiee. Par contre, la solution $\bar{x}_2$, obtenue avec la norme ℓ_2, rend toutes les composantes de $A\bar{x}_2$ différentes de celles de b. Par ailleurs, la résolution en norme ℓ_2, qui ne demande que la résolution d'un système linéaire, est moins coûteuse que la résolution en norme ℓ_1, laquelle requiert la résolution d'un problème d'optimisation linéaire.

20 Optimisation semi-définie positive

L'optimisation semi-définie positive (optimisation SDP ou OSDP) a connu un essor important dans les années 1990 pour au moins quatre raisons. D'abord, bien qu'ils soient non linéaires, les problèmes d'optimisation SDP peuvent être résolus en un nombre d'itérations polynomial. Ensuite, un grand nombre de problèmes convexes ont pu être exprimés dans ce formalisme, ce qui montre par conséquent qu'ils sont polynomialement résolubles. Puis, certains problèmes non convexes (en particulier des problèmes combinatoires) ont une relaxation SDP précise, qui permet de leur trouver rapidement une solution approchée de qualité. Enfin, certains problèmes non convexes peuvent être approché par une hiérarchie de problèmes SDP dont les valeurs optimales convergent vers celle du problèmes original (mais se pose dans des espaces vectoriels de dimension de plus en plus grande).

Les problèmes d'optimisation SDP ont une structure assez semblable à celle de l'optimisation linéaire (OL), ce qui les rend très vite familiers. Les résultats que l'on connaît en OL ne s'étendent cependant pas tous tels quels à l'optimisation SDP ; on ne peut guère en être étonné, vu la généralité du formalisme. Notre compte-rendu s'attachera à relever quelques subtilités qui ne peuvent être ignorées.

Connaissances supposées. La dualité par perturbation (section 14.2) pour établir l'existence de solutions primale et duale ; la notion de fonction asymptotique (section 3.3.4) pour l'existence du chemin central.

20.1 Définition du problème d'optimisation SDP

20.1.1 Les problèmes primal et dual

Les cônes $\mathcal{S}^n_+$ et $\mathcal{S}^n_{++}$

On note $\mathcal{S}^n$ l'espace vectoriel des matrices d'ordre n symétriques, qui est de dimension $n(n+1)/2$, $\mathcal{S}^n_+$ le cône de $\mathcal{S}^n$ formé des matrices semi-définies positives et $\mathcal{S}^n_{++}$ le cône de $\mathcal{S}^n$ formé des matrices définies positives. On note aussi $A \succcurlyeq B$ [resp. $A \succ B$] au lieu de $A-B \in \mathcal{S}^n_+$ [resp. $A-B \in \mathcal{S}^n_{++}$]. La relation $\succcurlyeq$ introduit un ordre partiel sur $\mathcal{S}^n$, dit de *Löwner*. Les relations $\preccurlyeq$ et $\prec$ s'introduisent naturellement : $A \preccurlyeq B$ équivaut à $B - A \succcurlyeq 0$ et $A \prec B$ équivaut à $B - A \succ 0$. On munit $\mathcal{S}^n$ du produit scalaire $\langle A, B\rangle = \operatorname{tr} AB = \sum_{i,j=1}^n A_{ij}B_{ij}$ ($\operatorname{tr} A$ désigne la trace de la matrice A). La norme associée à ce produit scalaire est le norme de Frobenius $\|A\|_F = (\sum_{i,j=1}^n A_{ij}^2)^{1/2}$.

Voici quelques propriétés des cônes $\mathcal{S}^n_+$ et $\mathcal{S}^n_{++}$ qui nous seront utiles.

Lemme 20.1 (cônes $\mathcal{S}_+^n$ et $\mathcal{S}_{++}^n$)
1) $A \succcurlyeq 0 \iff \forall\, B \succcurlyeq 0$, *on a* $\langle A, B\rangle \geqslant 0$.
2) $A \succ 0 \iff \forall\, B \succcurlyeq 0$ *non nulle, on a* $\langle A, B\rangle > 0$.
3) *Si* A *et* $B \in \mathcal{S}_+^n$, *on a* $\langle A, B\rangle = 0 \iff AB = 0$.

DÉMONSTRATION. 1) C'est l'exercice 2.35.

2) Soit $A \succ 0$. Si $B \succcurlyeq 0$ est non nulle, elle a une factorisation spectrale $B = \sum_{i=1}^n \lambda_i v_i v_i^\mathsf{T}$ avec au moins un $\lambda_i > 0$ (les autres valeurs propres sont $\geqslant 0$). Alors $\langle A, B\rangle = \sum_{i=1}^n \lambda_i \operatorname{tr}(A v_i v_i^\mathsf{T}) = \sum_{i=1}^n \lambda_i v_i^\mathsf{T} A v_i > 0$, puisque tous les $v_i^\mathsf{T} A v_i > 0$ et au moins un des $\lambda_i > 0$.

Inversement, supposons que $\langle A, B\rangle > 0$ pour toute matrice $B \succcurlyeq 0$ non nulle. En prenant $B = vv^\mathsf{T} \succcurlyeq 0$, avec v quelconque non nul, on doit avoir $0 < \langle A, B\rangle = v^\mathsf{T} A v$, ce qui montre que $A \succ 0$.

3) Il est clair que $AB = 0$ implique que $\langle A, B\rangle = 0$. Réciproquement, avec la factorisation spectrale de $A = \sum_{i=1}^n \lambda_i v_i v_i^\mathsf{T}$ (les λ_i sont $\geqslant 0$) et $\langle A, B\rangle = 0$, on a $\sum_i \lambda_i v_i B v_i^\mathsf{T} = 0$. Comme $B \in \mathcal{S}_+^n$, on en déduit que $Bv_i = 0$ lorsque $\lambda_i > 0$. Alors $AB = \sum_{i=1}^n \lambda_i v_i v_i^\mathsf{T} B = 0$. □

La propriété du point 1, attribuée à Fejér, exprime l'*autodualité* du cône $\mathcal{S}_+^n$:

$$(\mathcal{S}_+^n)^+ = \mathcal{S}_+^n.$$

Si v est un vecteur, on notera $\operatorname{Diag}(v)$ la matrice diagonale dont les éléments diagonaux sont les composantes v_i de v.

Le problème primal

On se donne $C \in \mathcal{S}^n$, $\mathcal{A} : \mathcal{S}^n \to \mathbb{R}^m$ une application linéaire, et $b \in \mathbb{R}^m$. Le problème primal consiste à trouver $X \in \mathcal{S}^n$ solution de

$$(P) \quad \begin{cases} \inf \langle C, X\rangle \\ \mathcal{A}(X) = b \\ X \succcurlyeq 0. \end{cases} \tag{20.1}$$

Ce problème a une structure très semblable à la forme standard (17.1) d'un problème d'optimisation linéaire. À première vue, les seules différences sont que l'on travaille sur l'espace des matrices $\mathcal{S}^n$ plutôt que $\mathbb{R}^n$ (mais cela reste un espace vectoriel) et que la contrainte de positivité est remplacée par la contrainte de semi-définie positivité de la matrice X (d'où le nom du problème). Il n'y en a pas d'autres en effet, mais cela implique que certaines propriétés de l'optimisation linéaire ne seront pas conservées en optimisation SDP. Examinons le problème plus en détail.

Le critère et la contrainte d'égalité sont linéaires (ou affines), mais la contrainte d'appartenance au cône $\mathcal{S}_+^n$ est « très » non linéaire, éventuellement non différentiable. Elle exprime en effet que $v^\mathsf{T} X v \geqslant 0$ pour tout $v \in \mathbb{R}^n$; de ce point de vue, (20.1) est un problème d'optimisation semi-infinie (il a un nombre infini de contraintes). Elle exprime aussi que toutes les valeurs propres de X (des fonctions non linéaires et non

différentiables de X) doivent être positives ; de ce point de vue, le problème primal est non convexe et non différentiable. Elle exprime enfin que tous les mineurs principaux de X (des polynômes des éléments de X) doivent être positifs ; de ce point de vue, le problème primal est non linéaire, non convexe, à données polynomiales.

On note val(P) la valeur optimale de (P) et

$$\mathcal{F}_{\text{P}} := \{X \in \mathcal{S}^n : \mathcal{A}(X) = b,\ X \succcurlyeq 0\} \tag{20.2a}$$
$$\mathcal{F}^s_{\text{P}} := \{X \in \mathcal{S}^n : \mathcal{A}(X) = b,\ X \succ 0\}, \tag{20.2b}$$

les ensembles des matrices admissibles et strictement admissibles de (P).

On peut représenter $\mathcal{A}$ au moyen de m matrices $A_i \in \mathcal{S}^n$ (théorème A.3 de Riesz-Fréchet). On a

$$\mathcal{A}(X) = \begin{pmatrix} \langle A_1, X\rangle \\ \vdots \\ \langle A_m, X\rangle \end{pmatrix}. \tag{20.3}$$

Dans cette représentation, l'application $\mathcal{A}$ est surjective si, et seulement si, les matrices A_i sont linéairement indépendantes dans $\mathcal{S}^n$.

Le problème dual

On se donne un produit scalaire sur $\mathbb{R}^m$, également noté $\langle\cdot,\cdot\rangle$, et l'on introduit l'opérateur $\mathcal{A}^* : \mathbb{R}^m \to \mathcal{S}^n$ adjoint de $\mathcal{A}$, qui est défini par

$$\forall\, X \in \mathcal{S}^n,\ \forall\, y \in \mathbb{R}^m\ :\quad \langle \mathcal{A}(X), y\rangle = \langle X, \mathcal{A}^*(y)\rangle.$$

La méthode la plus simple pour obtenir un dual de (P) est d'utiliser la dualisation lagrangienne de sa contrainte d'égalité. On utilise le lagrangien $\ell : \mathcal{S}^n_+ \times \mathbb{R}^m \to \mathbb{R}$ défini par

$$\ell(X, y) = \langle C, X\rangle - \langle y, \mathcal{A}(X) - b\rangle \tag{20.4}$$

et l'on écrit (P) comme un inf-sup :

$$\inf_{X\in\mathcal{S}^n_+} \sup_{y\in\mathbb{R}^m} \ell(X, y).$$

Le problème dual s'obtient alors en inversant l'infimum et le supremum (c'est la dualité min-max de la section 14.1). En utilisant le lemme 20.1, on obtient

$$\begin{aligned} \sup_{y\in\mathbb{R}^m} \inf_{X\in\mathcal{S}^n_+} \ell(X, y) &= \sup_{y\in\mathbb{R}^m} \inf_{X\in\mathcal{S}^n_+} \langle C - \mathcal{A}^*(y), X\rangle + \langle b, y\rangle \\ &= \sup_{y\in\mathbb{R}^m} \begin{cases} \langle b, y\rangle & \text{si } C - \mathcal{A}^*(y) \succcurlyeq 0 \\ -\infty & \text{sinon.} \end{cases} \end{aligned}$$

Le problème dual de (P) obtenu par ce procédé consiste donc à trouver $(y, S) \in \mathbb{R}^m \times \mathcal{S}^n$ solution de

$$\boxed{(D) \quad \begin{cases} \sup \langle b, y\rangle \\ \mathcal{A}^*(y) + S = C \\ S \succcurlyeq 0. \end{cases}} \tag{20.5}$$

On note val(D) sa valeur optimale et

$$\mathcal{F}_{\text{D}} := \{(y,S) \in \mathbb{R}^m \times \mathcal{S}^n : \mathcal{A}^*(y) + S = C,\ S \succcurlyeq 0\} \tag{20.6a}$$
$$\mathcal{F}^s_{\text{D}} := \{(y,S) \in \mathbb{R}^m \times \mathcal{S}^n : \mathcal{A}^*(y) + S = C,\ S \succ 0\}, \tag{20.6b}$$

les ensembles des couples admissibles et strictement admissibles de (D). On notera enfin

$$\mathcal{F} := \mathcal{F}_{\text{P}} \times \mathcal{F}_{\text{D}} \quad \text{et} \quad \mathcal{F}^s := \mathcal{F}^s_{\text{P}} \times \mathcal{F}^s_{\text{D}}.$$

Si l'on utilise le produit scalaire euclidien dans $\mathbb{R}^m$ et si l'on prend la représentation (20.3) de $\mathcal{A}$, son adjoint $\mathcal{A}^*$ s'écrit

$$\mathcal{A}^*(y) = \sum_{i=1}^m A_i y_i.$$

Dans ce cas, le problème dual peut se voir comme celui cherchant à maximiser une forme linéaire en y sur $\mathcal{S}^n$, tout en imposant qu'une combinaison affine de matrices (avec coefficients y_i) soit semi-définie positive :

$$C - \sum_{i=1}^m A_i y_i \succcurlyeq 0. \tag{20.7}$$

Rappelons que cette dernière relation est ce que l'on appelle une *inégalité matricielle linéaire* (on devrait dire affine) ; IML en abrégé.

La proposition ci-dessous donne quelques conséquences simples de la dualisation lagrangienne de (P). Le point 1 est connu sous le nom de *dualité faible*. L'écart val(P) − val(D) entre valeurs optimales primale et duale est appelé le *saut de dualité*. Le point 2 montre que $\langle X, S\rangle$ est l'écart entre valeurs primale et duale pour un triplet admissible $(X, y, S) \in \mathcal{F}$. Le point 3 donne une condition suffisante d'optimalité élémentaire, mais bien utile.

Proposition 20.2 (conséquences de la dualisation lagrangienne)
1) val$(D) \leqslant$ val(P).
2) $(X,y,S) \in \mathcal{F} \implies \langle C, X\rangle - \langle b, y\rangle = \langle X, S\rangle \geqslant 0$.
3) $(X,y,S) \in \mathcal{F}$ *et* $\langle X, S\rangle = 0 \implies X \in \mathcal{S}_{\text{P}}$ *et* $(y,S) \in \mathcal{S}_{\text{D}}$.

Démonstration. 1) C'est l'inégalité de dualité faible, issue de la technique de dualité min-max utilisée pour définir (D) ; voir la proposition 14.2.

2) Si $(X,y,S) \in \mathcal{F}$, alors $\langle C, X\rangle = \langle \mathcal{A}^*(y) + S, X\rangle = \langle y, \mathcal{A}(X)\rangle + \langle X, S\rangle = \langle y, b\rangle + \langle X, S\rangle$. De plus $\langle X, S\rangle \geqslant 0$ car $X \succcurlyeq 0$ et $S \succcurlyeq 0$.

3) Si $(X,y,S) \in \mathcal{F}$, on a $\langle b, y\rangle \leqslant$ val$(D) \leqslant$ val$(P) \leqslant \langle C, X\rangle$. Comme $\langle X, S\rangle = 0$, $\langle b, y\rangle = \langle C, X\rangle$ et donc on a égalité partout dans la relation précédente, ce qui implique que $X \in \mathcal{S}_{\text{P}}$ et $(y,S) \in \mathcal{S}_{\text{D}}$. □

La réciproque du point 3 est fausse en général (exemple 20.7-4), mais on verra qu'elle a lieu si $\mathcal{F}^s \neq \varnothing$ et $\mathcal{A}$ est surjective (corollaire 20.10). Losrqu'elle a lieu, (X, y) est un point-selle du lagrangien ℓ défini en (20.4) sur $\mathcal{S}^n_+ \times \mathbb{R}^m$ (exercice 20.3).

Dualisation par perturbation

On peut obtenir le même dual (20.5) en utilisant une technique de perturbation de la contrainte d'égalité. La fonction valeur associée s'écrit

$$v_{\text{SDP}} : \mathbb{R}^m \to \overline{\mathbb{R}} : p \mapsto v_{\text{SDP}}(p) = \inf_{\substack{X \in \mathcal{S}^n \\ \mathcal{A}(X) = b + p \\ X \succcurlyeq 0}} \langle C, X \rangle. \tag{20.8}$$

Sa conjuguée est la fonction $v^*_{\text{SDP}} : \mathbb{R}^m \to \overline{\mathbb{R}}$ définie par

$$\begin{aligned}
v^*_{\text{SDP}}(y) &= \sup_{p \in \mathbb{R}^m} \left(\langle y, p \rangle - \inf_{X \succcurlyeq 0} \left[\langle C, X \rangle + \mathcal{I}_{\{X : \mathcal{A}(X) = b + p\}}(X) \right] \right) \\
&= \sup_{X \succcurlyeq 0} \left(-\langle C, X \rangle + \sup_{p \in \mathbb{R}^m} \left[\langle y, p \rangle - \mathcal{I}_{\{X : \mathcal{A}(X) = b + p\}}(X) \right] \right) \\
&= \sup_{X \succcurlyeq 0} \Big(-\langle C, X \rangle + \langle y, \mathcal{A}(X) - b \rangle \Big) \\
&= -\langle b, y \rangle - \inf_{X \succcurlyeq 0} \langle C - \mathcal{A}^*(y), X \rangle \\
&= \begin{cases} -\langle b, y \rangle & \text{si } C - \mathcal{A}^*(y) \succcurlyeq 0 \\ +\infty & \text{sinon.} \end{cases}
\end{aligned}$$

Avec cette technique de perturbation, le dual consiste à maximiser la fonction concave $y \mapsto -v^*_{\text{SDP}}(y)$, si bien que l'on retrouve le même problème dual que par la dualisation lagrangienne.

De même, on peut retrouver le primal (20.1) par perturbation la contrainte d'égalité du dual (20.5). La fonction valeur associée s'écrit

$$w_{\text{SDP}} : \mathcal{S}^n \to \overline{\mathbb{R}} : P \mapsto w_{\text{SDP}}(P) = \inf_{\substack{(y,S) \in \mathbb{R}^m \times \mathcal{S}^n \\ \mathcal{A}^*(y) + S + P = C \\ S \succcurlyeq 0}} -\langle b, y \rangle. \tag{20.9}$$

Sa conjuguée est la fonction $w^*_{\text{SDP}} : \mathcal{S}^n \to \overline{\mathbb{R}}$ définie par

$$\begin{aligned}
w^*_{\text{SDP}}(X) &= \sup_{P \in \mathcal{S}^n} \left(\langle X, P \rangle - \inf_{\substack{(y,S) \\ S \succcurlyeq 0}} \left[-\langle b, y \rangle + \mathcal{I}_{\{(y,S) : \mathcal{A}^*(y) + S + P = C\}}(y, S) \right] \right) \\
&= \sup_{\substack{(y,S) \\ S \succcurlyeq 0}} \left(\langle b, y \rangle + \sup_{P \in \mathcal{S}^n} \left[\langle X, P \rangle - \mathcal{I}_{\{(y,S) : \mathcal{A}^*(y) + S + P = C\}}(y, S) \right] \right) \\
&= \sup_{\substack{(y,S) \\ S \succcurlyeq 0}} \left(\langle b, y \rangle + \langle X, C - \mathcal{A}^*(y) - S \rangle \right) \\
&= \langle C, X \rangle - \inf_{\substack{(y,S) \\ S \succcurlyeq 0}} \left(\langle \mathcal{A}(X) - b, y \rangle + \langle X, S \rangle \right) \\
&= \begin{cases} \langle C, X \rangle & \text{si } \mathcal{A}(X) = b \text{ et } X \succcurlyeq 0 \\ +\infty & \text{sinon.} \end{cases}
\end{aligned}$$

Avec cette technique de perturbation, le primal consiste à minimiser la fonction convexe $X \mapsto w^*_{\text{SDP}}(X)$, qui est bien le problème (20.1).

20.1.2 Réalisabilités primale et duale

Intéressons-nous à présent à la question de la réalisabilité des problèmes primal et dual. Quand peut-on trouver une matrice $X \succcurlyeq 0$ telle que $\mathcal{A}(X) = b$? Quand peut-on trouver un vecteur $y \in \mathbb{R}^m$ tel que $\mathcal{A}^*(y) \preccurlyeq C$? Les contraintes sont-elles qualifiées et dans quel sens ?

Dans $(\mathbb{R}^n, \geqslant)$, une partie de ces questions trouve une réponse dans le lemme de Farkas (proposition 2.45). Une application de ce lemme à l'optimisation SDP s'obtient en prenant $\mathbb{E} := \mathcal{S}^n$, $\mathbb{F} := \mathbb{R}^m$, l'application linéaire $\mathcal{A} : \mathcal{S}^n \to \mathbb{R}^m$ et $K := \mathcal{S}^n_+$; cela donne

$$\{y \in \mathbb{R}^m : \mathcal{A}^*(y) \succcurlyeq 0\}^+ = \operatorname{adh} \mathcal{A}(\mathcal{S}^n_+). \tag{20.10}$$

On ne peut pas se passer de l'adhérence à droite, car $\mathcal{A}(\mathcal{S}^n_+)$ n'est pas nécessairement un fermé, alors que le cône dual à gauche est toujours fermé (point 1 de la proposition 2.43) ; voir l'exercice 2.23 pour un exemple de cône $\mathcal{A}(\mathcal{S}^n_+)$ non fermé.

L'identité (20.10) nous donnera de l'information sur l'admissibilité du problème primal (20.1). Pour avoir de l'information sur l'admissibilité du problème dual (20.5), on utilise le lemme de Farkas sous la forme du point 1 de l'exercice 2.38, avec $\mathbb{E} := \mathbb{R}^m$, $\mathbb{F} := \mathcal{S}^n$, l'application linéaire $\mathcal{A}^*$, $K = \mathbb{R}^m$ et $L = \mathcal{S}^n_+$; cela donne

$$\{X \in \mathcal{S}^n_+ : \mathcal{A}(X) = 0\}^+ = \operatorname{adh}\left(\mathcal{A}^*(\mathbb{R}^m) + \mathcal{S}^n_+\right), \tag{20.11}$$

Le résultat suivant est une conséquence directe des identités (20.10) et (20.11). La proposition s'énonce de manière plus concise si l'on introduit les définitions suivantes. On dit que le système en $X \in \mathcal{S}^n$,

$$\mathcal{A}(X) = b \quad \text{et} \quad X \succcurlyeq 0, \tag{20.12}$$

est *quasi-réalisable* si $b \in \operatorname{adh} \mathcal{A}(\mathcal{S}^n_+)$, c'est-à-dire si, pour tout $\varepsilon > 0$, il existe $b_\varepsilon \in \mathbb{R}^m$ et $X_\varepsilon \in \mathcal{S}^n_+$ tels que $\|b_\varepsilon - b\| \leqslant \varepsilon$ et $\mathcal{A}(X_\varepsilon) = b_\varepsilon$. De même, on dit que le système en $(y, S) \in \mathbb{R}^m \times \mathcal{S}^n$,

$$\mathcal{A}^*(y) + S = C \quad \text{et} \quad S \succcurlyeq 0, \tag{20.13}$$

est *quasi-réalisable* si $C \in \operatorname{adh}\left(\mathcal{A}^*(\mathbb{R}^m) + \mathcal{S}^n_+\right)$, c'est-à-dire si, pour tout $\varepsilon > 0$, il existe $C_\varepsilon \in \mathcal{S}^n$, $y_\varepsilon \in \mathbb{R}^m$ et $S_\varepsilon \in \mathcal{S}^n_+$ tels que $\|C_\varepsilon - C\| \leqslant \varepsilon$ et $\mathcal{A}^*(y_\varepsilon) + S_\varepsilon = C_\varepsilon$. Évidemment, ces systèmes sont quasi-réalisables s'ils sont réalisables.

Proposition 20.3 (quasi-réalisabilités primale et duale)
1) *Le système primal* (20.12) *est quasi-réalisable si, et seulement si,* $\langle b, y\rangle \geqslant 0$ *pour tout* $y \in \mathbb{R}^m$ *tel que* $\mathcal{A}^*(y) \succcurlyeq 0$.
2) *Le système dual* (20.13) *est quasi-réalisable si, et seulement si,* $\langle C, X\rangle \geqslant 0$ *pour tout* $X \in \mathcal{S}^n_+$ *tel que* $\mathcal{A}(X) = 0$.

La quasi-réalisabilité des systèmes (20.12) et (20.13) n'est pas une propriété très forte (elle n'assure même pas leur réalisabilité !). Numériquement, il est certainement préférable d'avoir une réalisabilité plus robuste, insensible à de petites perturbations du second membre. On définit donc les concepts suivants. On dit que le système (20.12) est *fortement réalisable* si

$$b \in \operatorname{int} \mathcal{A}(\mathcal{S}^n_+).$$

De même, on dit que le système (20.13) est *fortement réalisable* si

$$C \in \operatorname{int} (\mathcal{A}^*(\mathbb{R}^m) + \mathcal{S}^n_+).$$

Comme le montre la proposition suivante, ces concepts ont un lien avec la condition de qualification de Robinson (4.100), notée (QC-R), pour les ensembles admissibles écrits sous la forme (4.65), à savoir

$$\mathcal{F}_{\text{P}} = \{X \in \mathcal{S}^n : c_P(X) \in G_P\} \tag{20.14}$$
$$\mathcal{F}_{\text{D}} = \{(y, S) \in \mathbb{R}^m \times \mathcal{S}^n : c_D(y, S) \in G_D\}, \tag{20.15}$$

avec

$$c_P : X \in \mathcal{S}^n \to (\mathcal{A}(X) - b, X) \in \mathbb{R}^m \times \mathcal{S}^n \quad \text{et} \quad G_P = \{0_{\mathbb{R}^m}\} \times \mathcal{S}^n_+,$$
$$c_D : (y, S) \in \mathbb{R}^m \times \mathcal{S}^n \to (\mathcal{A}^*(y) + S - C, S) \in \mathcal{S}^n \times \mathcal{S}^n \quad \text{et} \quad G_D = \{0_{\mathcal{S}^n}\} \times \mathcal{S}^n_+.$$

Ce lien n'est guère surprenant si l'on se remémore le corollaire 4.60 assurant la stabilité de l'ensemble admissible pour de petites perturbations (ce que l'on exprime dans ces conditions de réalisabilité fortes) lorsque la condition de Robinson est vérifiée. Ce corollaire a donc une réciproque dans le cas qui nous occupe ici. On notera bien que si la surjectivité est de rigueur dans la conditions (ii) de la proposition ci-dessous, elle ne l'est pas dans sa condition (ii').

Proposition 20.4 (réalisabilités primale et duale fortes)
1) *Si $\mathcal{F}_{\text{P}} \neq \varnothing$, les conditions suivantes sont équivalentes :*
   - (i) *le système primal (20.12) est fortement réalisable,*
   - (ii) *$\mathcal{F}^s_{\text{P}} \neq \varnothing$ et $\mathcal{A}$ est surjective,*
   - (iii) *(QC-R) a lieu en un/tout point de $\mathcal{F}_{\text{P}}$, exprimé sous la forme (20.14),*
   - (iv) *$\{(d, D) \in \mathbb{R}^m \times \mathcal{S}^n : \langle b, d\rangle \geqslant 0,\ \mathcal{A}^*(d) + D = 0,\ D \succcurlyeq 0\} = \{0\}$.*

2) *Si $\mathcal{F}_{\text{D}} \neq \varnothing$, les conditions suivantes sont équivalentes :*
   - (i') *le système dual (20.13) est fortement réalisable,*
   - (ii') *$\mathcal{F}^s_{\text{D}} \neq \varnothing$,*
   - (iii') *(QC-R) a lieu en un/tout point de $\mathcal{F}_{\text{D}}$, exprimé sous la forme (20.15),*
   - (iv') *$\{D \in \mathcal{S}^n : \langle C, D\rangle \leqslant 0,\ \mathcal{A}(D) = 0,\ D \succcurlyeq 0\} = \{0\}$.*

Démonstration. 1) $[(i) \Rightarrow (ii)]$ Montrons la surjectivité de $\mathcal{A}$: soit $b' \in \mathbb{R}^m$. Comme $b \in \operatorname{int} \mathcal{A}(\mathcal{S}^n_+)$, il existe $X \in \mathcal{S}^n_+$ tel que $\mathcal{A}(X) = b$ et $(1-t)b + tb' = \mathcal{A}(X_t)$ avec $X_t \in \mathcal{S}^n_+$ pour $t > 0$ assez petit. Alors $b' = \mathcal{A}(X_t - (1-t)X)/t$, qui est donc bien dans $\mathcal{R}(\mathcal{A})$.

D'autre part, comme $b \in \operatorname{int} \mathcal{A}(\mathcal{S}^n_+)$, on peut trouver un $t > 0$ tel que $b - t\mathcal{A}(I) \in \mathcal{A}(X_t)$ pour un $X_0 \succcurlyeq 0$, ce qui s'écrit $\mathcal{A}(X_0 + tI) = b$. Comme $X_0 + tI \succ 0$, $\mathcal{F}^s_{\text{P}}$ est non vide.

$[(ii) \Rightarrow (iii)]$ Soit $X_0 \in \mathcal{F}_{\text{P}}$. On choisit de montrer que (QC-R) a lieu en X_0 dans son expression (4.101c), qui s'écrit ici sous « forme matricielle »

$$\begin{pmatrix}\mathcal{A}\\ I\end{pmatrix}\mathcal{S}^n - \begin{pmatrix}0_{\mathbb{R}^m}\\ \mathrm{T}_{X_0}\,\mathcal{S}^n_+\end{pmatrix} = \begin{pmatrix}\mathbb{R}^m\\ \mathcal{S}^n\end{pmatrix}. \tag{20.16}$$

Il s'agit de montrer que, grâce à (ii), tout élément du membre de droite dans (20.16) peut s'exprimer sous la forme donnée dans le membre de gauche de cette identité. Soit $(v,S)\in\mathbb{R}^m\times\mathcal{S}^n$. Par la surjectivité de $\mathcal{A}$, supposée dans (ii), il existe une matrice $X\in\mathcal{S}^n$ telle que $\mathcal{A}(X)=v$. Par ailleurs, par (ii), il existe une matrice $\hat{X}\succ 0$ telle que $\mathcal{A}(\hat{X})=b$. Alors la matrice $X_t=X+t(\hat{X}-X_0)$ vérifie la première équation de (20.16), à savoir $\mathcal{A}(X_t)=v$ quel que soit $t\in\mathbb{R}$, ainsi que la seconde équation pour $t>0$ assez grand. En effet, grâce à (2.57), pour montrer que $X_t-S\in\mathrm{T}_{X_0}\,\mathcal{S}^n_+$, il suffit de constater que, pour tout $v\in\mathcal{N}(X_0)$, on a $v^{\mathsf{T}}(X_t-S)v=v^{\mathsf{T}}(X-S+t\hat{X})v\geqslant 0$ pour t assez grand.

$[(iii)\Rightarrow(iv)]$ Si (iii) a lieu en $X_0\in\mathcal{F}_{\mathrm{P}}$, on a (20.16). Par ailleurs, la condition (iv) peut s'écrire

$$\left\{\begin{pmatrix}d\\ D\end{pmatrix}\in\begin{pmatrix}\mathbb{R}^m\\ \mathcal{S}^n\end{pmatrix} : \begin{pmatrix}b^{\mathsf{T}} & 0\\ \mathcal{A}^* & I\\ 0 & I\end{pmatrix}\begin{pmatrix}d\\ D\end{pmatrix}\in\begin{pmatrix}\mathbb{R}_+\\ \{0_{\mathcal{S}^n}\}\\ \mathcal{S}^n_+\end{pmatrix}\right\} = \{0\}. \tag{20.17}$$

En prenant le dual de chaque membre et en utilisant le lemme de Farkas, on trouve (on peut ôter l'adhérence à gauche, voir le point 3 de l'exercice 2.12)

$$\begin{pmatrix}b & \mathcal{A} & 0\\ 0 & I & I\end{pmatrix}\begin{pmatrix}\mathbb{R}_+\\ \mathcal{S}^n\\ \mathcal{S}^n_+\end{pmatrix} = \begin{pmatrix}\mathbb{R}^m\\ \mathcal{S}^n\end{pmatrix}. \tag{20.18}$$

Pour démontrer (20.18), on prend $(v,S)\in\mathbb{R}^m\times\mathcal{S}^n$ dans son membre de droite et on cherche à le représenter par un élément de son membre de gauche. En prenant $-(v+\mathcal{A}(I),S)\in\mathbb{R}^m\times\mathcal{S}^n$ dans le second membre de (20.16), on obtient $X_1\in\mathcal{S}^n$ tel que

$$\mathcal{A}(X_1)=-v-\mathcal{A}(I)\qquad\text{et}\qquad X_1+S\succcurlyeq 0\ \text{dans}\ \mathcal{N}(X_0).$$

Alors $X_1+S+I\succ 0$ dans $\mathcal{N}(X_0)$, si bien qu'il existe $t>0$ tel que $X_1+S+I+tX_0\succ 0$ (lemme B.3 de Finsler). En définissant $X:=-(X_1+I+tX_0)$, on a

$$tb+\mathcal{A}(X)=v\qquad\text{et}\qquad S-X\succcurlyeq 0,$$

si bien que (20.18) est démontré.

$[(iv)\Rightarrow(i)]$ Il suffit de montrer que b est un point absorbant de $\mathcal{A}(\mathcal{S}^n_+)$ (proposition 2.15). Soit $d\in\mathbb{R}^m$. Il s'agit de trouver un scalaire $t>0$ tel que $b+td\in\mathcal{A}(\mathcal{S}^n_+)$. On utilise l'expression (20.18) de (iv) et on prend $(-b-d,0_{\mathcal{S}^n})\in\mathbb{R}^m\times\mathcal{S}^n$ dans son membre de droite. Il existe donc $(\alpha,X_1)\in\mathbb{R}_+\times\mathcal{S}^n$ tel que

$$\alpha b+\mathcal{A}(X_1)=-b-d\qquad\text{et}\qquad -X_1\succcurlyeq 0.$$

Alors $X:=-X_1/(1+\alpha)$ et $t=1/(1+\alpha)>0$ vérifie $\mathcal{A}(X)=b+td$ et $X\succcurlyeq 0$, si bien que $b+td\in\mathcal{A}(\mathcal{S}^n_+)$.

2) La démonstration est proposée à l'exercice 20.4. □

Notons que l'on peut obtenir des caractérisations de réalisabilité en termes de géométrie algébrique [355 ; 2013].

20.2 Exemples de modélisation SDP

Beaucoup de problèmes *convexes* ont une formulation SDP. L'intérêt d'exhiber une telle formulation est de montrer que l'on peut les résoudre par des algorithmes polynomiaux, souvent efficaces. Les exemples ont été choisis pour leur diversité. Certains problèmes *non convexes* ont une *relaxation SDP* précise, c'est-à-dire qu'il existe un problème d'optimisation SDP dont la valeur optimale est proche de celle du problème original. Nous en présentons quelques-unes parmi les plus citées. On trouvera d'autres exemples dans [456, 84, 626, 44, 158].

Commençons par quelques remarques qui nous permettrons de voir plus facilement quand est-ce qu'un problème d'optimisation peut se mettre sous forme SDP.

1) On peut toujours *faire passer en contrainte* le critère d'un problème d'optimisation, comme à la proposition 1.13. On veut dire par là, que le problème

$$\inf\{f(x) : c(x) \in G\},$$

où $f : E \to \mathbb{R}$ et $c : E \to F$ sont des fonctions définies sur un ensemble E et G est une partie de l'ensemble F, est équivalent au problème

$$\inf\{t : f(x) \leqslant t,\ c(x) \in G\}.$$

Par conséquent, il s'agira de voir si les variables du problème d'optimisation dont on cherche une formulation SDP peuvent s'interpréter comme des variables matricielles et si les fonctions définissant le problème (dans le critère ou les contraintes) peuvent s'écrire comme des fonctions matricielles à valeurs dans un $\mathbb{R}^m$ (formulation primale) ou des IML (formulation duale). Souvent, dans les exemples qui suivent, on ne s'intéressera qu'à ce second aspect de la formulation SDP.

2) Une contrainte qui s'écrit sous la forme $F(y) \succcurlyeq 0$, avec $F : \mathbb{R}^m \to \mathcal{S}^n$ affine, est bien une IML puisqu'elle s'écrit sous la forme (20.7), avec

$$C = F(0) \in \mathcal{S}^n \quad \text{et} \quad A_i := -\frac{\partial F}{\partial y_i}(0) \in \mathcal{S}^n.$$

3) Dans les exemples que nous donnons ci-dessous, on utilise parfois l'équivalence suivante

$$K := \begin{pmatrix} A & B \\ B^{\mathsf{T}} & C \end{pmatrix} \succ 0 \qquad \Longleftrightarrow \qquad A \succ 0 \quad \text{et} \quad A^s_K \succ 0, \tag{20.19}$$

dans laquelle A et C sont symétriques, B est de type adéquat et la matrice $A^s_K := C - B^{\mathsf{T}} A^{-1} B$ est le complément de Schur de A dans K.

Démonstration. D'abord, il est clair que la définie positivité de K entraîne celle de A. On peut donc montrer l'équivalence en supposant que $A \succ 0$. On a alors la *factorisation gaussienne* par blocs de K suivante

$$\begin{pmatrix} A & B \\ B^{\mathsf{T}} & C \end{pmatrix} = \begin{pmatrix} I & 0 \\ B^{\mathsf{T}} A^{-1} & I \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & A^s_K \end{pmatrix} \begin{pmatrix} I & A^{-1} B \\ 0 & I \end{pmatrix}.$$

Comme les deux matrices extrêmes dans le membre de droite sont inversibles et que $A \succ 0$, on a l'équivalence (20.19). □

Des conditions équivalentes à la semi-définie positivité de K sont données à l'exercice 20.7.

4) Si $B \in \mathbb{R}^{m\times n}$ et $t \in \mathbb{R}$, l'équivalence (20.44) avec $C = tI_n$ et $A = tI_m$ permet de voir que

$$\|B\| \leqslant t \iff t^2 - \lambda_{\max}(B^{\mathsf{T}}B) \succcurlyeq 0 \iff \begin{pmatrix} tI_m & B \\ B^{\mathsf{T}} & tI_n \end{pmatrix} \succcurlyeq 0, \tag{20.20}$$

où $\|B\| := \max\{\|Bx\|_2 : \|x\|_2 \leqslant 1\}$, qui est aussi la plus grande valeur singulière de B.

5) Certains problèmes s'expriment au moyen de plusieurs matrices X^i, $i \in [1:s]$, d'ordre respectif n_i, qui doivent être semi-définies positives et d'un vecteur $x^l \in \mathbb{R}^{n_l}$ qui est requis d'être positif. Matrices et vecteurs sont reliés par une contrainte affine de la forme $\mathcal{A}(X^1, \ldots, X^s, x) = b$. On peut se ramener au problème SDP primal (20.1) en introduisant une matrice X d'ordre $\sum_{i=1}^{s} n_i + n_l$, contenant les matrices X^i et les composantes du vecteur x^l sur sa diagonale et en ajoutant des contraintes linéaires imposant la nullité des éléments en dehors de cette diagonale par blocs. Cependant, dans un but numérique, il est préférable de développer la théorie et l'algorithmique du problème avec les variables $(X^1, \ldots, X^s, x^l)$ considérées séparément [575, 591, 240].

20.2.1 Optimisation linéaire

Le plus simple est d'obtenir une formulation SDP équivalente du problème d'optimisation linéaire dual

$$\begin{cases} \sup\ b^{\mathsf{T}}y \\ A^{\mathsf{T}}y \leqslant c. \end{cases}$$

Il suffit en effet d'exprimer la contrainte sous la forme d'une IML équivalente :

$$\begin{aligned} A^{\mathsf{T}}y \leqslant c \quad &\iff \quad \mathrm{Diag}\left(c - \sum_{i=1}^{m} A_{i:}^{\mathsf{T}} y_i\right) \succcurlyeq 0 \\ &\iff \quad \mathrm{Diag}(c) - \sum_{i=1}^{m} \mathrm{Diag}(A_{i:}^{\mathsf{T}}) y_i \succcurlyeq 0. \end{aligned}$$

20.2.2 Optimisation quadratique convexe

On considère un problème d'optimisation quadratique *convexe*, écrit sous la forme suivante (de manière à renforcer l'analogie avec le problème d'optimisation SDP équivalent que l'on obtient) :

$$\begin{cases} \sup\ b^{\mathsf{T}}y \\ y^{\mathsf{T}}C_i y - d_i^{\mathsf{T}}y - e_i \leqslant 0, \qquad \text{pour } i = [1:m], \end{cases}$$

où les $C_i \succcurlyeq 0$ (ce qui assure la convexité du problème). On aura une formulation SDP de ce problème (sous la forme du problème dual), si l'on montre que la contrainte peut s'écrire sous forme d'IML.

Première formulation SDP

On factorise $C_i = G_i^\mathsf{T} G_i$ (via une factorisation spectrale ou de Cholesky, par exemple), puis on observe que

$$y^\mathsf{T} C_i y - d_i^\mathsf{T} y - e_i \leqslant 0 \quad \Longleftrightarrow \quad \begin{pmatrix} I & G_i y \\ y^\mathsf{T} G_i^\mathsf{T} & d_i^\mathsf{T} y + e_i \end{pmatrix} \succcurlyeq 0.$$

Cette équivalence vient du fait que $-(y^\mathsf{T} C_i y - d_i^\mathsf{T} y - e_i)$ est le complément de Schur de I dans la matrice à droite. Par ailleurs, la condition de droite est bien une IML (une combinaison affine de matrices doit être semi-définie positive).

Seconde formulation SDP

On conserve la factorisation $C_i = G_i^\mathsf{T} G_i$ et l'on observe que

$$y^\mathsf{T} C_i y - d_i^\mathsf{T} y - e_i \leqslant 0 \quad \Longleftrightarrow \quad \begin{pmatrix} (1 + d_i^\mathsf{T} y + e_i) I & \begin{pmatrix} 1 - d_i^\mathsf{T} y - e_i \\ 2 G_i y \end{pmatrix} \\ \begin{pmatrix} 1 - d_i^\mathsf{T} y - e_i \\ 2 G_i y \end{pmatrix}^\mathsf{T} & 1 + d_i^\mathsf{T} y + e_i \end{pmatrix} \succcurlyeq 0.$$

En effet, chacun des deux membres de cette équivalence implique que $1 + d_i^\mathsf{T} y + e_i > 0$: celui de gauche parce que $C_i \succcurlyeq 0$, celui de droite parce qu'autrement $1 + d_i^\mathsf{T} y + e_i$ serait nul et cela impliquerait que les sous-matrices sur la diagonale gauche devraient être nulles, ce qui est impossible dans ce cas. L'équivalence se déduit alors en utilisant le complément de Schur.

20.2.3 Minimisation de la valeur propre maximale

On se donne une application matricielle *affine* $x \in \mathbb{R}^n \mapsto M(x) = A_0 + \sum_{i=1}^m x_i A_i \in \mathcal{S}^n$ et l'on cherche à résoudre le problème d'optimisation

$$\inf_{x \in \mathbb{R}^n} \ \lambda_{\max}(M(x)).$$

On sait qu'il s'agit d'un problème *convexe non différentiable* ($\lambda_{\max}(M(\cdot))$ n'est pas différentiable lorsque son argument x est tel que $\lambda_{\max}(M(x))$ n'est pas simple).

La transformation se fait en plaçant dans un premier temps le critère en contrainte

$$\begin{cases} \inf_{(x,t)} \ t \\ \lambda_{\max}(M(x)) \leqslant t, \end{cases}$$

puis en exprimant la contrainte par une inégalité matricielle linéaire

$$\lambda_{\max}(M(x)) \leqslant t \quad \Longleftrightarrow \quad tI - M(x) \succcurlyeq 0.$$

20.2.4 Minimisation de la norme matricielle ℓ_2

On se donne une application matricielle *affine* $x \in \mathbb{R}^n \mapsto M(x) = A_0 + \sum_{i=1}^m x_i A_i \in \mathcal{S}^n$ et l'on cherche à résoudre le problème d'optimisation

$$\inf_{x\in\mathbb{R}^n} \|M(x)\|_2.$$

Ce problème est équivalent au précédent si $M(x) \succcurlyeq 0$, ce que l'on ne suppose pas ici.

La transformation en problème d'optimisation SDP s'obtient à nouveau en faisant d'abord passer le critère en contrainte

$$\begin{cases} \inf_{(x,t)} \ t \\ \|M(x)\|_2 \leqslant t, \end{cases}$$

puis en exprimant la contrainte par une inégalité matricielle linéaire

$$\begin{aligned} \|M(x)\|_2 \leqslant t \quad &\iff \quad v^\mathsf{T} M(x)^2 v \leqslant t^2 \|v\|_2^2, \quad \forall\, v \\ &\iff \quad t^2 I - M(x)^2 \succcurlyeq 0 \\ &\iff \quad \begin{pmatrix} tI & M(x) \\ M(x) & tI \end{pmatrix} \succcurlyeq 0. \end{aligned}$$

La dernière équivalence se vérifie directement si $t = 0$ (car alors on doit avoir $M(x) = 0$) et en passant par le complément de Schur si $t > 0$.

20.2.5 Minimisation globale de polynômes

On note $\mathbb{R}[x]$ l'ensemble des *polynômes* à n indéterminées $x = (x_1, \ldots, x_n)$ et à coefficients dans $\mathbb{R}$. Un tel prolynôme s'écrit

$$p(x) = \sum_{\alpha\in\mathbb{N}^n} p_\alpha x^\alpha,$$

où $\alpha = (\alpha_1, \ldots, \alpha_n)$ est un n-uple d'entiers, $x^\alpha := x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}$ est appelé un *monôme* et il n'y a qu'un nombre fini de coefficients $p_\alpha \in \mathbb{R}$ non nuls. Le *degré* de p est défini par $\deg p := \max_\alpha\{|\alpha| : p_\alpha \neq 0\}$, où $|\alpha| = \sum_{i=1}^n \alpha_i$.

Le problème de minimiser p sur $\mathbb{R}^n$ peut s'écrire comme un problème d'optimisation à une inconnue et un nombre infini de contraintes :

$$\begin{cases} \sup_{\lambda\in\mathbb{R}} \ \lambda \\ p(x) - \lambda \geqslant 0, \quad \forall\, x \in \mathbb{R}^n. \end{cases} \tag{20.21}$$

On cherche donc un moyen d'exprimer la contrainte de ce problème, à savoir de caractériser la positivité de $p - \lambda$, si possible au moyen d'un nombre fini de conditions de manière à pouvoir le résoudre numériquement.

Un nombre réel positif est le carré d'un autre nombre réel (sa racine carrée). Mais un polynôme réel positif n'est pas nécessairement le carré d'un autre polynôme. Par exemple, même s'il n'y a qu'une seule indéterminé x, $x^2 + 1$ n'est pas de la forme $(ax + b)^2$, car il faudrait que $a^2 = 1$, $ab = 0$ et $b^2 = 1$ qui sont des conditions

incompatibles entre elles. Mais un polynôme réel positif *à une indéterminée* peut s'écrire comme une *somme de deux carrés* de polynômes réels.

Proposition 20.5 (positivité sur $\mathbb{R}$) *Un polynôme $p \in \mathbb{R}[x]$ à une indéterminée x est positif sur $\mathbb{R}$ si, et seulement si, il est de degré pair, disons $2m$, et s'écrit $p = q^2 + r^2$, avec q, $r \in \mathbb{R}[x]$, $\deg q = m$ et $\deg r \leqslant m-1$.*

Démonstration. Les conditions énoncées sont clairement suffisantes. Montrons qu'elles sont nécessaires.

Étant positif sur $\mathbb{R}$ le polynôme est nécessairement de degré pair, disons $2m$, et le coefficient dominant est positif. Il n'y a donc pas de perte de généralité à supposer que ce dernier vaut 1. Alors le polynôme se décompose en m facteurs de la forme

$$(x-a)^2 + b^2.$$

C'est en effet la forme que prend $(x-r)(x-\bar{r})$ lorsque r et $\bar{r}$ sont les deux racines complexes conjuguées $a \pm ib$. Par ailleurs toute racine réelle est de multiplicité paire sinon le polynôme changerait de signe dans le voisinage de cette racine. Chaque doublon de racines réelles a la forme ci-dessus avec $b = 0$.

On multiplie alors successivement les m facteurs de la forme $q^2 + r^2$ ci-dessus en utilisant la formule

$$(q_j^2 + r_j^2)(q^2 + r^2) = (q_j q + r_j r)^2 + (q_j r - r_j q)^2 =: q_{j+1}^2 + r_{j+1}^2.$$

Par récurrence, on voit que $\deg q_j = j$ et $\deg r_j \leqslant j-1$. C'est en effet le cas pour $j = 1$ car $\deg q_1 = 1$ et $\deg r_1 \leqslant 0$. Que r soit nul ou non, $\deg q_{j+1} = \deg q_j + 1$. Enfin, si $r = 0$, $\deg r_{j+1} = \deg r_j + 1 \leqslant j$ et, si $r \neq 0$, $\deg r_{j+1} \leqslant \max(\deg q_j, \deg r_j + 1) \leqslant j$. □

Ce résultat surprenant ne tient plus pour des polynômes à plus d'une indéterminée, même si l'on s'autorise à l'exprimer comme une somme de *plusieurs* carrés de polynômes ; autrement dit, un polynôme positif n'est pas nécessairement un élément de

$$\Sigma[x] := \left\{ \sum_{i=1}^{N} p_i^2 : p_i \in \mathbb{R}[x],\ N \in \mathbb{N} \right\}, \tag{20.22}$$

l'ensemble des polynômes qui peuvent s'écrire comme une somme d'un nombre fini de carrés de polynômes. On verra à la proposition 20.6 que l'on peut prendre $N = \binom{n+d}{d}$. Même si cela était connu dès 1920, le premier exemple de polynôme positif n'étant pas dans $\Sigma[x]$ n'a été trouvé qu'en 1967 par Motzkin ; il s'agit du polynôme à deux indéterminées suivant

$$p(x_1, x_2) = 1 + x_1^2 x_2^2 (x_1^2 + x_2^2 - 3). \tag{20.23}$$

On peut toutefois trouver des résultats de densité de $\Sigma[x]$ dans l'ensemble des polynômes positifs, pourvu que l'on se restreigne à un compact [48, 379]. L'intérêt de $\Sigma[x]$ vient du fait que l'appartenance à cet ensemble peut s'exprimer par une IML, alors que l'on ne connait pas de méthode simple pour exprimer la positivité de polynômes. On note $\Sigma_{2d}[x]$ la partie de $\Sigma[x]$ formée des polynômes de degré $\leqslant 2d$.

Proposition 20.6 (caractérisation des polynômes de $\Sigma_{2d}[x]$) *Un polynôme $p \in \mathbb{R}[x]$ de degré $\leqslant 2d$ est une somme de r carrés de polynômes si, et seulement si, il existe une matrice $S \succcurlyeq 0$ d'ordre $N := \binom{n+d}{d}$ et de rang $\leqslant r$ telle que $p(x) = v_d(x)^\mathsf{T} S v_d(x)$, où $v_d(x)$ est le vecteur des monômes à n variables de degré $\leqslant d$. En particulier, on peut restreindre la somme dans (20.22) à N termes.*

Démonstration. Observons d'abord que N est la dimension du vecteur des monômes $v_d(x)$, puisque le nombre de monôme de degré d est égal au nombre de combinaisons avec répétitions de n éléments pris d à d, c'est-à-dire à $\binom{n+d-1}{d}$. Alors la dimension de $v_d(x)$ vaut

$$\tbinom{n-1}{0} + \tbinom{n}{1} + \cdots + \tbinom{n+d-1}{d} = \tbinom{n+d}{d} = N.$$

Si p est une somme de r carrés de polynômes de degré $\leqslant d$, on a pour des vecteurs $s_i \in \mathbb{R}^N$:

$$p(x) = \sum_{i=1}^{r} (s_i^\mathsf{T} v_d(x))^2 = \sum_{i=1}^{r} v_d(x)^\mathsf{T} s_i s_i^\mathsf{T} v_d(x) = v_d(x)^\mathsf{T} S v_d(x),$$

où $S := \sum_{i=1}^{r} s_i s_i^\mathsf{T} \succcurlyeq 0$ est de rang $\leqslant r$.

Inversement, supposons que $p(x) = v_d(x)^\mathsf{T} S v_d(x)$, avec $S \succcurlyeq 0$ de rang $\leqslant r$. La décomposition spectrale de $S = \sum_{i=1}^{r} s_i s_i^\mathsf{T}$ permet d'écrire

$$p(x) = \sum_{i=1}^{r} v_d(x)^\mathsf{T} s_i s_i^\mathsf{T} v_d(x) = \sum_{i=1}^{r} (s_i^\mathsf{T} v_d(x))^2.$$

Comme S est d'ordre N, son rang r est nécessairement plus petit que N, c'est-à-dire qu'un polynôme de $\Sigma[x]$ peut s'exprimer comme somme d'au plus N carrés de polynômes. □

Montrons que le résultat précédent permet d'exprimer l'appartenance à $\Sigma_{2d}[x]$ par une IML. La proposition affirme que $p \in \Sigma_{2d}[x]$ si, et seulement si, on peut trouver une matrice $S \in \mathcal{S}_+^N$ telle que

$$\begin{aligned} p(x) &= v_d(x)^\mathsf{T} S v_d(x) \\ &= \sum_{|\alpha|,\,|\beta| \leqslant d} S_{\alpha,\beta} x^{\alpha+\beta} \\ &= \sum_{|\gamma| \leqslant 2d} \left(\sum_{\substack{|\alpha|,\,|\beta| \leqslant d \\ \alpha+\beta=\gamma}} S_{\alpha,\beta} \right) x^\gamma \\ &= \sum_{|\gamma| \leqslant 2d} \langle B_\gamma, S \rangle x^\gamma, \end{aligned}$$

où on a noté

$$B_\gamma := \sum_{\substack{|\alpha|,\,|\beta|\leqslant d \\ \alpha+\beta=\gamma}} E^{\alpha,\beta},$$

$E^{\alpha,\beta}$ étant la matrice élémentaire d'ordre N dont tous les éléments sont nuls sauf celui en positition (α,β), qui vaut 1. Il reste à identifier les coefficients de même puissance :

$$\boxed{p \in \Sigma_{2d}[x] \qquad \Longleftrightarrow \qquad \exists\, S \in \mathcal{S}_+^N,\ \forall\, |\gamma| \leqslant 2d : \quad \langle B_\gamma, S\rangle = p_\gamma.} \tag{20.24}$$

On a bien une IML à droite, dans laquelle on recherche une matrice symétrique semi-définie positive S d'ordre $\binom{n+d}{d}$ soumise à $\binom{n+2d}{2d}$ contraintes affines.

Comme $\Sigma[x]$ est inclus dans l'ensemble des polynômes positifs, on obtient une *relaxation* du problème de minimisation globale de p, exprimé par (20.21), en transformant sa contrainte $p-\lambda \geqslant 0$ par $p-\lambda \in \Sigma_{2d}[x]$: la valeur optimale du problème relaxé sera donc plus petite. Grâce à (20.24), on obtient le problème relaxé suivant [375] :

$$\begin{cases} \sup_{(S,\lambda)\in\mathcal{S}^N\times\mathbb{R}} \ \lambda \\ S_{00} = p_0 - \lambda \\ \langle B_\gamma, S\rangle = p_\gamma, \quad \forall\, \gamma \in \mathbb{N}^n : 0 < |\gamma| \leqslant 2d \\ \mathcal{S} \succcurlyeq 0. \end{cases}$$

On peut facilement le transformer en un problème SDP primal de la forme (20.1).

L'approche décrite brièvement ci-dessus peut s'étendre à la minimisation globale de polynôme sur un ensemble défini par des inégalités polynomiales. Cette extention peut se faire, par exemple, au moyen de notions fines de géométrie algébrique (théorème de Putinar [501]) [474, 375, 376] ou pas [543].

20.2.6 Relaxation SDP de problèmes non convexes

Relaxation de rang d'un problème quadratique en variables binaires

On considère un problème quadratique ($Q \in \mathcal{S}^n$) en nombres $\{-1,1\}$:

$$\begin{cases} \min\ x^{\mathsf{T}} Q x \\ x \in \{-1,1\}^n. \end{cases} \tag{20.25}$$

Ce problème est NP-ardu [626 ; chapitre 12, page 344]. La technique de relaxation SDP présentée ci-dessous s'utilise dans des contextes variés.

On peut écrire $x^{\mathsf{T}} Q x = \langle Q, xx^{\mathsf{T}}\rangle$. D'autre part $x_i \in \{-1,1\}$ si, et seulement si, $x_i^2 = 1$, dès lors

$$X = xx^{\mathsf{T}} \text{ et } x_i \in \{-1,1\} \quad \Longleftrightarrow \quad X \succcurlyeq 0,\ \mathrm{diag}(X) = e \text{ et } \mathrm{rg}(X) = 1.$$

La contrainte $\mathrm{rg}(X) = 1$ n'est pas convexe (le rang est à valeurs entières !). On obtient cependant une relaxation assez précise du problème (20.25) en abandonnant cette contrainte et en considérant à la place le problème d'optimisation SDP :

$$\begin{cases} \min\ \langle Q, X\rangle \\ \mathrm{diag}(X) = e \\ X \succcurlyeq 0. \end{cases}$$

Relaxation de rang du problème tout quadratique

Nous allons donner ici la *relaxation de rang*, parfois dite *de Shor* et que l'on pourrait qualifier de *lagrangienne* (voir l'exercice 20.8), du problème tout quadratique ; c'est une relaxation SDP. Le terme *relaxation* veut dire que la valeur optimale du problème relaxé est plus petite que celle du problème original.

Le *problème tout quadratique*, s'écrit

$$\begin{cases} \inf_x \ q_0(x) \\ q_i(x) = 0, \quad \forall\, i \in E \\ q_i(x) \leqslant 0, \quad \forall\, i \in I, \end{cases} \tag{20.26a}$$

où E et I sont des ensembles d'indices, finis et disjoints, et les $q_i : \mathbb{R}^n \to \mathbb{R}$ sont des fonctions quadratiques définies en $x \in \mathbb{R}^n$ par

$$q_i(x) = x^{\mathsf{T}} A_i x + 2b_i^{\mathsf{T}} x + c_i, \tag{20.26b}$$

avec des matrices $A_i \in \mathcal{S}^n$, des vecteurs $b_i \in \mathbb{R}^n$ et des scalaires $c_i \in \mathbb{R}$. Comme les matrices symétriques A_i ne sont pas nécessairement semi-définies positives, le problème n'est pas nécessairement convexe. Cette formulation inclut l'optimisation linéaire (matrices A_i nulles), les problèmes en variables $(0,1)$ (la contrainte $x_i \in \{0,1\}$ peut s'exprimer par la contrainte quadratique $x_i^2 - x_i = 0$) et plus généralement toute l'optimisation polynomiale (on peut en effet réduire d'une unité le degré d'un monôme de degré > 2 en remplaçant un produit $x_i x_j$ par une nouvelle variable y_k et en ajoutant la contrainte quadratique $x_i x_j = y_k$ [548]). Il s'agit donc de problèmes très difficiles, en tout cas NP-ardus (chapitre **??**).

Observons qu'une fonction quadratique peut s'écrire comme suit

$$x^{\mathsf{T}} A_i x + 2b_i^{\mathsf{T}} x + c_i = \begin{pmatrix} x \\ 1 \end{pmatrix}^{\mathsf{T}} \begin{pmatrix} A_i & b_i \\ b_i^{\mathsf{T}} & c_i \end{pmatrix} \begin{pmatrix} x \\ 1 \end{pmatrix} = \left\langle \begin{pmatrix} A_i & b_i \\ b_i^{\mathsf{T}} & c_i \end{pmatrix}, \begin{pmatrix} x \\ 1 \end{pmatrix} \begin{pmatrix} x \\ 1 \end{pmatrix}^{\mathsf{T}} \right\rangle.$$

où $\langle \cdot, \cdot \rangle$ est le produit scalaire dans $\mathcal{S}^{n+1}$. Si l'on note

$$\tilde{A}_i = \begin{pmatrix} A_i & b_i \\ b_i^{\mathsf{T}} & c_i \end{pmatrix}, \tag{20.27}$$

on peut récrire le problème d'optimisation quadratique (20.26) de la manière suivante

$$\begin{cases} \inf_{(X,x) \in \mathcal{S}^n \times \mathbb{R}^n} \ \langle \tilde{A}_0, X \rangle \\ \langle \tilde{A}_i, X \rangle = 0, \quad \forall\, i \in E \\ \langle \tilde{A}_i, X \rangle \leqslant 0, \quad \forall\, i \in I \\ X = \binom{x}{1} \binom{x}{1}^{\mathsf{T}}. \end{cases}$$

La dernière condition n'est pas linéaire en (X, x) et la relaxation proposée consiste à n'en garder qu'une partie, qui peut s'exprimer linéairement. On constate en effet que l'élément $(n+1, n+1)$ de $\binom{x}{1}(x^{\mathsf{T}}\ 1)$ vaut 1, si bien que les (X, x) vérifiant $X = \binom{x}{1}(x^{\mathsf{T}}\ 1)$ vérifient aussi $\operatorname{rg} X = 1$ et $X_{n+1,n+1} = 1$ (on agrandit l'ensemble admissible). La *relaxation de rang* consiste à se débarrasser de la contrainte de rang (à valeurs entières, donc non continue et certainement très désagréable) et à ne garder

que $X_{n+1,n+1} = 1$ (on agrandit encore l'ensemble admissible). La formulation relaxée obtenue ainsi s'écrit :

$$\begin{cases} \inf_{(X,x)\in\mathcal{S}^n\times\mathbb{R}^n} \langle \tilde{A}_0, X\rangle \\ \langle \tilde{A}_i, X\rangle = 0, \quad \forall\, i \in E \\ \langle \tilde{A}_i, X\rangle \leqslant 0, \quad \forall\, i \in I \\ X_{n+1,n+1} = 1 \\ X \succcurlyeq 0. \end{cases} \tag{20.28}$$

C'est bien un problème SDP (généralisé, à cause des inégalités pour les indices dans I, que l'on peut transformer en égalités avec des variables d'écart : $\langle \tilde{A}_i, X\rangle + s_i = 0$ et $s_i \geqslant 0$, pour $i \in I$ – voir la remarque 5 à la page 726). On peut montrer que c'est le bidual lagrangien du problème (20.26), voir l'exercice 20.8.

L'erreur commise sur la valeur optimale par cette relaxation a été estimée dans [255, 40, 454, 634, 635].

20.3 Existence de solution et CNS d'optimalité

En optimisation linéaire, les résultats d'existence de solution et de dualité forte (théorèmes 17.4 et 17.11) assurent que les propriétés suivantes sont équivalentes :

- (P) et (D) sont réalisables,
- (P) est réalisable et borné,
- (D) est réalisable et borné,
- (P) a une solution,
- (D) a une solution.

De plus, dans chacun de ces cas, il n'y a pas de saut de dualité. Bien que l'OL et l'OSDP aient une structure très semblable, ces résultats ne tiennent plus pour les problèmes d'optimisation SDP. Voici quelques contre-exemples qui pourront servir de garde-fous.

Exemples 20.7 On peut avoir les situations suivantes.

(1) (P) est fortement réalisable (par la matrice identité) et borné, mais n'a pas de solution ; (D) a une unique solution ; il n'y a pas de saut de dualité. Voici un exemple avec $n = 2$ et $m = 1$:

$$C = \begin{pmatrix} 2 & -1 \\ -1 & 0 \end{pmatrix}, \quad A_1 = \begin{pmatrix} 0 & -1 \\ -1 & 2 \end{pmatrix}, \quad b = 2.$$

(2) (P) a une solution ; (D) est réalisable et borné, mais n'a pas de solution ; il n'y a pas de saut de dualité (cas symétrique du précédent). Voici un exemple avec $n = 2$ et $m = 2$:

$$C = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad A_1 = \begin{pmatrix} -1 & 0 \\ 0 & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix}, \quad b = \begin{pmatrix} -1 \\ 0 \end{pmatrix}.$$

(3) (P) n'est pas réalisable ; (D) a une solution. Voici un exemple avec $n = 2$ et $m = 2$:

$$C = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad A_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad b = \begin{pmatrix} 0 \\ 2 \end{pmatrix}.$$

(4) (P) et (D) ont une solution ; il y a un saut de dualité. Voici un exemple avec $n = 3$ et $m = 2$:

$$C = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad A_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 2 \end{pmatrix}, \quad b = \begin{pmatrix} 0 \\ 2 \end{pmatrix}. \qquad \square$$

Les difficultés viennent en partie du fait observé dans l'exercice 2.23 que l'image par une application linéaire d'un cône convexe fermé, comme $\mathcal{S}^n_+$, n'est pas nécessairement fermée (alors que l'image par une application linéaire de l'orthant positif est fermée, ce qui rend l'OL plus aisée). Si l'image de $\mathcal{S}^n_+$ par l'application linéaire $X \mapsto (\langle C, X\rangle, A(X))$ était fermée, le seul fait d'avoir (P) réalisable et borné entraînerait l'existence d'une solution de (P) (même démonstration qu'à la proposition 17.4). Mais en toute généralité, on ne peut pas étendre la proposition 17.4 à l'optimisation SDP.

On fera bien attention au fait que l'absence de saut de dualité au point 1 (resp. au point 2) de la proposition ci-dessous n'implique pas que le problème dual (resp. primal) a une solution, mais seulement que sa valeur optimal est identique à celle du problème primal (resp. dual).

Proposition 20.8 (existence de solution)
1) *Si $\mathcal{F}_{\text{P}} \times \mathcal{F}^s_{\text{D}} \neq \varnothing$, alors $\mathcal{S}_{\text{P}}$ est un compact non vide.*
2) *Si $\mathcal{F}^s_{\text{P}} \times \mathcal{F}_{\text{D}} \neq \varnothing$ et si $\mathcal{A}$ est surjective, alors $\mathcal{S}_{\text{D}}$ est un compact non vide.*
3) *Si $\mathcal{F}^s \neq \varnothing$ et si $\mathcal{A}$ est surjective, alors $\mathcal{S}_{\text{P}}$ et $\mathcal{S}_{\text{D}}$ sont des compacts non vides.*
Dans chacun de ces cas, il n'y a pas de saut de dualité : $\text{val}(P) = \text{val}(D)$.

Démonstration. Les problèmes primal et dual pouvant être obtenus par une technique de perturbation, les points 1 et 2 peuvent se déduire de la proposition 14.11, point 1.*d* (on pourrait aussi utiliser la dualité de Fenchel dont le résultat de dualité forte dérive en fait de la même proposition).

1) La fonction valeur w_{SDP} définie en (20.9) est convexe, parce qu'elle est la fonction marginale de la fonction convexe

$$(y, S, P) \in \mathbb{R}^m \times \mathcal{S}^n \times \mathcal{S}^n \mapsto -\langle b, y\rangle + \mathcal{I}_{\{(y,S):\mathcal{A}^*(y)+S+P=C\}}(y, S) + \mathcal{I}_{\mathcal{S}^n_+}(S) \in \overline{\mathbb{R}}.$$

Il suffit alors de montrer que $0 \in \text{int}(\text{dom}\, w_{\text{SDP}})$ et que w_{SDP} ne prend pas la valeur $-\infty$. On voit facilement que $\text{dom}\, w_{\text{SDP}} = C - \mathcal{S}^n_+ - \mathcal{A}^*(\mathbb{R}^m)$, si bien que la première condition s'écrit

$$C \in \text{int}(\mathcal{A}^*(\mathbb{R}^m) + \mathcal{S}^n_+).$$

Cette relation exprime la réalisabilité duale forte qui, par la proposition 20.4, est équivalente à $\mathcal{F}^s_{\text{D}} \neq \varnothing$, qui fait partie des hypothèses. Par ailleurs, comme $\mathcal{F}_{\text{P}} \neq \varnothing$, $\text{val}(P) < +\infty$ et donc, par dualité faible, $w_{\text{SDP}}(0) = -\,\text{val}(D) > -\infty$. Ensuite, le fait que $0 \in \text{int}(\text{dom}\, w_{\text{SDP}})$ empêche w_{SDP} de prendre la valeur $-\infty$ (exercice 3.3).

2) On s'y prend de la même manière. La fonction valeur v_{SDP} définie en (20.8) est convexe comme fonction marginale de la fonction convexe

$$(X,p) \in \mathcal{S}^n \times \mathbb{R}^m \mapsto \langle C, X\rangle + \mathcal{I}_{\{X:\mathcal{A}(X)=b+p\}}(X) + \mathcal{I}_{\mathcal{S}^n_+}(X) \in \overline{\mathbb{R}}.$$

Il reste à montrer que $0 \in \operatorname{int}(\operatorname{dom} v_{\text{SDP}})$ et que v_{SDP} ne prend pas la valeur $-\infty$. On a $\operatorname{dom} v_{\text{SDP}} = \mathcal{A}(\mathcal{S}^n_+) - b$, si bien que la première condition s'écrit

$$b \in \operatorname{int} \mathcal{A}(\mathcal{S}^n_+).$$

Cette relation exprime la réalisabilité primale forte qui, par la proposition 20.4, est équivalente à la surjectivité de $\mathcal{A}$ et à $\mathcal{F}^s_{\text{P}} \neq \varnothing$. Par ailleurs, comme $\mathcal{F}_{\text{D}} \neq \varnothing$, $\operatorname{val}(D) > -\infty$ et donc, par dualité faible, $v_{\text{SDP}}(0) = \operatorname{val}(P) > -\infty$. Ensuite, le fait que $0 \in \operatorname{int}(\operatorname{dom} v_{\text{SDP}})$ empêche v_{SDP} de prendre la valeur $-\infty$.

3) C'est une conséquence immédiate des points 1 et 2. □

Il est intéressant de comparer les résultats de la proposition précédente avec ce que donnent les conditions d'optimalité des problèmes primal et dual, vus comme des problèmes « abstraits » de la section 4.5.1. Le théorème 4.56 sur lequel repose les conditions d'optimalité ci-dessous suppose d'emblée que le problème considéré a une solution pour pouvoir énoncer ses conditions d'optimalité et d'autres informations sur son dual. L'hypothèse de départ est donc plus forte que dans la proposition précédente, qui ne suppose sur le problème considéré qu'une hypothèse sur sa contrainte (rien sur l'existence de solution), mais elle a aussi une conséquence supplémentaire, à savoir la relation de complémentarité $\langle X, S\rangle = 0$, laquelle n'est pas assurée aux points 1 et 2 de la proposition précédente (et ne peut l'être comme le montre l'exemple 20.7 (1)).

Proposition 20.9 (CN d'optimalité)
1) *Si* $\mathcal{S}_{\text{D}} \neq \varnothing$, *alors* $\mathcal{F}^s_{\text{D}} \neq \varnothing \iff \mathcal{S}_{\text{P}} \neq \varnothing$ *et compact.*
2) *Si* $\mathcal{S}_{\text{P}} \neq \varnothing$, *alors* $\mathcal{F}^s_{\text{P}} \neq \varnothing$ *et* $\mathcal{A}$ *est surjective* $\iff \mathcal{S}_{\text{D}} \neq \varnothing$ *et compact.*
Dans chacun de ces cas, il n'y a pas de saut de dualité et l'on a complémentarité entre les solutions primale X *et duale* (y,S), *c'est-à-dire* $\langle X, S\rangle = 0$.

Démonstration. 1) Supposons que le problème dual ait une solution (y,S). Si $\mathcal{F}^s_{\text{D}}$ est non vide, ses contraintes sont qualifiées (équivalence $(ii')\Leftrightarrow(iii')$ de la proposition 20.4). En utilisant le lagrangien du problème dual $(y,S,X,X') \mapsto \langle b,y\rangle - \langle X, \mathcal{A}^*(y) + S - C\rangle + \langle X', S\rangle$, on obtient l'existence d'un $X \succcurlyeq 0$ tel que $\mathcal{A}(X) = b$ et $\langle X, S\rangle = 0$ (théorème 4.56), si bien que $\mathcal{S}_{\text{P}} \neq \varnothing$. Par ailleurs $\mathcal{S}_{\text{P}}$ est borné par la propriété de Gauvin (proposition 4.70). Inversement, si $\mathcal{S}_{\text{P}}$ est non vide et borné, (QC-R) a lieu par cette même propriété et donc $\mathcal{F}^s_{\text{D}} \neq 0$ par l'équivalence $(ii')\Leftrightarrow(iii')$ de la proposition 20.4.

2) On raisonne de la même manière que pour le point 1, mais sur le problème primal cette fois. □

On est à présent en mesure de donner des conditions nécessaires et suffisantes d'optimalité et d'avoir ainsi une réciproque du point 3 de la proposition 20.2.

Corollaire 20.10 (CNS d'optimalité) *Si $\mathcal{F}_{\mathrm{P}}^s \times \mathcal{F}_{\mathrm{D}}^s \neq \varnothing$, alors (X, y, S) est solution primale-duale de (20.1) si, et seulement si,*

$$\begin{cases} \mathcal{A}^*(y) + S = C, \quad S \succcurlyeq 0 \\ \mathcal{A}(X) = b, \quad X \succcurlyeq 0 \\ \langle X, S \rangle = 0. \end{cases} \tag{20.29}$$

Démonstration. *Les conditions (20.29) sont nécessaires.* En effet, on peut supposer que l'opérateur $\mathcal{A}$ est représenté par des matrices $A_i \in \mathcal{S}^n$ comme en (20.3) et éliminer les contraintes redondantes sans modifier le problème ni l'hypothèse $\mathcal{F}_{\mathrm{P}}^s \neq \varnothing$. L'opérateur $\mathcal{A}$ est alors remplacé par un opérateur $\tilde{\mathcal{A}}$ surjectif. En vertu du point 3 de la proposition 20.8, $\mathcal{S}_{\mathrm{P}} \times \mathcal{S}_{\mathrm{D}} \neq \varnothing$. Alors la proposition 20.9 assure la relation de complémentarité $\langle X, S \rangle = 0$. On retrouve la relation d'admissibilité duale $\mathcal{A}^*(y)+S = C$ avec l'opérateur $\mathcal{A}$ original en complétant les y_i trouvés par des zéros.

Les conditions (20.29) sont suffisantes d'après le point 3 de la proposition 20.2. □

Soit $z = (X, y, S)$ une solution primale-duale, sans saut de dualité : $\langle X, S \rangle = 0$. Alors $XS = SX = 0$ (lemme 20.1), en particulier X et S commutent. Il existe donc une matrice orthogonale V qui diagonalise à la fois X et S :

$$V^{\mathsf{T}} X V = \operatorname{diag}(\xi_1, \ldots, \xi_n) \quad \text{et} \quad V^{\mathsf{T}} S V = \operatorname{diag}(\sigma_1, \ldots, \sigma_n).$$

De plus

$$\xi_i \sigma_i = 0, \quad \xi_i \geqslant 0, \quad \sigma_i \geqslant 0, \quad \text{pour tout } i = [1\,{:}\,n], \tag{20.30}$$

ce qui rappelle les conditions de complémentarité en optimisation linéaire. Il y a n relations de complémentarité et non pas $\dim \mathcal{S}^n = n(n+1)/2$. On peut rapprocher cette observation avec le fait que la complexité itérative des algorithmes de points intérieurs en OSDP sera proportionnelle à n^ω et non pas à $(\dim \mathcal{S}^n)^\omega$.

Les relations de complémentarité (20.30) impliquent que $\operatorname{rg}(X) + \operatorname{rg}(S) \leqslant n$. On dit qu'une solution primale-duale (X, y, S) sans saut de dualité est *strictement complémentaire* si

$$\operatorname{rg}(X) + \operatorname{rg}(S) = n.$$

20.4 Algorithmes de points intérieurs

20.4.1 Le chemin central

Définition, existence et régularité

Le *chemin central* est l'ensemble des solutions de

$$\begin{cases} \mathcal{A}^*(y) + S = C, \quad S \succ 0 \\ \mathcal{A}(X) = b, \quad X \succ 0 \\ XS = \mu I. \end{cases} \tag{20.31}$$

pour des $\mu > 0$. La relaxation de la condition $XS = 0$, qui est équivalente à $\langle X, S\rangle = 0$ (voir le lemme 20.1), sous la forme $XS = \mu I$ est moins naturelle qu'en OL. Elle est motivée par un argument de pénalisation intérieure (comme en OL d'ailleurs, voir la proposition 20.12).

On introduit la *fonction log-déterminant* $\mathrm{ld} : \mathcal{S}^n \to \mathbb{R} \cup \{+\infty\}$, définie par

$$\mathrm{ld}(X) := \begin{cases} -\log\det(X) & \text{si } X \succ 0 \\ +\infty & \text{sinon.} \end{cases}$$

C'est la *barrière logarithmique* pour le cône $\mathcal{S}^n_{++}$. On se rappelle que si $X \succ 0$ et $H \in \mathcal{S}^n$:

$$\det{}'(X) \cdot H = (\det X)(\mathrm{tr}\, X^{-1}H).$$

On en déduit que, pour $X \succ 0$ et H, K, $L \in \mathcal{S}^n$, on a

$$\begin{aligned} \mathrm{ld}'(X) \cdot H &= -\langle X^{-1}, H\rangle \\ \mathrm{ld}''(X) \cdot (H,K) &= \langle X^{-1}HX^{-1}, K\rangle \\ \mathrm{ld}'''(X) \cdot (H,K,L) &= -\langle X^{-1}HX^{-1}KX^{-1} + X^{-1}KX^{-1}HX^{-1}, L\rangle. \end{aligned}$$

On a donc $\nabla \mathrm{ld}(X) = -X^{-1}$ et $\nabla^2 \mathrm{ld}(X) = X^{-1} \otimes X^{-1}$, si $(X \otimes Y)$ est l'application linéaire de $\mathcal{S}^n \to \mathcal{S}^n$ définie pour X et Y dans $\mathcal{S}^n$ par

$$(X \otimes Y)H = \frac{1}{2}(XHY + YHX). \tag{20.32}$$

Il s'agit d'un produit tensoriel (exercice 20.9). On montre que $\mathrm{ld} \in \mathrm{C}\overline{\mathrm{onv}}(\mathcal{S}^n)$ et que ld est strictement convexe. Il sera aussi utile de calculer la fonction asymptotique de ld (exercice 20.10)

$$\mathrm{ld}^\infty = \mathcal{I}_{\mathcal{S}^n_+}. \tag{20.33}$$

Pour chaque $\mu > 0$, on associe à (P) le problème-barrière suivant

$$(P_\mu) \quad \begin{cases} \inf_{X \in \mathcal{S}^n} \langle C, X\rangle + \mu\, \mathrm{ld}(X) \\ \mathcal{A}(X) = b \quad (X \succ 0). \end{cases}$$

On vérifie aisément (en utilisant un lagrangien) que X est solution de (P_μ) si, et seulement si, il existe $(y, S) \in \mathbb{R}^m \times \mathcal{S}^n$ tel que que l'on ait (20.31). Il s'agit de conditions nécessaires et suffisantes, du fait de la convexité de (P_μ). Ce fait justifie a posteriori la perturbation du système d'optimalité (20.29) adoptée dans (20.31).

De même, pour chaque $\mu > 0$, on associe à (D) le problème-barrière suivant

$$(D_\mu) \quad \begin{cases} \sup_{(y,S) \in \mathbb{R}^m \times \mathcal{S}^n} \langle b, y\rangle - \mu\, \mathrm{ld}(S) \\ \mathcal{A}^*(y) + S = C \quad (S \succ 0). \end{cases}$$

On montre qu'à la constante additive $n\mu(1 - \log\mu)$ près dans le critère de (D_μ), ce dernier est le dual de (P_μ), obtenu par relaxation lagrangienne de sa contrainte.

Le lemme suivant donne des conditions pour que le système (20.31) linéarisé (après d'éventuelles modifications) ait une solution unique. Cette solution est le pas de Newton.

Lemme 20.11 *Soient $\mathcal{P}$ et $\mathcal{Q}$ deux opérateurs linéaires de $\mathcal{S}^n \to \mathcal{S}^n$ tels que $\mathcal{P}$ soit inversible et $\mathcal{P}^{-1}\mathcal{Q}$ soit défini positif (sans être nécessairement auto-adjoint). On suppose que $\mathcal{A}$ est surjective. Alors pour $r_D \in \mathcal{S}^n$, $r_P \in \mathbb{R}^m$ et $r_C \in \mathcal{S}^n$, le système*

$$\begin{pmatrix} 0 & \mathcal{A}^* & I \\ \mathcal{A} & 0 & 0 \\ \mathcal{P} & 0 & \mathcal{Q} \end{pmatrix} \begin{pmatrix} d_X \\ d_y \\ d_S \end{pmatrix} = \begin{pmatrix} r_D \\ r_P \\ r_C \end{pmatrix} \tag{20.34}$$

a une solution unique, donnée par

$$\begin{aligned} d_y &= (\mathcal{A}\mathcal{P}^{-1}\mathcal{Q}\mathcal{A}^*)^{-1}(r_P - \mathcal{A}\mathcal{P}^{-1}(r_C - \mathcal{Q}(r_D))) \\ d_S &= r_D - \mathcal{A}^*(d_y) \\ d_X &= \mathcal{P}^{-1}(r_C - \mathcal{Q}d_S). \end{aligned}$$

Démonstration. Il suffit d'exhiber la solution. La troisième équation et l'inversibilité de $\mathcal{P}$ permet d'écrire

$$d_X + \mathcal{P}^{-1}\mathcal{Q}d_S = \mathcal{P}^{-1}(r_C).$$

Alors, comme $d_X \in \mathcal{S}^n$ et $\mathcal{P}^{-1}\mathcal{Q}d_S \in \mathcal{S}^n$, on peut appliquer $\mathcal{A}$ aux deux membres, ce qui donne en tenant compte de la seconde équation

$$\mathcal{A}\mathcal{P}^{-1}\mathcal{Q}d_S = \mathcal{A}\mathcal{P}^{-1}(r_C) - r_P.$$

En appliquant $\mathcal{A}\mathcal{P}^{-1}\mathcal{Q}$ à la première équation, on trouve alors

$$\mathcal{A}\mathcal{P}^{-1}\mathcal{Q}\mathcal{A}^*(d_y) = \mathcal{A}\mathcal{P}^{-1}\mathcal{Q}(r_D) + r_P - \mathcal{A}\mathcal{P}^{-1}(r_C) = r_P - \mathcal{A}\mathcal{P}^{-1}(r_C - \mathcal{Q}(r_D)). \tag{20.35}$$

D'après les hypothèses, il existe des constantes $\gamma_1 > 0$ et $\gamma_2 > 0$ telles que $\langle \mathcal{P}^{-1}\mathcal{Q}(X), X\rangle \geqslant \gamma_1 \|X\|_F^2$ pour tout $X \in \mathcal{S}^n$ et $\|\mathcal{A}^*(y)\| \geqslant \gamma_2\|y\|$ pour tout $y \in \mathbb{R}^m$. On en déduit que $\langle \mathcal{A}\mathcal{P}^{-1}\mathcal{Q}\mathcal{A}^*(y), y\rangle \geqslant \gamma_1\|\mathcal{A}^*(y)\|^2 \geqslant \gamma_1\gamma_2^2\|y\|^2$, si bien que $\mathcal{A}\mathcal{P}^{-1}\mathcal{Q}\mathcal{A}^*$ est inversible. Alors (20.35) permet de déterminer d_y. La première équation donne alors d_S et la troisième d_X. □

Observons que $X \in \mathcal{S}^n_{++}$ est solution de (P_μ) si, et seulement si, il existe $(y, S) \in \mathbb{R}^m \times \mathcal{S}^n_{++}$ tel que l'on ait (20.31). De même $(y, S) \in \mathbb{R}^m \times \mathcal{S}^n_{++}$ est solution de (D_μ) si, et seulement si, il existe $X \in \mathcal{S}^n_{++}$ tel que l'on ait (20.31). On peut maintenant montrer l'existence du chemin central et en étudier la régularité, sous l'hypothèse suivante :

$$\mathcal{F}^s := \mathcal{F}^s_{\mathrm{P}} \times \mathcal{F}^s_{\mathrm{D}} \neq \varnothing. \tag{20.36}$$

Proposition 20.12 (existence et régularité du chemin central) *Si $\mathcal{F}^s_{\mathrm{D}} \neq \varnothing$ et $\mu > 0$, alors le système (20.31) a une solution ; celle-ci est unique en X et S.*

Si, de plus, $\mathcal{A}$ est surjective, la solution est également unique en y et l'application qui à $\mu > 0$ fait correspondre cette solution (X, y, S) est de classe C^∞.

Démonstration. Le lien entre les solutions X de (P_μ) et celles (X, y, S) de (20.31), nous permet de nous intéresser à (P_μ). L'unicité de (X, S) est due à la stricte convexité de ld dans le critère de (P_μ). Pour l'existence, on récrit le problème (P_μ) comme celui qui consiste à minimiser la fonction $\varphi : \mathcal{S}^n \to \mathbb{R} \cup \{+\infty\}$ définie en $X \in \mathcal{S}^n$ par

$$\varphi(X) = \langle C, X\rangle + \mu\,\mathrm{ld}(X) + \mathcal{I}_{\mathscr{A}}(X).$$

where $\mathscr{A} := \{X \in \mathcal{S}^n : \mathcal{A}(X) = b\}$. Il suffit de vérifier que $\varphi^\infty(D) > 0$ lorsque $D \in \mathcal{S}^n \setminus \{0\}$ (proposition 3.29). En utilisant (20.33), on obtient

$$\varphi^\infty(D) = \langle C, D\rangle + \mathcal{I}_{\mathcal{S}^n_+}(D) + \mathcal{I}_{\mathcal{N}(\mathcal{A})}(D).$$

Dès lors, si $\varphi^\infty(D) \leqslant 0$, on a

$$\langle C, D\rangle \leqslant 0, \quad D \succcurlyeq 0 \quad \text{et} \quad \mathcal{A}(D) = 0.$$

Comme $\mathcal{F}_{\mathrm{D}}^s \neq \varnothing$, on peut trouver $(y_0, S_0) \in \mathbb{R}^m \times \mathcal{S}^n_{++}$ tel que $\mathcal{A}^*(y_0) + S_0 = C$. En multipliant scalairement cette équation par D, on trouve $\langle S_0, D\rangle \leqslant 0$ et donc $D = 0$ car $S_0 \succ 0$ et $D \succcurlyeq 0$ (proposition 20.1).

Supposons maintenant que $\mathcal{A}$ soit surjective. Alors la solution de (20.31) est clairement également unique en y. Pour la régularité du chemin central, on récrit (20.31) sous la forme d'un système d'équations non linéaires $F(X, y, S, \mu) = 0$, paramétrées par $\mu > 0$. La fonction F est définie par

$$F(X, y, S, \mu) = \begin{pmatrix} \mathcal{A}^*(y) + S - C \\ \mathcal{A}(X) - b \\ -\mu X^{-1} + S \end{pmatrix}.$$

Prendre $XS - \mu I$ comme dernière composante pose des difficultés par la suite, car XS n'est pas symétrique et l'on ne peut pas appliquer $\mathcal{A}$. Comme F est de classe C^∞, la régularité du chemin central sera une simple conséquence du théorème des fonctions implicites (théorème C.14), si l'on montre que

$$F'_{(X,y,S)}(X, y, S, \mu) = \begin{pmatrix} 0 & \mathcal{A}^* & I \\ \mathcal{A} & 0 & 0 \\ \mu X^{-1} \otimes X^{-1} & 0 & I \end{pmatrix}$$

est inversible. Le lemme 20.11 répond à cette question par l'affirmative, puisque $X^{-1} \otimes X^{-1}$ est un opérateur défini positif sur $\mathcal{S}^n$. □

20.4.2 Remarques générales sur les algorithmes

Opérateur de Lyapounov

L'*opérateur de Lyapounov* associé à une matrice $A \in \mathcal{S}^n$ est l'application linéaire $L_A : \mathcal{S}^n \to \mathcal{S}^n$ dont la valeur en $X \in \mathcal{S}^n$ est donnée par

$$L_A(X) = AX + XA.$$

Étant donné $B \in \mathcal{S}^n$, on s'intéresse au système linéaire en $X \in \mathcal{S}^n$ suivant, appelé *système de Lyapounov*

$$AX + XA = B. \tag{20.37}$$

C'est un cas particulier d'*équation de Sylvester*, laquelle est une équation en X de la forme $A_1X + XA_2 = B$. Le lemme ci-dessous donne une condition sur A pour que ce système ait une solution unique dans $\mathcal{S}^n$. On y a noté $\lambda(A)$ l'ensemble des valeurs propres de A. L'hypothèse $0 \notin \lambda(A) + \lambda(A)$ est donc vérifiée si $A \in \mathcal{S}^n_{++}$.

Lemme 20.13 *Si $A \in \mathcal{S}^n$ vérifie $0 \notin \lambda(A) + \lambda(A)$ et si $B \in \mathcal{S}^n$, alors le système de Lyapounov (20.37) a une solution et une seule dans $\mathcal{S}^n$.*

Démonstration. Il suffit d'exhiber la solution. Soit $A = V\Lambda V^\mathsf{T}$ la factorisation spectrale de A. Le système (20.37) devient alors

$$\Lambda Y + Y\Lambda = C,$$

où $Y = V^\mathsf{T}XV$ et $C = V^\mathsf{T}BV$. En prenant l'élément (i, j), on trouve $(\lambda_i + \lambda_j)Y_{ij} = C_{ij}$, ce qui permet de déterminer Y (car $\lambda_i + \lambda_j \neq 0$ lorsque $0 \notin \lambda(A) + \lambda(A)$) donc $X = VYV^\mathsf{T}$. □

Si A vérifie la condition du lemme, A est inversible. Dans ce cas, la solution du système de Lyapounov avec $B = A^p$, $p \in \mathbb{N}$, se calcule aisément :

$$L_A^{-1}(A^p) = \tfrac{1}{2}A^{p-1}.$$

Mesures du saut de dualité et du centrage

La *mesure du saut de dualité* est la quantité

$$\bar{\mu} \equiv \bar{\mu}(z) := \frac{1}{n}\langle X, S\rangle.$$

On observe que $z \in \mathcal{F}^s$ est sur le chemin central si, et seulement si, $X^{1/2}SX^{1/2} = \bar{\mu}(z)I$. On préfère cette expression à $XS = \bar{\mu}(z)I$, car $X^{1/2}SX^{1/2}$ est symétrique, ce qui n'est pas nécessairement le cas de XS. Ceci conduit à la définition de divers voisinages du chemin central, permettant de mesurer le centrage de $z \in \mathcal{F}^s$.

Pour $\theta \in]0, 1[$, on définit le « *petit voisinage* »

$$V_F(\theta) := \{z \in \mathcal{F}^s : \|X^{1/2}SX^{1/2} - \bar{\mu}(z)I\|_F \leqslant \theta\bar{\mu}(z)\},$$

où $\|\cdot\|_F$ désigne la norme de Frobenius : pour une matrice A symétrique, elle est définie par $\|A\|_F := \langle A, A\rangle^{1/2} = (\sum_{i=1}^n \lambda_i^2(A))^{1/2}$.

Pour $\theta \in]0, 1[$, on définit le « *grand voisinage* »

$$\begin{aligned} V_\infty^-(\theta) &:= \{z \in \mathcal{F}^s : \|X^{1/2}SX^{1/2} - \bar{\mu}(z)I\|_{-\infty} \leqslant \theta\bar{\mu}(z)\} \\ &= \{z \in \mathcal{F}^s : \lambda_{\min}(XS) \geqslant (1-\theta)\bar{\mu}(z)\}, \end{aligned}$$

où $\|\cdot\|_{-\infty}$ est définie en $A \in \mathcal{S}^n$ par $\|A\|_{-\infty} := -\lambda_{\min}(A)$. La dernière identité vient du fait que $z \in \mathcal{F}^s$ est dans $V_\infty^-(\theta)$ si, et seulement si, $\bar{\mu}(z) - \lambda_{\min}(X^{1/2}SX^{1/2}) \leqslant \theta\bar{\mu}(z)$ ou encore si, et seulement si, $(1-\theta)\bar{\mu}(z) \leqslant \lambda_{\min}(XS)$ (on a utilisé le fait que XS, SX et $X^{1/2}SX^{1/2}$ ont le même spectre).

Schéma algorithmique

Il est temps de concrétiser la structure d'un algorithme de points-intérieurs primal-dual de suivi de chemin pour résoudre le problème SDP (20.1). C'est l'objectif du schéma algorithmique ci-dessous, qui laisse cependant beaucoup de points dans le vague : le voisinage du chemin central n'est pas présisé et le système sur lequel est calculé la direction de Newton n'est pas donné.

Schéma algorithmique 20.14 (PI en OSDP, approche primale-duale) On se donne un voisinage du chemin central $V(\theta)$, avec $\theta \in]0,1[$. L'itéré courant z est supposé dans $V(\theta)$. L'itéré suivant $z_+ \in V(\theta)$ s'obtient par les étapes suivantes.

1. Calcul d'une direction de Newton d, pour une valeur de $\mu = \sigma\bar{\mu}(z)$ avec $\sigma \in]0.1[$.
2. Calcul d'un pas $\alpha \in]0,1]$ tel que $z + \alpha d \in V(\theta)$.
3. Nouvel itéré : $z_+ = z + \alpha d$.

Les directions de Newton primales-duales ▲

Si en optimisation linéaire, la direction de Newton primale-duale (18.34) s'impose comme solution des équations (18.3) linéarisées, il n'en va pas de même en optimisation SDP. La difficulté vient du fait que XS n'est pas nécessairement symétrique si l'équation

$$XS = \mu I \tag{20.38}$$

n'est pas vérifiée par l'itéré courant (ce qui est bien le cas, on le verra), si bien que le système (20.31) a plus d'équations que d'inconnues. D'ailleurs, si le système (20.31) linéarisé a une solution (d_X, d_y, d_S), la matrice d_S sera symétrique par la première équation linéarisée (car $d_S = C - \mathcal{A}^*(y + d_y) - S \in \mathcal{S}^n$), mais on connait des exemples qui dans lesquels la troisième équation linéarisée (à savoir $d_X\, S + X\, d_S = \mu I - XS$) peut avoir une solution d_X non symétrique [395]. La question centrale est donc de savoir comment linéariser cette équation.

Une première possibilité serait de prendre comme troisième équation

$$-\mu X^{-1} + S = 0,$$

comme on l'a fait dans la démonstration de la proposition 20.12. On retrouve alors la méthode primale : la direction de Newton est identique à celle que l'on aurait en appliquant l'algorithme de Newton sur (P_μ). Des algorithmes utilisant cette direction ont été proposés, mais uniquement avec de petits déplacements. En pratique, les pas qu'elle autorise ne sont pas aussi grands que ceux acceptables par les méthodes primales-duales que nous considérons plus loin, ce qui ralentit la convergence.

Les algorithmes primaux-duaux cherchent à symétriser l'équation (20.38), sans utiliser d'inverse de matrice. La première approche de ce type est due à Alizadeh,

Haeberly et Overton [9] et la direction de Newton qu'elle calcule porte le nom de *direction d'AHO*. La dernière équation est dans un premier temps symétrisée[1],

$$\frac{1}{2}(XS+SX)=\mu I,$$

et ensuite linéarisée :

$$\frac{1}{2}(Sd_X+d_XS+Xd_S+d_SX)=\mu I-\frac{1}{2}(XS+SX). \tag{20.39}$$

Le membre de gauche est de la forme $\mathcal{P}d_X+\mathcal{Q}d_S$ avec $\mathcal{P}=S\otimes I$ et $\mathcal{Q}=X\otimes I$. Malheureusement ce système peut être mal défini en certain point de $\mathcal{F}^s$ [590] (il est cependant bien défini dans un petit voisinage du chemin central [432]).

Une approche plus générale consiste à procéder en trois étapes.

1) *Transformation par similitude.* On se donne une matrice R inversible, dépendant de X et S, et l'on remplace l'équation (20.38) par l'équation similaire

$$RXSR^{-1}=\mu I.$$

2) *Symétrisation.* L'équation précédente devient

$$\frac{1}{2}(RXSR^{-1}+R^{-\mathsf{T}}SXR^{\mathsf{T}})=\mu I.$$

Si l'on introduit l'opérateur linéaire de symétrisation $\mathcal{S}_R:\mathbb{R}^{n\times n}\to\mathcal{S}^n$ défini en $U\in\mathbb{R}^{n\times n}$ par

$$\mathcal{S}_R(U)=\frac{1}{2}(RUR^{-1}+R^{-\mathsf{T}}U^{\mathsf{T}}R^{\mathsf{T}}),$$

l'équation symétrisée précédente s'écrit

$$\mathcal{S}_R(XS)=\mu I.$$

3) *Pseudo-linéarisation.* On linéarise l'équation précédente, sans linéariser R (qui dépend pourtant de (X,S)) :

$$\mathcal{S}_R(d_XS+Xd_S)=\mu I-\mathcal{S}_R(XS). \tag{20.40}$$

Cette procédure en trois étapes conduit à ce que l'on appelle la *famille de directions de MZ* (Monteiro et Zhang).

Que prendre pour matrice R ? Dans les trois propositions suivantes, qui sont les plus implémentées, R est choisi de telle sorte que $RXSR^{-1}$ soit symétrique.

- *Direction de HKM primale* [309, 361, 428] : $R=S^{1/2}$. Par ce choix, (20.40) devient

$$\boxed{d_X+\tfrac{1}{2}\left(Xd_SS^{-1}+S^{-1}d_SX\right)=\mu S^{-1}-X.} \tag{20.41}$$

Ceci correspond à $\mathcal{P}=\mathrm{I}_{\mathcal{S}^n}$ et $\mathcal{Q}=X\otimes S^{-1}$ dans (20.34).

[1] L'opération $\circ:(X,S)\in\mathcal{S}^n\mapsto X\circ S:=\frac{1}{2}(XS+SX)$ fait de $(\mathcal{S}^n,\circ)$ une algèbre de Jordan [201, 357, 8 ; 1994-2012].

- *Direction de HKM duale* [309, 361, 428] : $R = X^{-1/2}$. Par ce choix, (20.40) devient

$$\boxed{\tfrac{1}{2}\left(X^{-1}d_X S + S d_X X^{-1}\right) + d_S = \mu X^{-1} - S.} \tag{20.42}$$

  Ceci correspond à $\mathcal{P} = X^{-1} \otimes S$ et $\mathcal{Q} = \mathrm{I}_{\mathcal{S}^n}$ dans (20.34).

- *Direction de NT* [457, 590] : $R = W^{-1/2}$, où W est l'unique matrice symétrique définie positive telle que $WSW = X$, c'est-à-dire

$$W = X^{1/2}(X^{1/2}SX^{1/2})^{-1/2}X^{1/2} = S^{-1/2}(S^{1/2}XS^{1/2})^{1/2}S^{-1/2}.$$

  Par ce choix, (20.40) devient

$$\boxed{d_X + W d_S W = \mu S^{-1} - X.} \tag{20.43}$$

  Ceci correspond à $\mathcal{P} = \mathrm{I}_{\mathcal{S}^n}$ et $\mathcal{Q} = W \otimes W$ dans (20.34).

Complexité : NT en $O(nL)$ itérations, HKM en $O(n^{3/2}L)$ itérations.

Notes

Les équivalences $(ii) \Leftrightarrow (iv)$ et $(ii') \Leftrightarrow (iv')$ proposition 20.4 sont des expressions équivalentes aux alternatives de Trnovská [595]. Nos présentation et démonstration (laquelle utilisent directement le lemme de Farkas) sont toutefois différentes.

La dualité lagrangienne présentée à la section 20.1.1 remonte à Bellman et Fan [41 ; 1963] qui utilisaient déjà une hypothèse de qualification à la Slater pour garantir l'absence de saut de dualité. Ramana [504, 505 ; 1997] a proposé une théorie de la dualité en optimisation SDP, différente de la dualité lagrangienne ordinaire proposée ici, sans saut de dualité même lorsque $\mathcal{F}^s$ est vide. Cependant, la taille du problème dual le rend moins accessible au calcul.

Ce chapitre doit beaucoup aux revues d'Alizadeh [7 ; 1995], de Vandenberghe et Boyd [603 ; 1996], de Monteiro et Todd [431 ; 2000], de Todd [589 ; 2001] et au livre de Nemirovskii et Nesterov [456 ; 1994]. Les monographies de Saigal, Vandenberghe et Wolkowicz [626 ; 2000], de de Klerk [158 ; 2002] et d'Anjos et Lasserre [15 ; 2012] sont riches en information sur la théorie, l'algorithmique et les applications. Monteiro [429 ; 2003] passe en revue les méthodes récentes de résolution des problèmes d'OSDP, incluant des approches n'utilisant pas la notion de points intérieurs.

Exercices

20.1. *Faces du cône* $\mathcal{S}^n_+$. Pour une partie F de $\mathcal{S}^n$, les propriétés suivantes sont équivalentes :

(i) F est une face de $\mathcal{S}^n_+$,

(ii) (caractérisation par les matrices de rang 1) $F = \mathrm{cone}\{vv^\mathsf{T} : v \in E\}$, où E est un sous-espace vectoriel de $\mathbb{R}^n$; dans ce cas,

- $\mathrm{aff}\, F = \mathrm{vect}\{vv^\mathsf{T} : v \in E\}$,
- $\dim F = d(d+1)/2$ où $d := \dim E$;

(*iii*) (caractérisation par l'image) $F = \{A \in \mathcal{S}^n_+ : \mathcal{R}(A) \subseteq E\}$, où E est un sous-espace vectoriel de $\mathbb{R}^n$; dans ce cas,
- aff $F = \{A \in \mathcal{S}^n : \mathcal{R}(A) \subseteq E\}$,
- $\dim F = d(d+1)/2$ où $d := \dim E$;

(*iv*) (caractérisation par le noyau) $F = \{A \in \mathcal{S}^n_+ : E \subseteq \mathcal{N}(A)\}$, où E est un sous-espace vectoriel de $\mathbb{R}^n$; dans ce cas,
- aff $F = \{A \in \mathcal{S}^n : E \subseteq \mathcal{N}(A)\}$,
- $\dim F = d(d+1)/2$ où $d := n - \dim E$.

Ce résultat a pour corollaires :

1) toute face F de $\mathcal{S}^n_+$ est exposée : il existe $S \in \mathcal{S}^n$ telle que $F = \arg\min\{\langle S, A\rangle : A \in \mathcal{S}^n_+\}$,
2) si $A \in \mathcal{S}^n_+$, la face engendrée par A s'écrit $F(A) = \{B \in \mathcal{S}^n_+ : \mathcal{R}(B) \subseteq \mathcal{R}(A)\} = \{B \in \mathcal{S}^n_+ : \mathcal{N}(A) \subseteq \mathcal{N}(B)\}$.

20.2. *Dualisation de la contrainte d'appartenance au cône* $\mathcal{S}^n_+$. Montrez que l'on peut retrouver le dual (20.5) en dualisant la contrainte $X \succcurlyeq 0$ de (20.1), plutôt que la contrainte $\mathcal{A}(X) = b$.

20.3. *Point-selle du lagrangien.* Le couple (X, y) est un point-selle du lagrangien ℓ défini en (20.4) sur $\mathcal{S}^n_+ \times \mathbb{R}^m$ si, et seulement si, $X \in \mathcal{F}_{\text{P}}$, $(y, S) \in \mathcal{F}_{\text{D}}$ et $\langle X, S\rangle = 0$ (on a noté $S := C - \mathcal{A}^*(y)$).

20.4. Démontrez le point 2 de la proposition 20.4.

20.5. *Semi-définie positivité et rang.* Pour $X \in \mathcal{S}^n$, on a l'équivalence suivante :

$$X \succcurlyeq 0 \qquad \Longleftrightarrow \qquad \exists\, A \in \mathbb{R}^{n\times n} : \ \operatorname{rg}\begin{pmatrix} I_n & A \\ A^{\mathsf{T}} & X \end{pmatrix} \leqslant n.$$

20.6. *Formulation autoduale.* On suppose que $b \in \mathcal{R}(\mathcal{A})$. Montrez que l'on peut écrire les problèmes primal et dual de l'optimisation SDP de manière similaire, comme ci-dessous :

$$(P) \quad \begin{cases} \inf \langle C, X\rangle \\ X \in D + \mathcal{L}^\perp \\ X \succcurlyeq 0 \end{cases} \qquad \text{et} \qquad (D) \quad \begin{cases} \inf \langle D, S\rangle \\ S \in C + \mathcal{L} \\ S \succcurlyeq 0, \end{cases}$$

où $D \in \mathcal{S}^n$, $\mathcal{L}$ est un sous espace vectoriel de $\mathcal{S}^n$ et $\mathcal{L}^\perp$ son orthogonal dans $\mathcal{S}^n$.

20.7. *Complement de Schur singulier.* Soient $A \in \mathcal{S}^n$, $C \in \mathcal{S}^m$ et $B \in \mathbb{R}^{n\times m}$. On note $A^\dagger$ le pseudo-inverse de A. Montrez que

$$\begin{pmatrix} A & B \\ B^{\mathsf{T}} & C \end{pmatrix} \succcurlyeq 0 \quad \Longleftrightarrow \quad A \succcurlyeq 0, \ \ \mathcal{R}(B) \subseteq \mathcal{R}(A) \ \ \text{et} \ \ C - B^{\mathsf{T}} A^\dagger B \succcurlyeq 0. \tag{20.44}$$

20.8. *Relaxation de rang du problème tout quadratique.* Montrez que le bidual lagrangien du problème (20.26) est le problème (20.28).

20.9. *Produit tensoriel.* Montrez que l'application de $\mathcal{S}^n \times \mathcal{S}^n \to L(\mathcal{S}^n, \mathcal{S}^n) : (X, Y) \mapsto X \otimes Y$, où $X \otimes Y$ est définie par (20.32), est un produit tensoriel.

20.10. *Fonction asymptotique de* ld. Montrez que ld^∞ est donné par la formule (20.33).

Annexes

A Analyse ▲

A.1 Topologie

During the 20th century, real and complex analysis relied heavily on the concepts of open set, closed set, and limit point of a set. The earliest idea was that of the limit point of a set, due to Weierstrass, while that of a closed set (due to Cantor) arose somewhat later. The idea of an open set came latest of all. The very slow diffusion of the concept of open set is surprising in view of its importance now.

G. H. Moore [435 ; 2008].

Une *topologie* sur un ensemble $\mathbb{E}$ est une famille de parties de $\mathbb{E}$, appelées *ouverts*, vérifiant les propriétés suivantes : (i) $\mathbb{E}$ et l'ensemble vide (noté $\varnothing$) sont des ouverts, (ii) une intersection finie d'ouverts est un ouvert et (iii) une réunion quelconque d'ouverts est un ouvert. Un *espace topologique* $\mathbb{E}$ est un ensemble $\mathbb{E}$ muni d'une topologie.

Une partie d'un espace topologique $\mathbb{E}$ est un *fermé* si son complémentaire est ouvert. Par complémentarité des propriétés axiomatiques des ouverts, on voit que $\mathbb{E}$ et $\varnothing$ sont aussi des fermés, qu'une réunion finie de fermés est un fermé et qu'une intersection quelconque de fermés est un fermé.

Un *voisinage* d'un point $x \in \mathbb{E}$ est une partie V de $\mathbb{E}$ qui contient un ouvert contenant x. Une partie est ouverte si, et seulement si, elle est voisinage de tous ses points (c'est une caractérisation souvent utilisée). On dit qu'une topologie est *séparée* si deux points distincts quelconques de $\mathbb{E}$ ont des voisinages disjoints (d'intersection vide). Dans une topologie séparée, les singletons (ensembles formés d'un unique élément) sont des fermés.

Une *suite* d'un ensemble $\mathbb{E}$ est une application de l'ensemble des entiers $\mathbb{N}$ dans $\mathbb{E}$, $k \in \mathbb{N} \mapsto x_k \in \mathbb{E}$, que l'on désigne souvent par son image $\{x_k\}_{k\in\mathbb{N}}$ ou simplement $\{x_k\}$. On dit qu'une suite $\{x_k\}_{k\in\mathbb{N}}$ de $\mathbb{E}$ *converge* vers $x \in \mathbb{E}$, ou encore que x est *limite* de cette suite, si pour tout voisinage V de x, il existe un indice $k_V \in \mathbb{N}$, tel que pour tout $k \geqslant k_V$, on ait $x_k \in V$. On peut exprimer cela en utilisant des quantificateurs :

$$\forall\, V \text{ voisinage de } x, \quad \exists\, k_V \in \mathbb{N}, \quad \forall\, k \geqslant k_V : \qquad x_k \in V.$$

Dans une topologie séparée, les suites ont au plus une limite. Un exemple trivial de suite convergente est une *suite stationnaire* qui, par définition, est telle que $x_{k+1} = x_k$ pour tout indice k supérieur à un indice donné.

L'*intérieur* d'une partie A de $\mathbb{E}$, notée A° ou $\mathrm{int}(A)$, est la réunion de tous les ouverts contenus dans A. C'est donc un ouvert et même le plus grand ouvert contenu dans A. L'*adhérence* d'une partie A de $\mathbb{E}$, notée $\overline{A}$ ou $\mathrm{adh}(A)$, est l'intersection de tous les fermés contenant A. C'est donc un fermé et même le plus petit fermé contenant A. Un point est dans l'adhérence de A si, et seulement si, tout voisinage de ce point rencontre A. On dit qu'une partie A de $\mathbb{E}$ est *dense* dans $\mathbb{E}$ si $\overline{A} = \mathbb{E}$. La *frontière* d'une partie A de $\mathbb{E}$, notée ∂A, est l'ensemble des points de $\overline{A}$ qui ne sont pas dans A° : $\partial A = \overline{A} \setminus A^\circ$.

Si $\mathbb{E}_0$ est une partie d'un espace topologique $\mathbb{E}$, on peut mettre une topologie sur $\mathbb{E}_0$ à partir de celle de $\mathbb{E}$. On appelle *topologie induite* de $\mathbb{E}_0$, celle qui est définie comme suit : on dit qu'une partie de $\mathbb{E}_0$ est ouverte dans $\mathbb{E}_0$ si elle est l'intersection avec $\mathbb{E}_0$ d'un ouvert de $\mathbb{E}$.

A.2 Compacité

Pour étudier un problème, on est souvent amené à examiner une suite de problèmes « voisins », plus faciles à traiter. On pourra étendre les propriétés de ces problèmes au problème original si le passage à la limite peut se faire. La notion de compacité est très utile dans ce cas car elle permet d'extraire des sous-suites convergentes à partir de suites bornées (en tout cas dans les espaces normés), facilitant ainsi le passage à la limite.

On dit qu'un espace topologique $\mathbb{E}$ est *compact* s'il est séparé et si la *propriété de Heine-Borel-Lebesgue* suivante est vérifiée : de tout recouvrement de $\mathbb{E}$ par des ouverts on peut extraire un sous-recouvrement fini. Une partie d'un espace topologique est dite *compacte* si, munie de la topologie induite, cet ensemble est compact. La propriété de Heine-Borel-Lebesgue ne donne pas une idée claire de ce qu'est un compact, mais c'est un outil très utile pour démontrer la compacité d'un espace topologique. Ce seront plutôt les propriétés qu'ont les ensembles compacts qui permettront de se familiariser avec le concept de compacité.

Une partie compacte est fermée [obtenir un recouvrement en utilisant la propriété de séparation entre les points de cette partie et un point qui ne lui appartient pas]. Un fermé dans un espace compact est aussi compact (pour la topologie induite) [c'est clair en utilisant la propriété de Heine-Borel-Lebesgue].

On dit que x est un *point d'accumulation* d'une suite $\{x_k\}_{k\geqslant 1}$ si tout voisinage de x contient des points x_k pour une infinité d'indices k. Toute suite d'une espace compact admet un point d'accumulation [car l'intersection de la suite décroissante de fermés non vides $\overline{\{x_i\}_{i\geqslant k}}$ est non vide].

A.3 Fonction injective, surjective, continue

Soient $\mathbb{E}$ et $\mathbb{F}$ deux ensembles et $f : \mathbb{E} \to \mathbb{F}$ une application. On définit comme suit l'*image directe* (resp. l'*image réciproque*) d'une partie A de $\mathbb{E}$ (resp. B de $\mathbb{F}$) par f :

$$f(A) := \{f(x) \in \mathbb{F} : x \in A\} \quad (\text{resp. } f^{-1}(B) := \{x \in \mathbb{E} : f(x) \in B\}).$$

On dit que f est *injective* si l'image réciproque d'un singleton n'a pas plus d'un élément (il peut n'en avoir aucun), ce qui revient à dire que $x_1 = x_2$ chaque fois que $f(x_1) = f(x_2)$. On dit que f est *surjective* si $f(\mathbb{E}) = \mathbb{F}$.

Si maintenant $\mathbb{E}$ et $\mathbb{F}$ sont des espaces topologiques, on dit que f est *continue en un point* $x \in \mathbb{E}$, si pour tout voisinage V de $f(x)$ dans $\mathbb{F}$, il existe un voisinage U de x dans $\mathbb{E}$ tel que $f(U) \subseteq V$. Il revient au même de dire que l'image réciproque $f^{-1}(V)$ de tout voisinage V de $f(x)$ est un voisinage de x. On dit que f est *continue* si elle est continue en tout point de $\mathbb{E}$. Il en sera ainsi si, et seulement si, l'image réciproque d'un ouvert de $\mathbb{F}$ est un ouvert de $\mathbb{E}$ ou encore si, et seulement si, l'image réciproque d'un fermé de $\mathbb{F}$ est un fermé de $\mathbb{E}$. D'autre part, si $f : \mathbb{E} \to \mathbb{F}$ est continue et si $K \subseteq \mathbb{E}$ est une partie compacte de $\mathbb{E}$, alors l'image directe $f(K)$ de K par f est compacte [utiliser la propriété de Heine-Borel-Lebesgue].

On note $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, +\infty\}$ la *droite achevée*. Soit $\mathbb{E}$ un ensemble et $\varphi : \mathbb{E} \to \overline{\mathbb{R}}$ une fonction pouvant prendre les valeurs $-\infty$ et $+\infty$. On appelle *épigraphe* de φ, l'ensemble

$$\operatorname{epi} \varphi := \{(x, \alpha) \in \mathbb{E} \times \mathbb{R} : \varphi(x) \leqslant \alpha\}.$$

C'est donc la partie de $\mathbb{E} \times \mathbb{R}$ qui est « au-dessus » du graphe de φ (du grec *epi* signifiant *sur*).

Soit $\{a_k\}_{k\in\mathbb{N}}$ une suite d'éléments de $\overline{\mathbb{R}}$. Comme la suite $\{\inf_{k'\geqslant k} a_{k'}\}_{k\in\mathbb{N}}$ est croissante, elle a une limite dans $\overline{\mathbb{R}}$ que l'on appelle la *limite inférieure* de la suite $\{a_k\}_{k\in\mathbb{N}}$ et que l'on note

$$\liminf_{k\to\infty} a_k := \lim_{k\to\infty} \left(\inf_{k'\geqslant k} a_{k'} \right) \in \overline{\mathbb{R}}.$$

De manière similaire, on définit la *limite supérieure* de $\{a_k\}_{k\in\mathbb{N}}$ par

$$\limsup_{k\to\infty} a_k := \lim_{k\to\infty} \left(\sup_{k'\geqslant k} a_{k'} \right) \in \overline{\mathbb{R}}.$$

Comme $\sup_k a_k = -\inf_k (-a_k)$, on a

$$\limsup_{k\to\infty} a_k = -\liminf_{k\to\infty} (-a_k). \tag{A.1}$$

Si $\{a_k\}_{k\in\mathbb{N}}$ et $\{b_k\}_{k\in\mathbb{N}}$ sont deux suites de $\overline{\mathbb{R}}$, alors

$$\liminf_{k\to\infty} (a_k + b_k) \quad \geqslant \quad \liminf_{k\to\infty} a_k + \liminf_{k\to\infty} b_k, \tag{A.2a}$$

$$\limsup_{k\to\infty} (a_k + b_k) \quad \leqslant \quad \limsup_{k\to\infty} a_k + \limsup_{k\to\infty} b_k, \tag{A.2b}$$

pourvu que l'une des limites inférieures ou supérieures des membres de droite ne vaille pas $-\infty$ et l'autre $+\infty$. On a des égalités dans (A.2a) et (A.2b) si l'une des deux suites converge. Si $\{a_k\}_{k\in\mathbb{N}}$ et $\{b_k\}_{k\in\mathbb{N}}$ sont des suites positives de $\overline{\mathbb{R}}$, alors

$$\liminf_{k\to\infty} a_k b_k \geqslant \left(\liminf_{k\to\infty} a_k \right) \left(\liminf_{k\to\infty} b_k \right), \tag{A.3}$$

pourvu qu'un des facteurs à droite ne soit pas nul et l'autre $+\infty$.

Soient $\mathbb{E}$ est un espace topologique et $\varphi : \mathbb{E} \to \overline{\mathbb{R}}$ une fonction. On dit que φ est *semi-continue inférieurement* (*s.c.i.* en abrégé) en $x \in \mathbb{E}$ si

$$\forall\, \{x_k\} \to x : \qquad \varphi(x) \leqslant \liminf_{k\to\infty} \varphi(x_k).$$

On dit que φ est *s.c.i.* si elle est s.c.i. en tout point de $\mathbb{E}$. On dit que φ est *fermée* si son épigraphe est fermé dans $\mathbb{E}\times\mathbb{R}$. Comme l'affirme la proposition suivante, dont la démonstration est proposée à l'exercice A.2, il s'agit de deux notions équivalentes.

Proposition A.1 *Soient $\mathbb{E}$ est un espace topologique et $\varphi : \mathbb{E} \to \overline{\mathbb{R}}$ une fonction. Les trois propriétés suivantes sont équivalentes:*
(i) φ est fermée,
(ii) φ est s.c.i.,
(iii) $\forall\, \alpha \in \mathbb{R}$, l'ensemble $\{x \in \mathbb{E} : \varphi(x) \leqslant \alpha\}$ est fermé.

On dit que φ est *semi-continue supérieurement* (*s.c.s.* en abrégé) [en $x \in \mathbb{E}$] si $-\varphi$ est s.c.i [en $x \in \mathbb{E}$]. Clairement, φ est *continue* [en $x \in \mathbb{E}$] si, et seulement si, elle est s.c.i. et s.c.s. [en $x \in \mathbb{E}$].

A.4 Espace métrique

Définition

Soit $\mathbb{E}$ un ensemble. Une *distance* sur $\mathbb{E}$ est une application $\mathrm{d} : \mathbb{E}\times\mathbb{E} \to \mathbb{R}$ ayant les propriétés suivantes : pour tout x, y, $z \in \mathbb{E}$, *(i)* $\mathrm{d}(x,y) = 0$ si, et seulement si, $x = y$, *(ii) symétrie* : $\mathrm{d}(x,y) = \mathrm{d}(y,x)$, *(iii) inégalité triangulaire* : $\mathrm{d}(x,z) \leqslant \mathrm{d}(x,y) + \mathrm{d}(y,z)$. On a tout de suite les propriétés suivantes : pour tout x, y, $z \in \mathbb{E}$,

$$\mathrm{d}(x,y) \geqslant 0 \qquad \text{et} \qquad |\,\mathrm{d}(x,y) - \mathrm{d}(y,z)| \leqslant \mathrm{d}(x,z).$$

Un *espace métrique* est un couple $(\mathbb{E},\mathrm{d})$ formé d'un ensemble $\mathbb{E}$ muni d'une distance d. Dans un espace métrique $(\mathbb{E},\mathrm{d})$, on définit la *distance d'un point $x \in \mathbb{E}$ à un ensemble $P \subseteq \mathbb{E}$* par

$$\mathrm{d}_P(x) = \inf_{y\in P} \mathrm{d}(x,y). \tag{A.4}$$

Un point $\bar{x} \in P$ qui réalise l'infimum ci-dessus est appelé une *projection* de x sur P. Celle-ci peut ne pas exister ou exister mais ne pas être unique (voir la section 2.5.1).

Topologie

La *topologie* canonique d'un espace métrique $(\mathbb{E},\mathrm{d})$ est introduite comme suit. On définit d'abord la *boule ouverte* [resp. *boule fermée*] de *centre* x et de *rayon* r, comme l'ensemble

$$B(x,r) := \{y \in \mathbb{E} : \mathrm{d}(x,y) < r\} \qquad [\text{resp. } \bar{B}(x,r) := \{y \in \mathbb{E} : \mathrm{d}(x,y) \leqslant r\}].$$

On dit alors que $\Omega \subseteq \mathbb{E}$ est *ouvert* dans $\mathbb{E}$ si pour tout $x \in \Omega$, il existe un rayon $r > 0$ tel que $B(x,r) \subseteq \Omega$. La topologie d'un espace métrique est séparée ; en particulier, les singletons sont des fermés et les suites ont au plus une limite.

Fonctions lipschitzienne et contractante

Soient $(\mathbb{E}, \mathrm{d}_{\mathbb{E}})$ et $(\mathbb{F}, \mathrm{d}_{\mathbb{F}})$ deux espaces métriques et Ω un ouvert de $\mathbb{E}$.

On dit qu'une fonction $f : \Omega \to \mathbb{F}$ est *lipschitzienne* de *module* $L > 0$, si pour tout x, $y \in \Omega$,

$$\mathrm{d}_{\mathbb{F}}(f(x), f(y)) \leqslant L\ \mathrm{d}_{\mathbb{E}}(x, y).$$

Ce sont des fonctions dont la variation peut être estimée par celle de leur argument. On dit que $f : \Omega \to \mathbb{F}$ est *localement lipschitzienne* si tout point $x \in \Omega$ possède un voisinage V contenu dans Ω sur lequel f est lipschitzienne sur V.

On dit qu'une fonction $f : \Omega \to \mathbb{F}$ est *contractante*, si elle est lipschitzienne de module $L = 1$ et qu'elle est *strictement contractante*, si elle est lipschitzienne de module $L < 1$.

A.5 Espace normé

Définition

Soit $\mathbb{E}$ un espace vectoriel sur $\mathbb{R}$, une notion supposée connue. Si A et B sont des parties de $\mathbb{E}$ et $\alpha \in \mathbb{R}$ on définit la *somme de Minkowski* $A + B$ de A et B, leur différence $A - B$ et la multiplication de A par un scalaire α par

$$\begin{aligned} A + B &:= \{x + y : x \in A, y \in B\} \\ A - B &:= \{x - y : x \in A, y \in B\} \\ \alpha A &:= \{\alpha x : x \in A\}. \end{aligned}$$

On se gardera bien de penser que $A - A$ se réduit à $\{0\}$. Par exemple, avec $A := [0, 1] \subseteq \mathbb{R}$, on a $A - A = [-1, 1]$ (voir l'exercice A.6).

Une *norme* sur un espace vectoriel $\mathbb{E}$ est une application $x \in \mathbb{E} \mapsto \|x\| \in \mathbb{R}$ qui vérifie les propriétés suivantes : (i) $\|x\| = 0$ si, et seulement si, $x = 0$, (ii) pour tout $\alpha \in \mathbb{R}$ et tout $x \in \mathbb{E}$: $\|\alpha x\| = |\alpha|\, \|x\|$ et (iii) pour tout $x, y \in \mathbb{E}$: $\|x + y\| \leqslant \|x\| + \|y\|$ (*inégalité triangulaire*). Un *espace normé* est un espace vectoriel muni d'une norme. L'inégalité triangulaire implique que

$$\forall\, x, y \in \mathbb{E} : \qquad \Big|\|x\| - \|y\|\Big| \leqslant \|x - y\|.$$

On vérifie aisément que l'application $(x, y) \in \mathbb{E} \times \mathbb{E} \mapsto \|x - y\|$ est une distance sur $\mathbb{E}$; un espace normé est donc aussi un espace métrique. On définit alors la topologie canonique d'un espace normé comme la topologie canonique de l'espace métrique associé. Ses boules ouvertes [resp. fermées] sont donc de la forme

$$B(x, r) := \{y \in \mathbb{E} : \|y - x\| < r\} \qquad \text{et} \qquad \bar{B}(x, r) := \{y \in \mathbb{E} : \|y - x\| \leqslant r\}.$$

On dit qu'une partie A d'un espace normé est *bornée* s'il existe $r > 0$ tel que $A \subseteq B(0, r)$.

Sur $\mathbb{R}^n$ et pour $p \in [1, +\infty[\subseteq \mathbb{R}$, la fonction qui à $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ fait correspondre

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}} \tag{A.5a}$$

est une norme, dite *norme* ℓ_p ou *norme de Minkowski*. De même la fonction qui à $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ fait correspondre

$$\|x\|_\infty = \max_{1 \leqslant i \leqslant n} |x_i| \tag{A.5b}$$

est une norme, dite *norme* ℓ_∞. La norme ℓ_2 est aussi appelée *norme euclidienne*. On notera, qu'à $x \in \mathbb{R}^n$ fixé, l'application

$$p \in [1, +\infty[\mapsto \|x\|_p \text{ est décroissante}$$

et $\lim_{p \to \infty} \|x\|_p = \|x\|_\infty$.

Une partie compacte d'un espace normé est fermée (vrai dans tout espace topologique) et bornée. En dimension finie, la réciproque est vraie : une partie fermée et bornée est compacte. Dans un espace normé, toute suite incluse dans un compact admet une sous-suite convergente [on a vu qu'une telle suite avait un point d'accumulation ; la sous-suite s'obtient en considérant des boules centrées au point d'accumulation et de rayons décroissants].

On dit qu'une suite $\{x_k\}_{k \geqslant 1}$ est une *suite de Cauchy* si $\|x_k - x_l\|$ tend vers zéro lorsque k et l tendent vers l'infini ; de manière plus précise, elle est de Cauchy si

$$\forall\, \epsilon > 0, \quad \exists\, K \in \mathbb{N}, \quad \forall\, k, l \geqslant K : \qquad \|x_k - x_l\| < \epsilon.$$

Un espace normé est dit *complet* (on parle alors d'*espace de Banach*) si toutes ses suites de Cauchy convergent. L'intérêt de la notion de suite de Cauchy est de pouvoir, dans les espaces complets, reconnaître les suites convergentes sans en connaître la limite. Par exemple $\mathbb{R}^n$ muni d'une norme quelconque est un espace de Banach.

Soit $\{x_k\}_{k \geqslant 1}$ une suite d'un espace normé $\mathbb{E}$. On dit que la série $\sum_{k \geqslant 1} x_k$ est *convergente* si la suite des sommes partielles $\{\sum_{1 \leqslant k \leqslant l} x_k\}_{l \geqslant 1}$ converge dans $\mathbb{E}$. On dit que la série $\sum_{k \geqslant 1} x_k$ est *absolument convergente* si la série des normes $\sum_{k \geqslant 1} \|x_k\|$ est convergente dans $\mathbb{R}$. Dans un espace de Banach, une série $\sum_{k \geqslant 1} x_k$ absolument convergente est convergente et vérifie $\| \sum_{k \geqslant 1} x_k \| \leqslant \sum_{k \geqslant 1} \|x_k\|$.

On dit qu'une partie A d'un espace vectoriel $\mathbb{E}$ est un *sous-espace affine* de $\mathbb{E}$, s'il existe un point $x \in \mathbb{E}$ et un sous-espace vectoriel $\mathbb{E}(A)$ de $\mathbb{E}$ tels que $A = x + \mathbb{E}(A)$. Il revient au même de dire que A est non vide et que $tx + (1-t)y \in A$ lorsque $x, y \in A$ et $t \in \mathbb{R}$ (voir l'exercice A.7 pour une autre définition équivalente). On dit que $\mathbb{E}(A)$ est le sous-espace vectoriel *parallèle* au sous-espace affine A (il est déterminé de manière unique ; c'est en fait $A - A$) et sa *dimension* définit celle de A. Une intersection quelconque de sous-espaces affines est un sous-espace affine (dont le sous-espace vectoriel parallèle est l'intersection des sous-espaces vectoriels parallèles aux sous-espaces affines intersectés).

On dit qu'une fonction $T : \mathbb{E} \to \mathbb{F}$ entre deux espaces vectoriels $\mathbb{E}$ et $\mathbb{F}$ est *linéaire* si $T(\alpha x + \beta y) = \alpha T(x) + \beta T(y)$ pour tout α, $\beta \in \mathbb{R}$ et tout x, $y \in \mathbb{E}$. Une fonction linéaire T est injective si, et seulement si, $T(x) = 0$ implique que $x = 0$.

On dit qu'une fonction $f : \mathbb{E} \to \mathbb{F}$ entre deux espaces vectoriels $\mathbb{E}$ et $\mathbb{F}$ est *affine* s'il existe une fonction linéaire $T : \mathbb{E} \to \mathbb{F}$ telle que $f(y) - f(x) = T(y - x)$ pour tout x, $y \in \mathbb{E}$. Il revient au même de dire que $f(x) = f(0) + Tx$ pour tout $x \in \mathbb{E}$ ou encore que $f((1-t)x + ty) = (1-t)f(x) + tf(y)$ pour tout $t \in [0, 1]$ ou pour tout $t \in \mathbb{R}$.

Application linéaire continue

Soient $\mathbb{E}$ et $\mathbb{F}$ deux espaces normés, munis des normes $\|\cdot\|_{\mathbb{E}}$ et $\|\cdot\|_{\mathbb{F}}$ respectivement. On note $\mathcal{L}(\mathbb{E},\mathbb{F})$ l'ensemble des *applications linéaires continues* (ou *opérateurs continus*) de $\mathbb{E}$ dans $\mathbb{F}$. On vérifie qu'une application linéaire $T:\mathbb{E}\to\mathbb{F}$ est continue si, et seulement si,

$$\|T\| := \sup_{\|x\|_{\mathbb{E}}\leqslant 1} \|Tx\|_{\mathbb{F}}$$

est fini. L'application $T\mapsto\|T\|$ définit une norme sur $\mathcal{L}(\mathbb{E},\mathbb{F})$. Le *noyau* d'une application linéaire $T\in\mathcal{L}(\mathbb{E},\mathbb{F})$ est la partie de $\mathbb{E}$ définie par

$$\mathcal{N}(T) := \{x\in\mathbb{E} : Tx = 0\}.$$

L'*image* de T est la partie de $\mathbb{F}$ définie par[1]

$$\mathcal{R}(T) := T(\mathbb{E}) = \{Tx : x\in\mathbb{E}\}.$$

On dit qu'une application de $\mathcal{L}(\mathbb{E},\mathbb{F})$ est *inversible* si elle est bijective et si son application réciproque est continue (elle est nécessairement linéaire).

Lemme A.2 (de perturbation de Banach) *Soient $\mathbb{E}$ et $\mathbb{F}$ deux espaces de Banach. Si R et $T\in\mathcal{L}(\mathbb{E},\mathbb{F})$ sont tels que R est inversible et $\|R^{-1}T\|<1$, alors*
1) *$R-T$ est inversible,*
2) *$(R-T)^{-1} = \sum_{k\geqslant 0}(R^{-1}T)^k R^{-1}$ (série normalement convergente),*
3) *$\|(R-T)^{-1}\| \leqslant \|R^{-1}\|/(1-\|R^{-1}T\|)$.*

Ce résultat montre en particulier que *l'ensemble des applications linéaires continues inversibles de $\mathbb{E}$ dans $\mathbb{F}$ est un ouvert de $\mathcal{L}(\mathbb{E},\mathbb{F})$.*

Opérateur adjoint

On appelle *forme* une application à valeurs dans $\mathbb{R}$. Les formes linéaires continues sont donc les éléments de $\mathbb{E}' := \mathcal{L}(\mathbb{E},\mathbb{R})$, appelé *espace dual* (*topologique*) de $\mathbb{E}$. Si $f\in\mathbb{E}'$, on note

$$\langle f, x\rangle_{\mathbb{E}',\mathbb{E}} := f(x).$$

Le *crochet de dualité* $\langle\cdot,\cdot\rangle_{\mathbb{E}',\mathbb{E}}$ entre $\mathbb{E}'$ et $\mathbb{E}$ se notera aussi $\langle\cdot,\cdot\rangle$ s'il n'y a pas d'ambiguïté. On ne le confondra pas avec un produit scalaire (voir ci-dessous), qui est lui un opérateur bilinéaire sur un même espace.

Soient $\mathbb{E}$ et $\mathbb{F}$ deux espaces de Banach. L'*adjoint* d'un opérateur $T\in\mathcal{L}(\mathbb{E},\mathbb{F})$ est l'opérateur $T^*\in\mathcal{L}(\mathbb{F}',\mathbb{E}')$, défini de la manière suivante. Soit $g\in\mathbb{F}'$. L'application $x\in\mathbb{E}\mapsto\langle g, Tx\rangle_{\mathbb{F}',\mathbb{F}}\in\mathbb{R}$ est linéaire continue ; c'est donc un élément de $\mathbb{E}'$, que l'on note T^*g. On a donc

$$\forall\, x\in\mathbb{E},\quad \forall\, g\in\mathbb{F}' : \qquad \langle g, Tx\rangle_{\mathbb{F}',\mathbb{F}} = \langle T^*g, x\rangle_{\mathbb{E}',\mathbb{E}}.$$

On remarque finalement que l'application $T^* : g\in\mathbb{F}'\mapsto T^*g\in\mathbb{E}'$ est linéaire continue. C'est l'application linéaire adjointe de T.

Voici quelques propriétés des opérateurs adjoints.

[1] $\mathcal{N}$ pour « null space » et $\mathcal{R}$ pour « range space ».

- On a l'égalité des normes :
$$\|T^*\| = \|T\|.$$
- *Théorème de l'image fermée.* Si $\mathcal{R}(T)$ est fermée (c'est toujours le cas en dimension finie), le noyau de T et l'image de T^* sont reliés par
$$\mathcal{R}(T^*) = \mathcal{N}(T)^\perp. \tag{A.6}$$
Dans ce cas,
$$T \text{ est injective} \iff T^* \text{ est surjective.}$$
- Si T est inversible, alors T^* est aussi inversible et
$$(T^*)^{-1} = (T^{-1})^*.$$
On note alors souvent $(T^*)^{-1} = (T^{-1})^*$ par T^{-*} .

Application bilinéaire

Soient $\mathbb{E}_1$, $\mathbb{E}_2$ et $\mathbb{F}$ trois espaces vectoriels sur $\mathbb{R}$. On dit qu'une application $b : \mathbb{E}_1 \times \mathbb{E}_2 \to \mathbb{F}$ est *bilinéaire* si elle est linéaire par rapport à chacun de ses arguments, l'autre étant fixé à une valeur arbitraire :

$$\begin{aligned} &\forall x_2 \in \mathbb{E}_2 : \quad x_1 \in \mathbb{E}_1 \mapsto b(x_1, x_2) \in \mathbb{F} \text{ est linéaire,} \\ &\forall x_1 \in \mathbb{E}_1 : \quad x_2 \in \mathbb{E}_2 \mapsto b(x_1, x_2) \in \mathbb{F} \text{ est linéaire.} \end{aligned}$$

Si $\mathbb{E}_1$, $\mathbb{E}_2$ et $\mathbb{F}$ sont maintenant des espaces normés sur $\mathbb{R}$, on sait qu'une application bilinéaire $b : \mathbb{E}_1 \times \mathbb{E}_2 \to \mathbb{F}$ est *continue*, si, et seulement si, elle est continue en zéro, ou encore si, et seulement si, la quantité

$$\|b\| := \sup_{\substack{\|x_1\| \leqslant 1 \\ \|x_2\| \leqslant 1}} \|b(x_1, x_2)\| \equiv \sup_{\substack{x_1 \neq 0 \\ x_2 \neq 0}} \frac{\|b(x_1, x_2)\|}{\|x_1\| \, \|x_2\|}$$

est finie. En réalité, $b \mapsto \|b\|$ est une *norme* sur l'ensemble des applications binliéaires continues de $\mathbb{E}_1 \times \mathbb{E}_2 \to \mathbb{F}$.

Soient $\mathbb{E}_1$ et $\mathbb{E}_2$ deux ensembles. On dit qu'une application $a : \mathbb{E}_1^2 \to \mathbb{E}_2$ est *symétrique* si
$$\forall (x_1, x_2) \in \mathbb{E}_1^2 : \quad a(x_1, x_2) = a(x_2, x_1).$$

A.6 Espace de Hilbert

Espace pré-hilbertien

Soit $\mathbb{E}$ un espace vectoriel. Un *produit scalaire* [275 ; 1847] sur $\mathbb{E}$ est une application

$$(x, y) \in \mathbb{E} \times \mathbb{E} \mapsto \langle x, y \rangle \in \mathbb{R}$$

bilinéaire ($x \mapsto \langle x, y \rangle$ et $y \mapsto \langle x, y \rangle$ sont linéaires) symétrique ($\langle x, y \rangle = \langle y, x \rangle$) et définie positive ($\langle x, x \rangle > 0$, si $x \neq 0$). L'application $x \in \mathbb{E} \mapsto \|x\| = \langle x, x \rangle^{1/2} \in \mathbb{R}$

est une norme, dite *norme associée au produit scalaire* $\langle\cdot,\cdot\rangle$. On appelle *espace pré-hilbertien* un espace vectoriel muni d'un produit scalaire. S'il est de dimension finie, on dit qu'il s'agit d'un *espace euclidien*.

Par exemple, sur $\mathbb{R}^n$, l'application qui à $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ et $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$ fait correspondre

$$x^{\mathsf{T}} y := \sum_{i=1}^{n} x_i y_i$$

est un produit scalaire, appelé *produit scalaire euclidien*. La norme associée est la norme euclidienne ou ℓ_2.

Dans un espace pré-hilbertien, le produit scalaire et la norme associée sont liés par l'*inégalité de Cauchy-Schwarz*:

$$\forall\, x, y \in \mathbb{E} : \qquad |\langle x, y\rangle| \leqslant \|x\|\, \|y\|. \tag{A.7}$$

On en déduit que (on peut aussi enlever la valeur absolue)

$$\|x\| = \sup_{\|y\|\leqslant 1} |\langle x, y\rangle|.$$

Plus généralement, si $\|\cdot\|_{\text{P}}$ est une norme sur $\mathbb{E}$, pouvant être différente de la norme associée au produit scalaire $\langle\cdot,\cdot\rangle$, on définit sa *norme duale* $\|\cdot\|_{\text{D}}$ par (voir l'exercice A.11)

$$\|y\|_{\text{D}} := \sup_{\|x\|_{\text{P}}\leqslant 1} |\langle y, x\rangle|. \tag{A.8}$$

Sa valeur ne change pas si l'on enlève les valeurs absolues. En prenant la norme duale de la norme duale, on retrouve la norme primale (voir le point 2 de l'exercice 3.30). On a l'*inégalité de Cauchy-Schwarz généralisée*:

$$\forall\, x, y \in \mathbb{E} : \qquad |\langle x, y\rangle| \leqslant \|x\|_{\text{P}}\, \|y\|_{\text{D}}. \tag{A.9}$$

Ainsi sur $\mathbb{R}^n$, muni du produit scalaire euclidien, si on se donne des *nombres conjugués* p et $p' \in [1, \infty]$, c'est-à-dire vérifiant $\frac{1}{p} + \frac{1}{p'} = 1$ ($p = 1$ si, et seulement si, $p' = +\infty$), les normes ℓ_p et $\ell_{p'}$ sont duales l'une de l'autre et l'inégalité

$$\forall\, x, y \in \mathbb{R}^n : \qquad |x^{\mathsf{T}} y| \leqslant \|x\|_p\, \|y\|_{p'},$$

porte le nom d'*inégalité de Hölder* (pour une démonstration utilisant les conditions d'optimalité, voir l'exercice 4.11).

Si $(\mathbb{E}_1, \langle\cdot,\cdot\rangle_1)$ et $(\mathbb{E}_2, \langle\cdot,\cdot\rangle_2)$ sont deux espaces pré-hilbertiens, l'espace produit $\mathbb{E} := \mathbb{E}_1 \times \mathbb{E}_2 := \{(x_1, x_2) : x_1 \in \mathbb{E}_1,\ x_2 \in \mathbb{E}_2\}$ est muni canoniquement du *produit scalaire produit* suivant

$$\langle\cdot,\cdot\rangle : ((x_1, x_2), (y_1, y_2)) \in \mathbb{E} \times \mathbb{E} \mapsto \langle x_1, y_1\rangle_1 + \langle x_2, y_2\rangle_2. \tag{A.10}$$

Espace hilbertien

On dit que $\mathbb{E}$ est un *espace de Hilbert* ou un *espace hilbertien* si c'est un espace pré-hilbertien tel que, pour la norme associée, $\mathbb{E}$ est complet (c'est donc un espace de Banach).

Théorème A.3 (de représentation de Riesz-Fréchet) *Si $\mathbb{E}$ est un espace de Hilbert et si $f \in \mathbb{E}'$ alors il existe un unique $\xi \in \mathbb{E}$ tel que*

$$\forall\, x \in \mathbb{E} : \qquad f(x) = \langle \xi, x \rangle.$$

DÉMONSTRATION. Donnons une démonstration en dimension finie. Pour une démonstration en dimension infinie, on pourra par exemple consulter [91 ; Théorème V.5] ou suivre l'indication donnée au cours de la démonstration.

Soit $\xi \in \mathbb{E}$. Alors l'application

$$J_\xi : \mathbb{E} \to \mathbb{R} : x \mapsto \langle \xi, x \rangle$$

est linéaire continue (sa norme vaut $\|\xi\|$). On a donc défini une application

$$J : \xi \in \mathbb{E} \mapsto J_\xi \in \mathbb{E}'.$$

Cette application J est également linéaire continue (de norme 1). Elle est aussi injective puisque si $J_\xi = 0$, c'est-à-dire $\langle \xi, x \rangle = 0$ pour tout $x \in \mathbb{E}$, on a nécessairement $\xi = 0$ (prendre $x = \xi$). En dimension finie, $\dim \mathbb{E}' = \dim \mathbb{E}$ et donc J est aussi surjective (en dimension infinie, la surjectivité de J s'obtient en montrant que l'image de J est fermée et dense dans $\mathbb{E}'$). Ceci implique que pour tout $f \in \mathbb{E}'$ il existe un unique $\xi \in \mathbb{E}$ tel que $f = J_\xi$. C'est ce que l'on cherchait à démontrer. □

Comme son nom l'indique, le théorème précédent sert à représenter des formes linéaires continues sur $\mathbb{E}$ par des vecteurs de $\mathbb{E}$. Ce résultat nous sera très utile pour définir le gradient d'une fonction différentiable et la hessienne d'une fonction deux fois différentiable.

Par le théorème de représentation de Riesz-Fréchet, le crochet de dualité entre $f \in \mathbb{E}'$ et $x \in \mathbb{E}$ n'est autre que le produit scalaire entre le représentant ξ de f et x. Alors, si $\mathbb{E}$ et $\mathbb{F}$ sont deux espaces de Hilbert et $T \in \mathcal{L}(\mathbb{E}, \mathbb{F})$, l'opérateur dual de T est identifiable à un opérateur $T^* \in \mathcal{L}(\mathbb{F}, \mathbb{E})$. Supposons à présent que $T \in \mathcal{L}(\mathbb{E}, \mathbb{E})$, où $\mathbb{E}$ est un espace de Hilbert. On dit que T est *auto-adjoint* si $T = T^*$. Cette relation a bien un sens puisque T et T^* appartiennent au même espace $\mathcal{L}(\mathbb{E}, \mathbb{E})$.

Supposons à présente que l'on prenne sur des espaces de Hilbert $\mathbb{E}$ et $\mathbb{F}$ des produits scalaires $\langle\!\langle \cdot, \cdot \rangle\!\rangle_\mathbb{E}$ et $\langle\!\langle \cdot, \cdot \rangle\!\rangle_\mathbb{F}$ différents de ceux $\langle \cdot, \cdot \rangle_\mathbb{E}$ et $\langle \cdot, \cdot \rangle_\mathbb{F}$ donnés dans leur définition. Alors, par le théorème de représentation de Riesz-Fréchet, il existe des opérateurs auto-adjoints définis positifs (donc inversibles) $S_\mathbb{E} : \mathbb{E} \to \mathbb{E}$ et $S_\mathbb{F} : \mathbb{F} \to \mathbb{F}$ tels que

$$\begin{aligned} \forall\, (x, x') \in \mathbb{E} \times \mathbb{E} : \qquad & \langle\!\langle x, x' \rangle\!\rangle_\mathbb{E} = \langle S_\mathbb{E} x, x' \rangle_\mathbb{E}, \\ \forall\, (y, y') \in \mathbb{F} \times \mathbb{F} : \qquad & \langle\!\langle y, y' \rangle\!\rangle_\mathbb{F} = \langle S_\mathbb{F} y, y' \rangle_\mathbb{F}. \end{aligned}$$

Si on note $T^\star \in \mathcal{L}(\mathbb{F}, \mathbb{E})$ l'opérateur adjoint de $T \in \mathcal{L}(\mathbb{E}, \mathbb{F})$ pour ces nouveaux produits scalaires, on a

$$T^\star = S_\mathbb{E}^{-1} T^* S_\mathbb{F}. \tag{A.11}$$

A.7 Multifonction

Soient $\mathbb{E}$ et $\mathbb{F}$ deux ensembles. On note $\mathcal{P}(\mathbb{F})$ l'ensemble des parties de $\mathbb{F}$. Une *multifonction* T de $\mathbb{E}$ dans $\mathbb{F}$, aussi appelée *fonction multivoque*, est une application de $\mathbb{E}$ dans $\mathcal{P}(\mathbb{F})$. On la note

$$T : \mathbb{E} \multimap \mathbb{F}$$

pour la distinguer d'une fonction $f : \mathbb{E} \to \mathbb{F}$, qui prend ses valeurs dans $\mathbb{E}$ et pas dans $\mathcal{P}(\mathbb{E})$.

On définit le *domaine*, l'*image* et le *graphe* d'une multifonction T par

$$\begin{aligned} \operatorname{dom} T &:= \{x \in \mathbb{E} : T(x) \neq \varnothing\}, \\ \mathcal{R}(T) &:= \cup\{T(x) : x \in \operatorname{dom} T\}, \\ \mathcal{G}(T) &:= \{(x, y) \in \mathbb{E} \times \mathbb{F} : y \in T(x)\}. \end{aligned}$$

On notera bien que l'on a choisi de définir le graphe comme une partie de $\mathbb{E} \times \mathbb{F}$ et pas de $\mathbb{E} \times \mathcal{P}(\mathbb{F})$. L'image d'une partie $P \subseteq \mathbb{E}$ par T est définie par

$$T(P) := \bigcup_{x \in P} T(x).$$

Une multifonction est entièrement déterminée par son graphe et il y a une bijection entre l'ensemble des multifonctions est l'ensemble des parties de $\mathbb{E} \times \mathbb{F}$. Ainsi si G est une partie de $\mathbb{E} \times \mathbb{F}$, la multifonction $T : \mathbb{E} \multimap \mathbb{F}$ telle que $\mathcal{G}(T) = G$ est définie en $x \in \mathbb{E}$ par

$$T(x) := \{y \in \mathbb{F} : (x, y) \in G\}.$$

Une multifonction est donc identique à ce que l'on appelle une *relation binaire* entre deux ensembles, laquelle est spécifiée par la donnée d'une partie de leur produit cartésien. Cependant, en analyse multifonctionnelle, on donne de la structure aux ensembles $\mathbb{E}$ et $\mathbb{F}$ et on insiste sur les propriétés que l'on peut donner ainsi à la multifonction.

La *multifonction réciproque* d'une multifonction $T : \mathbb{E} \multimap \mathbb{F}$ est la multifonction notée $T^{-1} : \mathbb{F} \multimap \mathbb{E}$ définie par

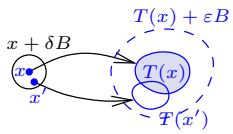
$$T^{-1}(y) = \{x \in \mathbb{E} : y \in T(x)\}.$$

Clairement, pour $x \in \mathbb{E}$ et $y \in \mathbb{F}$, on a

$$\begin{aligned} y \in T(x) &\iff x \in T^{-1}(y), \\ (x, y) \in \mathcal{G}(T) &\iff (y, x) \in \mathcal{G}(T^{-1}). \end{aligned}$$

On voit aussi que, pour une partie $P \subseteq \mathbb{E}$, $T(P) = \{y \in \mathbb{F} : T^{-1}(y) \cap P \neq \varnothing\}$.

Une multifonction $T : \mathbb{E} \multimap \mathbb{F}$ est *semi-continue supérieurement* en $x \in \mathbb{E}$ si pour tout $\varepsilon > 0$, il existe un $\delta > 0$ tel que $\forall\, x' \in x + \delta B$ on a $T(x') \subseteq T(x) + \varepsilon B$. Cette notion est illustrée à la figure ci-dessous.



Lorsque $\mathbb{E}$ est un espace euclidien de produit scalaire $\langle\cdot,\cdot\rangle$ et que $\mathbb{F}=\mathbb{E}$, on dit que T est *monotone* si

$$\forall\,(x_1,y_1)\in\mathcal{G}(T),\quad \forall\,(x_2,y_2)\in\mathcal{G}(T):\qquad \langle y_2-y_1,x_2-x_1\rangle\geqslant 0.$$

et que T est *fortement monotone* de module $\alpha>0$ si

$$\forall\,(x_1,y_1)\in\mathcal{G}(T),\quad \forall\,(x_2,y_2)\in\mathcal{G}(T):\qquad \langle y_2-y_1,x_2-x_1\rangle\geqslant \alpha\|x_1-x_2\|^2.$$

On dit que T est *monotone maximale* si T est monotone et si son graphe n'est pas strictement contenu dans le graphe d'un autre opérateur monotone. On vérifie facilement que cette dernière propriété s'écrit aussi

$$\Big[\langle y_2-y_1,x_2-x_1\rangle\geqslant 0,\quad \forall\,(x_1,y_1)\in\mathcal{G}(T)\Big]\quad\Longrightarrow\quad (x_2,y_2)\in\mathcal{G}(T).$$

Notes

Tous les sujets abordés dans ce chapitre sont largement développés dans de nombreux ouvrages ; citons Choquet [124 ; 1969], Schwartz [541 ; 1991] et Mawhin [413 ; 1992].

Exercices

A.1. Si $\{a_k\}_{k\in\mathbb{N}}$ et $\{b_k\}_{k\in\mathbb{N}}$ sont des suites de $\mathbb{R}$ avec $a_k\geqslant 0$, $a_k\to a\in\mathbb{R}$ et $\liminf_k b_k>-\infty$, alors

$$\liminf_{k\to\infty} a_k b_k\geqslant a\left(\liminf_{k\to\infty} b_k\right), \tag{A.12}$$

pourvu que $\liminf_k b_k<+\infty$ si $a=0$.

A.2. Démontrez la proposition A.1.

A.3. Soient f et $g:\mathbb{E}\to\mathbb{R}$ deux fonctions s.c.i. en un point x d'un espace topologique $\mathbb{E}$. Alors $f+g$ est s.c.i. en x.

A.4. Soient $\mathbb{E}$ et $\mathbb{F}$ deux espaces topologiques et $\varphi:\mathbb{E}\times\mathbb{F}\to\mathbb{R}$ une fonction. On considère la *fonction marginale* $f:\mathbb{E}\to\overline{\mathbb{R}}$ définie en $x\in\mathbb{E}$ par

$$f(x)=\inf_{y\in\mathbb{F}}\varphi(x,y).$$

- Si pour tout $y\in\mathbb{F}$, $\varphi(\cdot,y)$ est s.c.s., alors f est s.c.s..
- Si $\mathbb{E}$ est un espace métrique, $\mathbb{F}$ est compact et φ est continue, alors f est continue.

A.5. *Contractilité d'une distance.* Soient $\mathbb{E}$ un espace métrique et P une partie de $\mathbb{E}$. Montrez que d_P, la distance à P, est contractante.

A.6. *Somme d'ensembles et inclusion.* Soient A, B et C trois parties non vides d'un espace vectoriel sur $\mathbb{R}$. Montrez $A+C\subseteq B\Rightarrow A\subseteq B-C$; mais $A\subseteq B+C\not\Rightarrow A-C\subseteq B$.

A.7. *Autres définitions d'une sous-espace affine.* Soit A une partie non vide d'un espace vectoriel $\mathbb{E}$. Montrez que les propriétés suivantes sont équivalentes :

(i) A est un sous-espace affine,

(*ii*) pour tout $x, y \in A$ et tout $t \in \mathbb{R}$, on a $(1-t)x + ty \in A$,
(*iii*) $A - A$ est un sous-espace vectoriel de $\mathbb{E}$ et $A + (A - A) = A$.

A.8. *Intersection de sous-espaces affines.* On note $\mathbb{E}(A) := A - A$ le sous-espace vectoriel parallèle à un sous-espace affine A d'un espace vectoriel $\mathbb{E}$. Soit $\{A_i\}_{i\in I}$ une famille quelconque de sous-espaces affines, d'intersection non vide. Alors $\cap_{i\in I} A_i$ est un sous-espace affine et $\mathbb{E}(\cap_{i\in I} A_i) = \cap_{i\in I}\mathbb{E}(A_i)$.

A.9. *Norme associée à un produit scalaire.* Soit $\mathbb{E}$ un espace pré-hilbertien, de produit scalaire $\langle\cdot,\cdot\rangle$ et de norme associée $\|\cdot\|$. Montrez que les identités suivantes ont lieu, quels que soient x et $y \in \mathbb{E}$:

$$\begin{aligned} \|x+y\|^2 + \|x-y\|^2 &= 2\left(\|x\|^2 + \|y\|^2\right) \qquad (\textit{identité du parallélogramme}) \\ \|x+y\|^2 - \|x-y\|^2 &= 4\langle x, y\rangle \qquad (\textit{relation de polarisation}). \end{aligned}$$

A.10. *Généralisation de l'inégalité de Cauchy-Schwarz.* Montrez que l'inégalité de Cauchy-Schwarz (A.7) reste vraie même si $\langle\cdot,\cdot\rangle$ n'est que semi-définie positive (pour tout x, on a « $\langle x, x\rangle \geqslant 0$ » au lieu de « $\langle x, x\rangle > 0$ »).

A.11. *Norme duale.* Montrez que la norme duale définie par (A.8) est effectivement une norme.

B Algèbre linéaire ▲

L'algèbre linéaire est la branche des mathématiques qui étudie les espaces vectoriels et les applications linéaires entre ceux-ci. Nous supposerons toujours que les espaces vectoriels considérés sont réels (l'espace des scalaires est $\mathbb{R}$) et de dimension finie (notion définie ci-dessous).

B.1 Espaces vectoriels

On dit que des vecteurs $u_1, \ldots, u_m$ en nombre fini d'un espace vectoriel $\mathbb{E}$ sur $\mathbb{R}$ sont *linéairement indépendants* si quels que soient les scalaires $\alpha_1, \ldots, \alpha_m \in \mathbb{R}$, on a l'implication

$$\sum_{i=1}^{m} \alpha_i u_i = 0 \qquad \Longrightarrow \qquad \forall\, i \in [1:m]: \quad \alpha_i = 0.$$

On dit aussi que la famille $\mathcal{U} := (u_1, \ldots, u_n)$ est *libre*.

Supposons que l'espace vectoriel considéré $\mathbb{E}$ ait une *base* $\mathcal{U} := (u_1, \ldots, u_n)$, c'est-à-dire une famille finie, libre et génératrice d'éléments u_i de $\mathbb{E}$. On dit alors que $\mathbb{E}$ est de *dimension finie*. Toute base de $\mathbb{E}$ a alors un nombre identique d'éléments, appelé la *dimension de l'espace vectoriel*. Cette base permet d'identifier $\mathbb{E}$ à $\mathbb{R}^n$ au moyen de l'*application-coordonnée* $\gamma_{\mathcal{U}} : \mathbb{E} \to \mathbb{R}^n$ qui donne les coordonnées $x = (x_1, \ldots, x_n) = \gamma_{\mathcal{U}}(u)$ d'un vecteur $u = \sum_{i=1}^{n} x_i u_i$ dans la base $\mathcal{U}$.

Théorème B.1 (de la base incomplète) *Toute famille libre d'un espace vectoriel de dimension finie peut être complétée de manière à en former une base.*

Si $\mathbb{E}$ est un espace vectoriel de dimension finie, on peut toujours le munir d'une structure d'espace de Hilbert. Il suffit en effet de se donner une base de $\mathbb{E}$. Le produit scalaire euclidien sur $\mathbb{R}^n$ induit alors un produit scalaire sur $\mathbb{E}$: $\langle u, v\rangle = \gamma(u)^{\mathsf{T}}\gamma(v)$, où $\gamma : \mathbb{E} \to \mathbb{R}^n$ est l'application-coordonnée. Comme l'espace normé $\mathbb{R}^n$ est complet, il en est de même de $\mathbb{E}$ (γ est en effet une isométrie pour les normes associées aux deux produits scalaires : $\|\gamma(u)\|_2 = \|u\|$, pour tout $u \in \mathbb{E}$).

Soient $\mathbb{E}_1$ et $\mathbb{E}_2$ deux sous-espaces vectoriels d'un espace vectoriel $\mathbb{E}$. On définit leur *somme* par

$$\mathbb{E}_1 + \mathbb{E}_2 := \{x_1 + x_2 : x_1 \in \mathbb{E}_1, x_2 \in \mathbb{E}_2\}.$$

On dit que $\mathbb{E}_1 + \mathbb{E}_2$ est une *somme directe* si tout élément x de $\mathbb{E}_1 + \mathbb{E}_2$ se décompose de manière unique en $x_1 + x_2$ avec $x_1 \in \mathbb{E}_1$ et $x_2 \in \mathbb{E}_2$. On note $\mathbb{E}_1 \oplus \mathbb{E}_2$ une somme

directe. On dit que $\mathbb{E}_1$ et $\mathbb{E}_2$ sont *supplémentaires* dans $\mathbb{E}$ si $\mathbb{E} = \mathbb{E}_1 \oplus \mathbb{E}_2$, ce qui est le cas si, et seulement si, $\mathbb{E}_1 \cap \mathbb{E}_2 = \{0\}$. En dimension finie, tout sous-espace vectoriel $\mathbb{E}_1$ admet un sous-espace vectoriel supplémentaire $\mathbb{E}_2$ (donc tel que $\mathbb{E}_1 \oplus \mathbb{E}_2 = \mathbb{E}$) ; on dit alors que $\dim \mathbb{E}_2$ est la *codimension* de $\mathbb{E}_1$.

B.1.1 Orthogonalité

Soit $\langle \cdot, \cdot \rangle : \mathbb{E} \times \mathbb{E} \to \mathbb{R}$ un produit scalaire sur $\mathbb{E}$ et $\| \cdot \|$ la norme associée. On dit que deux vecteurs x et y de $\mathbb{E}$ sont *orthogonaux* si $\langle x, y \rangle = 0$. Si $\mathbb{E}_0$ est un sous-espace vectoriel de $\mathbb{R}^n$, on définit son sous-espace vectoriel *orthogonal* par

$$\mathbb{E}_0^\perp := \{x \in \mathbb{R}^n : \langle x, y \rangle = 0 \text{ pour tout } y \in \mathbb{E}_0\}.$$

On a $\dim \mathbb{E} = \dim \mathbb{E}_0 + \dim \mathbb{E}_0^\perp$ et donc $(\mathbb{E}_0^\perp)^\perp = \mathbb{E}_0$.

Algorithme d'orthogonalisation de Gram-Schmidt

Soient $x_1, \ldots, x_p$ des vecteurs linéairement indépendants dans $\mathbb{E}$. L'algorithme de Gram-Schmidt construit à partir de ces vecteurs des vecteurs $u_1, \ldots, u_p$ *orthonormaux*, tels que

$$\text{pour tout } j = 1, \ldots, p : \quad \text{vect}\{u_1, \ldots, u_j\} = \text{vect}\{x_1, \ldots, x_j\}, \tag{B.1}$$

où $\text{vect}\{u_1, \ldots, u_p\}$ désigne le sous-espace vectoriel engendré par les vecteurs $u_1, \ldots, u_p$, c'est-à-dire l'ensemble des vecteurs $u \in \mathbb{E}$ de la forme

$$u = \sum_{i=1}^{p} \alpha_i u_i$$

où $\alpha_1, \ldots, \alpha_p \in \mathbb{R}$.

Algorithme B.2 (Gram-Schmidt)

1. $u_1 = x_1 / \|x_1\|$;
2. Pour $i = 2, \ldots, p$ faire :
   2.1. $\alpha_{ij} = \langle x_i, u_j \rangle$, pour $j = 1, \ldots, i-1$;
   2.2. $\tilde{u}_i = x_i - \sum_{j=1}^{i-1} \alpha_{ij} u_j$;
   2.3. $u_i = \tilde{u}_i / \|\tilde{u}_i\|$.

L'algorithme B.2 est bien défini si les vecteurs $x_1, \ldots, x_p$ sont linéairement indépendants. En effet, à l'étape 1, $x_1 \neq 0$ et à l'étape 2.3, $\tilde{u}_i \neq 0$, sinon, d'après l'étape 2.2, on aurait

$$x_i = \sum_{j=1}^{i-1} \alpha_{ij} u_j.$$

Or, par récurrence, on voit que u_i est combinaison linéaire des $x_1, \ldots, x_i$. Alors (B.1) impliquerait que les vecteurs $x_1, \ldots, x_i$ sont linéairement dépendants, ce qui est contraire à l'hypothèse.

On vérifie également que les vecteurs $u_1, \ldots, u_p$ sont orthonormaux. Ils sont normalisés aux étapes 1 et 2.3 de l'algorithme. Quant à l'orthogonalité, elle se montre par récurrence : si $k < i$, on a

$$\begin{aligned}\langle u_i, u_k\rangle &= \frac{1}{\|\tilde{u}_i\|}\left(\langle x_i, u_k\rangle - \sum_{j=1}^{i-1} \alpha_{ij}\langle u_j, u_k\rangle\right)\\ &= \frac{1}{\|\tilde{u}_i\|}\left(\langle x_i, u_k\rangle - \alpha_{ik}\right) \qquad \text{[hypothèse de récurrence]}\\ &= 0.\end{aligned}$$

Enfin, la relation (B.1) est bien vérifiée puisqu'à chaque étape du calcul, u_i est combinaison linéaire de $x_1, \ldots, x_i$.

B.2 Applications linéaires

Supposons que $\mathbb{E}$ et $\mathbb{F}$ soient deux espaces vectoriels réels, de dimension finie n et m respectivement. Si L est une application linéaire de $\mathbb{E}$ dans $\mathbb{F}$, le *théorème du rang* affirme que

$$\boxed{\dim \mathbb{E} = \dim \mathcal{N}(L) + \dim \mathcal{R}(L).} \tag{B.2}$$

Soient $\mathcal{U} := (u_1, \ldots, u_n)$ une base de $\mathbb{E}$ et $\mathcal{V} := (v_1, \ldots, v_m)$ une base de $\mathbb{F}$. On peut associer à $L \in \mathcal{L}(\mathbb{E}, \mathbb{F})$ une *matrice* A, tableau de dimension $m \times n$, dont l'élément (i, j) est le nombre réel

$$A_{ij} := \text{composante } i \text{ de } L(u_j) \text{ dans la base } \mathcal{V}.$$

On dit que A est *de type* $m \times n$ et on note $\mathbb{R}^{m\times n}$ l'ensemble des matrices de type $m \times n$. On dit que A est *carrée* d'*ordre* n si $n = m$. On note $\mathcal{M}_{\mathcal{U},\mathcal{V}}(L) = A$, pour signifier que A est le représentant matriciel de L dans les bases $\mathcal{U}$ et $\mathcal{V}$. La *matrice identité* est la matrice carrée notée I vérifiant $I_{ij} = 0$ si $i \neq j$ et $I_{ii} = 1$ pour tout i.

La *trace* d'une matrice carrée $A \in \mathbb{R}^{n\times n}$ est le scalaire noté et défini par

$$\operatorname{tr} A = \sum_{i=1}^{n} A_{ii}.$$

L'ensemble $\mathbb{R}^{m\times n}$ des matrices de type $m \times n$ forme un espace vectoriel réel que l'on peut munir d'une norme. On dit que la norme matricielle est *sous-multiplcative* si lorsque $A \in \mathbb{R}^{m\times n}$ et $B \in \mathbb{R}^{p\times n}$, on a

$$\|AB\| \leqslant \|A\|\,\|B\|. \tag{B.3}$$

On dit que la norme matricielle $\|\cdot\|$ est *subordonnée* si elle peut être définie à partir de normes vectorielles dans $\mathbb{R}^n$ et $\mathbb{R}^m$ (notées de manière identique) par la formule

$$\|A\| = \sup_{\|x\| \leqslant 1} \|Ax\|. \tag{B.4}$$

Alors, pour tout x

$$\|Ax\| \leqslant \|A\| \, \|x\|.$$

On en déduit qu'une norme subordonnée est sous-multiplicative et vérifie $\|I\| = 1$. À chaque jeu de normes vectorielles correspond une norme matricielle subordonnée. On peut ainsi définir les normes matricielles ℓ_p à partir des normes vectorielles ℓ_p. On montre que pour $A = (A_{ij}) \in \mathbb{R}^{m \times n}$, on a

$$\|A\|_1 = \max_{1 \leqslant j \leqslant n} \|A_{\cdot j}\|_1, \tag{B.5}$$

$$\|A\|_2 = \|\sigma(A)\|_\infty, \tag{B.6}$$

$$\|A\|_\infty = \max_{1 \leqslant i \leqslant m} \|A_{i \cdot}\|_1. \tag{B.7}$$

En (B.6), on a noté $\sigma(A)$ le vecteur des valeurs singulières non nulles de A. Le calcul des normes matricielles ℓ_1 et ℓ_∞ peut se faire exactement en un temps linéaire en n. Celui de la norme ℓ_2 peut se faire à ε près en un temps polynomial en n et $\log \varepsilon^{-1}$. Celui des autres normes matricielles ℓ_p pour $p \in [1, \infty] \setminus \{1, 2, \infty\}$, de matrice avec des éléments rationnels, est NP-ardu, même de manière approchée [310].

On peut aussi munir $\mathbb{R}^{m \times n}$ du produit scalaire suivant

$$\langle \cdot, \cdot \rangle : (A, B) \in \mathbb{R}^{m \times n} \times \mathbb{R}^{m \times n} \mapsto \langle A, B \rangle := \operatorname{tr}(AB^{\mathsf{T}}) = \sum_{i,j} A_{ij} B_{ij}, \tag{B.8}$$

où $\operatorname{tr}(AB^{\mathsf{T}})$ désigne la trace de la matrice carrée AB^{T}. Il s'agit donc du produit scalaire euclidien de A et B vus comme éléments de $\mathbb{R}^{mn}$. La norme associée à ce produit scalaire est la *norme de Frobenius*, que l'on note

$$\|A\|_F := \left(\sum_{i,j} A_{ij}^2 \right)^{\frac{1}{2}}. \tag{B.9}$$

Ce n'est pas une norme subordonnée si $n > 1$, puisque $\|I\|_F = \sqrt{n}$, mais c'est une norme sous-multiplcative. Pour une matrice orthogonale Q d'ordre n, on a $\|AQ\|_F = \|A\|_F$. On en déduit que si $\sigma(A)$ désigne le vecteurs des valeurs singulières non nulles de A, on a $\|A\|_F = \|\sigma(A)\|_2$ et donc aussi $\|A\|_2 \leqslant \|A\|_F$.

On note $\mathcal{N}(A) := \{x \in \mathbb{R}^n : Ax = 0\}$ le *noyau* d'une matrice A et $\mathcal{R}(A) := \{Ax \in \mathbb{R}^m : x \in \mathbb{R}^n\}$ son *image*. On dit qu'une matrice A est *injective* si $\mathcal{N}(A) = \{0\}$ (dans ce cas $n \leqslant m$), qu'elle est *surjective* si $\mathcal{R}(A) = \mathbb{R}^m$ (dans ce cas $n \geqslant m$) et qu'elle est *bijective* ou *inversible* si elle est injective et surjective (dans ce cas $n = m$). Dire qu'une matrice A est d'ordre n est inversible revient à dire qu'il existe une matrice d'ordre n, notée A^{-1} et appelée *matrice inverse* de A, telle que

$$AA^{-1} = A^{-1}A = I.$$

On appelle *rang* de A la dimension de son image $\mathcal{R}(A)$.

Ayant choisi des bases sur $\mathbb{E}$ et $\mathbb{F}$, les produits scalaires euclidiens sur $\mathbb{R}^n$ et $\mathbb{R}^m$ induisent des produits scalaires sur $\mathbb{E}$ et $\mathbb{F}$ (voir section B.1). Alors, si $L \in \mathcal{L}(\mathbb{E}, \mathbb{F})$,

on peut définir l'application linéaire adjointe $L^* \in \mathcal{L}(\mathbb{F}, \mathbb{E})$. La matrice associée à L^* est donnée par

$$\mathcal{M}_{\mathcal{V},\mathcal{U}}(L^*) = A^\mathsf{T},$$

où la composante (i, j) de la matrice A^T vaut

$$(A^\mathsf{T})_{ij} = A_{ji}.$$

La matrice A^T est appelée *matrice transposée* de A[1]. Les matrices A et A^T ont le même rang (le nombre de colonnes linéairement indépendantes est identique au nombre de lignes linéairement indépendantes). D'autre part, on a

$$\boxed{\mathcal{N}(A^\mathsf{T}) = \mathcal{R}(A)^\perp} \tag{B.10}$$

et donc aussi $\mathcal{N}(A)^\perp = \mathcal{R}(A^\mathsf{T})$ (l'orthogonal est pris ici pour le produit scalaire euclidien). On en déduit que A est injective si, et seulement si, A^T est surjective.

Une matrice carrée est dite *symétrique* si $A^\mathsf{T} = A$. Une telle matrice est dite *définie positive* si $x^\mathsf{T} A x > 0$ pour tout vecteur x non nul dans $\mathbb{R}^n$ et elle est dite *semi-définie positive* si $x^\mathsf{T} A x \geqslant 0$ pour tout vecteur $x \in \mathbb{R}^n$. On dit que A est *définie négative* (resp. *semi-définie négative*) si $-A$ est définie positive (resp. semi-définie positive). On note respectivement

$$\mathcal{S}^n, \quad \mathcal{S}^n_+ \quad \text{et} \quad \mathcal{S}^n_{++}$$

l'ensemble des matrices d'ordre n qui sont symétriques, symétriques semi-définies positives et symétriques définies positives. On notera aussi

$$\begin{aligned} A \succcurlyeq 0 \quad &\iff \quad A \in \mathcal{S}^n_+ \\ A \succ 0 \quad &\iff \quad A \in \mathcal{S}^n_{++}. \end{aligned}$$

Le lemme suivant nous sera utile, en particulier dans les méthodes de pénalisation. On trouvera à l'exercice B.15 une généralisation de ce résultat.

Lemme B.3 (Finsler, 1937) *Soient M et P deux matrices d'ordre n symétriques, avec M définie positive dans le noyau de P (c.-à-d., $u^\mathsf{T} M u > 0$ pour tout vecteur non nul $u \in \mathcal{N}(P)$) et P semi-définie positive. Alors il existe $\bar{r} \in \mathbb{R}$ tel que, pour tout $r \geqslant \bar{r}$, $M + rP$ est définie positive.*

Démonstration. La démonstration la plus simple se fait par l'absurde. Si le résultat est faux, on peut trouver une suite $\{r_k\}$ de réels tendant vers l'infini tels que $M + r_k P$ n'est pas définie positive. Il existe alors une suite de vecteurs non nuls $\{u_k\} \subseteq \mathbb{R}^n$, tels que

$$u_k^\mathsf{T} M u_k + r_k (u_k^\mathsf{T} P u_k) \leqslant 0. \tag{B.11}$$

On peut supposer que $\|u_k\| = 1$ (il suffit en effet de diviser l'inégalité par $\|u_k\|^2 \neq 0$). En extrayant une sous-suite au besoin, on peut supposer que $u_k \to u$. En divisant l'inégalité (B.11) par r_k et en prenant la limite lorsque $k \to \infty$ on trouve $u^\mathsf{T} P u = 0$,

[1] Certains auteurs, en particulier francophones mais pas seulement, note ${}^t A$ la transposée de A. Nous avons préféré la notation A^T qui est plus répandue internationalement.

si bien que $u \in \mathcal{N}(P)$ (car $P \succcurlyeq 0$). L'inégalité (B.11) montre aussi que $u_k^\mathsf{T} M u_k \leqslant 0$, qui à la limite en k fournit $u^\mathsf{T} M u \leqslant 0$. On a obtenu une contradiction, puisque par hypothèse on doit avoir $u^\mathsf{T} M u > 0$ ($u \in \mathcal{N}(P)$ et $u \neq 0$). □

Le lemme de Finsler est illustré à la figure B.1 au moyen des courbes de niveau

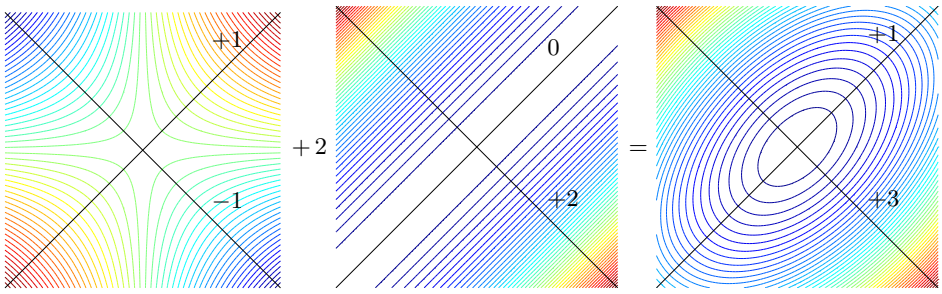


Fig. B.1. Illustration du lemme B.3 de Finsler, par les courbes de niveau des formes quadratiques associées aux matrices $M = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ (à gauche), $P = \left(\begin{smallmatrix} +1 & -1 \\ -1 & +1 \end{smallmatrix}\right)$ (au milieu) et $M + 2P = \left(\begin{smallmatrix} +2 & -1 \\ -1 & +2 \end{smallmatrix}\right)$ (à droite) ; M est définie positive sur $\mathcal{N}(P)$, P est semi-définie positive et $M + 2P$ définie positive.

de trois formes quadratiques : $x \mapsto x^\mathsf{T} M x$ qui est indéfinie mais définie positive sur $\mathcal{N}(P) = \mathbb{R}\binom{1}{1}$ (à gauche), $x \mapsto x^\mathsf{T} P x$ qui est semi-définie positive (au centre) et $x \mapsto x^\mathsf{T}(M + 2P)x$ qui est définie positive (à droite).

Soit A une matrice de type $m \times n$. Le *pseudo-inverse* de Moore-Penrose [434, 479] de A est la matrice $A^\dagger$ de type $n \times m$ telle que, quel que soit $b \in \mathbb{R}^m$, $A^\dagger b$ est la solution de norme minimale du problème de moindres-carrés linéaire

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|_2.$$

On montre qu'il s'agit de l'unique matrice $A^\dagger$ de type $n \times m$ qui vérifie

$$AA^\dagger A = A, \quad A^\dagger A A^\dagger = A^\dagger, \quad (AA^\dagger)^\mathsf{T} = AA^\dagger \quad \text{et} \quad (A^\dagger A)^\mathsf{T} = A^\dagger A. \tag{B.12}$$

Si $A = V \Sigma U^\mathsf{T}$ est la factorisation en valeurs singulières de A, on a

$$A^\dagger = U \Sigma^{-1} V^\mathsf{T},$$

si bien que $AA^\dagger = VV^\mathsf{T}$ et $A^\dagger A = UU^\mathsf{T}$, qui sont les projecteurs orthogonaux sur $\mathcal{R}(A) = \mathcal{R}(V)$ et $\mathcal{R}(A^\mathsf{T}) = \mathcal{R}(U)$ respectivement.

On appelle *déterminant* l'unique application $\det : \mathbb{R}^{n \times n} \to \mathbb{R} : A \mapsto \det A$ qui considérée comme fonction des colonnes de A est multilinéaire et alternée (elle change de signe si on permute deux colonnes) et qui vérifie $\det I = 1$. De cette définition, on déduit la formule du déterminant :

$$\det A = \sum_{\sigma \in \mathcal{S}_n} s_\sigma \, A_{\sigma_1 1} A_{\sigma_2 2} \cdots A_{\sigma_n n}.$$

On a noté $\mathcal{S}_n$ l'ensemble des *permutations* de $[1:n]$, c'est-à-dire des applications bijectives $\sigma : [1:n] \to [1:n]$; $\sigma(i) = \sigma_i$ la valeur en $i \in [1:n]$ de la permutation σ et s_σ la signature de σ. On sait que A est inversible si, et seulement si, son déterminant est non nul.

B.3 Analyse spectrale

Soit $A \in \mathcal{S}^n$. Alors toutes ses valeurs propres sont réelles. On les supposera ordonnées comme suit :

$$\lambda_1 \leqslant \cdots \leqslant \lambda_n.$$

On note $\Lambda = \mathrm{Diag}(\lambda_1, \ldots, \lambda_n)$ et $v_1, \ldots, v_n$ les vecteurs propres associés aux λ_i. On peut supposer qu'ils forment une base orthonormale de $\mathbb{R}^n$; alors $V = (v_1 \cdots v_n)$ est une matrice orthogonale et on a $V^\mathsf{T} A V = \Lambda$, ainsi que la *factorisation spectrale*

$$A = V \Lambda V^\mathsf{T} = \sum_{i=1}^{n} \lambda_i v_i v_i^\mathsf{T}.$$

La première égalité se déduit de $AV = V\Lambda$ et de l'orthogonalité de V ; la seconde se vérifie en prenant l'élément (k, l) de chaque membre.

Théorème B.4 (principe du min-max de Courant-Fisher) *Soit $A \in \mathcal{S}^n$. Alors pour tout $i = 1, \ldots, n$, pour tout sous-espace vectoriel $\mathbb{E}'_i$ de dimension $n-i+1$ et pour tout sous-espace vectoriel $\mathbb{E}_i$ de dimension i, on a*

$$\min_{\substack{x \in \mathbb{E}'_i \\ \|x\|_2 = 1}} x^\mathsf{T} A x \leqslant \lambda_i \leqslant \max_{\substack{x \in \mathbb{E}_i \\ \|x\|_2 = 1}} x^\mathsf{T} A x. \tag{B.13}$$

On a égalité à gauche en prenant $\mathbb{E}'_i = \mathrm{vect}\{v_i, \ldots, v_n\}$ et égalité à droite en prenant $\mathbb{E}_i = \mathrm{vect}\{v_1, \ldots, v_i\}$. Dès lors, pour tout $i = 1, \ldots, n$:

$$\max_{\substack{s.e.v.\ \mathbb{E}' \\ \dim \mathbb{E}' = n-i+1}} \min_{\substack{x \in \mathbb{E}' \\ \|x\|_2 = 1}} x^\mathsf{T} A x = \lambda_i = \min_{\substack{s.e.v.\ \mathbb{E} \\ \dim \mathbb{E} = i}} \max_{\substack{x \in \mathbb{E} \\ \|x\|_2 = 1}} x^\mathsf{T} A x. \tag{B.14}$$

Démonstration. On peut écrire $\mathbb{E}'_i = \mathcal{R}(U)$, où U est une matrice $n \times (n-i+1)$ orthonormale ($U^\mathsf{T} U = I_{n-i+1}$). La première inégalité de (B.13) sera établie si l'on montre que $\min\{x^\mathsf{T} U^\mathsf{T} A U x : \|x\|_2 = 1\} \leqslant \lambda_i$. Comme $\dim \mathcal{R}(U) + \dim \mathrm{vect}\{v_1, \ldots, v_i\} > n$, on peut trouver un vecteur unitaire x' tel que $Ux' \in \mathrm{vect}\{v_1, \ldots, v_i\}$ (exercice B.1), ce qui s'écrit $Ux' = \sum_{k=1}^{i} \alpha_k v_k$ avec des α_k tels que $\sum_{k=1}^{i} \alpha_k^2 = 1$. On voit que $\min\{x^\mathsf{T} U^\mathsf{T} A U x : \|x\|_2 = 1\} \leqslant (x')^\mathsf{T} U^\mathsf{T} A U x' = \sum_{k=1}^{i} \alpha_k^2 \lambda_k \leqslant \lambda_i$.

On obtient la première égalité de (B.14) en prenant $\mathbb{E}' = \mathrm{vect}\{v_i, \ldots, v_n\}$ et $x = v_i$.

La seonde inéaglité de (B.13) et la seconde égalité de (B.14) se démontrent de la même manière. □

Corollaire B.5 *Soient A et $B \in \mathcal{S}^n$. Alors, pour tout $i = 1, \ldots, n$:*

$$\lambda_i(A) + \lambda_1(B) \leqslant \lambda_i(A+B) \leqslant \lambda_i(A) + \lambda_n(B),$$

ce qui implique que $\lambda_i(\cdot)$ est une application 1-lipschitzienne :

$$|\lambda_i(B) - \lambda_i(A)| \leqslant \|B - A\|_2.$$

Ce corollaire montre en particulier qu'en ajoutant à A une matrice semi-définie positive, on accroît ses valeurs propres. De plus, si B est définie positive, l'accroissement est strict et vaut au moins $\lambda_1(B)$.

Corollaire B.6 (entrelacement des valeurs propres) *Soient $A \in \mathcal{S}^n$ et $b \in \mathbb{R}^n$. Alors pour tout $i = 1, \ldots, n-1$:*

$$\lambda_i(A) \leqslant \lambda_i(A + bb^\mathsf{T}) \leqslant \lambda_{i+1}(A) \leqslant \lambda_{i+1}(A + bb^\mathsf{T}),$$

$$\lambda_i(A - bb^\mathsf{T}) \leqslant \lambda_i(A) \leqslant \lambda_{i+1}(A - bb^\mathsf{T}) \leqslant \lambda_{i+1}(A).$$

Théorème B.7 (principe variationnel de Ky Fan) *Soit $A \in \mathcal{S}^n$. Alors pour tout $p = 1, \ldots, n$, on a*

$$\lambda_{n-p+1} + \cdots + \lambda_n = \max_{\substack{X \in \mathbb{R}^{n \times p} \\ X^\mathsf{T} X = I_p}} \langle AX, X \rangle.$$

Théorème B.8 (inégalité de trace de Ky Fan) *Pour tout A et $B \in \mathcal{S}^n$, on a*

$$\langle A, B \rangle \leqslant \lambda(A)^\mathsf{T} \lambda(B), \tag{B.15}$$

où $\langle \cdot, \cdot \rangle$ désigne le produit scalaire (B.8). On a égalité ci-dessus si, et seulement si, l'on peut obtenir les factorisations spectrales « ordonnées » $\lambda(A)$ et $\lambda(B)$ de A et B par la même matrice orthogonale, c'est-à-dire si, et seulement si,

$$\exists V \text{ orthogonale} : \quad A = V \ \mathrm{Diag}(\lambda(A))\ V^\mathsf{T} \quad \text{et} \quad B = V \ \mathrm{Diag}(\lambda(B))\ V^\mathsf{T}.$$

On sait que A et $B \in \mathcal{S}^n$ sont simultanément diagonalisables si, et seulement si, elles commutent. La condition énoncée dans le théorème B.8 ci-dessus pour avoir l'égalité dans (B.15) est plus forte, car elle requiert que les matrices diagonales

obtenues soient *ordonnées*. Ainsi $A = \mathrm{Diag}(1,2)$ et $B = \mathrm{Diag}(2,1)$ commutent alors que $\langle A, B\rangle = 4$ diffère de $\lambda(A)^{\mathsf{T}}\lambda(B) = 5$.

L'inégalité de Ky Fan est un raffinement de l'inégalité de Cauchy-Schwarz sur $\mathcal{S}^n$, dans le sens où cette dernière peut se déduire de la première. En effet, si $A = V\,\mathrm{Diag}(\lambda(A))\,V^{\mathsf{T}}$ avec V orthogonale, on a

$$\|\lambda(A)\|_2 = \|\,\mathrm{Diag}(\lambda(A))\|_F = \|V\,\mathrm{Diag}(\lambda(A))\,V^{\mathsf{T}}\|_F = \|A\|_F.$$

Dès lors, l'inégalité de Ky Fan et l'inégalité de Cauchy-Schwarz sur $\mathbb{R}^n$ donnent

$$\langle A, B\rangle \leqslant \lambda(A)^{\mathsf{T}}\lambda(B) \leqslant \|\lambda(A)\|_2\,\|\lambda(B)\|_2 = \|A\|_F\,\|B\|_F.$$

On en déduit l'inégalité de Cauchy-Schwarz sur $\mathcal{S}^n$ en tenant compte également de celle obtenue en remplaçant ci-dessus B par $-B$.

En appliquant l'inégalité de Ky Fan à des matrices diagonales, on trouve une *inégalité de Hardy, Littlewood et Pólya* [305], simple à démontrer directement, selon laquelle le produit scalaire euclidien de deux vecteurs x et y est majoré par celui des vecteurs $[x]$ et $[y]$ obtenus à partir des vecteurs précédents en ordonnant leurs composantes par ordre décroissant :

$$\forall\, x, y \in \mathbb{R}^n : \qquad x^{\mathsf{T}} y \leqslant [x]^{\mathsf{T}}[y].$$

Un *mineur principal* d'une matrice carré A est le déterminant d'une sous-matrice carrée de A dont la diagonale est une partie de celle de A : si A est d'ordre n, il s'agit d'un

$$m_I := \det(A_{ij})_{i,j\in I}, \quad \text{où } I \subseteq [1\,{:}\,n].$$

Les *mineurs principaux de tête* sont les $m_{[1\,:\,j]}$ pour $j = 1, \ldots, n$.

Proposition B.9 (critères de définie positivité) *Les propriétés suivantes sont équivalentes :*

(*i*) *A est définie positive,*
(*ii*) *toutes les valeurs propres de A sont strictement positives,*
(*iii*) *tous les mineurs principaux de tête de A sont strictement positifs.*

Proposition B.10 (critères de semi-définie positivité) *Les propriétés suivantes sont équivalentes :*

(*i*) *A est semi-définie positive,*
(*ii*) *toutes les valeurs propres de A sont positives,*
(*iii*) *tous les mineurs principaux de A sont positifs.*

Démonstration. [$(i) \Leftrightarrow (ii)$] Vient du fait qu'une matrice symétrique s'écrit $A = \sum_i \lambda_i v_i v_i^{\mathsf{T}}$ où les λ_i sont ses valeurs prorpres et les v_i sont les vecteurs propres orthogonaux associés.

[$(i) \Rightarrow (iii)$] Car alors, pour tout $\varepsilon > 0$, $A + \varepsilon I \succ 0$. Tous les mineurs principaux de $A + \varepsilon I$ sont donc > 0. On passe alors à la limite lorsque $\varepsilon \downarrow 0$, en utilisant la continuité du déterminant.

[$(iii) \Rightarrow (i)$] On observe d'abord que, dans ce cas, lorsqu'on ajoute $\varepsilon > 0$ à un élément de la diagonale de A on accroît (pas nécessairement strictement) ses mineurs principaux. Par récurrence, il en est de même si l'on ajoute $\varepsilon > 0$ à plusieurs éléments diagonaux. Montrons maintenant que si l'on ajoute $\varepsilon > 0$ à *tous* les éléments diagonaux de A, les mineurs principaux de tête augmentent strictement ; plus précisément, si on note m_I^ε les mineurs de $A + \varepsilon I$, on a $m^\varepsilon_{[1\,:\,j]} \geqslant \varepsilon^j$ pour tout j. Montrons cela par récurrence. L'inégalité est claire pour $j = 1$. Puis

$$m^\varepsilon_{[1\,:\,j]} = \underbrace{\det\left(A_{[1\,:\,j],[1\,:\,j]} + \varepsilon \sum_{i=1}^{j-1} (e^i)^{\mathsf{T}} e^i\right)}_{\geqslant 0} + \varepsilon\, m^\varepsilon_{[1\,:\,j-1]} \geqslant \varepsilon^j.$$

Par la proposition B.9, on en déduit que $A + \varepsilon I \succ 0$ et donc que $A \succcurlyeq 0$. □

Remarques B.11 1) Une matrice ne sera pas nécessairement semi-définie positive si seuls ses mineurs principaux *de tête* sont positifs ; contre-exemple : $\mathrm{Diag}(0, -1)$.

2) Le critère (iii) ne peut être inversé pour qualifier une matrice semi-définie *négative*. Ainsi

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

a tous ses mineurs principaux négatifs, mais n'est pas semi-définie négative (ni semi-définie positive d'ailleurs). La raison vient de ce que $\det(-A_{I,I}) = \det A_{I,I}$, si $|I|$ est pair. □

B.4 Matrices complexes

B.4.1 Nombre, vecteur et matrice complexes

On note $\mathrm{i} \in \mathbb{C}$ le *nombre imaginaire pur*, celui qui vérifie $\mathrm{i}^2 = -1$. Si $x = a + \mathrm{i}b \in \mathbb{C}$, avec a et $b \in \mathbb{R}$, on note $|x| = (a^2 + b^2)^{1/2}$ le *module* de x et $\bar{x} = a - ib \in \mathbb{C}$ son *nombre complexe conjugué*. Pour des nombres x et $y \in \mathbb{C}$, on a

$$|x + y| \leqslant |x| + |y| \qquad \text{et} \qquad |xy| = |x|\,|y|.$$

Pour une matrice $A \in \mathbb{C}^{m\times n}$, on note A^{H} sa *transposée-conjuguée*, qui est définie par

$$(A^{\mathsf{H}})_{ij} = \bar{A}_{ji}.$$

Si A et $B \in \mathbb{C}^{m\times n}$, on a

$$(AB)^{\mathsf{H}} = B^{\mathsf{H}} A^{\mathsf{H}}.$$

Une matrice carrée $A \in \mathbb{C}^{n\times n}$ est dite *unitaire* si $A^{\mathsf{H}} A = I$ ou, de manière équivalente, si $AA^{\mathsf{H}} = I$.

Dans sa notation matricielle, un vecteur $v \in \mathbb{C}^n$ est considéré comme une matrice $n \times 1$. On munit $\mathbb{C}^n$ de la norme

$$v \in \mathbb{C}^n \mapsto \|v\| = (v^{\mathsf{H}} v)^{1/2} = \left(\sum_{i=1}^{n} |v_i|^2 \right)^{1/2} \in \mathbb{R}.$$

On dit que $v \in \mathbb{C}^n$ est un *vecteur propre* de *valeur propre* $\lambda \in \mathbb{C}$ de $A \in \mathbb{C}^{n \times n}$ si

$$Av = \lambda v \qquad \text{et} \qquad v \neq 0. \tag{B.16}$$

Dans ce cas, quel que soit $\alpha \in \mathbb{C}$, non nul, αv est encore vecteur propre de valeur propre λ. On peut donc normaliser v en supposant que $\|v\| = 1$ (il suffit de multiplier v par $1/\|v\|$) ; on dit alors que le vecteur propre est *unitaire*. L'ensemble des valeurs propres de A est appelé le *spectre* de A et est noté $\lambda(A)$; c'est donc une partie de $\mathbb{C}$. Le *rayon spectral* est la quantité notée et définie par

$$\rho(A) := \max_{\lambda \in \lambda(A)} |\lambda|.$$

B.4.2 Matrice hermitienne

On dit qu'une matrice $A \in \mathbb{C}^{n \times n}$ est *hermitienne* si $A^{\mathsf{H}} = A$. On note

$$\mathcal{H}^n$$

l'espace vectoriel sur $\mathbb{R}$ formé des matrices complexes hermitiennes d'ordre n (ce n'est pas un espace vectoriel sur $\mathbb{C}$ car iA n'est pas hermitienne si A est hermitienne non nulle).

Si A, $B \in \mathcal{H}^n$ et $v \in \mathbb{C}^n$, on montre facilement que $v^{\mathsf{H}} A v \in \mathbb{R}$, que $\operatorname{Diag} A \in \mathbb{R}^n$, que $\operatorname{tr} A \in \mathbb{R}$ et que $\operatorname{tr} AB \in \mathbb{R}$. On peut donc munir $\mathcal{H}^n$ du produit scalaire

$$\langle \cdot, \cdot \rangle : (A, B) \in \mathcal{H}^n \times \mathcal{H}^n \mapsto \langle A, B \rangle = \operatorname{tr} AB \in \mathbb{R},$$

qui fait de $\mathcal{H}^n$ un espace vectoriel euclidien.

Une matrice hermitienne a des valeurs propres réelles : $\lambda(A) \subseteq \mathbb{R}$. Elle admet la décomposition spectrale

$$A = V \Lambda V^{\mathsf{H}} = \sum_{i=1}^{n} \lambda_i v_i v_i^{\mathsf{H}}, \tag{B.17}$$

où $V := (v_1 \ \cdots \ v_n) \in \mathbb{C}^{n \times n}$ est une matrice unitaire et Λ est la matrice diagonale réelle $\operatorname{Diag}(\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^{n \times n}$, formée des valeurs propres de A. On peut donc définir la *semi-définie positivité* d'une matrice hermitienne A par l'une des propriétés équivalentes suivantes :

$$\forall\, x \in \mathbb{C}^n : \quad x^{\mathsf{H}} A x \geqslant 0,$$

toutes les valeurs propres de A sont positives.

Cette propriété est notée $A \succcurlyeq 0$.

La valeur propre maximale $\lambda_{\max}(A)$ d'une matrice hermitienne $A \in \mathcal{H}^n$ vérifie la propriété

$$\lambda_{\max}(A) = \max_{\substack{x \in \mathbb{C}^n \\ \|x\| = 1}} x^{\mathsf{H}} A x, \tag{B.18}$$

le maximum étant atteint si, et seulement si, x est dans l'espace propre associé à la valeur propre maximale.

B.5 Factorisations

Une matrice A est dite *triangulaire inférieure* si $A_{ij} = 0$ lorsque $i < j$. Elle est dite *triangulaire supérieure* si A^T est triangulaire inférieure. Une matrice triangulaire (inférieure ou supérieure) A est dite *unitaire* si $A_{ii} = 1$ pour tout i. La résolution du système linéaire $Ax = b$, où A est une matrice carrée triangulaire inférieure inversible se fait sans difficulté en commençant par la détermination de x_1 (par le première équation), puis de x_2 par la seconde (car on connaît maintenant x_1) et ainsi de suite. Si A est triangulaire supérieure, on déterminera successivement x_n, x_{n-1}, *etc.*

Une matrice carrée Q est dit *orthogonale* si $QQ^\mathsf{T} = I$, c'est-à-dire si Q^T est l'inverse de Q. Il est donc aussi aisé de résoudre un système linéaire $Qx = b$ lorsque Q est orthogonale.

Les factorisations matricielles rappelées ci-dessous permettent d'écrire une matrice donnée sous la forme d'un produit de matrices triangulaire ou orthogonale.

B.5.1 Factorisation QR

La *factorisation* QR d'une patrice A de type $m \times n$ est la factorisation suivante

$$A = QR,$$

où Q est orthogonale d'ordre m et R est triangulaire supérieure de type $m \times n$. Elle est le plus souvent obtenue par des réflexions de Houselholder ou des rotations de Givens.

Réflexions de Householder

Rotations de Givens

Un peu plus coûteuses que les réflexions de Householder pour une matrice pleine, les rotations de Givens sont en général utilisées lorsque la matrice a une structure creuse, en particulier pour des matrices bandes ou des matrices de Hessenberg. L'approche a alors un coût proportionnel au nombre d'éléments non nuls à gauche de la diagonale dans le *profil ligne* de la matrice (tout élément ayant sur la même ligne et à sa gauche un élément non nul est pris dans le profil).

La *rotation de Givens* G^{kl}, $k < l$, est une matrice qui ne diffère de la matrice identité que par ses lignes et colonnes k et l et de manière plus précise, par la sous-matrice d'ordre 2 :

$$G^{kl}_{\{k,l\},\{k,l\}} = \begin{pmatrix} c & s \\ -s & c \end{pmatrix},$$

où $c := \cos\theta$ et $s := \sin\theta$. C'est une matrice antisymétrique et orthogonale. On se sert de G^{kl} pour annuler la composante k ou l de $G^{kl}v$. Par exemple, pour annuler la composante l, il faut que $sv_k = cv_l$ ou $\operatorname{tg}\theta = v_l/v_k$. Comme $\cos\theta = (1 + \operatorname{tg}^2\theta)^{-1/2}$ et $\sin\theta = (1 - \cos^2\theta)^{1/2}$, il suffit donc de prendre

$$c = \frac{v_k}{(v_k^2 + v_l^2)^{1/2}} \qquad \text{et} \qquad s = \frac{v_l}{(v_k^2 + v_l^2)^{1/2}}.$$

Il n'est donc pas nécessaire de calculer θ.

Si A est une matrice, il est clair que $G^{kl}A$ ne modifie que les lignes k et l de A et si ces lignes ont des éléments nuls sur la même colonne, ces éléments restent nuls après multiplication à gauche par la matrice de Givens. Pour obtenir la factorisation QR d'une matrice pleine A, on procède en général comme suit. On annule successivement les éléments sous la diagonale, colonne par colonne et de bas en haut, comme l'illustre l'exemple 4×3 ci-dessous, dans lequel les éléments modifiés par les opérateurs G^{kl} sont soulignés :

$$A = \begin{pmatrix} \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \end{pmatrix} \xrightarrow{G^{34}} \begin{pmatrix} \times & \times & \times \\ \times & \times & \times \\ \underline{\times} & \underline{\times} & \underline{\times} \\ \underline{0} & \underline{\times} & \underline{\times} \end{pmatrix} \xrightarrow{G^{23}} \begin{pmatrix} \times & \times & \times \\ \underline{\times} & \underline{\times} & \underline{\times} \\ \underline{0} & \underline{\times} & \underline{\times} \\ 0 & \times & \times \end{pmatrix} \xrightarrow{G^{12}} \begin{pmatrix} \underline{\times} & \underline{\times} & \underline{\times} \\ \underline{0} & \underline{\times} & \underline{\times} \\ 0 & \times & \times \\ 0 & \times & \times \end{pmatrix}$$
$$\xrightarrow{G^{34}} \begin{pmatrix} \times & \times & \times \\ 0 & \times & \times \\ 0 & \underline{\times} & \underline{\times} \\ 0 & \underline{0} & \underline{\times} \end{pmatrix} \xrightarrow{G^{23}} \begin{pmatrix} \times & \times & \times \\ 0 & \underline{\times} & \underline{\times} \\ 0 & \underline{0} & \underline{\times} \\ 0 & 0 & \times \end{pmatrix} \xrightarrow{G^{34}} \begin{pmatrix} \times & \times & \times \\ 0 & \times & \times \\ 0 & 0 & \underline{\times} \\ 0 & 0 & \underline{0} \end{pmatrix} = R.$$

Quant à la matrice orthogonale Q^{T}, elle est le produit de droite à gauche des matrices de Givens successivement utilisées.

B.5.2 Factorisation gaussienne (ou LU)

> *Newton, in notes that he would rather not have seen published, described a process for solving simultaneous equations that later authors applied specifically to linear equations. This method — which Euler did not recommend, which Legendre called "ordinary," and which Gauss called "common" — is now named after Gauss: "Gaussian" elimination.*
>
> J. F. Grcar [278 ; 2011].

On dit que A admet une *factorisation gaussienne ou LU* si l'on peut écrire

$$A = LU,$$

avec L triangulaire inférieure unitaire et U triangulaire supérieure. Un intérêt de cette factorisation est que, si A est inversible (donc L et U le sont, car $0 \neq \det A = (\det L)(\det U)$), alors on peut résoudre le système linéaire $Ax = b$ en résolvant successivement les systèmes triangulaires $Ly = b$ (en y) et $Ux = y$ (en x).

Proposition B.12 (factorisation gaussienne) *Une matrice A d'ordre n a une unique factorisation gaussienne si, et seulement si,*

$$\text{pour tout } k = 1, \dots, n-1, \quad A_k := A^{(1,\dots,k)}_{(1,\dots,k)} \text{ est inversible.} \tag{B.19}$$

Si (B.19) n'a pas lieu, la factorisation gaussienne peut exister, mais elle n'est pas unique.

DÉMONSTRATION. Supposons que (B.19) ait lieu. L'existence et l'unicité de la factorisation gaussienne peut se montrer en factorisant successivement les matrices principales de tête A_k, pour $k = 1, \dots, n-1$. Soit $1 \leqslant k \leqslant n-1$ et supposons que A_k ait une unique factorisation gaussienne : $A_k = L_k U_k$ (c'est clairement le cas pour $k = 1$). Alors

$$A_{k+1} = \begin{pmatrix} L_k U_k & b \\ c^{\mathsf{T}} & \alpha \end{pmatrix} = \begin{pmatrix} L_k & 0 \\ l^{\mathsf{T}} & 1 \end{pmatrix} \begin{pmatrix} U_k & u \\ 0 & \beta \end{pmatrix}, \tag{B.20}$$

où b, $c \in \mathbb{R}^k$ et $\alpha \in \mathbb{R}$. Par (B.19), $0 \neq \det A_k = (\det L_k)(\det U_k)$, si bien que L_k et U_k sont inversibles. Alors, l, u et β sont déterminés de manière unique par

$$U_k^{\mathsf{T}} l = c, \quad L_k u = b \quad \text{et} \quad \beta = \alpha - l^{\mathsf{T}} u.$$

Donc A_{k+1} a une unique factorisation gaussienne.

Réciproquement, supposons que $A = LU$ soit l'unique factorisation gaussienne de A. Du fait de la structure triangulaire de L et U, on a $A_k = L_k U_k$, avec

$$L_k := L^{(1,\dots,k)}_{(1,\dots,k)} \quad \text{et} \quad U_k := U^{(1,\dots,k)}_{(1,\dots,k)}, \tag{B.21}$$

et donc $\det A_k = (\det L_k)(\det U_k) = \det U_k$ (car L_k est unitaire). Dès lors, si (B.19) n'a pas lieu, $U_{kk} = 0$ pour un indice $k \in [1 : n-1]$. On a aussi

$$A = \sum_{i=1}^{n} L^i U_i = \sum_{i=1}^{k-1} L^i U_i + (L^k + L^{k+1}) U_k + L^{k+1}(U_{k+1} - U_k) + \sum_{i=k+2}^{n} L^i U_i.$$

On obtient donc une factorisation de $A = \tilde{L}\tilde{U}$, avec

$$\tilde{L}^i = \begin{cases} L^i & \text{si } i \neq k \\ L^k + L^{k+1} & \text{si } i = k \end{cases} \qquad \text{et} \qquad \tilde{U}_i = \begin{cases} U_i & \text{si } i \neq k+1 \\ U_{k+1} - U_k & \text{si } i = k+1. \end{cases}$$

Il s'agit bien d'une nouvelle factorisation gaussienne puisque la matrice $\tilde{L}$ est triangulaire inférieure unitaire et est différente de L (car $L_{k+1,k+1} = 1$) et la matrice $\tilde{U}$ est triangulaire supérieure ($\tilde{U}_{k+1,j} = 0$ si $j \leqslant k$ parce que $U_{kk} = 0$). Cette contradiction avec l'hypothèse de départ montre que (B.19) est impliquée par celle-ci.

Pour la dernière partie du résultat, on constate en effet que la factorisation gaussienne de la matrice nulle existe pour une matrice triangulaire unitaire L arbitraire : $0 = L0$. □

Sous les conditions (B.19), il résulte de la démonstration précédente que les éléments diagonaux de U s'écrivent $U_{11} = A_{11}$ et pour $i = 2, \dots, n$:

$$U_{kk} = \frac{\det A^{(1,\dots,k)}_{(1,\dots,k)}}{\det A^{(1,\dots,k-1)}_{(1,\dots,k-1)}}.$$

En effet, $A_k = L_k U_k$, où L_k et U_k sont données par (B.21), si bien que $\det A_k = \det U_k$ (L_k est unitaire). Alors le calcul (B.20) conduit à $\det A_{k+1} = (\det U_k) U_{kk} = (\det A_k) U_{kk}$. En fait, tous les éléments de L et U peuvent s'exprimer par des déterminants [228, 321].

La manière habituelle de construire les facteurs L et U de A est de procéder par transformations successives de $A = A^{(1)}, \ldots, A^{(n)} = U$. Les éléments des $k-1$ premières colonnes de $A^{(k)}$ situés sous la diagonale sont nuls. Pour obtenir $A^{(k+1)}$ à partir de $A^{(k)}$, on annule la partie sous la diagonale de la colonne k de $A^{(k)}$, sans toucher aux $k-1$ premières colonnes :

$$A^{(k+1)} = M_k A^{(k)}, \quad \text{avec} \quad M_k := I - v_k e_k^{\mathsf{T}}, \tag{B.22}$$

où e_k est le k-ième vecteur de base de $\mathbb{R}^n$ et le vecteur v_k a ses composantes données par : $(v_k)_i = 0$ si $i \leqslant k$ et $(v_k)_i = A^{(k)}_{ik}/A^{(k)}_{kk}$ si $i > k$. Dans cette opération, l'élément $A^{(k)}_{kk}$ est appelé le *pivot*. Après $n-1$ étapes, on obtient la matrice triangulaire supérieure

$$U = M_{n-1} M_{n-2} \cdots M_1 A \tag{B.23}$$

Le facteur

$$L = M_1^{-1} \cdots M_{n-2}^{-1} M_{n-1}^{-1} \tag{B.24}$$

s'obtient facilement car $M_k^{-1} = I + v_k e_k^{\mathsf{T}}$. Comme produit de matrices triangulaires inférieures unitaires, L a aussi cette propriété. D'ailleurs, on vérifie aisément que

$$L = I + v_1 e_1^{\mathsf{T}} + \cdots + v_{n-1} e_{n-1}^{\mathsf{T}}.$$

Ce procédé n'est autre que l'*élimination gaussienne* (voir l'épigraphe de cette section), qui consiste à résoudre le système linéaire $Ax = b$ en éliminant l'une après l'autre les variables $x_1, x_2, \ldots, x_{n-1}$, jusqu'à l'obtension d'une équation en x_n seul, que l'on résout ; on inverse ensuite l'ordre pour calculer successivement $x_{n-1}, x_{n-2}, \ldots, x_1$. Explicitons cela. La première équation permet d'exprimer x_1 comme fonction affine de $x_2, \ldots, x_n$ par

$$x_1 = \frac{1}{A_{11}} \left(b_1 - \sum_{j=2}^{n} A_{1j} x_j \right),$$

et d'éliminer x_1 des équations d'indice $i \geqslant 2$, ce qui conduit au système

$$M_1 A x = M_1 b,$$

avec M_1 définie comme en (B.22). Ensuite, on peut utiliser la seconde équation de ce dernier système pour exprimer x_2 comme fonction affine de $x_3, \ldots, x_n$ et en éliminant x_2 des équations d'indice $i \geqslant 3$ de ce dernier système, on obtient le nouveau système

$$M_2 M_1 A x = M_2 M_1 b,$$

avec M_2 définie comme en (B.22). Si l'on poursuit de la sorte, on obtient finalement l'équation

$$U x = L^{-1} b,$$

avec U et L donnés par (B.23) et (B.24). Ce système se résout en calculant d'abord x_n par la dernière équation, puis x_{n-1} par l'équation $n-1$, et ainsi de suite jusqu'à la première équation qui donne x_1.

L'*élimination gaussienne par blocs* sur la matrice carrée

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix},$$

avec A carrée inversible, conduit à la matrice

$$\begin{pmatrix} I & 0 \\ CA^{-1} & I \end{pmatrix} M = \begin{pmatrix} A & B \\ 0 & A^s_M \end{pmatrix},$$

où

$$A^s_M := D - CA^{-1}B$$

est appelé le *complément de Schur* de A dans M [140]. On obtient ainsi la factorisation par blocs de M suivante

$$M = \begin{pmatrix} I & 0 \\ CA^{-1} & I \end{pmatrix} \begin{pmatrix} A & B \\ 0 & A^s_M \end{pmatrix}.$$

B.5.3 Factorisation de Cholesky

On suppose ci-dessous que A est une matrice symétrique d'ordre n.

Le cas défini positif

Considérons d'abord le cas où $A \succ 0$. Comme tous les mineurs principaux en tête de A sont non nuls, A a une factorisation gaussienne unique (proposition B.12). L'unicité implique que la factorisation est nécessairement de la forme

$$A = R^\mathsf{T} R = LDL^\mathsf{T},$$

où L est triangulaire inférieure unitaire, $D \succ 0$ est diagonale et $R = D^{1/2}L^\mathsf{T}$ est triangulaire supérieure. Ces factorisations sont dites *de Cholesky* [123 ; 1910] et sont déterminées de façon unique. La forme LDL^T permet de ne pas devoir calculer n racines carrées. Cette factorisation permet de montrer l'équivalence suivante

$$A \succ 0 \quad \Longleftrightarrow \quad \forall\, I = [1:k] : \quad \det A_{I,I} \geqslant 0. \tag{B.25}$$

Un premier algorithme calculant L et D s'obtient à partir de l'expression de l'élément (i,j) de A : pour tout $i \geqslant j$,

$$A_{ij} = \sum_{k=1}^{j} L_{ik} D_{kk} L_{jk}.$$

On peut donc calculer les éléments de L et D de proche en proche, en commençant par les petits indices :

Algorithme B.13 (Factorisation de Cholesky I)

Pour $i = 1, \ldots, n$:
  Pour $j = 1, \ldots, i-1$:
    $L_{ij} = (A_{ij} - \sum_{k=1}^{j-1} L_{ik} D_{kk} L_{jk})/D_{jj}$;
  $L_{ii} = 1$;
  $D_{ii} = A_{ii} - \sum_{k=1}^{i-1} L_{ik}^2 D_{kk}$;

On peut aussi procéder de la manière suivante. On a

$$LDL^{\mathsf{T}} = \sum_{k=1}^{n} D_{kk} L_{:k} L_{:k}^{\mathsf{T}}, \tag{B.26}$$

où $L_{:k}$ est la k-ième colonne de L. À l'étape k de l'algorithme ci-dessous, on obtient la matrice $A^{(k)}$ en retranchant de $A^{(k-1)}$ la matrice $D_{kk} L_{:k} L_{:k}^{\mathsf{T}}$, de manière à annuler ses ligne et colonne k :

$$A^{(k)} = A^{(k-1)} - D_{kk} L_{:k} L_{:k}^{\mathsf{T}} = \begin{pmatrix} 0 & 0 \\ 0 & B^{(k)} \end{pmatrix},$$

avec $L_{:k}^{\mathsf{T}} = (0, \ldots, 0, 1, L_{k+1,k}, \ldots, L_{nk})$ et $B^{(k)}$ est d'ordre $n-k$. Ceci conduit à l'algorithme suivant, dans lequel on suppose que les matrices successives $A^{(k)}$ sont récrites dans A, dont seule la partie triangulaire inférieure st utilisée.

Algorithme B.14 (Factorisation de Cholesky II)

Pour $k = 1, \ldots, n-1$:
  $D_{kk} = A_{kk}$;
  Pour $i = k+1, \ldots, n$:
    $L_{ik} = A_{ik}/D_{kk}$;
    Pour $j = k+1, \ldots, i$:
      $A_{ij} := A_{ij} - D_{kk} L_{ik} L_{jk}$;

Ce second algorithme a l'avantage de permettre le pivotage, ce qui n'est pas le cas du premier puisque l'élément diagonal D_{ii} (sur lequel porte pivotage) n'apparaît qu'après le calcul de la i-ième colonne de L. À l'étape k, on pourra choisir comme *pivot*, un élément maximal de la diagonale de $B^{(k)}$, à savoir celui dont l'indice s_k est déterminé par

$$s_k = \min\left(\arg\max\left\{B_{ii}^{(k)} : i \in [k:n]\right\}\right).$$

On permute ensuite les lignes k et s_k de $L^{(k)}$ et les lignes et colonnes d'indices k et s_k de $B^{(k)}$.

Le cas semi-défini positif

Une matrice A semi-définie positive, ce que l'on note $A \succcurlyeq 0$, admet également une factorisation de Cholesky [313 ; section 10.3]. Pour le voir, il suffit de partir de la

factorisation QR de $A^{1/2} = QR$, dans laquelle Q est orthogonale et R est triangulaire supérieure, et d'écrire

$$A = (A^{1/2})^{\mathsf{T}}(A^{1/2}) = R^{\mathsf{T}}R. \tag{B.27}$$

Cependant, le facteur R n'est pas nécessairement unique dans ce cas, comme le montre l'exemple suivant

$$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ \cos\theta & \sin\theta \end{pmatrix} \begin{pmatrix} 0 & \cos\theta \\ 0 & \sin\theta \end{pmatrix}.$$

On peut aussi avoir la factorisation

$$A = LDL^{\mathsf{T}}, \tag{B.28}$$

avec L triangulaire inférieure unitaire et $D \succcurlyeq 0$ diagonale. Pour voir cela, on peut par exemple partir de la factorisation $A = R^{\mathsf{T}}R = \sum_k (R_k)^{\mathsf{T}}R_k$ et modifier les colonnes R_k de R de manière à avoir soit $R_{kk} \neq 0$, soit $R_k = 0$. Dans ce but, on examine successivement les colonnes d'indice $k = 1, \ldots, n-1$: si $R_{kk} \neq 0$, on garde R_k inchangé ; sinon on l'annule après l'avoir ajoutée à R_{k+1}. On garde donc la factorisation $A = R^{\mathsf{T}}R$. On obtient alors (B.28) en prenant $D_{kk} = R_{kk}^2$, tandis que la colonne k de L est le vecteur unité e_k si $R_{kk} = 0$ et $L_k = R_k/R_{kk}$ sinon.

La factorisation (B.28) peut s'obtenir par l'algorithme B.13, en annulant les éléments L_{ij} lorsque $D_{jj} = 0$, mais cet algorithme est instable lorsque A n'est pas définie positive. Mieux vaut utiliser l'algorithme B.14 avec pivotages.

La caractérisation (B.25) ne se généralise pas aux matrices semi-définies positives. Cependant, on peut montrer que A est semi-définie positive si, et seulement si, tous ses mineurs principaux sont positifs :

$$A \succcurlyeq 0 \quad \Longleftrightarrow \quad \forall\, I \subseteq [1:n] : \quad \det A_{I,I} \geqslant 0. \tag{B.29}$$

Mise à jour des facteurs pour une correction de rang un

Supposons qu'une matrice $\bar{A}$ soit obtenue par une correction de rang 1 additive d'une matrice A d'ordre n symétrique définie positive :

$$\bar{A} = A + \alpha\, vv^{\mathsf{T}},$$

où $\alpha > 0$ est un réel et $v \in \mathbb{R}^n$. Bien sûr, $\bar{A}$ est encore d'ordre n symétrique définie positive. Si on connaît les facteurs de Cholesky de $A = LDL^{\mathsf{T}}$, on peut obtenir ceux de $\bar{A} = \bar{L}\bar{D}\bar{L}^{\mathsf{T}}$ en $O(n^2)$ opérations plutôt qu'avec les $O(n^3)$ opérations que requiert une factorisation directe de $\bar{A}$. On peut s'y prendre comme suit.

Avec p solution de $Lp = v$ (ceci requiert $O(n^2)$ opérations), on a

$$\bar{A} = L(D + \alpha\, pp^{\mathsf{T}})L^{\mathsf{T}}.$$

On se ramène ainsi à la factorisation de Cholesky de $D + \alpha\, pp^{\mathsf{T}} = \tilde{L}\tilde{D}\tilde{L}^{\mathsf{T}}$ et au produit $\bar{L} = L\tilde{L}$. Bien sûr, il suffit de prendre $\bar{D} = \tilde{D}$. Nous allons montrer que cela ne demande pas plus de $O(n^2)$ opérations.

Considérons d'abord la factorisation de Cholesky de $D + \alpha\, pp^{\mathsf{T}} = \tilde{L}\tilde{D}\tilde{L}^{\mathsf{T}}$. En examinant les colonnes des deux membres de cette identité, on remarque par récurrence

sur $j = 1, \ldots, n$, que la partie de la colonne j de $\tilde{L}$ sous la diagonale est parallèle à la partie correspondante de p : en notant $\tilde{L}^j$ la j-ième colonne de $\tilde{L}$ et e^k le k-ième vecteur de base de $\mathbb{R}^n$, on a

$$\tilde{L}^j = e^j + \beta_j \left(p_j e^j + \cdots + p_n e^n \right). \tag{B.30}$$

On vérifie que le coefficient de proportionnalité β_j s'écrit

$$\beta_j = \frac{p_j s_{j-1}}{\tilde{D}_j}, \qquad \text{où } s_j := \alpha - \textstyle\sum_{k=1}^{j} \beta_k^2 \tilde{D}_k \text{ et } s_0 = \alpha.$$

En prenant l'élément (j, j) de l'identité $D + \alpha\, pp^{\mathsf{T}} = \tilde{L}\tilde{D}\tilde{L}^{\mathsf{T}}$, on trouve alors

$$D_j + s_{j-1} p_j^2 = \tilde{D}_j.$$

Ces relations permettent de montrer que $s_j > 0$:

$$s_j = s_{j-1} - \beta_j^2 \tilde{D}_j = s_{j-1} - \frac{p_j^2 s_{j-1}^2}{\tilde{D}_j} = s_{j-1} \left(\frac{D_j}{D_j + s_{j-1} p_j^2} \right).$$

On en déduit en particulier que $\tilde{D}_j \geqslant D_j$ pour tout j. On peut alors introduire $t_j := 1/s_j$, qui vérifie la relation de récurrence numériquement stable :

$$t_j = t_{j-1} + \frac{p_j^2}{D_j}. \tag{B.31}$$

On a aussi $\tilde{D}_j / D_j = 1 + s_{j-1} p_j^2 / D_j = t_j / t_{j-1}$, si bien que le calcul de $\tilde{D}_j$ et β_j peut se faire par

$$\tilde{D}_j = \frac{D_j t_j}{t_{j-1}} \quad \text{et} \quad \beta_j = \frac{p_j}{D_j t_j}. \tag{B.32}$$

On peut à présent écrire l'algorithme qui calcule les facteurs de Cholesky de $D + \alpha\, pp^{\mathsf{T}} = \tilde{L}\tilde{D}\tilde{L}^{\mathsf{T}}$.

1. $t_0 = 1/\alpha$;
2. Pour $j = 1, \ldots, n$:
   2.1. Calcul de t_j, $\tilde{D}_j$ et β_j par (B.31) et (B.32) ;
   2.2. $\tilde{L}_{jj} = 1$;
   2.3. Pour $k = j+1, \ldots, n$: $\tilde{L}_{kj} = p_k \beta_j$;

Cet algorithme ne demande pas plus de $O(n^2)$ opérations élémentaires.

Intéressons nous à présent au produit $\bar{L} = L\tilde{L}$. En utilisant (B.30) et $Lp = v$:

$$\bar{L}^j = L \left(e^j + \beta_j (p - p_1 e^1 - \cdots - p_j e^j) \right) = L^j + \beta_j \left(v - p_1 L^1 - p_2 L^2 - \cdots - p_j L^j \right).$$

Dans l'algorithme suivant, on construit en même temps p et $\bar{L}$. Au début de la j-ième étape, le vecteur p a sa valeur correcte dans les j premières positions et a la valeur $v - p_1 L^1 - p_2 L^2 - \cdots - p_{j-1} L^{j-1}$ dans le $n - j$ suivantes. Comme les formules (B.31) et (B.32) n'utilisent que p_j, on obtient comme algorithme final de calcul de $\bar{L}$ et $\bar{D}$:

Algorithme B.15 (Mise à jour des facteurs de Cholesky I)

1. $t_0 = 1/\alpha$;

2. $p = v$;
3. Pour $j = 1, \ldots, n$:
   3.1. Calcul de t_j, $\bar{D}_j \equiv \tilde{D}_j$ et β_j par (B.31) et (B.32) ;
   3.2. $\bar{L}_{jj} = 1$;
   3.3. Pour $k = j+1, \ldots, n$:
      3.3.1. $p_k = p_k - p_j L_{kj}$;
      3.3.2. $\bar{L}_{kj} = L_{kj} + \beta_j p_k$;

On pourra vérifier par un raisonnement analogue que, si l'on connaît la factorisation de Cholesky $A = LL^\mathsf{T}$, les facteurs de Cholesky $\bar{L}\bar{L}^\mathsf{T}$ de $A + \alpha vv^\mathsf{T}$ sont donnés par l'algorithme suivant.

Algorithme B.16 (Mise à jour des facteurs de Cholesky II)

1. $t_0 = 1/\alpha$;
2. $p = v$;
3. Pour $j = 1, \ldots, n$:
   3.1. $p_j = p_j / L_{jj}$;
   3.2. $\delta = (1 + p_j^2 / t_{j-1})^{1/2}$;
   3.3. $\beta_j = p_j / (t_{j-1}\delta)$;
   3.4. $t_j = t_{j-1} + p_j^2$;
   3.5. Pour $k = j, \ldots, n$:
      3.5.1. $p_k = p_k - p_j L_{kj}$;
      3.5.2. $\bar{L}_{kj} = \delta L_{kj} + \beta_j p_k$;

Une réalisation informatique de l'algorithme précédent ne doit pas utiliser deux matrices L et $\bar{L}$, mais peut stocker $\bar{L}$ dans L, car lorsque $\bar{L}_{kj}$ est calculé, on n'a plus besoin de L_{kj}.

Une autre possibilité, qui fonctionne aussi si A est singulière (mais semi-définie positive), est fondée sur l'observation suivante (on a intégré le facteur $\alpha > 0$ dans le vecteur v) [314 ; 2008] :

$$\bar{A} = A + vv^\mathsf{T} = LL^\mathsf{T} + vv^\mathsf{T} = \begin{pmatrix} v & L \end{pmatrix} \begin{pmatrix} v & L \end{pmatrix}^\mathsf{T}.$$

La matrice $\begin{pmatrix} v & L \end{pmatrix}^\mathsf{T}$ de type $(n+1) \times n$ est de rang $\leqslant n$, si bien qu'elle peut être transformée en une matrice triangulaire supérieure de rang $\leqslant n$. Comme les éléments non nuls sous la diagonale sont placés directement sous celle-ci, n rotations de Givens suffisent à la transformer en une matrice triangulaire supérieure de la forme $\begin{pmatrix} \bar{L} & 0_{n\times 1} \end{pmatrix}^\mathsf{T}$. Si l'on note G le produit de ces rotations, qui est une matrice orthogonale, on a

$$\bar{A} = \begin{pmatrix} v & L \end{pmatrix} G^\mathsf{T} G \begin{pmatrix} v & L \end{pmatrix}^\mathsf{T} = \begin{pmatrix} \bar{L} & 0_{n\times 1} \end{pmatrix} \begin{pmatrix} \bar{L} & 0_{n\times 1} \end{pmatrix}^\mathsf{T} = \bar{L}\bar{L}^\mathsf{T},$$

qui est la factorisatiion de Cholesky souhaitée.

Mise à jour des facteurs pour une augmentation d'ordre

Considérons à présent le cas où une matrice $\bar{A}$ d'ordre $n+1$ symétrique définie positive est obtenue à partir de A par l'ajout d'une ligne et d'une colonne en position $n+1$:

$$\bar{A} = \begin{pmatrix} A & v \\ v^\mathsf{T} & \alpha \end{pmatrix},$$

où $v \in \mathbb{R}^n$ et $\alpha > 0$ est un nombre réel. Si les facteurs de Cholesky de $A = LDL^\mathsf{T}$ sont connus, on peut déterminer les facteurs de Cholesky de $\bar{A} = \bar{L}\bar{D}\bar{L}^\mathsf{T}$ en résolvant un seul système linéaire avec la matrice L.

Grâce à l'unicité de la factorisation de Cholesky, on vérifie en effet qu'il faut prendre

$$\bar{L} = \begin{pmatrix} L & 0 \\ l^\mathsf{T} & 1 \end{pmatrix} \quad \text{et} \quad \bar{D} = \begin{pmatrix} D & 0 \\ 0 & \delta \end{pmatrix},$$

où $l \in \mathbb{R}^n$ et $\delta \in \mathbb{R}$ sont donnés par

$$LD\,l = v \quad \text{et} \quad \delta = \alpha - l^\mathsf{T} D\,l.$$

Pour déterminer l, on résout d'abord le système linéaire triangulaire $Lw = v$ qui donne $w \in \mathbb{R}^n$, puis $l = D^{-1}w$. Au total, cette mise à jour des facteurs de Cholesky requiert $O(n^2)$ opérations au lieu des $O((n+1)^3)$ que demanderait une factorisation directe de $\bar{A}$.

On note au passage que la matrice symétrique $\bar{A}$ est définie positive si, et seulement si, $\delta > 0$, c'est-à-dire $\alpha > v^\mathsf{T}\bar{A}^{-1}v$ (on savait déjà que cette condition était nécessaire, car avec $\overline{v} = \begin{pmatrix} v^\mathsf{T}A^{-\mathsf{T}} & -1 \end{pmatrix}^\mathsf{T}$ on doit avoir $\overline{v}^\mathsf{T}\bar{A}\overline{v} > 0$).

Mise à jour des facteurs pour une diminution d'ordre

On suppose maintenant que $\bar{A}$ est obtenue à partir de A en lui ôtant la ligne i et la colonne i. On cherche les facteurs de Cholesky de $\bar{A} = \bar{L}\bar{D}\bar{L}^\mathsf{T}$ à partir de ceux de $A = LDL^\mathsf{T}$. Ici aussi un seul système linéaire triangulaire, utilisant une sous-matrice principale de L, doit être résolu. Il ne faut donc pas plus de $O(n^2)$ opérations.

Les matrices A et $\bar{A}$ sont reliées par l'identité

$$A = \begin{pmatrix} \bar{A}_{11} & \vdots & \bar{A}_{21}^\mathsf{T} \\ \cdots & & \cdots \\ \bar{A}_{21} & \vdots & \bar{A}_{22} \end{pmatrix},$$

où les pointillés correspondent à la i-ième ligne et à la i-ième colonne. On a une partition similaire pour les facteurs L, D, $\bar{L}$ et $\bar{D}$:

$$L = \begin{pmatrix} L_{11} & 0 & 0 \\ \cdots & 1 & 0 \\ L_{21} & v & L_{22} \end{pmatrix}, \quad D = \begin{pmatrix} D_1 & 0 & 0 \\ 0 & \delta & 0 \\ 0 & 0 & D_2 \end{pmatrix},$$

$$\bar{L} = \begin{pmatrix} \bar{L}_{11} & 0 \\ \bar{L}_{21} & \bar{L}_{22} \end{pmatrix} \quad \text{et} \quad \bar{D} = \begin{pmatrix} \bar{D}_1 & 0 \\ 0 & \bar{D}_2 \end{pmatrix}.$$

On trouve facilement

$$\begin{aligned}\bar{A}_{11} &= L_{11}D_1L_{11}^\mathsf{T} = \bar{L}_{11}\bar{D}_1\bar{L}_{11}^\mathsf{T},\\ \bar{A}_{21} &= L_{21}D_1L_{11}^\mathsf{T} = \bar{L}_{21}\bar{D}_1\bar{L}_{11}^\mathsf{T},\\ \bar{A}_{22} &= L_{21}D_1L_{21}^\mathsf{T} + \delta vv^\mathsf{T} + L_{22}D_2L_{22}^\mathsf{T}\\ &= \bar{L}_{21}\bar{D}_1\bar{L}_{21}^\mathsf{T} + \bar{L}_{22}\bar{D}_2\bar{L}_{22}^\mathsf{T}.\end{aligned}$$

La factorisation de Cholesky étant unique et les facteurs inversibles, on en déduit successivement que

$$\bar{L}_{11} = L_{11}, \quad \bar{D}_1 = D_1, \quad \bar{L}_{21} = L_{21}, \quad \text{et} \quad \bar{L}_{22}\bar{D}_2\bar{L}_{22}^\mathsf{T} = L_{22}D_2L_{22}^\mathsf{T} + \delta vv^\mathsf{T}.$$

Donc $\bar{L}_{22}\bar{D}_2\bar{L}_{22}^\mathsf{T}$ est la factorisation de Cholesky d'une correction de rang 1 de $L_{22}D_2L_{22}^\mathsf{T}$. Elle peut donc être obtenue en résolvant un unique système linéaire avec la matrice L_{22}.

B.5.4 Factorisation en valeurs singulières (SVD)

Soient $\mathbb{E}$ et $\mathbb{F}$ deux espaces euclidiens sur $\mathbb{R}$ (produits scalaires notés $\langle\cdot,\cdot\rangle$ et normes associées notées $\|\cdot\|$), de dimension respective

$$n := \dim\mathbb{E} \qquad \text{et} \qquad m := \dim\mathbb{F}.$$

On considère une application linéaire $A : \mathbb{E} \to \mathbb{F}$ de rang $r \geqslant 1$ (le cas où A est nulle n'est guère intéressant). On note $A^* : \mathbb{F} \to \mathbb{E}$ son adjointe. On sera aussi utile d'introduire l'opérateur $\hat{A} : (x, y) \in \mathbb{E}\times\mathbb{F} \mapsto (A^*y, Ax)$, qui est noté matriciellement par

$$\hat{A} = \begin{pmatrix} 0 & A^* \\ A & 0 \end{pmatrix}.$$

Si $\mathbb{E}_0$ est un sous-espace vectoriel d'un certain espace euclidien $\mathbb{V}$, on note $\mathrm{P}_{\mathbb{E}_0} : \mathbb{V} \to \mathbb{E}_0$ le projecteur orthogonal sur $\mathbb{E}_0$. Enfin, on munit $\mathbb{R}^r$ du produit scalaire euclidien canonique.

Proposition B.17 (factorisation en valeurs singulières) *Dans le cadre défini ci-dessus, il existe des applications* injectives $U : \mathbb{R}^r \to \mathbb{E}$ *et* $V : \mathbb{R}^r \to \mathbb{F}$ *et un opérateur diagonal* $\Sigma = \mathrm{Diag}(\sigma_1, \ldots, \sigma_r) : \mathbb{R}^r \to \mathbb{R}^r : (z_1, \ldots, z_r) \mapsto (\sigma_1 z_1, \ldots, \sigma_r z_r)$ *avec des* $\sigma_i > 0$, *tels que*

$$U^*U = I_{\mathbb{R}^r}, \quad UU^* = \mathrm{P}_{\mathcal{R}(A^*)}, \quad V^*V = I_{\mathbb{R}^r}, \quad VV^* = \mathrm{P}_{\mathcal{R}(A)} \quad \textit{et} \quad A = V\Sigma U^*. \tag{B.33}$$

On a aussi

$$AA^*V = V\Sigma^2 \qquad \textit{et} \qquad A^*AU = U\Sigma^2, \tag{B.34}$$

qui montrent que les σ_i^2, $i \in [1\,{:}\,r]$, *sont aussi les valeurs propres non nulles de* AA^* *et de* A^*A.

Démonstration. L'opérateur $\hat{A}$ est autoadjoint. Il a donc un système de $n + m = \dim\mathbb{E} + \dim\mathbb{F}$ vecteurs propres orthogonaux. Si $r = \mathrm{rg}(A)$ est le rang de A, alors

$\mathrm{rg}(\hat{A}) = 2r$ car $\mathrm{rg}(A^*) = \mathrm{rg}(A)$ et donc $\hat{A}$ a $2r$ valeurs propres non nulles. Celles-ci sont opposées deux à deux, car si

$$\begin{pmatrix} 0 & A^* \\ A & 0 \end{pmatrix} \begin{pmatrix} u_i \\ v_i \end{pmatrix} = \sigma_i \begin{pmatrix} u_i \\ v_i \end{pmatrix}$$

alors

$$\begin{pmatrix} 0 & A^* \\ A & 0 \end{pmatrix} \begin{pmatrix} -u_i \\ v_i \end{pmatrix} = -\sigma_i \begin{pmatrix} -u_i \\ v_i \end{pmatrix}.$$

Soient $\{(u_i, v_i)\}_{i\in[1\,:\,r]}$ les vecteurs propres correspondant aux valeurs propres $\sigma_i > 0$. On introduit les opérateurs linéaires $U : \mathbb{R}^r \to \mathbb{E} : \alpha \mapsto \sum_{i=1}^r \alpha u_i$ et $V : \mathbb{R}^r \to \mathbb{E} : \alpha \mapsto \sum_{i=1}^r \alpha v_i$ qui sous forme matricielle s'écrivent

$$U = \begin{pmatrix} u_1 & \cdots & u_r \end{pmatrix} \qquad \text{et} \qquad V = \begin{pmatrix} v_1 & \cdots & v_r \end{pmatrix}.$$

On notant $\Sigma = \mathrm{Diag}(\sigma_1, \ldots, \sigma_r)$, on a alors

$$AU = V\Sigma \qquad \text{et} \qquad A^*V = U\Sigma. \tag{B.35}$$

Montrons que l'on peut supposer que les colonnes de U et de V sont orthonormales :

$$U^*U = I_{\mathbb{R}^r} \qquad \text{et} \qquad V^*V = I_{\mathbb{R}^r}. \tag{B.36}$$

En effet, d'une part, on peut normaliser la matrice des vecteurs propres orthogonaux $\begin{pmatrix} U^* & V^* \end{pmatrix}^*$ de $\hat{A}$ par

$$U^*U + V^*V = \begin{pmatrix} U \\ V \end{pmatrix}^* \begin{pmatrix} U \\ V \end{pmatrix} = 2I_{\mathbb{R}^r}. \tag{B.37}$$

Par ailleurs, (B.35) montre que $V^*V\Sigma = \Sigma U^*U$. En utilisant l'identité (B.37), on trouve que $U^*U\Sigma + \Sigma U^*U = 2I_{\mathbb{R}^r}$. En regardant l'élément (i, j) des deux membres, on voit que $\langle u_i, u_j\rangle = 0$ si $i \neq j$ et que $2\sigma_i\langle u_i, u_i\rangle = 2\sigma_i$. Ceci montre que $U^*U = I_{\mathbb{R}^r}$. Alors $V^*V = I_{\mathbb{R}^r}$ se déduit de (B.37).

On peut compléter les matrices $U \in \mathbb{R}^{n\times r}$ et $V \in \mathbb{R}^{m\times r}$, dont les colonnes sont orthogonales par (B.36), par des matrices $\tilde{U} \in \mathbb{R}^{n\times(n-r)}$ et $\tilde{V} \in \mathbb{R}^{m\times(m-r)}$, respectivement, de manière à ce que $\begin{pmatrix} U & \tilde{U} \end{pmatrix}$ soit orthogonale. Mais alors

$$\begin{pmatrix} \pm U \\ V \end{pmatrix}^* \begin{pmatrix} \tilde{U} \\ 0 \end{pmatrix} = \pm U^*\tilde{U} = 0,$$

si bien que les colonnes de $\begin{pmatrix} \tilde{U}^* & 0 \end{pmatrix}^*$ sont des vecteurs propres de $\hat{A}$ associé à la vecteur propre nulle, ce qui montre que

$$A \begin{pmatrix} U & \tilde{U} \end{pmatrix} = V \begin{pmatrix} \Sigma & 0 \end{pmatrix}.$$

Comme $\begin{pmatrix} U & \tilde{U} \end{pmatrix}$ est orthogonale, on trouve finalement

$$A = V \begin{pmatrix} \Sigma & 0 \end{pmatrix} \begin{pmatrix} U^* \\ \tilde{U}^* \end{pmatrix} = V\Sigma U^*.$$

Il reste à montrer dans (B.33) que $UU^* = \mathrm{P}_{\mathcal{R}(A^*)}$ et $VV^* = \mathrm{P}_{\mathcal{R}(A)}$. On sait que l'opérateur idempotent et auto-adjoint UU^* est le projecteur orthogonal sur $\mathcal{R}(U)$. On obtient alors la première affirmation en observant que $A^* = U\Sigma V^*$ et que l'inversibilité de Σ et la surjectivité de V^* impliquent que $\mathcal{R}(U) = \mathcal{R}(A^*)$. On s'y prend de la même manière pour montrer la seconde affirmation.

La première identité de (B.34) s'obtient en multipliant à droite les deux membres de $AA^* = (V\Sigma U^*)(U\Sigma V^*) = V\Sigma^2 V^*$ par V et en utilisant (B.36). La seconde identité de (B.34) se montre de la même manière, en utilisant $A^*A = U\Sigma^2 U^*$. □

Les $\sigma_i \equiv \sigma_i(A)$ sont appelés les *valeurs singulières* de A et la factorisation $A = V\Sigma U^*$ est appelée la *factorisation en valeurs singulières* de A. On note $\sigma(A)$ le vecteur des valeurs singulières non nulles de A, rangées en ordre décroissant :

$$\sigma_1(A) \geqslant \cdots \geqslant \sigma_r(A).$$

Si $\{e^i\}_{i=1}^r$ est la base orthonormale canonique de $\mathbb{R}^r$, les vecteurs $u_i = Ue^i$, $i = 1, \ldots, r$, sont orthonormaux dans $\mathbb{E}$, les vecteurs $v_i = Ve^i$, $i = 1, \ldots, r$, sont orthonormaux dans $\mathbb{F}$ et

$$A = \sum_{i=1}^{r} \sigma_i \, (v_i \otimes u_i),$$

où $v \otimes u$ est le produit tensoriel de $v \in \mathbb{F}$ et $u \in \mathbb{E}$, qui est l'opérateur linéaire $x \in \mathbb{E} \mapsto v\langle u, x\rangle \in \mathbb{F}$.

Les valeurs singulières sont aussi les r valeurs propres strictement positives de l'application auto-adjointe $\hat{A} : \mathbb{E} \times \mathbb{F} \to \mathbb{E} \times \mathbb{F} : (x, y) \mapsto (A^*y, Ax)$, qui sous forme matricelle s'écrit

$$\hat{A} = \begin{pmatrix} 0 & A^* \\ A & 0 \end{pmatrix}.$$

Cette opérateur a aussi r valeurs propres strictement négatives, les $-\sigma_i$, et $n + m - 2r$ valeurs propres nulles. Le vecteur propre correspondant à σ_i est (u_i, v_i), celui correspondant à $-\sigma_i$ est $(-u_i, v_i)$.

Les valeurs singulières sont aussi les racines carrées des r valeurs propres strictement positives des applications auto-adjointes semi-définies positives

$$A^*A \qquad \text{et} \qquad AA^*,$$

qui ont aussi respectivement $n - r$ et $m - r$ valeurs propres nulles.

On peut aussi obtenir la valeur singulière maximale en résolvant un problème d'optimisation (exercice 4.13).

Théorème B.18 (inégalité de trace de von Neumann) *Si A et $B \in \mathbb{R}^{m\times n}$, on a*

$$\langle A, B\rangle \leqslant \sigma(A)^{\mathsf{T}} \sigma(B). \tag{B.38}$$

où $\langle \cdot, \cdot \rangle$ désigne le produit scalaire (B.8).

Notes

Pour l'étude de la résolution des systèmes linéaires, on pourra consulter [320, 373, 267, 164, 594]. On trouvera des inégalités matricielles intéressantes dans le livre de Zhan [641]. L'imposant ouvrage de Higham [313] est consacré à la précision et à la stabilité des schémas numériques de l'algèbre linéaire.

L'inégalité de trace de Ky Fan (théorème B.8), publiée en 1949 [200], est étroitement reliée à un travail antérieur de von Neumann [613 ; 1937] (l'inégalité de trace du théorème B.18). La condition pour avoir l'égalité est due à Teobald [577 ; 1975]. L'inégalité de trace de von Neumann [613 ; 1937] est plus générale que dans le théorème B.18 puisqu'elle permet aux matrices d'être complexes et qu'elle prend en compte des transformations unitaires supplémentaires.

L'équivalence $(i) \Leftrightarrow (iii)$ de la proposition B.10 est reprise de [44 ; exercice 4.1].

André-Louis Cholesky (1875-1918) était un officier français, originaire de la région Poitou-Charentes, affecté à des opérations de géodésie, d'abord en France, puis en Crète (1907-08) et enfin en Algérie et Tunisie (1910-14). Il est décédé au combat, durant la première guerre mondiale. La factorisation $R^{\mathsf{T}}R$ qui porte son nom a été publiée à titre posthume en 1924 par Benoit [47], puis retrouvée par C. Brezinski [89 ; 2006] dans les archives que sa famille a déposées à l'École Polytechnique ; le manuscrit date du 2 décembre 1910 [123]. La description de la mise à jour des facteurs de Cholesky de la section B.5.3 est reprise de [251 ; page 42-43]. On pourra aussi consulter [46, 247, 215, 249, 267].

Exercices

B.1. *Dimension de sous-espaces vectoriels.* Soient $\mathbb{E}$ un espace vectoriel de dimension finie, $\mathbb{E}_0$ et $\mathbb{E}_1$ deux sous-espaces vectoriels de $\mathbb{E}$. Démontrez les affirmations suivantes :

1) $\dim(\mathbb{E}_0 + \mathbb{E}_1) + \dim(\mathbb{E}_0 \cap \mathbb{E}_1) = \dim \mathbb{E}_0 + \dim \mathbb{E}_1$ (*formule de Grassmann*),
2) $\mathbb{E}_0 \cap \mathbb{E}_1 \neq \varnothing$ si $\dim \mathbb{E}_0 + \dim \mathbb{E}_1 > \dim \mathbb{E}$.

B.2. *Dimensions d'images et de noyaux.* Soient $\mathbb{E}$, $\mathbb{F}$ et $\mathbb{G}$ trois espaces vectoriels de dimension finie et $A : \mathbb{E} \to \mathbb{F}$ et $B : \mathbb{E} \to \mathbb{G}$ deux applications linéaires. Montrez que

1) $\dim \mathcal{R}\binom{A}{B} \leqslant \dim \mathcal{R}(A) + \dim \mathcal{R}(B)$,
2) $\dim \mathcal{N}(A) \leqslant \dim \mathcal{N}\binom{A}{B} + \dim \mathcal{R}(B)$.

B.3. *Théorème de l'application ouverte.* Soit $A : \mathbb{E} \to \mathbb{F}$ une application linéaire continue entre deux espaces normés, avec $\mathbb{F}$ de dimension finie (le résultat reste vrai sans cette hypothèse mais est plus difficile à démontrer [91]). On note $\bar{B}_{\mathbb{E}}$ et $\bar{B}_{\mathbb{F}}$ les boules unités fermées de $\mathbb{E}$ et $\mathbb{F}$ respectivement. Alors A est surjective si, et seulement si, il existe $r > 0$ tel que $r\bar{B}_{\mathbb{F}} \subseteq A(\bar{B}_{\mathbb{E}})$.

B.4. *Semi-continuité inférieure du rang.* L'application rang : $\mathbb{R}^{m\times n} \to \mathbb{N}$ est s.c.i..

B.5. *Orthogonalité.* Soient $\mathbb{E}$ un espace euclidien de dimension finie, muni d'un produit scalaire, $\mathbb{E}_0$ et $\mathbb{E}_1$ des sous-espaces vectoriels. Montrez les relations suivantes.

1) $\mathbb{E} = \mathbb{E}_0 \oplus \mathbb{E}_0^{\perp}$ (somme directe).
2) $(\mathbb{E}_0^{\perp})^{\perp} = \mathbb{E}_0$.
3) $\mathbb{E}_0 \subseteq \mathbb{E}_1 \Longrightarrow \mathbb{E}_1^{\perp} \subseteq \mathbb{E}_0^{\perp}$.
4) $(\mathbb{E}_0 \cap \mathbb{E}_1)^{\perp} = \mathbb{E}_0^{\perp} + \mathbb{E}_1^{\perp}$.

B.6. *Rang-ligne et rang-colonne.* Soit A une matrice $m \times n$. Montrez que le nombre de lignes linéairement indépendantes de A (son *rang-ligne*) est égal à son nombre de colonnes linéairement indépendantes (son *rang-colonne*).

B.7. *Dérivée du déterminant.* On considère l'application déterminant $\det : \mathbb{R}^{n\times n} \to \mathbb{R} : A \mapsto \det A$, avec $\mathbb{R}^{n\times n}$ muni du produit scalaire $\langle A, B\rangle = \operatorname{tr} AB$. Montrez que son gradient s'écrit

$$\nabla \det(A) = \operatorname{cof}(A),$$

où $\operatorname{cof}(A)$ est la *matrice des cofacteurs* ou *comatrice* de $A \in \mathbb{R}^{n\times n}$.

B.8. *Théorème d'entrelacement de Cauchy.* Soient $A \in \mathcal{S}^n$ et U une matrice orthogonale de type $n \times r$, avec $1 \leqslant r \leqslant n$ (c.-à-d., $U^\mathsf{T}U = I_r$). On note $\lambda_i(M)$ les valeurs propres d'une matrice symétrique M, rangées par ordre croissant. Alors pour $1 \leqslant i \leqslant r$, on a

$$\lambda_i(A) \leqslant \lambda_i(U^\mathsf{T}AU) \leqslant \lambda_{n-r+i}(A).$$

B.9. Trouvez une matrice M d'ordre n symétrique *non* définie positive telle que sur deux sous-espaces vectoriels supplémentaires $\mathbb{E}_1$ et $\mathbb{E}_2$ de $\mathbb{R}^n$, on ait $x^\mathsf{T}Mx > 0$, pour tout $x \in \mathbb{E}_1$ et pour tout $x \in \mathbb{E}_2$.

Remarque. Une matrice peut donc être définie positive dans deux sous-espaces supplémentaires sans être définie positive.

B.10. 1) Si A est matrice symétrique définie positive, il existe $\alpha > 0$ tel que $x^\mathsf{T}Ax \geqslant \alpha\|x\|_2^2$, pour tout x. Donnez la meilleure constante α.

2) Si A est matrice quelconque, il existe $\alpha > 0$ tel que $\|A^\mathsf{T}Ax\|_2 \geqslant \alpha\|Ax\|_2$, pour tout x. Donnez la meilleure constante α. En déduire que $x \mapsto \|A^\mathsf{T}x\|_2$ est une *norme* sur $\mathcal{R}(A)$.

B.11. *Produit de matrices définies positives.* Si A et $B \in \mathcal{S}_{++}^n$, alors la matrice non nécessairement symétrique AB a n valeurs propres qui sont toutes réelles et strictement positives.

B.12. *Représentation $\mathcal{S}_{++}^n$ d'un produit scalaire.* Soit $\langle\cdot,\cdot\rangle$ un produit scalaire sur $\mathbb{R}^n$. Montrez qu'il existe une unique matrice Q d'ordre n telle que $\langle x, y\rangle = x^\mathsf{T}Qy$, pour tout $x, y \in \mathbb{R}^n$. De plus, Q est symétrique définie positive. [Indication : procédez comme à la section C.2.2.]

B.13. *Deux propriétés des matrices semi-définies positives.* Soient A et $B \in \mathcal{S}_+^n$. Alors

1) $\mathcal{R}(A + B) = \mathcal{R}(A) + \mathcal{R}(B)$ et $\mathcal{N}(A + B) = \mathcal{N}(A) \cap \mathcal{N}(B)$,
2) si $\mathcal{R}(A) \subseteq \mathcal{R}(B)$, alors $(1 - t)A + tB \in \mathcal{S}_+^n$ pour $t > 1$ suffisamment proche de 1.

B.14. *Conséquences des factorisations.*

1) Soient A et $B \in \mathcal{S}^n$ avec $A \succ 0$. Montrez qu'il existe une matrice inversible R et une matrice diagonale Λ telles que $A = RR^\mathsf{T}$ et $B = R\Lambda R^\mathsf{T}$.
2) Montrez que $A \in \mathbb{R}^{n\times n}$ et $B \in \mathbb{R}^{n\times n}$ commutent (c.-à-d., $AB = BA$) si, et seulement si, il existe $V \in \mathbb{R}^{n\times n}$ orthogonale telle que $V^\mathsf{T}AV$ et $V^\mathsf{T}BV$ soient diagonales.

B.15. *Propriété d'augmentabilité* [16]. Soit M une matrice symétrique d'ordre n, semi-définie positive dans le noyau d'une autre matrice A de type $m \times n$ (c.-à-d., $u^\mathsf{T}Mu \geqslant 0$ pour tout $u \in \mathcal{N}(A)$). Montrez que les propriétés suivantes sont équivalentes :

(i) $v \in \mathcal{N}(A)$ et $v^\mathsf{T}Mv = 0$ impliquent que $Mv = 0$,
(ii) il existe $r \in \mathbb{R}$ tel que $M + rA^\mathsf{T}A \succcurlyeq 0$ (*propriété d'augmentabilité*),
(iii) $\inf\{(v^\mathsf{T}Mv)/\|Av\|_2^2 : Av \neq 0\} = \inf\{v^\mathsf{T}Mv : \|Av\|_2 = 1\} < +\infty$.

Trouvez une matrice M, semi-définie positive dans le noyau d'une autre matrice A, pour laquelle les propriétés ci-dessus ne sont pas vérifiées.

B.16. *Formules de SMW (Sherman-Morrison-Woodbury).*

(*i*) Soient I_n la matrice identité d'ordre n et u et $v \in \mathbb{R}^n$. Alors

$$\det(I_n + uv^\mathsf{T}) = 1 + v^\mathsf{T}u,$$

si bien que $I_n + uv^\mathsf{T}$ est inversible si, et seulement si, $1 + v^\mathsf{T}u \neq 0$. Dans ce cas,

$$\left(I_n + uv^\mathsf{T}\right)^{-1} = I_n - \frac{uv^\mathsf{T}}{1+v^\mathsf{T}u}.$$

[Indication : Montrez que les seules valeurs propres de I_n+uv^T sont 1 et $1+v^\mathsf{T}u$, et que cette dernière est simple.]

(*ii*) Soient A un matrice d'ordre n inversible et u et $v \in \mathbb{R}^n$. Alors

$$\det(A + uv^\mathsf{T}) = (\det A)(1 + v^\mathsf{T}A^{-1}u),$$

si bien que $A + uv^\mathsf{T}$ est inversible si, et seulement si, $1 + v^\mathsf{T}A^{-1}u \neq 0$. Dans ce cas, on a la formule de Sherman et Morrison (1949):

$$\left(A + uv^\mathsf{T}\right)^{-1} = A^{-1} - \frac{A^{-1}uv^\mathsf{T}A^{-1}}{1+v^\mathsf{T}A^{-1}u}.$$

Remarque. En pratique, on peut donc résoudre $(A+uv^\mathsf{T})x = b$ en x en résolvant deux systèmes linéaires avec la matrice A: $Ay = b$ en y et $Az = u$ en z. Alors $x = y - (v^\mathsf{T}y)/(1+v^\mathsf{T}z)z$. Cette approche est très utile lorsqu'il est aisé de résoudre un système linéaire avec la matrice A.

(*iii*) Soient u_1, v_1, u_2 et v_2 des vecteurs de $\mathbb{R}^n$. Alors

$$\det(I_n + u_1v_1^\mathsf{T} + u_2v_2^\mathsf{T}) = (1 + v_1^\mathsf{T}u_1)(1 + v_2^\mathsf{T}u_2) - (v_1^\mathsf{T}u_2)(v_2^\mathsf{T}u_1).$$

(*iv*) Soient A un matrice d'ordre n inversible, U une matrice $n \times m$, V une matrice $m \times n$ et R une matrice d'ordre m inversible (n et m peuvent être des entiers non nuls quelconques). Alors

$$\det(I_n + UV^\mathsf{T}) = \det(I_m + V^\mathsf{T}U)$$

et plus généralement

$$\det(A + URV^\mathsf{T}) = \det(A)\det(R^{-1} + V^\mathsf{T}A^{-1}U)\det(R).$$

Dès lors, $A + URV^\mathsf{T}$ est inversible si, et seulement si, $X := R^{-1} + V^\mathsf{T}A^{-1}U$ est inversible et, dans ce cas, on a la *formule de Woodbury*:

$$\boxed{\left(A + URV^\mathsf{T}\right)^{-1} = A^{-1} - A^{-1}UX^{-1}V^\mathsf{T}A^{-1}.}$$

B.17. *Inverse de matrices triangulaires.* Soit T une matrice inversible triangulaire supérieure (resp. inférieure). Montrez que T^{-1} est triangulaire supérieure (resp. inférieure).

C Calcul différentiel ▲

First the derivative was used, then discovered, explored and developed, and only then, defined.

J.V. Grabiner [273 ; 1983].

Dans cette annexe, nous rappelons les notions essentielles du calcul différentiel. Celles-ci sont présentées dans le cadre des espaces normés de dimension infinie. Tous les résultats énoncés sont démontrés. Nous pensons en effet qu'une bonne connaissance du calcul différentiel est nécessaire pour se sentir à l'aise en optimisation. Le cadre de la dimension infinie a été choisi car il est utile dans bien des domaines des mathématiques appliquées et ne présente pas de difficultés supplémentaires par rapport à la dimension finie. Nous supposons connues les notions de différentiabilité des fonctions réelles d'une variable réelle (voir par exemple [542 ; 1992]).

Sont également présentées deux notions essentielles pour la conception des algorithmes d'optimisation : celles de gradient (section C.1) et de hessienne (section C.2.2).

Soient $\mathbb{E}$ et $\mathbb{F}$ deux espaces normés (pouvant être de dimension infinie), munis des normes $\|\cdot\|_{\mathbb{E}}$ et $\|\cdot\|_{\mathbb{F}}$ respectivement. Dans cette section, nous étudions la différentiabilité d'une application f définie sur une partie Ω de $\mathbb{E}$ à valeurs dans $\mathbb{F}$:

$$f : \Omega \subseteq \mathbb{E} \to \mathbb{F}.$$

En dehors de l'étude de la dérivabilité directionnelle, nous supposerons que Ω est ouvert.

C.1 Dérivées directionnelle et au sens de Gâteaux

Dérivée directionnelle

En optimisation, beaucoup de fonctions ne sont pas différentiables dans le sens classique de Fréchet (section C.2), mais sont toutefois directionnellement différentiables, une notion que nous présentons dans cette section.

On dit que f a une *dérivée directionnelle* en $x \in \Omega$ (Ω ne doit pas nécessairement être ouvert ici) dans la direction $h \in \mathbb{E}$ si $x + th \in \Omega$ pour $t > 0$ suffisamment petit (ce sera le cas si Ω est ouvert) et si la limite

$$f'(x;h) := \lim_{t \downarrow 0} \frac{f(x+th) - f(x)}{t} \tag{C.1}$$

existe. La notation « $t \downarrow 0$ » signifie « $t \to 0$ avec $t > 0$ ». On dit que f a *des dérivées directionnelles* en x (sans spécifier les directions) si elle a des dérivées directionnelles en x suivant toutes les directions de $\mathbb{E}$. On notera que la dérivée directionnelle $f'(x;h)$ est un élément de $\mathbb{F}$.

On n'a pas besoin de mettre une topologie sur $\mathbb{E}$ pour que cette définition ait un sens. En regardant le comportement de f le long de la direction h, on se ramène à l'étude d'une fonction définie dans un voisinage de 0 de $\mathbb{R}_+$, muni de sa topologie canonique. Par contre, on a besoin d'une topologie sur $\mathbb{F}$ pour donner un sens à la limite dans (C.1).

Pour les fonctions d'une variable réelle, la dérivée directionnelle $f'_+(\cdot) := f'(\cdot;1)$ est appelée *dérivée à droite* et la dérivée directionnelle $f'_-(\cdot) := f'(\cdot;-1)$ est appelée *dérivée à gauche*.

Pour les fonctions à valeurs dans $\mathbb{R}$, il arrivera que l'on admette des dérivées directionnelles valant $-\infty$ ou $+\infty$, c'est-à-dire que l'on admettra parfois de prendre la limite de (C.1) dans $\overline{\mathbb{R}}$ plutôt que dans $\mathbb{R}$. Ainsi, on pourra parler de la dérivée à droite de $f(x) = \sqrt{x}$ en 0, celle-ci valant $+\infty$.

Exemple C.1 (dérivabilité directionnelle de la fonction max) La *fonction max* est définie par

$$\mu : x \in \mathbb{R}^n \mapsto \mu(x) := \max_{i \in [1:n]} x_i. \tag{C.2}$$

On note

$$I(x) := \{i \in [1:n] : \mu(x) = x_i\}. \tag{C.3}$$

La fonction max est convexe et lipschitzienne. Elle est différentiable dans le sens classique de Fréchet en x si, et seulement si, $I(x)$ est un singleton. Mais, μ est directionnellement différentiable en tout point et l'on a pour x et $h \in \mathbb{R}^n$:

$$\mu'(x;h) = \max_{i \in I(x)} h_i. \tag{C.4}$$

L'exercice C.2 propose de démontrer ces affirmations. □

En général, la composition de deux fonctions ayant des dérivées directionnelles n'a pas nécessairement de dérivées directionnelles. Voici un contre-exemple [546 ; p. 484].

Contre-exemple C.2 (composition non directionnellement différentiable) Soient $\varphi : \mathbb{R} \to \mathbb{R}^2$ et $\psi : \mathbb{R}^2 \to \mathbb{R}$ définies par

$$\varphi(x) = \begin{cases} (x, x^2 \sin(1/x)) & \text{si } x \neq 0 \\ 0 & \text{si } x = 0 \end{cases} \qquad \text{et} \qquad \psi(y_1, y_2) = \begin{cases} y_1 & \text{si } y_2 = 0 \\ 0 & \text{si } y_2 \neq 0. \end{cases}$$

Il est simple de voir que φ est différentiable en zéro (voir la section C.2.1 au besoin), que ψ est positivement homogène de degré un (c'est-à-dire $\psi(ty) = t\psi(y)$, pour tout $y \in \mathbb{R}^2$ et tout $t > 0$) et donc directionnellement différentiable en zéro, mais que $\psi \circ \varphi$ n'est pas directionnellement différentiable en zéro. □

Cependant, en exigeant un peu plus de régularité pour la fonction ψ, on peut obtenir la dérivabilité directionnelle de la composition $\psi \circ \varphi$. Ci-dessous, on demande une

propriété un peu forte, la lipschitzianité de ψ, mais celle-ci est facilement vérifiable et a souvent lieu.

Proposition C.3 (dérivabilité directionnelle d'une composition) *Soient $\mathbb{E}$, $\mathbb{F}$ et $\mathbb{G}$ trois espaces vectoriels normés. On suppose que $\varphi : \mathbb{E} \to \mathbb{F}$ a une dérivée directionnelle en $x \in \mathbb{E}$ dans la direction $h \in \mathbb{E}$ et que $\psi : \mathbb{F} \to \mathbb{G}$ est lipschitzienne dans un voisinage de $\varphi(x)$ et a une dérivée directionnelle en $\varphi(x)$ dans la direction $\varphi'(x;h)$. Alors $(\psi \circ \varphi)$ a une dérivée directionnelle en x dans la direction h et l'on a*

$$(\psi \circ \varphi)'(x;h) = \psi'(\varphi(x);\varphi'(x;h)). \tag{C.5}$$

Démonstration. Pour $t \downarrow 0$, on a

$$\begin{aligned}(\psi \circ \varphi)(x+th) &= \psi(\varphi(x) + t\varphi'(x;h) + o(t)) && [\varphi \text{ est DD}]\\ &= \psi(\varphi(x) + t\varphi'(x;h)) + o(t) && [\psi \text{ est lipschitzienne}]\\ &= (\psi \circ \varphi)(x) + t\psi'(\varphi(x);\varphi'(x;h)) + o(t) && [\psi \text{ est DD}],\end{aligned}$$

où DD abrège « directionnellement différentiable ». C'est le résultat recherché. □

Cette proposition et l'exemple C.1 permettent d'obtenir des conditions de dérivabilité directionnelle d'un maximum de fonctions directionnellement différentiables. Soient $f_i : \mathbb{E} \to \mathbb{R}$, $i \in [1:m]$, des fonctions, *en nombre fini*, ayant une dérivée directionnelle en $x \in \mathbb{E}$ dans la direction $h \in \mathbb{E}$. Alors leur maximum $f : \mathbb{E} \to \mathbb{R}$, définie en $x \in \mathbb{E}$ par

$$f(x) := \max_{i \in [1:m]} f_i(x),$$

a une dérivée directionnelle en x dans la direction h, qui est donnée par

$$f'(x;h) = \max_{i \in I(x)} f_i'(x;h), \tag{C.6}$$

où $I(x) := \{i \in [1:m] : f(x) = f_i(x)\}$. Le cas du supremum d'un *nombre infini* de fonctions est beaucoup plus complexe à analyser et joue un rôle dans l'étude de la perturbation de problèmes d'optimisation [74]. Le théorème de Danskin [152] [74 ; théorème 4.13] est alors souvent utilisé.

Dérivée au sens de Gâteaux

Dans cette notion de différentiabilité directionnelle, on regarde le comportement de f suivant les demi-droites $\{x + th : t > 0\}$ et on ne demande pas qu'il y ait un lien entre les dérivées directionnelles suivant différentes directions. En demandant un tel lien, on obtient une notion de différentiabilité un peu plus forte. On dit que f est *Gâteaux-différentiable* (on dit aussi *Gâteaux-dérivable* ou *G-différentiable* ou encore *G-dérivable*) en $x \in \Omega$ si elle admet une dérivée directionnelle en x suivant toutes les directions $h \in \mathbb{E}$ et si l'application

$$h \in \mathbb{E} \mapsto f'(x;h) \in \mathbb{F}$$

est linéaire continue. On note $f'(x) \in \mathcal{L}(\mathbb{E},\mathbb{F})$ cet opérateur. Donc

$$f'(x) \cdot h = f'(x;h), \quad \forall h \in \mathbb{E}.$$

Gradient d'une fonction réelle

Supposons que $\mathbb{E}$ soit un espace de Hilbert, muni du produit scalaire $\langle \cdot, \cdot \rangle$. Si la fonction $f : \Omega \subseteq \mathbb{E} \to \mathbb{R}$ *à valeurs scalaires* est G-différentiable en $x \in \Omega$, $f'(x)$ est une forme linéaire continue sur $\mathbb{E}$. Par le théorème de représentation de Riesz-Fréchet (théorème A.3), il existe un unique vecteur de $\mathbb{E}$, noté $\nabla f(x)$ et appelé *gradient* de f en x, tel que

$$\langle \nabla f(x), h \rangle = f'(x) \cdot h, \quad \forall h \in \mathbb{E}.$$

Cette relation définit $\nabla f(x) \in \mathbb{E}$. Elle relie la notion de dérivée, apparaissant à droite dans la relation, qui est de nature topologique (elle n'utilise que la notion de limite), à la notion de gradient, apparaissant à gauche dans la relation. Cette dernière est de nature géométrique, puisqu'elle fait intervenir un produit scalaire. Par conséquent, en changeant de produit scalaire, on ne modifie pas $f'(x)$ qui n'en dépend pas, tandis que le gradient $\nabla f(x)$ sera généralement modifié.

Le gradient va jouer un rôle essentiel dans les algorithmes de minimisation. En effet, la direction $-\nabla f(x)$ est une direction suivant laquelle la fonction f décroît localement en x. En effet, si $\nabla f(x) \neq 0$ et si α est suffisamment petit, par définition de la dérivée au sens de Gâteaux, $f(x - \alpha \nabla f(x)) - f(x)$ est proche de

$$\alpha f'(x) \cdot (-\nabla f(x)) = -\alpha \|\nabla f(x)\|^2 < 0.$$

Donc

$$f(x - \alpha \nabla f(x)) < f(x), \quad \text{pour } \alpha > 0 \text{ petit}.$$

Le gradient est modifié par un changement de produit scalaire. Si l'on prend sur $\mathbb{E}$ un autre produit scalaire :

$$(x,y) \in \mathbb{E} \times \mathbb{E} \mapsto \langle x, y \rangle_{\mathbb{E}} = \langle Sx, y \rangle,$$

où $S : \mathbb{E} \to \mathbb{E}$ est un opérateur auto-adjoint et défini positif pour le produit scalaire $\langle \cdot, \cdot \rangle$ et si l'on note $\nabla_{\mathbb{E}} f(x)$ le gradient de f en x pour ce nouveau produit scalaire, on a

$$\nabla_{\mathbb{E}} f(x) = S^{-1} \nabla f(x). \tag{C.7}$$

En effet, pour h arbitraire dans $\mathbb{E}$, on a

$$\langle \nabla f(x), h \rangle = f'(x) \cdot h = \langle \nabla_{\mathbb{E}} f(x), h \rangle_{\mathbb{E}} = \langle S \nabla_{\mathbb{E}} f(x), h \rangle.$$

On en déduit (C.7).

C.2 Dérivée au sens de Fréchet

C.2.1 Dérivée première

Définition

On dit que f est *Fréchet-différentiable* [222 ; 1911] (on dit aussi *Fréchet-dérivable* ou *F-différentiable* ou *F-dérivable* ou simplement *différentiable* ou *dérivable*) en $x \in \Omega$ s'il existe un opérateur linéaire continu L de $\mathbb{E}$ dans $\mathbb{F}$ tel que

$$\lim_{\|h\|_{\mathbb{E}} \downarrow 0} \frac{1}{\|h\|_{\mathbb{E}}} \Big(f(x+h) - f(x) - Lh \Big) = 0. \tag{C.8}$$

On dit que $f : \Omega \to \mathbb{F}$ est *F-différentiable sur* Ω si f est F-différentiable en tout point de Ω.

La limite dans (C.8) est prise dans $\mathbb{F}$. On a pris soin de prendre la limite pour des $h \neq 0$ afin que le quotient dans (C.8) ait un sens. L'opérateur L est appelé *dérivée* de f en x. On peut aussi récrire la condition (C.8) comme suit

$$f(x+h) = f(x) + Lh + o(h), \tag{C.9}$$

si le *petit o* de h, $o(h)$, désigne une fonction nulle en $h = 0$ et vérifiant la propriété

$$\lim_{\|h\|_{\mathbb{E}} \downarrow 0} \frac{1}{\|h\|_{\mathbb{E}}} o(h) = 0,$$

où la limite est prise dans un espace normé approprié (ici $\mathbb{F}$).

Lien avec la Gâteaux-différentiabilité

On voit clairement par (C.9) que si f est F-différentiable est $x \in \Omega$, alors f est continue en x. Ceci n'est pas vrai si f est seulement G-différentiable en x (exercice C.3). En fait, la F-différentiabilité est une notion plus forte que la G-différentiabilité, comme le montre la proposition suivante.

Proposition C.4 *Si $f : \Omega \subseteq \mathbb{E} \to \mathbb{F}$ est F-différentiable en $x \in \Omega$ avec une dérivée L, alors f est G-différentiable en x et $L = f'(x)$.*

Démonstration. Soit $h_0 \in \mathbb{E}$. En prenant $h = th_0$ dans (C.8) avec $t \downarrow 0$, on obtient

$$\frac{1}{t} \Big(f(x + th_0) - f(x) - tLh_0 \Big) \to 0.$$

On en déduit que pour $t \downarrow 0$

$$\frac{1}{t} \Big(f(x + th_0) - f(x) \Big) \to Lh_0.$$

Ceci montre que f admet une dérivée directionnelle en x suivant h_0. Comme h_0 est arbitraire dans $\mathbb{E}$, on en déduit que f est G-différentiable et que $f'(x) = L$. □

La réciproque de ce résultat n'a pas lieu : une fonction G-différentiable peut ne pas être F-différentiable, puisqu'une fonction G-différentiable en un point peut ne pas être continue en ce point (exercice C.3), ce qui n'est pas le cas d'une fonciton F-différentiable.

Dérivée partielle

Supposons que l'espace vectoriel $\mathbb{E}$ soit le produit de n espaces vectoriels $\mathbb{E}_1$, …, $\mathbb{E}_n$, c'est-à-dire $\mathbb{E} := \mathbb{E}_1 \times \cdots \times \mathbb{E}_n$, et que $f : \Omega \to \mathbb{F}$ soit une application définie sur un ouvert Ω de $\mathbb{E}$ à valeurs dans un espace vectoriel $\mathbb{F}$. En un point $x := (x_1, \ldots, x_n) \in \Omega$, avec des $x_i \in \mathbb{E}_i$, on peut considérer la i-ième *application partielle*

$$f_i : \Omega_i \to \mathbb{F} : v_i \mapsto f(x_1, \ldots, x_{i-1}, v_i, x_{i+1}, \ldots, x_n),$$

où $i \in [1:n]$ et $\Omega_i := \Omega \cap (\{x_1\} \times \cdots \times \{x_{i-1}\} \times \mathbb{E}_i \times \{x_{i+1}\} \times \cdots \times \{x_n\})$. Si cette application admet une dérivée en x_i, on l'appelle la i-ième *dérivée partielle* de f en x. C'est une application linéaire continue de $\mathbb{E}_i$ dans $\mathbb{F}$, un élément de $\mathcal{L}(\mathbb{E}_i, F)$ donc. Suivant le contexte, la note

$$f_i'(x) \qquad \text{ou} \qquad f_{x_i}'(x) \qquad \text{ou} \qquad \partial_i f(x) \qquad \text{ou} \qquad \partial_{x_i} f(x) \qquad \text{ou} \qquad \frac{\partial f}{\partial x_i}(x).$$

Si f est dérivable en $x \in \Omega$, alors ses dérivées partielles existent et $f'(x)$ peut se calculer à partir des $f_i'(x)$, comme l'affirme le théorème suivant.

Théorème C.5 (dérivée partielle) *Dans le cadre défini ci-dessus, si f est dérivable en $x \in \Omega$, alors elle possède en ce point des dérivées partielles $f_i'(x)$ pour tout $i \in [1:n]$ et, pour une direction $d = (d_1, \ldots, d_n) \in \mathbb{E}$, on a*

$$f'(x) \cdot d = \sum_{i=1}^{n} f_i'(x) \cdot d_i. \tag{C.10}$$

Démonstration. Comme $f'(a)$ est une application linéaire continue de $\mathbb{E}$ dans $\mathbb{F}$ et que $d = \sum_i (0, \ldots, 0, d_i, 0, \ldots, 0)$, on a

$$f'(x) \cdot d = \sum_{i=1}^{n} L_i(d_i), \tag{C.11}$$

où L_i est l'application linéaire continue $d_i \in \mathbb{E}_i \mapsto f'(x) \cdot (0, \ldots, 0, d_i, 0, \ldots, 0)$. Par la dérivabilité de f en x, on a

$$f\Big(x + (0, \ldots, 0, d_i, 0, \ldots, 0)\Big) = f(x) + f'(x) \cdot (0, \ldots, 0, d_i, 0, \ldots, 0) + o(\|d_i\|),$$

qui se récrit $f_i(x_i + d_i) = f_i(x_i) + L_i(d_i) + o(\|d_i\|)$: c'est la définition même de la dérivabilité de f_i en x_i et du fait que $f_i'(x) = L_i$. Alors l'expression (C.11) conduit à (C.10). □

Dérivée d'une fonction composée

Théorème C.6 (dérivée d'une fonction composée) *Soient $\mathbb{E}$, $\mathbb{F}$ et $\mathbb{G}$ trois espaces normés, Ω un ouvert de $\mathbb{E}$ et $\mathcal{O}$ un ouvert de $\mathbb{F}$. Si $f : \Omega \to \mathcal{O}$ est F-différentiable en $x \in \Omega$ et $g : \mathcal{O} \to \mathbb{G}$ est F-différentiable en $f(x)$, alors $(g \circ f)$ est F-différentiable en x et on a*

$$(g \circ f)'(x) = g'(f(x)) \circ f'(x). \tag{C.12}$$

Remarquons que l'identité (C.12) a bien un sens puisqu'à gauche on trouve un opérateur de $\mathcal{L}(\mathbb{E}, \mathbb{G})$ et qu'à droite on a la composition d'un opérateur $f'(x) \in \mathcal{L}(\mathbb{E}, \mathbb{F})$ et d'un opérateur linéaire $g'(f(x)) \in \mathcal{L}(\mathbb{F}, \mathbb{G})$.

Démonstration. Comme f est F-différentiable en x, on a

$$f(x+h) = f(x) + f'(x) \cdot h + \alpha_1(h),$$

où la fonction $h \mapsto \alpha_1(h)$ est un $o(h)$. On en déduit

$$(g \circ f)(x+h) = g(f(x+h)) = g\Big(f(x) + f'(x) \cdot h + \alpha_1(h)\Big). \tag{C.13}$$

D'autre part, g est F-différentiable en $f(x)$, si bien que l'on a

$$g(f(x)+k) = g(f(x)) + g'(f(x)) \cdot k + \alpha_2(k),$$

où la fonction $k \mapsto \alpha_2(k)$ est un $o(k)$. En prenant $k = f'(x) \cdot h + \alpha_1(h)$ et en combinant avec (C.13), on obtient

$$(g \circ f)(x+h) = g(f(x)) + g'(f(x)) \cdot \Big(f'(x) \cdot h + \alpha_1(h)\Big) + \alpha_2\Big(f'(x) \cdot h + \alpha_1(h)\Big).$$

On a

$$g'(f(x)) \cdot \alpha_1(h) = o(h)$$

et comme $f'(x) \cdot h + \alpha_1(h)$ est borné par une constante fois $\|h\|_{\mathbb{E}}$, on a aussi

$$\alpha_2\Big(f'(x) \cdot h + \alpha_1(h)\Big) = o\Big(f'(x) \cdot h + \alpha_1(h)\Big) = o(h).$$

On déduit de ces estimations

$$(g \circ f)(x+h) = (g \circ f)(x) + g'(f(x)) \cdot (f'(x) \cdot h) + o(h).$$

Donc $(g \circ f)$ est F-différentiable en x et on a la formule (C.12). □

Accroissements finis

La formule (C.9) définissant la dérivée de f en x nous dit comment varie f pour une variation infiniment petite de la variable x, mais ne donne pas d'estimation sur la différence $f(x+h)-f(x)$ pour un h non nul (on dit « fini » par opposition à « infiniment petit »). C'est cette information qu'apportent les théorèmes des accroissements finis, les théorèmes C.7 et C.12.

On appelle *segment ouvert* d'un espace vectoriel, un sous-ensemble de $\mathbb{E}$ défini à partir de la donnée de deux points *distincts* x et $y \in \mathbb{E}$ par

$$]x,y[:= \{(1-t)x + ty : t \in]0,1[\}.$$

On définit de la même manière le *segment fermé* (dans ce cas on peut avoir $x = y$) :

$$[x,y] := \{(1-t)x + ty : t \in [0,1]\}.$$

Si f est à valeurs scalaires on obtient aisément le résultat du théorème C.7 à partir du théorème des accroissements finis pour les fonctions réelles d'une variable réelle.

Théorème C.7 (des accroissements finis pour fonctions à valeurs scalaires) *Soient Ω un ouvert de $\mathbb{E}$, $x \in \Omega$ et $h \in \mathbb{E}$ tel que le segment fermé $[x, x+h] \subseteq \Omega$. On suppose que $f : \Omega \to \mathbb{R}$ à valeurs scalaires est continue sur Ω et différentiable sur le segment ouvert $]x, x+h[$. Alors il existe $\theta \in]0,1[$ tel que*

$$f(x+h) = f(x) + f'(x+\theta h)\cdot h.$$

Démonstration. On se ramène au cas d'une fonction $\xi : [0,1] \to \mathbb{R}$ en définissant

$$\xi(t) = f(x+th), \quad \forall t \in [0,1].$$

Cette fonction est continue sur $[0,1]$, différentiable sur $]0,1[$ et pour tout $t \in]0,1[$ le théorème C.6 donne

$$\xi'(t) = f'(x+th)\cdot h.$$

On applique le théorème des accroissements finis à ξ : il existe $\theta \in]0,1[$ tel que

$$\xi'(\theta) = \xi(1) - \xi(0).$$

On en déduit le résultat. □

Si f est à valeurs vectorielles, le théorème précédent n'est plus nécessairement vrai (un contre-exemple est donné dans l'exercice C.6). On a toutefois le résultat très utile du théorème C.12. La démonstration de celui-ci utilise le lemme d'intérêt général suivant.

Lemme C.8 *Soient $\xi : [0,1] \to \mathbb{F}$ et $\mu : [0,1] \to \mathbb{R}$ deux fonctions continues sur $[0,1]$ et différentiables à droite sur $]0,1[$ (μ'_+ pouvant prendre la valeur $+\infty$). On suppose que pour tout $t \in]0,1[$, on a $\|\xi'_+(t)\|_{\mathbb{F}} \leqslant \mu'_+(t)$. Alors*

$$\|\xi(1) - \xi(0)\|_{\mathbb{F}} \leqslant \mu(1) - \mu(0),$$

avec inégalité stricte s'il existe un $t_0 \in]0,1[$ tel que $\|\xi'_+(t_0)\|_{\mathbb{F}} < \mu'_+(t_0)$.

Démonstration. Supposons que μ'_+ ne prenne que des valeurs finies (on pourra aisément adapter la démonstration au cas où μ'_+ prend la valeur $+\infty$). Soit $\varepsilon > 0$. On va montrer que pour tout $t \in [0,1]$, on a

$$\|\xi(t) - \xi(0)\|_{\mathbb{F}} \leqslant \mu(t) - \mu(0) + \varepsilon t + \varepsilon. \tag{C.14}$$

On en déduit la première partie du résultat en prenant $t = 1$ et en faisant tendre $\varepsilon \to 0$.

Par continuité, l'inégalité (C.14) est vérifiée pour $t \geqslant 0$ petit. Soit alors

$$\bar{t} = \sup\{t \in [0,1] : \text{ (C.14) est vérifiée sur } [0,t[\}.$$

Par continuité, (C.14) est aussi vérifiée pour $t = \bar{t}$. Montrons que $\bar{t} = 1$, ce qui clôturera la première partie de la démonstration.

Supposons que $\bar{t} < 1$. Alors pour $\tau > 0$ tel que $\bar{t} + \tau \in [0,1]$, on a grâce à la différentiabilité à droite de ξ et μ en $\bar{t}$

$$\xi(\bar{t} + \tau) - \xi(\bar{t}) - \xi'_+(\bar{t})\tau = o(\tau),$$

$$\mu(\bar{t} + \tau) - \mu(\bar{t}) - \mu'_+(\bar{t})\tau = o(\tau).$$

Donc pour $\tau > 0$ assez petit, le fait que $\|\xi'_+(\bar{t})\|_{\mathbb{F}} \leqslant \mu'_+(\bar{t})$ implique

$$\|\xi(\bar{t} + \tau) - \xi(\bar{t})\|_{\mathbb{F}} \leqslant \|\xi'_+(\bar{t})\tau\|_{\mathbb{F}} + \frac{\varepsilon}{2}\tau \leqslant \mu'_+(\bar{t})\tau + \frac{\varepsilon}{2}\tau,$$

$$\mu'_+(\bar{t})\tau \leqslant \mu(\bar{t} + \tau) - \mu(\bar{t}) + \frac{\varepsilon}{2}\tau.$$

En sommant les deux dernières inégalités, on obtient

$$\|\xi(\bar{t} + \tau) - \xi(\bar{t})\|_{\mathbb{F}} \leqslant \mu(\bar{t} + \tau) - \mu(\bar{t}) + \varepsilon\tau.$$

Comme (C.14) est vérifiée en $t = \bar{t}$, on a en utilisant l'inégalité triangulaire

$$\|\xi(\bar{t} + \tau) - \xi(0)\|_{\mathbb{F}} \leqslant \mu(\bar{t} + \tau) - \mu(0) + \varepsilon(\bar{t} + \tau) + \varepsilon.$$

Ceci montre que (C.14) est vérifiée en $\bar{t} + \tau$, ce qui contredit la maximalité de $\bar{t}$.

Supposons à présent qu'il existe $t_0 \in]0,1[$ tel que $\delta := \mu'_+(t_0) - \|\xi'_+(t_0)\|_{\mathbb{F}} > 0$. Alors pour $\tau > 0$ suffisamment petit, on a en procédant comme ci-dessus

$$\mu'_+(t_0)\tau \leqslant \mu(t_0 + \tau) - \mu(t_0) + \frac{\delta}{3}\tau$$

et donc

$$\begin{aligned}\|\xi(t_0+\tau)-\xi(t_0)\|_{\mathbb{F}} &\leqslant \|\xi'_+(t_0)\|_{\mathbb{F}}\tau+\frac{\delta}{3}\tau\\ &= \mu'_+(t_0)\tau-\frac{2\delta}{3}\tau\\ &\leqslant \mu(t_0+\tau)-\mu(t_0)-\frac{\delta}{3}\tau\\ &< \mu(t_0+\tau)-\mu(t_0).\end{aligned}$$

D'autre part, d'après la première partie,

$$\|\xi(t_0)-\xi(0)\|_{\mathbb{F}}\leqslant\mu(t_0)-\mu(0)$$

$$\|\xi(1)-\xi(t_0+\tau)\|_{\mathbb{F}}\leqslant\mu(1)-\mu(t_0+\tau).$$

En sommant ces trois dernières inégalités, on obtient l'inégalité stricte désirée. □

Corollaire C.9 *Soient $a,b\in\mathbb{R}$, avec $a<b$, et $\varphi:[a,b]\to\mathbb{R}$ une fonction continue, dérivable à droite sur $]a,b[$, φ'_+ pouvant prendre la valeur $+\infty$. On suppose que pour tout $t\in]a,b[$, on a $\varphi'_+(t)\geqslant 0$. Alors*

$$\varphi(b)\geqslant\varphi(a),$$

avec inégalité stricte s'il existe un $t_0\in]a,b[$ tel que $\varphi'_+(t_0)>0$.

Démonstration. On applique le lemme C.8 avec $\xi\equiv 0\in\mathbb{R}$ et $\mu(s)=\varphi((1-s)a+sb)$. Pour tout $s\in]0,1[$, on a bien

$$\mu'_+(s)=\varphi'_+((1-s)a+sb)(b-a)\geqslant|\xi'_+(s)|=0.$$

On en déduit que $\mu(1)\geqslant\mu(0)$. On a aussi $\mu(1)>\mu(0)$ si $\mu'_+(s_0)>0$ en un seul point $s_0\in]0,1[$. □

Le résultat précédent s'utilise le plus souvent pour comparer les valeurs de deux fonctions continues φ et ψ dont les dérivées à droite se comparent.

Corollaire C.10 *Soient $a,b\in\mathbb{R}$, avec $a<b$, et $\varphi,\psi:[a,b]\to\mathbb{R}$ deux fonctions continues, dérivables à droite sur $]a,b[$. On suppose que pour tout $t\in]a,b[$, on a $\varphi'_+(t)\leqslant\psi'_+(t)$. Alors*

$$\varphi(b)-\varphi(a)\leqslant\psi(b)-\psi(a),$$

avec inégalité stricte s'il existe un $t_0\in]a,b[$ tel que $\varphi'_+(t_0)<\psi'_+(t_0)$.

Démonstration. On applique le corollaire C.9 sur $(\psi-\varphi)$. □

Le *théorème de Rolle* affirme que si $a, b \in \mathbb{R}$, $a < b$, et $\varphi : [a, b] \to \mathbb{R}$ est une fonction continue, dérivable sur $]a, b[$, alors il existe un point $c \in]a, b[$ tel que $\varphi'(c) = \frac{\varphi(b)-\varphi(a)}{b-a}$. Voici une variante de ce théorème qui peut être utilisée lorsque la fonction n'admet que des dérivées directionnelles à droite.

Corollaire C.11 *Soient $a, b \in \mathbb{R}$, $a < b$, et $\varphi : [a, b] \to \mathbb{R}$ une fonction continue et dérivable à droite sur $]a, b[$, φ'_+ pouvant prendre les valeurs $-\infty$ ou $+\infty$. Alors il existe $c_1, c_2 \in]a, b[$ tels que*

$$\varphi'_+(c_1) \leqslant \frac{\varphi(b) - \varphi(a)}{b - a} \leqslant \varphi'_+(c_2).$$

Démonstration. On raisonne par l'absurde. Si l'inégalité de gauche n'est jamais vérifiée sur $]a, b[$, alors pour tout $t \in]a, b[$, $\varphi'_+(t) > \psi'(t)$, où $\psi(t) := \frac{\varphi(b)-\varphi(a)}{b-a}t$. D'après le corollaire C.9, on en déduirait que $\varphi(b) - \varphi(a) > \psi(b) - \psi(a)$, ce qui est absurde. On s'y prend de la même manière pour l'inégalité de droite. □

Théorème C.12 (des accroissements finis pour fonctions à valeurs vectorielles) *Soient Ω un ouvert de $\mathbb{E}$, $x \in \Omega$ et $h \in \mathbb{E}$ tel que le segment fermé $[x, x + h] \subseteq \Omega$. On suppose que $f : \Omega \to \mathbb{F}$ est continue sur Ω et différentiable sur le segment ouvert $]x, x + h[$. Alors*

$$\|f(x + h) - f(x)\|_{\mathbb{F}} \leqslant \left(\sup_{z \in]x, x+h[} \|f'(z)\| \right) \|h\|_{\mathbb{E}}.$$

Démonstration. On prend $\xi(t) = f(x + th)$ et $\mu(t) = M\|h\|_{\mathbb{E}}t$, où

$$M = \sup_{z \in]x, x+h[} \|f'(z)\|.$$

Si $M = +\infty$, il n'y a rien à démontrer. Sinon, ces fonctions à valeurs dans $\mathbb{F}$ et $\mathbb{R}$ respectivement sont bien continues sur $[0, 1]$ et différentiables sur $]0, 1[$. On a aussi pour $t \in]0, 1[$,

$$\|\xi'(t)\| = \|f'(x + th)h\|_{\mathbb{F}} \leqslant M\|h\|_{\mathbb{E}} = \mu'(t).$$

L'application du lemme C.8 donne le résultat. □

Corollaire C.13 *Sous les hypothèses du théorème C.12, si $L \in \mathcal{L}(\mathbb{E}, \mathbb{F})$, on a*

$$\|f(x+h)-f(x)-Lh\|_{\mathbb{F}} \leqslant \left(\sup_{z\in]x,x+h[}\|f'(z)-L\|\right)\|h\|_{\mathbb{E}}.$$

DÉMONSTRATION. On applique le théorème à l'application $x \mapsto f(x) - Lx$. □

Lorsque f est dérivable en x, ce corollaire s'utilise souvent avec $L = f'(x)$.

Fonction implicite

Le cadre est le suivant. On suppose que $\mathbb{E}$ est un espace *topologique*, que $\mathbb{F}$ et $\mathbb{G}$ sont deux espaces vectoriels *normés*, que Ω est un ouvert de $\mathbb{E}\times\mathbb{F}$ et que F est une application de $\Omega \to \mathbb{G}$. Soit (a,b) est un point de Ω tel que

$$F(a,b) = 0.$$

On cherche une fonction $x \to \eta(x)$, dite *fonction implicite*, définie dans un voisinage ouvert U de a telle que pour tout $x \in U$, on ait $(x,\eta(x)) \in \Omega$ et

$$F(x,\eta(x)) = 0.$$

Lorsqu'elle existe, on note F'_y la dérivée partielle de F par rapport à y.

L'existence d'une fonction implicite est assurée dans les conditions spécifiées dans le résultat suivant. La construction de $\eta(x)$ dans $\mathbb{F}$ requiert la complétude de $\mathbb{F}$.

Théorème C.14 (existence d'une fonction implicite) *Dans le cadre ci-dessus, on suppose que $\mathbb{F}$ est complet, que $F : \Omega \to \mathbb{G}$ est continue, que $F'_y(x,y)$ existe pour tout $(x,y) \in \Omega$ et $F'_y : \Omega \to \mathcal{L}(\mathbb{F},\mathbb{G})$ est continue, et que $F'_y(a,b)$ est un isomorphisme de $\mathbb{F}$ dans $\mathbb{G}$ (donc $F'_y(a,b)^{-1} \in \mathcal{L}(\mathbb{G},\mathbb{F})$), Alors, il existe un voisinage ouvert U de a dans $\mathbb{E}$, un voisinage ouvert V de b dans $\mathbb{F}$ et une fonction $\eta : U \to V$ continue tels que*

1) *pour tout $x \in U$, $F(x,\eta(x)) = 0$,*
2) *pour $x \in U$, le seul $y \in V$ vérifiant $F(x,y) = 0$ est $\eta(x)$.*

Si $\mathbb{F}$ est complet et $F'_y(a,b)$ est un isomorphisme, alors $\mathbb{G}$ est nécessairement complet.

Intéressons nous à présent à la différentiabilité de la fonction implicite $\eta : U \to V$. Il faut alors supposer que $\mathbb{E}$ est un *espace vectoriel normé*. Par ailleurs il faut nécessairement une hypothèse de dérivabilité de F par rapport à x, car sans cela, il n'y pas de raisons pour que η soit dérivable, comme le montre le cas où $F : \mathbb{R}^2 \to \mathbb{R}$ est définie par $F(x,y) = |x| - y$. La fonction implicite $\eta(x) = |x|$ n'est pas dérivable en $x = 0$. Dans le résultat suivant, on suppose que la fonction implicite existe et on montre sa régularité sous des hypothèses plus faibles que celles du théorème d'existence.

Théorème C.15 (dérivabilité d'une fonction implicite) *On suppose que* $\mathbb{E}$, $\mathbb{F}$ *et* $\mathbb{G}$ *sont trois espaces normés et que* Ω *est un ouvert de* $\mathbb{E} \times \mathbb{F}$. *Si* U *est un voisinage ouvert de* a *dans* $\mathbb{E}$ *et* V *est un voisinage ouvert de* b *dans* $\mathbb{F}$ *tels que* $U \times V \subseteq \Omega$, $\eta : U \to V$ *est une fonction implicite de* F *passant par* (a, b), η *est* continue *en* a, F *est dérivable en* (a, b), $F'_y(a, b)$ *est un* isomorphisme *de* $\mathbb{F}$

dans $\mathbb{G}$. Alors η est dérivable en a et on a la formule

$$\eta'(a) = -\left(F'_y(a,b)\right)^{-1} \circ \left(F'_x(a,b)\right).$$

Dès lors, si l'on peut montrer l'existence et la continuité de la fonction implicite par un autre moyen que le théorème d'existence ci-dessus, sa dérivabilité ne dépendra que de la dérivabilité de F sur le graphe de η.

L'étude des fonctions implicites peut se faire dans un cadre beaucoup plus général que celui des systèmes d'équations différentiables considérés ici. Il s'agit toujours de « systèmes » dont on veut décrire l'évolution des solutions en fonction de paramètres. En ce qui concerne l'optimisation, la question est ébauchée à la section 4.6 (Analyse de sensibilité) et traitée plus systématiquement dans l'ouvrage [74]. Pour les inéquations variationnelles, on pourra consulter [173]. Dans ces deux derniers cas, le bon concept est celui de multifonction implicite.

C.2.2 Dérivée seconde

Définition

Soient $\mathbb{E}$ et $\mathbb{F}$ deux espaces normés, Ω un ouvert de $\mathbb{E}$ et

$$f : \Omega \to \mathbb{F}$$

une application F-différentiable sur Ω. Pour tout $x \in \Omega$, $f'(x)$ appartient à l'espace normé $\mathcal{L}(\mathbb{E},\mathbb{F})$. Il y a donc un sens à s'intéresser à la différentiabilité de l'application

$$f' : \Omega \to \mathcal{L}(\mathbb{E},\mathbb{F}) : x \mapsto f'(x).$$

On dit que f est *deux fois différentiable* en $x \in \Omega$ si f est F-différentiable dans un voisinage de x et si f' est F-différentiable en x. On note cette dérivée $f''(x)$, qui est donc un élément de $\mathcal{L}(\mathbb{E},\mathcal{L}(\mathbb{E},\mathbb{F}))$. Dès lors, pour tout h et $k \in \mathbb{E}$, on a

$$\begin{array}{rcl} f''(x)\cdot h & \in & \mathcal{L}(\mathbb{E},\mathbb{F}) \\ (f''(x)\cdot h)\cdot k & \in & \mathbb{F}. \end{array}$$

On note encore

$$f''(x)\cdot(h,k) := (f''(x)\cdot h)\cdot k \qquad \text{et} \qquad f''(x)\cdot h^2 := f''(x)\cdot(h,h).$$

Cette notation est sans danger car, comme le montre la proposition suivante, l'application $(h,k) \mapsto f''(x)\cdot(h,k)$ est symétrique.

Propriétés

Proposition C.16 *Supposons que f soit deux fois différentiable en $x \in \Omega$. Alors l'application*

$$f''(x) : (h,k) \in \mathbb{E}^2 \mapsto f''(x)\cdot(h,k) \in \mathbb{F}$$

est bilinéaire symétrique.

Démonstration. La bilinéarité de $f''(x)$ vient de la linéarité de

$$h \in \mathbb{E} \mapsto f''(x) \cdot h \in \mathcal{L}(\mathbb{E}, \mathbb{F})$$

et de la linéarité de

$$k \in \mathbb{E} \mapsto (f''(x) \cdot h) \cdot k \in \mathbb{F}.$$

Montrons à présent la symétrie. Fixons h et $k \in \mathbb{E}$ et définissons pour $t > 0$ assez petit

$$S_t(h,k) = \frac{1}{t^2}(f(x+th+tk) - f(x+th) - f(x+tk) + f(x)).$$

Nous allons montrer que

$$\lim_{t \downarrow 0} S_t(h,k) = f''(x) \cdot (h,k).$$

Comme S_t est symétrique, cela démontrera le résultat.

Remarquons d'abord qu'en introduisant

$$g(k) = f(x+th+tk) - f(x+tk),$$

on a

$$S_t(h,k) - f''(x) \cdot (h,k) = \frac{1}{t^2}\Big(g(k) - g(0) - t^2 f''(x) \cdot (h,k)\Big).$$

On va majorer la norme du membre de droite en utilisant le théorème des accroissements finis. Dans ce but, on calcule

$$g'(k) = t\Big(f'(x+th+tk) - f'(x+tk)\Big),$$

qui, avec

$$\begin{aligned} f'(x+th+tk) &= f'(x) + tf''(x) \cdot (h+k) + o(t)\|h+k\|_{\mathbb{E}} \\ f'(x+tk) &= f'(x) + tf''(x) \cdot k + o(t)\|h\|_{\mathbb{E}}, \end{aligned}$$

devient

$$g'(k) = t^2 f''(x) \cdot h + to(t)(\|h+k\|_{\mathbb{E}} - \|h\|_{\mathbb{E}}).$$

Par le corollaire C.13 du théorème des accroissements finis, on obtient alors

$$\begin{aligned} \|S_t(h,k) - f''(x) \cdot (h,k)\|_{\mathbb{F}} &\leqslant \frac{1}{t^2}\left(\sup_{\xi \in]0,k[} \|g'(\xi) - t^2 f''(x) \cdot h\|\right)\|k\|_{\mathbb{E}} \\ &= \frac{o(t)}{t}\left(\sup_{\xi \in]0,k[} (\|h+\xi\|_{\mathbb{E}} - \|h\|_{\mathbb{E}})\right)\|k\|_{\mathbb{E}}, \end{aligned}$$

qui tend vers 0 lorsque $t \downarrow 0$. □

Le résultat suivant établit le lien entre $f''(x) \cdot (h,k)$ et la dérivée directionnelle de $x \mapsto f'(x) \cdot h$. Il fournit ainsi un moyen simple de calculer $f''(x) \cdot (h,k)$.

Proposition C.17 *Si f est deux fois différentiable en $x \in \Omega$ et $(h,k) \in \mathbb{E}^2$, alors l'application*

$$y \mapsto f'(y) \cdot h \in \mathbb{F},$$

définie dans un voisinage de x, est différentiable en x et sa dérivée directionnelle en x suivant $k \in \mathbb{E}$ est $f''(x) \cdot (h,k)$.

Démonstration. L'application ξ définie dans un voisinage de $x \in \Omega$ par

$$\xi(y) = f'(y) \cdot h$$

est différentiable en x. C'est en effet la composée $e_h \circ f'$ de f' qui est différentiable en x et de

$$e_h : \mathcal{L}(\mathbb{E}, \mathbb{F}) \to \mathbb{F} : L \mapsto Lh$$

qui est linéaire continue, donc différentiable (exercice C.4). On a

$$\xi'(x) \cdot k = (e_h \circ f')'(x) \cdot k = e_h(f''(x) \cdot k) = f''(x) \cdot (k,h). \qquad \square$$

Développements au deuxième ordre

Si f est deux fois différentiable, $f(x+h)$ a un développement au deuxième ordre autour de x à un $o(\|h\|^2)$ près. Celui-ci étend en quelque sorte la condition de dérivabilité (C.8) au deuxième ordre. Réciproquement, un développement au deuxième ordre autour de x à un $o(\|h\|^2)$ près *d'une fonction deux fois différentiable* permet de déterminer les dérivées première et seconde. C'est ce que montre le théorème suivant. Pour comprendre sa démonstration, il sera utile d'avoir fait l'exercice C.4 donnant les formules de la dérivée de fonctions linéaire et bilinéaire.

Théorème C.18 *Soient Ω un ouvert de $\mathbb{E}$ et $x \in \Omega$. On suppose que $f : \Omega \to \mathbb{F}$ est différentiable sur Ω et deux fois différentiable en x. Alors*

$$f(x+h) = f(x) + f'(x) \cdot h + \frac{1}{2} f''(x) \cdot h^2 + o(\|h\|_{\mathbb{E}}^2). \tag{C.15}$$

Réciproquement, s'il existe $L_0 \in \mathbb{F}$, une application linéaire continue $L_1 \in \mathcal{L}(\mathbb{E}, \mathbb{F})$ et une application bilinéaire continue symétrique $L_2 \in \mathcal{L}_2(\mathbb{E}, \mathbb{F})$ telles que

$$f(x+h) = L_0 + L_1 \cdot h + \frac{1}{2} L_2 \cdot h^2 + o(\|h\|_{\mathbb{E}}^2), \tag{C.16}$$

alors $L_0 = f(x)$, $L_1 = f'(x)$ et $L_2 = f''(x)$.

Démonstration. [(i)] Soit V un voisinage de $0 \in \mathbb{E}$ tel que $x + V \subseteq \Omega$. On introduit la fonction $g : V \to \mathbb{F}$ définie en $h \in V$ par

$$g(h) := f(x+h) - f(x) - f'(x)\cdot h - \frac{1}{2}f''(x)\cdot h^2.$$

Il faut montrer que $g(h) = o(\|h\|_{\mathbb{E}}^2)$. Clairement, g est dérivable sur V (théorème C.6 et exercice C.4) et

$$g'(h) = f'(x+h) - f'(x) - f''(x)\cdot h.$$

Mais f' est dérivable en x et $(f')'(x)\cdot h = f''(x)\cdot h$, si bien que cette expression de $g'(h)$ montre que

$$g'(h) = o(\|h\|_{\mathbb{E}}).$$

On utilise alors le théorème des accroissements finis (théorème C.12) à g et le fait que $g(0) = 0$:

$$\|g(h)\|_{\mathbb{F}} \leqslant \left(\sup_{k\in]0,h[} \|g'(k)\|\right) \|h\|_{\mathbb{E}} \leqslant o(\|h\|_{\mathbb{E}}^2).$$

[(ii)] Démontrons la réciproque. En prenant $h = 0$ dans (C.16), on voit que $L_0 = f(x)$. Compte tenu de cette information, en retranchant (C.15) et (C.16), en prenant $h = th'$ où h' est fixé et $t \downarrow 0$, en divisant par t et en passant à la limite, on trouve $L_1\cdot h' = f'(x)\cdot h'$. Comme h' est arbitraire, on a $L_1 = f'(x)$. On réitère le procédé en tenant compte des identités $L_0 = f(x)$ et $L_1 = f'(x)$: on retranche (C.15) et (C.16), on prend $h = th'$ où h' est fixé et $t \downarrow 0$, on divise par t^2 et on passe à la limite, pour trouver que $L_2\cdot(h')^2 = f''(x)\cdot(h')^2$. Comme h' est arbitraire, on a $L_2 = f''(x)$ parce que la valeur en (h,k) d'une fonction bilinéaire symétrique L_2 peut s'obtenir à partir de ses valeurs unidirectionnelles :

$$L_2(h,k) = \frac{1}{4}\Big(L_2(h+k,h+k) - L_2(h-k,h-k)\Big).$$

□

Aux théorèmes des accroissements finis C.7 et C.12, qui donnaient des estimations de $f(x+h) - f(x)$, correspondent les théorèmes C.19 et C.20 qui donnent des estimations de $f(x+h) - f(x) - f'(x)\cdot h$. Dans le théorème C.19 la fonction f doit être à valeurs scalaires.

Théorème C.19 *Soient Ω un ouvert de $\mathbb{E}$, $x \in \Omega$ et $h \in \mathbb{E}$ tel que le segment fermé $[x, x+h] \subseteq \Omega$. On suppose que $f : \Omega \to \mathbb{R}$ à valeurs scalaires est différentiable sur Ω et est deux fois différentiable sur le segment ouvert $]x, x+h[$. Alors il existe $\theta \in]0,1[$ tel que*

$$f(x+h) = f(x) + f'(x)\cdot h + \frac{1}{2}f''(x+\theta h)\cdot h^2.$$

Démonstration. On se ramène au cas d'une fonction $\xi : [0,1] \to \mathbb{R}$ en définissant

$$\xi(t) = f(x+th), \quad \forall t \in [0,1].$$

Cette fonction est différentiable sur $[0,1]$, deux fois différentiable sur $]0,1[$ et

$$\begin{aligned}\xi'(t) &= f'(x+th)\cdot h, \quad \forall t\in[0,1],\\ \xi''(t) &= f''(x+th)\cdot h^2, \quad \forall t\in\,]0,1[.\end{aligned}$$

On obtient le résultat en appliquant la formule de Taylor à ξ : il existe $\theta\in\,]0,1[$ tel que $\xi(1)=\xi(0)+\xi'(0)+\frac{1}{2}\xi''(\theta)$. □

Théorème C.20 *Soient Ω un ouvert de $\mathbb{E}$, $x\in\Omega$ et $h\in\mathbb{E}$ tel que le segment fermé $[x,x+h]\subseteq\Omega$. On suppose que $f:\Omega\to\mathbb{R}$ est différentiable sur Ω et deux fois différentiable sur le segment ouvert $]x,x+h[$. Alors*

$$\|f(x+h)-f(x)-f'(x)\cdot h\|_{\mathbb{F}}\leqslant\frac{1}{2}\left(\sup_{z\in]x,x+h[}\|f''(z)\|\right)\|h\|_{\mathbb{E}}^2.$$

Démonstration. On se ramène au théorème des accroissements finis C.12 et à son lemme, en introduisant la fonction $\xi:[0,1]\to\mathbb{F}$ définie par

$$\xi(t):=f(x+th)-f(x)-f'(x)\cdot(th).$$

Cette fonction est bien définie sur $[0,1]$, continue sur cet intervalle et dérivable en tout point de $]0,1[$. Il s'agit de majorer $\|\xi(1)-\xi(0)\|_{\mathbb{F}}$, ce que l'on peut faire grâce au lemme C.8 en majorant la dérivée de ξ, qui s'écrit pour tout $t\in\,]0,1[$

$$\xi'(t)=f'(x+th)\cdot h-f'(x)\cdot h.$$

D'après la proposition C.17, $z\mapsto f'(z)\cdot h$ est différentiable en tout point de $]x,x+h[$. En utilisant le théorème des accroissements finis C.12 sur cette fonction, on obtient grâce à l'identité ci-dessus

$$\begin{aligned}\|\xi'(t)\|_{\mathbb{F}} &\leqslant \left(\sup_{z\in]x,x+th[}\|f''(z)\cdot h\|\right)t\|h\|_{\mathbb{E}}\\ &\leqslant \left(\sup_{z\in]x,x+h[}\|f''(z)\|\right)t\|h\|_{\mathbb{E}}^2\\ &=: Mt\|h\|_{\mathbb{E}}^2.\end{aligned}$$

Si on définit $\mu(t):=\frac{1}{2}Mt^2\|h\|_{\mathbb{E}}^2$, on a $\|\xi'(t)\|_{\mathbb{F}}\leqslant\mu'(t)$ pour tout $t\in\,]0,1[$. Grâce au lemme C.8, on en déduit que $\|\xi(1)-\xi(0)\|_{\mathbb{F}}\leqslant\mu(1)-\mu(0)$, qui est le résultat attendu. □

Hessienne d'une fonction réelle

Supposons que $\mathbb{E}$ soit un espace de Hilbert, muni du produit scalaire $\langle\cdot,\cdot\rangle$. Si la fonction $f:\Omega\subseteq\mathbb{E}\to\mathbb{R}$ *à valeurs scalaires* est deux fois différentiable en $x\in\Omega$, l'application

$$k \in \mathbb{E} \mapsto f''(x) \cdot (h,k) \in \mathbb{R}$$

est linéaire continue. D'après le théorème de représentation de Riesz-Fréchet (théorème A.3), il existe un unique vecteur $v_h(x) \in \mathbb{E}$ tel que

$$f''(x) \cdot (h,k) = \langle v_h(x), k \rangle, \quad \forall k \in \mathbb{E}.$$

De la bilinéarité et de la continuité de $f''(x)$, on déduit que $v_h(x)$ dépend linéairement et continûment de h. Il existe donc un unique opérateur linéaire continu, noté $\nabla^2 f(x) \in \mathcal{L}(\mathbb{E},\mathbb{E})$ et appelé *hessienne* de f en x, tel que

$$f''(x) \cdot (h,k) = \langle \nabla^2 f(x) h, k \rangle, \quad \forall h,k \in \mathbb{E}.$$

Par la symétrie de $f''(x)$, on voit que $\nabla^2 f(x)$ est auto-adjoint :

$$\langle \nabla^2 f(x) h, k \rangle = \langle h, \nabla^2 f(x) k \rangle, \quad \forall h,k \in \mathbb{E}.$$

Notes

Tous les sujets abordés dans ce chapitre sont largement développés dans de nombreux ouvrages. Nous nous sommes souvent inspiré de Schwartz [542 ; 1992].

Exercices

C.1. Soient $\mathbb{E}$, $\mathbb{F}$ et $\mathbb{G}$ trois espaces normés, Ω un ouvert de $\mathbb{E}$ et $\mathcal{O}$ un ouvert de $\mathbb{F}$. Montrez que si $f : \Omega \to \mathcal{O}$ a une dérivée directionnelle en $x \in \Omega$ dans la direction $h \in \mathbb{E}$ et si $g : \mathcal{O} \to \mathbb{G}$ est *lipschitzienne* et a une dérivée directionnelle en $f(x)$ dans la direction $f'(x;h) \in \mathbb{F}$, alors $(g \circ f)$ a une dérivée directionnelle en x dans la direction h et

$$(g \circ f)'(x;h) = g'\Big(f(x); f'(x;h)\Big).$$

C.2. *Fonction max.* On considère la fonction max, notée μ en (C.2), et l'ensemble d'indices $I(x)$ défini en (C.3).

1) Montrez que μ est convexe et lipschitzienne.
2) Donnez la dérivée de μ lorsque $I(x)$ est un singleton.
3) Montrez que μ n'est pas Fréchet différentiable en x si $I(x)$ contient au moins deux indices.
4) Démontrez la formule (C.4).

C.3. Montrez que la fonction $f : \mathbb{R}^2 \to \mathbb{R}$ définie par

$$f(x) = \begin{cases} \frac{x_1^4 + x_2^6}{(x_1 - x_2^2)^2 + x_1^5} & \text{si } x \neq 0 \\ 0 & \text{si } x = 0 \end{cases}$$

est G-différentiable en 0 avec $f'(0) = 0$ mais n'est pas continue en 0.

Par conséquent, *une fonction G-différentiable n'est pas nécessairement continue, donc pas nécessairement F-différentiable.* On verra (exercice **??**) qu'*une fonction convexe G-différentiable est F-différentiable.*

C.4. *Dérivées des fonctions linéaire et bilinéaire.*

1) Soient $\mathbb{E}$ et $\mathbb{F}$ deux espaces normés. Si $l : \mathbb{E} \to \mathbb{F}$ est linéaire continue alors l est différentiable en tout point x de $\mathbb{E}$ et on a
$$l'(x) = l \qquad \text{et} \qquad l''(x) = 0.$$

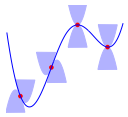
2) Soient $\mathbb{E}_1$, $\mathbb{E}_2$ et $\mathbb{F}$ des espaces normés. Si $b : \mathbb{E}_1 \times \mathbb{E}_2 \to \mathbb{F}$ est bilinéaire continue alors b est différentiable en tout point (x_1, x_2) de $\mathbb{E}_1 \times \mathbb{E}_2$ et pour tout (h_1, h_2) et $(k_1, k_2) \in \mathbb{E}_1 \times \mathbb{E}_2$, on a
$$b'(x_1, x_2) \cdot (h_1, h_2) = b(h_1, x_2) + b(x_1, h_2),$$
$$b''(x_1, x_2) \cdot ((h_1, h_2), (k_1, k_2)) = b(h_1, k_2) + b(k_1, h_2).$$

C.5. On considère la fonction définie sur $\mathbb{R}^n$ par $f(x) = \frac{1}{2}x^{\mathsf{T}}Ax + b^{\mathsf{T}}x$, où A est une matrice d'ordre n et $b \in \mathbb{R}^n$. Montrez que pour le produit scalaire euclidien
$$\nabla f(x) = \frac{1}{2}(A + A^{\mathsf{T}})x + b \quad \text{et} \quad \nabla^2 f(x) = \frac{1}{2}(A + A^{\mathsf{T}}).$$

Remarque. On notera que $f(x)$ ne dépend que de la partie symétrique $\frac{1}{2}(A + A^{\mathsf{T}})$ de A. Il est donc normal que $\nabla f(x)$ et $\nabla^2 f(x)$, qui ne dépendent que des valeurs $f(x)$ et non pas de la manière avec laquelle ces valeurs sont calculées, ne fasse intervenir A que par sa partie symétrique.

C.6. *Contre-exemple de l'hélice.* On considère l'application $f : \mathbb{R} \to \mathbb{R}^2$ définie par $f(x) = (\cos 2\pi x, \sin 2\pi x)$. Montrez que le résultat du théorème C.7 n'est pas vérifié pour cette fonction à valeurs vectorielles.

C.7. *Encadrement quadratique d'une fonction $\mathcal{C}^{1,1}_L$.* On dit que $f : \Omega \to \mathbb{R}$ est de classe $\mathcal{C}^{1,1}_L(\Omega)$, pour un ouvert Ω d'un espace euclidien $\mathbb{E}$, si elle est dérivable sur Ω et si sa dérivée y est L-lipschitzienne. Montrez que si Ω est convexe, une telle fonction vérifie pour tout x et $y \in \Omega$:
$$\left|f(y) - f(x) - \langle \nabla f(x), y - x\rangle\right| \leqslant \frac{L}{2}\,\|y - x\|^2.$$



C.8. *Gradient et hessien pour le produit scalaire euclidien.* Soient $f : \Omega \to \mathbb{R}$ une fonction deux fois dérivable définie sur un ouvert Ω de $\mathbb{R}^n$ et $x \in \Omega$. Montrez que si le gradient $\nabla f(x)$ et la hessienne $\nabla^2 f(x)$ sont calculés en utilisant le produit scalaire euclidien, on a
$$(\nabla f(x))_i = \frac{\partial f}{\partial x_i}(x) \qquad \text{et} \qquad (\nabla^2 f(x))_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}(x).$$

C.9. *Distance à un ensemble* [128]. Soient $\mathbb{E}$ un espace euclidien (produit scalaire $\langle \cdot, \cdot \rangle$ et norme associée $\|\cdot\|$) et P une partie non vide et fermée de $\mathbb{E}$. On considère la fonction $d_P : \mathbb{E} \to \mathbb{R}$, la *distance à P*, définie par $d_P(x) = \inf_{y \in P} \|x - y\|$. On appelle *projection* de x aur P, toute solution de $\inf_{y \in P} \|x - y\|$ (une telle projection existe car $\mathbb{E}$ est de dimension finie). Soit $x \in \mathbb{E}$ tel que d_P est différentiable en x et $d'_P(x) \neq 0$. Montrez que

1) $x \notin P$;
2) $\nabla d_P(x) = (x - \bar{x})/\|x - \bar{x}\|$, où $\bar{x}$ est une projection de x sur P;
3) x a une unique projection sur P.

D Solutions des exercices ▲

Ces mille questions
Qui se ramifient
N'amènent, au fond
Qu'ivresse et folie

A. Rimbaud (1872). Âge d'or.

There are two kinds of mathematicians. There are those who develop theories and those who are primarily problem solvers. I am the latter.

L. Nirenberg [507 ; 2015].

Get up in the morning with your problem before you. Breakfast with it. Go to the laboratory with it. Eat your lunch with it. Keep it before you after diner. Go to bed with it in your mind. Dream about it.

Attribué à Pavlov dans [483 ; p. 127].

S'il n'y a pas de solution, il n'y a pas de problème.

Les Shadoks.

Once a result is discovered, one may then look for a more direct proof. When first hunting for certainty it is reasonable to use whatever tools one possess.

Jonathan M. Borwein [77 ; 2014].

Nous n'avons pas essayé de rendre ce chapitre très lisible ! Les caractères utilisés sont plus petits et il y a peu de paragraphes structurant le texte. C'est tout à fait volontaire. D'une part, cela permet de raccourcir la longueur du chapitre et d'autre part, cela va dans le sens que l'on veut donner à ces notes : un moyen d'aider le lecteur qui s'efforce de faire les exercices, qui est à la recherche d'indications supplémentaires ou qui cherche à savoir si sa démonstration est correcte. C'est aussi pour cette raison que nous donnons dans un premier temps quelques indications sur les exercices les plus difficiles ou les plus longs (section D.1). Viennent ensuite des solutions possibles, données *in extenso* (section D.2). Dans celles-ci, des résultats obtenus dans des exercices précédents l'exercice considéré sont parfois utilisés.

D.1 Indications

Introduction

1.8. 1) Le problème peut se récrire sous forme différentiable en introduisant une variable vectorielle $t \in \mathbb{R}^m$: $\inf_{(x,t)\in\mathbb{R}^n\times\mathbb{R}^m} \{f(x) + \sum_{i=1}^m t_i : x \in X,\ -t_i \leqslant F_i(x) \leqslant t_i$ pour tout $i\}$, ce qu'il faut maintenant justifier.
On peut aussi écrire le problème sous la forme différentiable $\inf_{(x,u,v)\in\mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}^m} \{f(x) + \sum_{i=1}^m (u_i + v_i) : x \in X,\ F(x) = u - v,\ u \geqslant 0,\ v \geqslant 0\}$, ce qu'il faut maintenant justifier.
2) Le problème peut se récrire sous forme différentiable en introduisant une variable scalaire $t \in \mathbb{R}$: $\inf_{(x,t)\in\mathbb{R}^n\times\mathbb{R}} \{f(x) + t : x \in X,\ -t \leqslant F_i(x) \leqslant t$ pour tout $i\}$, ce qu'il faut maintenant justifier.

Ensembles convexes

2.1. 4) Il est instructif de considérer d'abord l'ensemble $\{x \in \mathbb{E} : \mathrm{d}_{\{p\}}(x) \leqslant \mathrm{d}_{\{q\}}(x)\}$, où $q \in Q$.

2.9. Pour le point 6, on pourra s'inspirer du contre-exemple de l'exercice 2.23 montrant que l'image par une application linéaire d'un cône convexe fermé n'est pas nécessairement fermée.

2.15. Utilisez les points 1 et 3 de la proposition 2.17.

2.13. 6) On pourra utiliser avec soin le théorème 2.38 de séparation et la notion de face exposée (exercice 2.11).

2.17. Les points extrêmes de la boule unité ℓ_1 ont toutes leurs composantes nulles sauf une valant ± 1 et ceux de la boule unité ℓ_∞ ont toutes leurs composantes valant ± 1.

2.18. L'ensemble $\mathrm{ext}(\mathbb{G}^n)$ est formé des matrices de permutation d'ordre n, celles dont les colonnes sont obtenues en permutant les colonnes de la matrice identité d'ordre n.

2.19. 3) On ne suppose pas que les x_i sont linéairement indépendants !

2.22. On pouura utiliser le lemme de Farkas pour montre qu'un élément de P_1 est dans la projetion de P sur $\mathbb{R}^{n_1}$.

2.28. On pourra utiliser la caractérisation de la projection sur un cône convexe fermé non vide donnée par l'exercice 2.26.

2.31. On se ramène à un problème de projection sur le convexe fermé non vide $C = P_1 - P_2$.

2.32. On pourra commencer par montrer que $\mathrm{aff}\, P = \mathrm{aff}(\mathrm{co}\, P)$ et que, si P est relativement ouvert, $P \subseteq (\mathrm{co}\, P)^{\bullet}$.

2.33. Pour démontrer (2.53) et (2.54), on pourra utiliser la proposition 2.40. L'égalité en 2.54 peut s'obtenir grâce au corollaire 2.4 du théorème de Carathéodory.

2.35. 5) On pourra commencer par montrer que $\mathcal{C}^{n+} = \mathrm{co}\{xx^{\mathsf{T}} : x \in \mathbb{R}^n_+\}$ et qu'une matrice $A \in \mathcal{C}^{n+}$ peut s'écrire $A = BB^{\mathsf{T}}$ avec une matrice $B \geqslant 0$ *dont le type est indépendant de* $A \in \mathcal{C}^{n+}$.

2.40. Exprimez l'incompatibilité du système sous une forme duale (lemme de Farkas), prendre un sommet du polyèdre convexe ainsi formé et revenir à la forme primale.

2.42. Le plus simple est de raisonner par l'absurde.

2.44. Formule de $\mathrm{T}_S(\mathcal{S}^n_+)$, inclusion $\supseteq$: on pourra commencer par considérer le cône $\{D \in \mathcal{S}^n : v^\mathsf{T} D v > 0, \text{ pour tout } v \in \mathcal{N}(S) \setminus \{0\}\}$.

Fonctions convexes

3.4. 2) La condition nécessaire se démontre aisément si f est sous-différentiable en $x := \int_\Omega F \,\mathrm{d}\mu \in \operatorname{dom} f$ car alors, il existe un sous-gradient $g \in \partial f(x)$ tel que, pour tout $z \in \Omega$, on a $f(F(z)) \geqslant f(x) + \langle g, F(z) - x\rangle$. Dans le cas contraire, on se ramène à ce premier cas en montrant que x est sur une partie convexe de la frontière relative du domaine de f, de dimension strictement plus petite que celle de $\operatorname{dom} f$, ce qui permet de poursuivre éventuellement par récurrence.

La condition suffisante se démontre en prenant des mesures de Dirac en certains points.

4) Utilisez la croissance et la concavité du logarithme.

5) On pourra considérer la fonction $a \in \mathbb{R}_{++} \mapsto \frac{1}{p} a^p$, qui est strictement convexe lorsque $p > 1$.

3.8. 3) Il s'écrit $\{d \in \mathcal{N}(H) : g^\mathsf{T} d \leqslant 0\}$, ce qu'il s'agit maintenant de démontrer.

3.11. On pourra commencer par montrer que la forte convexité de f est équivalente à la convexité de $x \mapsto f(x) - \frac{\alpha}{2}\|x\|^2$.

3.13. Pour la première partie, on pourra utiliser le point 6 de l'exercice 2.13. Pour la seconde partie, on pourra utiliser le point 3 de l'exercice 2.16.

3.14. 4) Montrez que $(f(x_*) - \mathbb{R}^m_{++}) \cap (f(X) + \mathbb{R}^m_+) = \varnothing$ et séparez ces deux ensembles convexes disjoints.

3.15. 2) On pourra utiliser l'exercice 2.21.

3) On pourra utiliser la représentation primale (3.4) d'une fonction convexe polyédrique.

3.17. 1) Commencez par évaluer $f^*(0)$.

2) Un contre-exemple est donné par $f : x \in \mathbb{R} \mapsto \sqrt{|x|}$.

3) On pourra raisonner par l'absurde, séparer un point de $\arg\min f^{**}$ et $\overline{\mathrm{co}}(\arg\min f)$ (point 3 du corollaire 2.37) et construire une direction non nulle dans $(\arg\min f^{**})^\infty$, ce qui contredit la bornitude de $\arg\min f^{**}$.

3.19. Calculez la biconjuguée de $\delta_x(\cdot) := f'(x;\cdot)$, en utilisant le fait que

$$\delta_x(d) = \inf_{t>0} \frac{1}{t}(f(x+td) - f(x)).$$

3.26. La condition nécessaire est immédiate en utilisant l'inégalité de Cauchy-Schwarz. Pour la condition suffisante, on pourra séparer avec soin les ensembles convexes $C_1 := \operatorname{epi} f$ et $C_2 := \{(y,\alpha) \in \mathbb{E} \times \mathbb{R} : \alpha < f(x) - L\|y - x\|\}$.

3.29. 1) On pourra utilisé le point 3 de l'exercice 3.28.

3.30. 3) Pour démontrer l'identité $\|\cdot\|_{\text{DD}} = \|\cdot\|$, il suffira de montrer que $\|\cdot\|_{\text{DD}}$ est la biconjuguée de $\|\cdot\|$ et d'appliquer le corollaire 3.47.

3.31. 2) On pourra commencer par montrer que $f^*(s) \leqslant \frac{1}{p'}\|s\|_{\text{D}}^{p'}$ en utilisant l'inégalité de Cauchy-Schwarz généralisée (A.9). Pour l'inégalité inverse, on choisira un argument particulier pour le supremum intervenant dans la définition de $f^*(s)$.

3) Le point 5 de l'exercice 3.4 pourra être utile.

3.32. On pourra commencer par montrer que, pour tout $x^* \in \mathbb{R}^n$, $f^*(x^*) \geqslant 0$ et $f^*(x^*) \leqslant (\|x^*\|_\infty - 1)^+$ (en utilisant l'inégalité de Hölder). On en déduira (3.86) puis (3.87).

3.34. On pourra s'inspirer de l'exercice 3.33, qui utilise des arguments semblables dans un contexte plus simple.

4.c) On pourra utiliser la formule du max.

3.35. On pourra s'inspirer de l'exercice 3.33, qui utilise des arguments semblables dans un contexte plus simple.

4.c) On pourra utiliser la formule du max.

3.36. 1) On pourra voir d_C comme une fonction marginale ou donner une démonstration directe.

2) On pourra commencer par montrer que $\frac{1}{t}\,\text{P}_C(y + tp) \to \text{P}_{C^\infty}(p)$, quand $t \to \infty$.

3.37. 1) Penser aux conditions assurant la convexité d'une composition.

2) Pour l'inégalité de droite, il faut faire apparaître h dans le facteur de gauche du produit scalaire, de manière à obtenir une majoration par $\|h\|^2$.

3) Faire un développement limité.

3.39. 2) On pourra démontrer $(ii) \Rightarrow (i)$ par l'absurde en montrant que si (i) n'a pas lieu, $\forall k \geqslant 1$, $\exists\, x_k \in \mathbb{E}$ tels que $\|x_k\|/k \to \infty$ et $f(x_k)/\|x_k\| \leqslant 1/k$. On considère alors la suite de points $y_k = x + (k/\|x_k\|)(x_k - x)$.

Pour l'équivalence $(i) \Leftrightarrow (iii)$, on pourra commencer par montrer que (i) est équivalent à $\exists\, \varepsilon > 0$, $\exists\, \alpha > 0$, $\forall x \in \mathbb{E}\,:\, f(x) \geqslant \varepsilon\|x\| - \alpha$. On en déduit une inégalité sur f^* qui, grâce à le lemme 3.13, permet de conclure.

3) Il pourra être utile de considérer l'application $x \in \mathbb{E} \mapsto f(x) - \langle x_0^*, x\rangle$, avec x_0^* voisin de zéro, et d'utiliser l'équivalence $(ii) \Leftrightarrow (iii)$.

3.40. On pourra raisonner par l'absurde.

3.41. L'équivalence $(i) \Rightarrow (ii)$ est claire.

L'implication $(ii) \Rightarrow (iii)$. Une minoration locale de f peut s'obtenir par une majoration locale de f^*. Pour avoir la constante $\alpha = 1/L$, on pourra utiliser un argument récursif. Il est utile d'examiner le cas quadratique strictement convexe.

L'implication $(iii) \Rightarrow (ii)$ s'obtient facilement avec $\alpha = 1/(2L)$.

Conditions d'optimalité

4.3. Pour le point 1, on pourra utiliser l'expression (2.46) $\text{T}_y\, a(C) = \overline{\mathbb{R}_{++}(a(C) - y)}$ du cône tangent. Pour le point 2, on pourra commencer par démontrer les identités

$$a^{-1}(C) - x = A^{-1}([C \cap \mathcal{R}(a)] - y) \quad \text{et} \quad A^{-1}((\mathbb{R}_{++}([C \cap \mathcal{R}(a)] - y))^\ominus) \neq \varnothing,$$

qui seront utiles dans le calcul de $\mathrm{T}_x(a^{-1}(C))$ par la formule $\overline{\mathbb{R}_{++}(a^{-1}(C)-x)}$.

4.6. 1, 2) On pourra utiliser la condition d'optimalité qui est nécessaire et suffisante lorsque $A \succcurlyeq 0$.

3) Distinguez les cas où la valeur optimale du problème quadratique vaut $-\infty$ ($A \not\succcurlyeq 0$ ou $b \notin \mathcal{R}(A)$) et celle où elle est finie ($A \succcurlyeq 0$ et $b \in \mathcal{R}(A)$). Dans ce dernier cas, on distinguera aussi le cas où $b = 0$ (valeur nulle) et $b \neq 0$ (on exprime le problème dans le membre de droite comme un problème d'optimisation avec une contrainte d'égalité).

4.7. En utilisant la condition d'optimalité du premier ordre, on peut montrer que v est soit le vecteur nul, soit un vecteur propre de M de valeur propre $\|v\|^2$. En utilisant la condition d'optimalité du second ordre, on peut montrer que v est un vecteur propre de M (donc non nul), associé à la valeur propre maximale de M.

4.11. 1) Dans le cas où $p = 1$, le problème est équivalent au suivant $\min\{c^\mathsf{T}x : \sum_{i=1}^n v_i = 1, -v \leqslant x \leqslant v\}$. On en déduit que ses solutions sont les $x \in \mathbb{R}^n$ tels que

$$x_{I^c} = 0, \quad \operatorname{sgn}(x_i) = -\operatorname{sgn}(c_i) \quad \text{et} \quad \|x\|_1 = 1.$$

2) Dans le cas où $1 < p < \infty$, la solution s'écrit

$$x_i = -\operatorname{sgn}(c_i)\left(\frac{|c_i|}{\|c\|_{p'}}\right)^{\frac{p'}{p}}.$$

3) Dans le cas où $p = \infty$, on a n problèmes indépendants.

4.14. On pourra s'aider des conditions d'optimalité.

4.19. Si (QC-MF) n'a pas lieu, on a l'identité ci-dessus avec $\lambda_0 = 0$; sinon on montre que l'on a les conditions d'optimalité (4.33), donc l'identité ci-dessus avec $\lambda_0 = 1$.

4.20. Procédez par l'absurde et construisez une direction critique non nulle comme dans la démonstration du théorème 4.49; en déduire une contradiction avec (4.63).

4.21. 2) On pourra utiliser la technique du développement du lagrangien $\ell(\cdot, \lambda_*)$ comme dans la démonstration du théorème 4.21.

4.23. Pour certaines questions, il pourra être utile de se ramener au problème (P_G) et à ses propriétés en écrivant (P_{QG}) comme suit

$$(\tilde{P}_{QG}) \quad \begin{cases} \min\ f(x) \\ \tilde{c}(x) \in \tilde{G}, \end{cases}$$

où $\tilde{c} : \mathbb{E} \to \mathbb{E} \times \mathbb{F}$ est définie en $x \in \mathbb{E}$ par

$$\tilde{c}(x) = (x, c(x)) \tag{D.1}$$

et $\tilde{G} := Q \times G$ est un convexe fermé non vide. En particulier pour le point 3, on pourra commencer par montrer que (4.142) est une condition équivalente à

$$0 \in \operatorname{int}\Big(\tilde{c}(x_0) + \tilde{c}'(x_0)\mathbb{E} - \tilde{G}\Big). \tag{D.2}$$

Prolégomènes à l'algorithmique

5.5. 1) $C(\{x_k\})$ est une intersection de demi-espaces fermés.

5.8. On pourra utiliser les formules établies à l'exercice 5.7.

Méthodes à directions de descente

6.3. On pourra s'inspirer de l'algorithme de Fletcher-Lemaréchal déterminant un pas de Wolfe (l'algorithme 6.6), ainsi que de la proposition 6.7 et de sa démonstration.

Premiers algorithmes

Optimisation quadratique

8.4. On note $\{u_k\} \subseteq \mathbb{R}^p$ et $\{d_k^r\} \subseteq \mathbb{R}^p$ les suites des itérés et des directions conjuguées générées par le GC réduit. De même, on note $\{x_k\} \subseteq \mathbb{R}^n$ et $\{d_k\} \subseteq \mathbb{R}^n$ les suites des itérés et des directions conjuguées générées par le GC préconditionné. On montre par récurrence que, pour tout $k \geqslant 1$, on a $Zd_{k-1}^r = d_{k-1}$ et $x_k = x_1 + Zu_k$.

Algorithme de Newton

10.1. On a $\varphi'(x) \cdot d = -\|\nabla f(x)\|$.

Problèmes non lisses

??. Montrez que x n'est pas un point isolé de X en utilisant le théorème 4.58.

Pénalisation

13.2. On a $\lim_{r\to+\infty} \lambda_{\max}(H_r)/r = \|A\|_2^2$ et $\lim_{r\to+\infty} \lambda_{\min}(H_r) = \sigma$.

Dualité

14.3. 1) Utilisez la définition de l'infimum et du supremum et l'inégalité de dualité faible (14.6).

14.4. Le dual de Fenchel s'obtient par perturbation de Ax dans (P_F). On peut donc retrouver le dual associé en récrivant le problème (P_F) en subtituant Ax par λ, en ajoutant la contrainte $Ax = \lambda$ et en faisant une dualisation lagrangienne de cette dernière.

14.5. 1) Le problème dual s'écrit $\sup_{\lambda\in\mathbb{R}} \delta(\lambda)$, où $\delta(\lambda) = \inf_{x\in X} \lambda x_1 + x_2$.

2) Les problèmes primal et dual ont une solution et ont la même valeur optimale (absence de saut de dualité). On utilise alors le théorème 14.3.

14.9. On pourra utiliser le théorème 9.2 pour montrer que le lagrangien a un point-selle.

14.10. 3) On peut commencer par montrer que, si $c(x) = Ax - b$, on peut trouver un $\mu \in \mathbb{R}_+^m$ tel que $A^\mathsf{T}\mu = 0$ et $b^\mathsf{T}\mu = -1$.

14.11. Écrivez (P) comme un « infsup », dualisez le problème par inversion de l'infimum et du supremum (section 14.1) et utilisez l'homogénéité de $\ell_h(\cdot, \bar{\lambda})$ pour obtenir la valeur du problème interne dans le dual.

14.13. L'expression du sous-différentiel $\partial\, \mathrm{d}_C(x)$ établie au point 4 de l'exercice 3.36 devrait être utile .

14.15. 1) La première identité peut se démontrer en utilisant $\lambda_{\max}(A^\mathsf{T} A) = [\sigma_1(A)]^2$ et la seconde en utilisant la décomposition en valeurs singulières de A. La dernière identité est plus difficile à démontrer ; on pourra suivre les étapes suivantes (on utilise la notation Matlab pour les matrices dans le texte) :

(*a*) $\|A\|_* \geqslant \|\sigma(A)\|_1$,
(*b*) $\|B\| \leqslant t$, si, et seulement si, $[tI_m \;\; B; B^\mathsf{T} \;\; tI_n] \succcurlyeq 0$,
(*c*) $\|A\|_*$ est le supremum en $B \in \mathbb{R}^{m\times n}$ de $\inf\{\langle A + 2Z, B\rangle + \langle X, I_m\rangle + \langle Y, I_n\rangle : X \in \mathbb{R}^{m\times m},\, Y \in \mathbb{R}^{n\times n},\, Z \in \mathbb{R}^{m\times n},\, [tI_m \;\; B; B^\mathsf{T} \;\; tI_n] \succcurlyeq 0\}$,
(*d*) $\|A\|_* \leqslant \inf\{\frac{1}{2}(\operatorname{tr} X + \operatorname{tr} Y) : X \in \mathbb{R}^{m\times m},\, Y \in \mathbb{R}^{n\times n},\, [X \;\; -A; -A^\mathsf{T} \;\; Y] \succcurlyeq 0\}$,
(*e*) $\|A\|_* \leqslant \|\sigma(A)\|_1$.

3) On montre que $\mathrm{rg}^* = \mathcal{I}_{\{0\}}$ en utilisant, en particulier, le fait que rg est bornée supérieurement.

4) On obtient une minoration de $f^*(A^*)$ en prenant la factorisation SVD de A^*. Pour la majoration de $f^*(A^*)$, on utilise l'inégalité de trace de von Neumann (B.38).

5) On pourra montrer successivement que $f^{**}(A) \leqslant \|A\|_*$ si $\|A\| \leqslant 1$, que $f^{**}(A) \geqslant \|A\|_*$ si $\|A\| \leqslant 1$, et que $f^{**}(A) = +\infty$ si $\|A\| > 1$.

14.17. On pourra utiliser la conjuguée de la norme (exercice 3.30).

Méthodes de points intérieurs

16.1. L'ellipsoïde $x + \mathscr{E}_A$ est dans $\mathbb{R}^n_+$ ssi[1] $\|A^{-1/2} e^i\|_2 \leqslant x_i$, pour tout $i = 1, \ldots, n$ (e^i désigne le i-ième vecteur de base de $\mathbb{R}^n$). Par ailleurs, le volume de $\mathscr{E}_A$ (ou de $x + \mathscr{E}_A$) est $\sigma_n \det A^{-1/2}$, où σ_n est le volume de $\bar{B}$ ($\bar{B}$ est la boule unité fermée de $\mathbb{R}^n$ pour la norme euclidienne).

Problèmes de moindres-carrés

19.1. Hormis pour le point 5, il n'est jamais nécessaire d'écrire les condtions d'optimalité des problèmes (P) et (P_ε), qui ne sont pas différentiables lorsque les normes ne sont pas associées à un produit scalaire. Les démonstrations sont courtes.

Optimisation linéaire : théorie et algorithme du simplexe

17.1. On ne peut pas prendre le système d'optimalité car celui-ci est non linéaire à cause des conditions de complémentarité, mais on peut remplacer ces dernières par une condition linéaire équivalente.

17.2. (*ii*) Montrez que $\inf\{c^\mathsf{T} d : d \in \mathrm{T}_{\bar{x}}\, X \cap S\} > 0$, où $\mathrm{T}_{\bar{x}}\, X$ est le cône tangent à X en $\bar{x}$ et S est la sphère unité de $\mathbb{R}^n$.

[1] Dans ce chapitre de solutions, nous utiliserons « ssi » comme abréviation de « si, et seulement si, ».

17.5. Pour l'implication $(i) \Rightarrow (ii)$, on peut traiter chaque ligne $b_i^\mathsf{T} \in \mathbb{R}^{1\times n}$, $i \in [1\,{:}\,m_B]$, de $B \in \mathbb{R}^{m_B\times n}$ séparément, en considérant le problème d'optimisation linéaire en $x \in \mathbb{R}^n$ suivant

$$\begin{cases} \inf_x \ -b_i^\mathsf{T} x \\ Ax \leqslant a. \end{cases}$$

17.7. 1) On pourra utiliser le point 5 de l'exercice 2.20.

17.8. 1) Considérez le problème d'optimisation linéaire $\inf\{0^\mathsf{T} x : Ax = b,\ x \geqslant 0\}$ et son dual.

2) Considérez le problème d'optimisation linéaire $\inf\{c^\mathsf{T} x - \alpha : Ax \geqslant b\}$ et son dual.

17.9. 1) Considérez le problème d'optimisation linéaire

$$\begin{cases} \inf_{(\alpha,\beta,\gamma)} a^\mathsf{T}\alpha + b^\mathsf{T}\beta + c^\mathsf{T}\gamma \\ A^\mathsf{T}\alpha + B^\mathsf{T}\beta + C^\mathsf{T}\gamma = 0 \\ \beta \geqslant 0 \\ \gamma \geqslant 0 \end{cases} \tag{D.3}$$

et son dual. Ensuite utilisez les résultats de dualité faible (proposition 17.10), de dualité forte (théorème 17.11) et les résultats d'existence de solution (théorème 17.4 et proposition 17.8).

17.12. Il existe beaucoup de démonstrations de ce résultat. Celle proposée ici se fonde sur la dualité forte en optimisation linéaire. On note α la valeur à gauche dans (17.26), β la valeur à droite et e le vecteur dont toutes les composantes valent 1 (il peut varier en dimension, suivant le contexte).

1) Justifiez l'utilisation des opérateurs « min » et « max » dans (17.26) (au lieu de « inf » et « sup », voir la remarque 1.1).
2) Montrez que $\min\{v^\mathsf{T} x : x \in \Delta_n\} = \min\{v_i : 1 \leqslant i \leqslant n\}$ et que $\max\{v^\mathsf{T} x : x \in \Delta_n\} = \max\{v_i : 1 \leqslant i \leqslant n\}$.
3) En ajoutant au besoin une même quantité à chaque élément de A, montrez que l'on peut supposer que $\alpha > 0$ et $\beta > 0$.
4) Montrez que $\alpha^{-1} = \min\{e^\mathsf{T} y : y \geqslant 0,\ A^\mathsf{T} y \geqslant e\}$ et que $\beta^{-1} = \max\{e^\mathsf{T} x : x \geqslant 0,\ Ax \leqslant e\}$.
5) Conclure.

17.14. [(i)] Fixez $x = (1,0) \in \mathbb{R}^2$, $b = 0$ et déterminez des $A_\varepsilon \in \mathbb{R}^{m\times n}$ pour $\varepsilon \downarrow 0$ tels que $P_\varepsilon(x) = 0$ (donc $\mathrm{d}_{P_\varepsilon}(x) = 1$) et $\|(A_\varepsilon x)^+\|_2 \downarrow 0$.

[(ii)] (a) On peut montrer l'inégalité $\mathrm{d}_P(x) \geqslant \sup\{(Ax-b)^\mathsf{T} y : y \geqslant 0,\ \|A^\mathsf{T} y\|_\mathrm{D} \leqslant 1\}$ par dualité faible. On peut montrer l'inégalité $\mathrm{d}_P(x) \leqslant \sup\{(Ax-b)^\mathsf{T} y : y \geqslant 0,\ A^\mathsf{T} y = u\}$, pour un certain $u \in \mathbb{R}^n$ tel que $\|u\|_\mathrm{D} = 1$ en séparant les ensembles convexes non vides disjoints $\{v \in \mathbb{R}^n : \|v\| < \mathrm{d}_P(x)\}$ et $\{v \in \mathbb{R}^n : Av \geqslant Ax - b\}$. L'affirmation (17.28) s'obtient alors en combinant ces deux inégalités.

(b) Choisir la solution sur un sommet de l'ensemble admissible.

Optimisation linéaire : algorithmes de points intérieurs

18.2. On pourra passer par un problème d'optimisation comme à la proposition 18.1.

Optimisation semi-définie

20.1. On peut montrer les implications dans le sens naturel $(i) \Rightarrow (ii) \Rightarrow (iii) \Rightarrow (iv) \Rightarrow (i)$. Pour les implications $(i) \Rightarrow (ii)$ et $(iv) \Rightarrow (i)$, on pourra utiliser l'exercice B.13.

1) On pourra utiliser l'expression de F donnée par le point (iii) et prendre $S = \sum_i w_i w_i^{\mathsf{T}}$, où $\{w_i\}$ est une base de $E^{\perp}$.

2) On pourra utiliser la définition de $F(A)$, le point (iii) et le point 2 de l'exercice B.13.

20.4. Il suffit d'imiter la démonstration du point 1 de la proposition 20.4. Celle-ci est plus simple d'ailleurs car elle ne doit pas faire intervenir la surjectivité de $\mathcal{A}$.

20.5. Faire une factorisation gaussienne par bloc de la matrice d'ordre $2n$.

20.8. En utilisant l'exercice 20.7, on peut montrer que le dual lagrangien de (20.26) s'écrit

$$\begin{cases} \sup_{(\lambda,t)\in\Lambda\times\mathbb{R}} -t \\ \begin{pmatrix} A(\lambda) & b(\lambda) \\ b(\lambda)^{\mathsf{T}} & c(\lambda)+t \end{pmatrix} \succcurlyeq 0. \end{cases} \tag{D.4}$$

20.9. Il suffit de montrer que l'application $(X, Y) \mapsto X \otimes Y$ est bilinéaire et que les opérateurs $E^s_{ij} \otimes E^s_{kl}$, avec $i \leqslant j$ et $k \leqslant l$, forment une base de $L(\mathcal{S}^n, \mathcal{S}^n)$. On a noté $E^s_{ij} \in \mathcal{S}^n$ la matrice dont les éléments (i, j) et (j, i) valent 1 et les autres valent zéro (ces matrices E^s_{ij}, avec $i \leqslant j$, forment une base de $\mathcal{S}^n$).

Espaces topologiques

A.11. Il suffit de montrer que les trois propriétés que doit vérifier une norme (voir la page 751) sont effectivement vérifiées par la norme duale (A.8). Pour l'inégalité triangulaire, l'exercice 1.2, point (ii), pourra être utile.

Algèbre linéaire

B.1. 1) On pourra utiliser le théorème B.1 de la base incomplète.

B.3. Pour l'implication "$\Rightarrow$", on pourra définir un inverse à droite linéaire continue de A. L'implication "$\Leftarrow$" est évidente.

B.4. On peut considérer le déterminant d'une sous-matrice carrée dont l'ordre est égal au rang.

B.8. On pourra utiliser le principe du min-max de Courant-Fisher (théorème B.4).

B.15. La difficulté principale est l'implication $(i) \Rightarrow (ii)$. On pourra raisonner par l'absurde en adaptant avec soin la démonstration du lemme de Finsler.

Calcul différentiel

C.7. Il suffit d'utiliser l'expression intégrale de l'écart $f(y) - f(x)$.

C.9. 2) Montrez que $\|\nabla d_P(x)\| \leqslant 1$ et que $\langle \nabla d_P(x), \bar{x} - x \rangle = -\|\bar{x} - x\|$.

D.2 Solutions

Introduction

1.1. 1) Quel que soit $x_0 \in X_0$, on a $\inf_{x\in X} f(x) \leqslant f(x_0)$, puisque $x_0 \in X$. Alors $\inf_{x\in X} f(x) \leqslant \inf_{x_0\in X_0} f(x_0)$.

2) Comme $X \subseteq \bar{X}$, $\inf_{x\in X} f(x) \geqslant \inf\{f(x) : x \in \bar{X}\} =: \bar{f}_{\min}$, par le point 1. Montrons l'inégalité inverse. On peut trouver une suite minimisante $\{x_k\} \subseteq \bar{X}$, qui vérifie donc $f(x_k) \to \bar{f}_{\min}$. Comme tout voisinage de x_k rencontre X et que f est semi-continue supérieurement sur $\bar{X}$, pour tout $k \geqslant 1$, on peut trouver $x'_k \in X$ tel que $f(x_k) \geqslant f(x'_k) - \frac{1}{k} \geqslant \bar{f}_{\min} - \frac{1}{k}$. Forcément, $f(x'_k) \to \bar{f}_{\min}$ et donc $\inf_{x\in X} f(x) \leqslant \bar{f}_{\min}$.

1.2. (i) Soient $\epsilon > 0$ et x_ϵ tel que $f(x_\epsilon) + g(x_\epsilon) \leqslant \inf(f+g) + \epsilon$. Comme $\inf f \leqslant f(x_\epsilon)$ et $\inf g \leqslant g(x_\epsilon)$, on a $(\inf f) + (\inf g) \leqslant \inf(f+g) + \epsilon$. Comme ϵ est arbitraire, on en déduit que $(\inf f) + (\inf g) \leqslant \inf(f+g)$.

(ii) On a $(\sup f) + (\sup g) = -[\inf(-f) + \inf(-g)] \geqslant -\inf(-f-g) = \sup(f+g)$.

1.3. $[(i) \Rightarrow (ii)]$ On démontre la contraposée en supposant qu'il existe une constante $\nu \in \mathbb{R}$ telle que $\{x \in X : f(x) \leqslant \nu\}$ ne soit pas borné. Alors on peut trouver une suite $\{x_k\} \subseteq X$ telle que $\|x_k\| \to +\infty$ et $f(x_k) \leqslant \nu$, si bien que (ii) n'a pas lieu.

$[(ii) \Rightarrow (i)]$ On démontre également la contraposée en supposant qu'il existe une suite $\{x_k\} \subseteq X$ et une constante $\nu \in \mathbb{R}$ telles que $\|x_k\| \to \infty$ et $f(x_k) \leqslant \nu$. Alors $\{x_k\} \subseteq \{x \in X : f(x) \leqslant \nu\}$ qui n'est pas borné.

1.4. Il est clair que a est coercive si elle vérifie (1.14). Réciproquement, par (1.5), si a est coercive, il existe $\alpha > 0$ tel que, pour tout x vérifiant $\|x\| \geqslant 1/\alpha^{1/2}$, on a $a(x,x) \geqslant 1$. Alors, quel que soit $x \neq 0$, on a $a(x/(\alpha^{1/2}\|x\|), x/(\alpha^{1/2}\|x\|)) \geqslant 1$ ou $a(x,x) \geqslant \alpha\|x\|^2$.

1.5. Comme une mesure de Dirac en $x \in X$, $\mu = \delta_x$, vérifie les contraintes du problème à droite dans (1.15), la valeur optimale de ce dernier est plus petite que $\inf_{x\in X} \int_X f \,\mathrm{d}\delta_x = \inf_{x\in X} f(x)$. Réciproquement, comme f prend des valeurs supérieures à sa valeur minimale $\inf_{x\in X} f(x)$, la valeur optimale du problème de droite dans (1.15) est plus grande que $[\inf_{x\in X} f(x)]\,\mu(X) = \inf_{x\in X} f(x)$.

1.6. Montrons d'abord (1.16a). Par l'inversibilité de A et la remarque 1.7, on a

$$\alpha := \min_{x\,:\,\|x\|=1} \|Ax\| = \min_{y\,:\,\|A^{-1}y\|=1} \|y\|. \tag{D.5a}$$

Observons que $\alpha > 0$ (car le minimum dans le membre de gauche de (D.5a) est atteint et A est inversible). Alors, $\|A^{-1}y\| \leqslant 1$ pour tout y vérifiant $\|y\| = \alpha$ (sinon $\tilde{y} := y/\|A^{-1}y\|$ vérifierait $\|A^{-1}\tilde{y}\| = 1$ et $\|\tilde{y}\| = \|y\|/\|A^{-1}y\| < \alpha$, ce qui contradirait la définition de α). De plus,

$$\max_{\|y\|=\alpha} \|A^{-1}y\| = 1,$$

car toute solution y du problème à droite dans (D.5a) est telle que $\|y\| = \alpha$ et $\|A^{-1}y\| = 1$. L'identité (1.16a) se déduit maintenant par

$$1 = \max_{\|y\|=\alpha} \|A^{-1}y\| = \alpha \max_{\|y\|=1} \|A^{-1}y\| = \alpha\,\|A^{-1}\|. \tag{D.5b}$$

Démontrons maintenant la seconde partie de l'exercice. Soit x_* une solution du problème dans le membre de gauche de (1.16a) et posons $y_* := Ax_*/\alpha$, où α est la valeur optimal du problème. Alors, $\|y_*\| = 1$ et $\|A^{-1}y_*\| = \|x_*\|/\alpha = 1/\alpha = \|A^{-1}\|$

par (D.5b). On a montré que y_* est solution du problème dans le membre de gauche de (1.16b).

Réciproquement, supposons que y_* soit solution du problème dans le membre de gauche de (1.16b) et posons $x_* := \|A^{-1}\|^{-1}A^{-1}y_*$. Alors, $\|x_*\| = 1$ et $\|Ax_*\| = \|A^{-1}\|^{-1}$, si bien que x_* résout le problème du membre de gauche dans (1.16a).

1.7. On applique la proposition 1.13 à la fonction $f = \sup_{i\in I} f_i$, en tenant compte du fait qu'alors, $f(x) \leqslant \alpha$ ssi $f_i(x) \leqslant \alpha$ pour tout $i \in I$.

1.8. 1) Soient

$$(P_1) \qquad \inf_{x\in\mathbb{R}^n} \{f(x) + \|F(x)\|_1 : x \in X\},$$

$$(P_1') \qquad \inf_{(x,t)\in\mathbb{R}^n\times\mathbb{R}^n} \left\{ f(x) + \sum_{i=1}^n t_i : x \in X,\ -t_i \leqslant F_i(x) \leqslant t_i \text{ pour tout } i \right\}.$$

D'après la proposition 1.13, le problème (P_1) est équivalent au problème

$$(P_1'') \qquad \inf_{(x,t)\in\mathbb{R}^n\times\mathbb{R}^m} \left\{ f(x) + \sum_i t_i : x \in X,\ |F_i(x)| \leqslant t_i \text{ pour tout } i \right\}.$$

Quant à l'équivalence entre (P_1'') et (P_1'), elle se déduit du fait que $|F_i(x)| \leqslant t_i$ ssi $-t_i \leqslant F_i(x) \leqslant t_i$.

La seconde expression

$$\inf_{(x,u,v)\in\mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}^m} \left\{ f(x) + \sum_{i=1}^m (u_i + v_i) : x \in X,\ F(x) = u - v,\ u \geqslant 0,\ v \geqslant 0 \right\}$$

s'obtient en écrivant $F(x) = u - v$, avec $u \geqslant 0$ et $v \geqslant 0$ et en notant que $\|F(x)\|_1 = \sum_i |F_i(x)| = \sum_i (u_i + v_i)$.

2) Soient

$$(P_2) \qquad \inf_{x\in\mathbb{R}^n} \{f(x) + \|F(x)\|_\infty : x \in X\},$$

$$(P_2') \qquad \inf_{(x,t)\in\mathbb{R}^n\times\mathbb{R}} \{f(x) + t : x \in X,\ -t \leqslant F_i(x) \leqslant t \text{ pour tout } i\}.$$

D'après la proposition 1.13, le problème (P_2) est équivalent au problème

$$(P_2'') \qquad \inf_{(x,t)\in\mathbb{R}^n\times\mathbb{R}} \{f(x) + t : x \in X,\ \|F(x)\|_\infty \leqslant t\}.$$

Quant à l'équivalence entre (P_2'') et (P_2'), elle se déduit du fait que $\|F(x)\|_\infty \leqslant t$ ssi $-t \leqslant F_i(x) \leqslant t$ pour tout i.

Ensembles convexes

2.1. 1) Soient $\lambda_1, \lambda_2 \in \Lambda$, $x_1, x_2 \in C$ et $t \in]0,1[$. Il s'agit de montrer que $z := (1-t)\lambda_1 x_1 + t\lambda_2 x_2 \in \Lambda C$. Si $\lambda_1 = \lambda_2 = 0$, alors $0 \in \Lambda$ et $z = 0 \in \Lambda C$. Sinon, $\lambda := (1-t)\lambda_1 + t\lambda_2 > 0$ et on a $z = \lambda x$, avec $\lambda \in \Lambda$ et $x = ((1-t)\lambda_1/\lambda)x_1 + (t\lambda_2/\lambda)x_2 \in C$. Donc $z \in \Lambda C$.

2) Quel que soit l'ensemble C, on a $2C \subseteq C + C$. Car si $x \in C$, $2x = x + x \in C + C$. Si C est convexe et x et $y \in C$, on a $\frac{1}{2}x + \frac{1}{2}y \in C$ et donc $x + y = 2(\frac{1}{2}x + \frac{1}{2}y) \in 2C$.

3) Si $K \subseteq \mathbb{E}$ est un cône convexe, alors $K+K = 2K$ [par le point 2] $= K$ (car K est un cône). Inversement, supposons que K soit un cône tel que $K+K=K$. Soient x et $y \in K$ et $t \in]0,1[$. Alors $(1-t)x+ty \in K+K$ [car $1-t>0$ et $x \in K$ impliquent que $(1-t)x \in K$] $= K$ [par hypothèse].

4) On peut écrire $V_{p,Q} := \{x \in \mathbb{E} : \mathrm{d}_{\{p\}}(x) \leqslant \mathrm{d}_{\{q\}}(x) \text{ pour tout } q \in Q\} = \cap_{q\in Q}\{x \in \mathbb{E} : \|x-p\| \leqslant \|x-q\|\}$, où $\|\cdot\|$ est la norme associée au produit scalaire de $\mathbb{E}$. En élevant les normes au carré et en simplifiant, on obtient $V_{p,Q} = \cap_{q\in Q}\{x \in \mathbb{E} : \langle q-p, x\rangle \leqslant (\|q\|^2 - \|p\|^2)/2\}$, si bien que $V_{p,Q}$ est un polyèdre convexe (une intersection d'un nombre fini de demi-espaces fermés).

Le *diagramme de Voronoï* d'un ensemble fini de points P est l'ensemble des $V_{p,P\setminus\{p\}}$ lorsque p parcourt P.

2.2. 1) Car les sous-espaces affines qui contiennent Q contiennent P.

2) Car $P \cap Q \subseteq P$ implique que $\mathrm{aff}(P\cap Q) \subseteq \mathrm{aff}\, P$ [point 1] ; de même $\mathrm{aff}(P\cap Q) \subseteq \mathrm{aff}\, Q$. On n'a pas nécessairement l'égalité, même si P et Q sont convexes, comme dans le cas où $P = \bar{B}+e^1$ et $Q = \mathbb{R}e^2$ dans $\mathbb{R}^2$ (on a noté $\bar{B} := \{x \in \mathbb{R}^2 : \|x\|_2 \leqslant 1\}$).

3) La relation $\sum_i t_i(x_i+y_i) = \sum_i t_i x_i + \sum_i t_i y_i$ montre que $\mathrm{aff}(P+Q) \subseteq (\mathrm{aff}\, P)+(\mathrm{aff}\, Q)$. Inversement, si $\sum_i t_i = 1$ et $\sum_j s_j = 1$, on a $\sum_i t_i x_i + \sum_j s_j y_j = \sum_{ij} t_i s_j (x_i+y_j)$ avec $\sum_{ij} t_i s_j = 1$, dès lors $(\mathrm{aff}\, P)+(\mathrm{aff}\, Q) \subseteq \mathrm{aff}(P+Q)$.

4) Clairement $\mathrm{aff}\, P \subseteq \mathrm{aff}\, \overline{P}$. On a aussi $\overline{P} \subseteq \mathrm{aff}\, P$ ($\mathrm{aff}\, P$ est un fermé qui contient P), donc $\mathrm{aff}\, \overline{P} \subseteq \mathrm{aff}\,\mathrm{aff}\, P = \mathrm{aff}\, P$ ($\mathrm{aff}\, P$ est affine).

5) On a toujours $\mathrm{aff}\, P \subseteq \mathrm{vect}\, P$ (comparer les définitions). Inversement, un élément de $\mathrm{vect}\, P$ est de la forme $\sum_i t_i x_i = (1-\sum_i t_i)0 + \sum_i t_i x_i$ avec des $t_i \in \mathbb{R}$; comme $0 \in \overline{P}$ et $(1-\sum_i t_i)+\sum_i t_i = 1$, la dernière expression de l'élément montre qu'il est dans $\mathrm{aff}\, \overline{P} = \mathrm{aff}\, P$ (point 4).

Par ailleurs les relations $\mathrm{aff}(\mathbb{R}_+P) \subseteq \mathrm{vect}\, P$ [voir les définitions] $= \mathrm{aff}\, P$ [on vient de le démontrer] $\subseteq \mathrm{aff}(\mathbb{R}_+P)$ permettent d'affirmer que $\mathrm{aff}\, P = \mathrm{aff}(\mathbb{R}_+P)$. Puis $P \subseteq \mathbb{R}_{++}P \subseteq \mathbb{R}_+P$ implique que $\mathrm{aff}\, P \subseteq \mathrm{aff}(\mathbb{R}_{++}P) \subseteq \mathrm{aff}(\mathbb{R}_+P) = \mathrm{aff}\, P$, donc $\mathrm{aff}\, P = \mathrm{aff}(\mathbb{R}_{++}P)$.

6) On a successivement $\mathrm{aff}(\mathrm{cone}\, P) = \mathrm{aff}(\overline{\mathrm{cone}\, P})$ [point 4] $= \mathrm{vect}(\overline{\mathrm{cone}\, P})$ [$0 \in \overline{\mathrm{cone}\, P}$ et point 5] $= \mathrm{vect}(\mathrm{cone}\, P)$ [raisonner comme au point 4] $= \mathrm{vect}\, P$.

7) D'après la proposition 2.1, on a clairement $\mathrm{aff}\, C \supseteq Z := \{tx+(1-t)y : t \in \mathbb{R},\ x,y \in C\}$. Inversement, d'après la même proposition, un élément $z \in \mathrm{aff}\, C$ est de la forme $z = \sum_{i=1}^m t_i x_i$ avec $m \in \mathbb{N}^*$, des $x_i \in C$ et des $t_i \in \mathbb{R}$ vérifiant $\sum_{i=1}^m t_i = 1$. Si tous les t_i sont positifs, $z \in C$ (proposition 2.2), donc $z \in Z$. Sinon, soient $I := \{i : t_i \geqslant 0\}$, $I^c := \{i : t_i < 0\}$ (non vide donc) et $\alpha := \sum_{i\in I} t_i = 1 + \sum_{i\in I^c} |t_i| > 1$. On a $\sum_{i=1}^m t_i x_i = \sum_{i\in I}^m t_i x_i - \sum_{i\in I^c}^m |t_i| x_i = \alpha[\sum_{i\in I}^m (t_i/\alpha)x_i] + (1-\alpha)[\sum_{i\in I^c}^m (|t_i|/(\alpha-1))x_i]$, qui est de la forme $\alpha x + (1-\alpha)y$ avec $x,y \in C$. Donc $z \in Z$.

8) [$\subseteq$] Observons d'abord que $(\mathrm{aff}\, P)\times(\mathrm{aff}\, Q)$ est un sous-espace affine de $\mathbb{E}\times\mathbb{F}$. En effet, on peut écrire $\mathrm{aff}\, P = x_0 + \mathbb{E}_0$ où $x_0 \in \mathbb{E}$ et $\mathbb{E}_0$ est un sous-espace vectoriel de $\mathbb{E}$. De même $\mathrm{aff}\, Q = y_0 + \mathbb{F}_0$ où $y_0 \in \mathbb{F}$ et $\mathbb{F}_0$ est un sous-espace vectoriel de $\mathbb{F}$. Alors $(\mathrm{aff}\, P)\times(\mathrm{aff}\, Q) = (x_0+\mathbb{E}_o)\times(y_0+\mathbb{F}_0) = (x_0,y_0)+\mathbb{E}_0\times\mathbb{F}_0$ qui est bien un sous-espace affine de $\mathbb{E}\times\mathbb{F}$. Par conséquent, $\mathrm{aff}(P\times Q) \subseteq (\mathrm{aff}\, P)\times(\mathrm{aff}\, Q)$, car $\mathrm{aff}(P\times Q)$ est le plus petit sous-espace affine contenant $P\times Q$ et $(\mathrm{aff}\, P)\times(\mathrm{aff}\, Q)$ est un sous-espace affine contenant $P\times Q$.

[$\supseteq$] Soit $x := \sum_{i=1}^p \alpha_i x_i$ et $y := \sum_{j=1}^q \beta_j y_j$ des éléments de $(\mathrm{aff}\, P)$ et $(\mathrm{aff}\, Q)$ respectivement (donc avec $\sum_{i=1}^p \alpha_i = 1$ et $\sum_{j=1}^q \beta_j = 1$). Si $\alpha_1 \leqslant \beta_1$ (on s'y prend de manière analogue si $\alpha_1 \geqslant \beta_1$), on a

$$\left(\sum_{i=1}^{p}\alpha_i x_i, \sum_{j=1}^{q}\beta_j y_j\right) = \alpha_1(x_1,y_1) + \left(\sum_{i=2}^{p}\alpha_i x_i, (\beta_1-\alpha_1)y_1 + \sum_{j=2}^{q}\beta_j y_j\right).$$

Le second terme du membre de droite est de la même forme que le membre de gauche, mais avec un terme en moins et avec des coefficients dont la somme vaut $1-\alpha_1$. Si on applique la même transformation à ce second terme, et ainsi de suite récurcivement, on finit par écrire (x,y) comme une somme pondérée d'éléments de $P\times Q$, avec des pondérations dont la somme vaut un. Ceci montre que $(x,y)\in \mathrm{aff}(P\times Q)$.

2.3. Soit $\overline{v} := \max_i v_i = v_j$. Alors $\alpha_i v_i \leqslant \alpha_i \overline{v}$ pour tout $i\in[1:p]$ tel que $\alpha_i\neq 0$ (ou $\alpha_i>0$ car $\alpha_i\geqslant 0$). En sommant, on obtient $\sum_{i:\alpha_i\neq 0}\alpha_i v_i \leqslant (\sum_{i:\alpha_i\neq 0}\alpha_i)\overline{v} = \overline{v}$ (car $\sum_i \alpha_{i:\alpha_i\neq 0}=1$). On a l'égalité en prenant $\alpha\in\Delta_p$ vérifiant $\alpha_j=1$ et $\alpha_i=0$ si $i\neq j$.

2.4. 1) Avec des notations standards, on peut écrire $a(x)=Ax+b$. L'identité énoncée résulte alors des égalités suivantes (avec $\sum_i\alpha_i=1$) : $\sum_i\alpha_i(a(x_i)) = \sum_i\alpha_i(Ax_i+b) = A(\sum_i\alpha_i x_i)+b = a(\sum_i\alpha_i x_i)$.

2) L'identité $\sum_i\alpha_i(x_i+y_i)=\sum_i\alpha_i x_i+\sum_i\alpha_i y_i$ montre que $\mathrm{co}(P+Q)\subseteq \mathrm{co}\,P+\mathrm{co}\,Q$. Inversement, si $\sum_i\alpha_i=1$ et $\sum_j\beta_j=1$, on a

$$\begin{array}{rcl}\sum_i\alpha_i x_i+\sum_j\beta_j y_j &=& \sum_j\beta_j\sum_i\alpha_i x_i+\sum_i\alpha_i\sum_j\beta_j y_j\\ &=& \sum_{ij}\alpha_i\beta_j x_i+\sum_{ij}\alpha_i\beta_j y_j\\ &=& \sum_{ij}\alpha_i\beta_j(x_i+y_j).\end{array}$$

Comme $\sum_{ij}\alpha_i\beta_j=1$, on en déduit que $\mathrm{co}\,P+\mathrm{co}\,Q\subseteq\mathrm{co}(P+Q)$.

3) L'identité $\sum_i\alpha_i(x_i,y_i)=(\sum_i\alpha_i x_i,\sum_i\alpha_i y_i)$ montre que $\mathrm{co}(P\times Q)\subseteq(\mathrm{co}\,P)\times(\mathrm{co}\,Q)$. Inversement, si $\sum_i\alpha_i=1$ et $\sum_j\beta_j=1$, on a

$$\begin{array}{rcl}(\sum_i\alpha_i x_i,\sum_j\beta_j y_j) &=& (\sum_j\beta_j\sum_i\alpha_i x_i,\sum_i\alpha_i\sum_j\beta_j y_j)\\ &=& (\sum_{ij}\alpha_i\beta_j x_i,\sum_{ij}\alpha_i\beta_j y_j)\\ &=& \sum_{ij}\alpha_i\beta_j(x_i,y_j).\end{array}$$

Comme $\sum_{ij}\alpha_i\beta_j=1$, on a $(\mathrm{co}\,P)\times(\mathrm{co}\,Q)\subseteq\mathrm{co}(P\times Q)$.

2.6. 1) $\mathrm{co}\{x_{i_1}+\cdots+x_{i_k} : 1\leqslant i_1\leqslant\cdots\leqslant i_k\leqslant n\}\subseteq \mathrm{co}\,X+\cdots+\mathrm{co}\,X$ (k fois) $= k\,(\mathrm{co}\,X)$, car $\mathrm{co}\,X$ est convexe. Réciproquement, un élément de $k\,(\mathrm{co}\,X)$ est de la forme $k\sum_i\alpha_i x_i$ avec des $\alpha_i\geqslant 0$ vérifiant $\sum_i\alpha_i=1$. Cet élément s'écrit aussi $\sum_i\alpha_i(x_i+\cdots+x_i)$ (k fois) et est donc un élément de $\mathrm{co}\{x_{i_1}+\cdots+x_{i_k} : 1\leqslant i_1\leqslant\cdots\leqslant i_k\leqslant n\}$.

2) L'inclusion « $\subseteq$ » est claire. Si x est dans l'ensemble de gauche, il s'écrit $x=\sum_{j=1}^{l}\alpha_j(\varepsilon_{1j}x_1+\cdots+\varepsilon_{nj}x_n)$, avec des $\alpha_j\geqslant 0$ vérifiant $\sum_{j=1}^{l}\alpha_j=1$ et des $\varepsilon_{ij}\in\{0,1\}$ vérifiant $\sum_{i=1}^{n}\varepsilon_{ij}=k$. Alors $x=\sum_{i=1}^{n}t_i x_i$ avec des $t_i=\sum_{j=1}^{l}\alpha_j\varepsilon_{ij}$. Clairement, les $t_i\in[0,1]$ (c'est une combinaison convexe de 0 et de 1) et $\sum_{i=1}^{n}t_i=\sum_{j=1}^{l}\sum_{i=1}^{n}\alpha_j\varepsilon_{ij}=k\sum_{j=1}^{l}\alpha_j=k$.

Nous montrons l'inclusion réciproque « $\supseteq$ » par récurrence sur n qui est pris $\geqslant k\geqslant 1$. Elle est vérifiée si $n=k$, car alors les deux ensembles sont réduits au singleton $\{x_1+\cdots+x_n\}$. Montrons l'inclusion lorsque $n>k$, sachant qu'elle est vérifiée pour $n-1$. Un élément de l'ensemble de droite est donc de la forme $x=\sum_{i=1}^{n}t_i x_i$ avec des $t_i\in[0,1]$ vérifiant $\sum_{i=1}^{n}t_i=k$. Sans perte de généralité, on peut supposer que $0<t_n\leqslant t_i$ pour tout $i=1,\ldots,n$ (si $t_n=0$ le résultat se déduit de l'hypothèse de récurrence). On note $\mathcal{C}_k^P$ l'ensemble des parties formées de k éléments distincts d'un ensemble P. On peut alors écrire

$$x = \sum_{i=1}^{n-1} t'_i x_i + \sum_{\{j_1,\dots,j_{k-1}\}\in\mathcal{C}_{k-1}^{[1\,:\,n-1]}} \alpha_{j_1\dots j_{k-1}}(x_{j_1}+\cdots+x_{j_{k-1}}+x_n), \tag{D.6}$$

pourvu que les t'_i vérifient

$$t'_i = t_i - \sum_{\{j_1,\dots,j_{k-2}\}\in\mathcal{C}_{k-2}^{[1\,:\,n-1]\setminus\{i\}}} \alpha_{ij_1\dots j_{k-2}}$$

et que les $\alpha_{j_1\dots j_{k-1}}$ soient choisis tels que (leur positivité sera utile tout de suite après)

$$\alpha_{j_1\dots j_{k-1}} \geqslant 0 \quad \text{et} \quad \sum_{\{j_1,\dots,j_{k-1}\}\in\mathcal{C}_{k-1}^{[1\,:\,n-1]}} \alpha_{j_1\dots j_{k-1}} = t_n. \tag{D.7}$$

Observons que la minimalité de t_n parmi les t_i et (D.7) impliquent que $t'_i \geqslant t_i - t_n \geqslant 0$. Par ailleurs

$$t' := \sum_{i=1}^{n-1} t'_i = \sum_{i=1}^{n-1} t_i - \sum_{i=1}^{n-1} \sum_{\{j_1,\dots,j_{k-2}\}\in\mathcal{C}_{k-2}^{[1\,:\,n-1]\setminus\{i\}}} \alpha_{ij_1\dots j_{k-2}} = k(1-t_n),$$

car la première somme vaut $k - t_n$ et la seconde vaut $(k-1)t_n$ puisque chaque ensemble d'indices $\{l_1,\dots,l_{k-1}\} \in \mathcal{C}_{k-1}^{[1\,:\,n-1]}$ y est apparaît $k-1$ fois (une fois lorsque $i = l_1$, une fois lorsque $i = l_2$, *etc*). Observons que l'on a $t' > 0$, car dans le cas contraire, $t' = 0$, tous les t_i vaudraient 1 et donc on aurait $n = k$, ce que l'on a exclu dans le cas présent. On récrit alors (D.6) comme suit

$$x = \frac{t'}{k}\left(\sum_{i=1}^{n-1} t''_i x_i\right) + \sum_{\{j_1,\dots,j_{k-1}\}\in\mathcal{C}_{k-1}^{[1\,:\,n-1]}} \alpha_{j_1\dots j_{k-1}}(x_{j_1}+\cdots+x_{j_{k-1}}+x_n), \tag{D.8}$$

avec $t''_i = kt'_i/t'$. On a $t''_i \geqslant 0$ et $\sum_{i=1}^{n-1} t''_i = k$. Supposons que l'on ait aussi les $t''_i \leqslant 1$, c'est-à-dire que pour $i = 1,\dots,n-1$:

$$\sum_{\{j_1,\dots,j_{k-2}\}\in\mathcal{C}_{k-2}^{[1\,:\,n-1]\setminus\{i\}}} \alpha_{ij_1\dots j_{k-2}} \geqslant t_i + t_n - 1. \tag{D.9}$$

Alors, on peut obtenir le résultat par récurrence, car la première somme dans (D.8) s'écrit (par l'hypothèse de récurrence) :

$$\sum_{i=1}^{n-1} t''_i x_i = \sum_{\{j_1,\dots,j_k\}\in\mathcal{C}_k^{[1\,:\,n-1]}} \beta_{j_1\dots j_k}(x_{j_1}+\cdots+x_{j_k}),$$

avec des $\beta_{j_1\dots j_k} \geqslant 0$ de somme 1. En tenant compte de (D.8), on voit que x s'écrit comme une combinaison conique d'éléments de la forme $x_{j_1}+\cdots+x_{j_k}$, avec des $\{j_1,\dots,j_k\} \in \mathcal{C}_k^{[1\,:\,n]}$. Par ailleurs, la somme des coefficients de la combinaison conique vaut $t'/k + t_n = 1$, si bien qu'il s'agit de la combinaison convexe sous la forme de laquelle on voulait mettre x. Il reste donc à montrer que l'on peut trouver des $\alpha_{j_1\dots j_{k-1}}$, pour $\{j_1,\dots,j_{k-1}\} \in \mathcal{C}_{k-1}^{[1\,:\,n-1]}$, qui vérifient (D.7) et (D.9). Il s'agit d'un ensemble d'égalité et d'inégalités linéaires dont la compatibilité peut être démontrée par le lemme de Farkas. En raisonnant comme à l'exercice 2.39, il revient au même de montrer que si $y \in \mathbb{R}$ et $z \in \mathbb{R}^{n-1}$ vérifient $z \geqslant 0$ et

$$\forall\, J \in \mathcal{C}_{k-1}^{[1\,:\,n-1]} \;:\quad y \geqslant \sum_{i\in J} z_i, \tag{D.10}$$

alors

$$t_n y \geqslant \sum_{i=1}^{n-1} (t_i + t_n - 1) z_i.$$

Mais, en sommant les inégalités de (D.10), on obtient

$$\tbinom{n-1}{k-1} y \geqslant \tbinom{n-2}{k-2} \sum_{i=1}^{n-1} z_i.$$

Cette inégalité est en fait une conséquence claire des prémisses, puisqu'en utilisant $z \geqslant 0$ et $t_i \leqslant 1$, on obtient

$$\sum_{i=1}^{n-1} (t_i + t_n - 1) z_i \leqslant t_n \sum_{i=2}^{n-1} z_i \leqslant t_n y.$$

2.7. De $\sum_i \alpha_i(x_i + y_i) = \sum_i \alpha_i x_i + \sum_i \alpha_i y_i$, on déduit que $\operatorname{cone}(P + Q) \subseteq \operatorname{cone} P + \operatorname{cone} Q$. Inversement, avec $\alpha = \sum_i \alpha_i$ et $\beta = \sum_j \beta_j$, on a $\sum_i \alpha_i x_i + \sum_j \beta_j y_j = \sum_{ij} \alpha_i \beta_j (x_i/\beta + y_j/\alpha)$, dont on déduit que $\operatorname{cone} P + \operatorname{cone} Q \subseteq \operatorname{cone}(\mathbb{R}_+ P + \mathbb{R}_+ Q)$.

Si P et Q sont des cônes convexes, $\mathbb{R}_+ P = P$, $\mathbb{R}_+ Q = Q$ et donc $P+Q = \operatorname{cone}(P+Q)$, qui est donc un cône convexe.

2.8. 1) Soient K un cône convexe, $\{x_i\}_{i=1}^m \subseteq K$ et des $t_i \geqslant 0$. Si $t := \sum t_i \neq 0$, alors $\sum (t_i/t) x_i \in K$ qui est convexe et $\sum t_i x_i = t[\sum (t_i/t) x_i] \in K$ qui est un cône. Si $t = 0$, tous les $t_i = 0$ et $\sum t_i x_i = 0 \in K$. Inversement, si K contient toutes les combinaisons coniques de ses éléments, on a $\operatorname{co} K \subseteq K$ et $tK \subseteq K$ pour tout $t \geqslant 0$; donc K est un cône convexe.

2) Comme l'ensemble à droite dans (2.6) est un cône convexe, il contient $\operatorname{cone} P$. Inversement, $\operatorname{cone} P$ étant un cône convexe, il contient toutes les combinaisons coniques de ses éléments, donc des éléments de P.

2.9. 1) Si K est une cône convexe fermé, $K^\infty = K^\infty(0) = \bigcap_{t>0} (K - 0)/t = K$. La réciproque est évidente par la proposition 2.7.

2) Soit $d \in C_1^\infty$. Alors pour $x \in C_1$, on a $x + \mathbb{R}_+ d \subseteq C_1 \subseteq C_2$. Comme $x \in C_2$, on en déduit que $d \in C_2^\infty$.

3) Soit $x \in \cap_{i\in I} C_i$ qui est non vide. Alors $d \in (\cap_{i\in I} C_i)^\infty \iff x + td \in \cap_{i\in I} C_i$, pour tout $t \geqslant 0 \iff x + td \in C_i$, pour tout $i \in I$ et tout $t \geqslant 0 \iff d \in C_i^\infty$, pour tout $i \in I \iff d \in \cap_{i\in I} C_i^\infty$.

On a $C_1 \cap C_2 = \varnothing$ et $C_1^\infty \cap C_2^\infty \neq \varnothing$ lorsque $\mathbb{E} = \mathbb{R}^2$, $C_1 = \mathbb{R} \times \{0\}$ et $C_2 = \{x \in \mathbb{R}^2 : x_1 x_2 \geqslant 1\}$.

4) Immédiat.

5) Par hypothèse il existe un $x \in \mathbb{E}$ tel que $Ax \in C$. Le résultat est alors une conséquence des équivalences suivantes : $d \in A^{-1}(C^\infty) \Leftrightarrow Ad \in C^\infty \Leftrightarrow Ax + \mathbb{R}_+ Ad \subseteq C \Leftrightarrow A(x + \mathbb{R}_+ d) \subseteq C \Leftrightarrow x + \mathbb{R}_+ d \subseteq A^{-1}(C) \Leftrightarrow d \in [A^{-1}(C)]^\infty$.

6.a) Soient $d \in C^\infty$ et $x \in C$. Alors $x + td \in C$ pour tout $t \geqslant 0$. Dès lors, $Ax + tAd \in A(C) \subseteq \operatorname{adh} A(C)$ pour tout $t \geqslant 0$, c'est-à-dire $Ad \in (\operatorname{adh} A(C))^\infty$. On a montré que $A(C^\infty) \subseteq (\operatorname{adh} A(C))^\infty$. Le résultat se déduit alors du fait que $(\operatorname{adh} A(C))^\infty$ est fermé.

6.b) Commençons par montrer que si l'on a une suite $\{x_k\} \subseteq C$ et une suite $\{t_k\} \subseteq \{t \in \mathbb{R}\colon t \geqslant 1\}$ telles que $\{Ax_k/t_k\}$ est bornée, alors on peut trouver une sous-suite $\{x'_k\} \subseteq C$ telle que $\{x'_k/t_k\}$ est bornée et $Ax'_k = Ax_k$.

On munit $\mathbb{E}$ d'un produit scalaire $\langle\cdot,\cdot\rangle$ avec norme associée $\|\cdot\|$. Si $\{x^1_k/t_k\} \equiv \{x_k/t_k\}$ n'est pas bornée, pour une sous-suite d'indices, $(x^1_k/t_k)/\|x^1_k/t_k\| = x^1_k/\|x^1_k\| \to d^1 \in \mathcal{N}(A) \cap C^\infty$ (nécessairement $\{x^1_k\}$ n'est pas bornée, car $t_k \geqslant 1$); donc $-d^1 \in C^\infty$ (par l'hypothèse) et la suite $\{x^2_k\}$ construite par $x^2_k = x^1_k - \langle x^1_k, d^1\rangle d^1$ est dans $C \cap \{d^1\}^\perp$ et vérifie $Ax^2_k = Ax_k$. De même, si $\{x^2_k/t_k\}$ n'est pas bornée alors, pour une sous-suite d'indices, $(x^2_k/t_k)/\|x^2_k/t_k\| = x^2_k/\|x^2_k\| \to d^2 \in \mathcal{N}(A) \cap C^\infty \cap \{d^1\}^\perp$ (nécessairement $\{x^2_k\}$ n'est pas bornée); donc $-d^2 \in C^\infty$ et la suite $\{x^3_k\}$ construite par $x^3_k = x^2_k - \langle x^2_k, d^2\rangle d^2$ est dans $C \cap \operatorname{vect}\{d^1, d^2\}^\perp$ et vérifie $Ax^3_k = Ax_k$. Si on poursuit de cette manière, on finit par trouver une suite $\{x^i_k/t_k\}$ bornée ou la suite identiquement nulle (car les d^i sont orthogonaux entre eux), qui est aussi bornée.

Pour montrer que $A(C)$ est fermé, on suppose que l'on a une suite $\{x_k\} \subseteq C$ telle que $Ax_k \to y$ et on montre que $y \in A(C)$. Par le résultat préliminaire (avec $t_k = 1$ pour tout k), on peut trouver une sous-suite bornée $\{x'_k\} \subseteq C$ telle que $Ax'_k = Ax_k \to y$. En extrayant une sous-suite convergente de $\{x'_k\}$, disons vers $x \in C$, on trouve $y = Ax \in A(C)$.

Montrons à présent que $A(C^\infty) = A(C)^\infty$. Par le point 1 et le fait que $A(C)$ est fermé, on a l'inclusion $A(C^\infty) \subseteq A(C)^\infty$. Il reste à montrer l'inclusion inverse. Soit $h \in A(C)^\infty$. Alors il existe une suite $\{x_k\} \subseteq C$ et une suite $\{t_k\} \uparrow \infty$ telles que $Ax_k/t_k \to h$. En appliquant le résultat préliminaire, on voit qu'il existe une sous-suite $\{x'_k\} \subseteq C$ telle que $\{x'_k/t_k\}$ est bornée et $Ax'_k/t_k = Ax_k/t_k \to h$. En extrayant une sous-suite convergente de $\{x'_k/t_k\}$, disons vers $d \in C^\infty$, on trouve $h = Ad \in A(C^\infty)$.

Dans le cas où $A : (x_1, x_2) \in \mathbb{R}^2 \mapsto x_1 \in \mathbb{R}$ et $C := \{x \in \mathbb{R}^2 : x_2 \geqslant x_1^2\}$, on a $C^\infty = \{d \in \mathbb{R}^2 : d_1 = 0,\ d_2 \geqslant 0\}$ et $A(C) = \mathbb{R}$, si bien que $A(C^\infty) = \{0\} \neq \mathbb{R} = [A(C)]^\infty$.

2.10. 1) [$\subseteq$] Soit $(z_1, z_2) \in \operatorname{ext}(C_1 \times C_2)$. Par symétrie, il suffit de montrer que $z_1 \in C_1$. Supposons que l'on ait $z_1 = (1-t)x_1 + ty_1$ avec $t \in]0,1[$ et $x_1,\ y_1 \in C_1$. Il s'agit de montrer que $x_1 = y_1 = z_1$. Or on a alors $(z_1, z_2) = (1-t)(x_1, z_2) + t(y_1, z_2)$, avec $(x_1, z_2) \in C_1 \times C_2$ et $(y_1, z_2) \in C_1 \times C_2$. Comme $(z_1, z_2) \in \operatorname{ext}(C_1 \times C_2)$, il vient $(x_1, z_2) = (y_1, z_2) = (z_1, z_2)$, d'où $x_1 = y_1 = z_1$.

[$\supseteq$] Soient $z_1 \in \operatorname{ext}(C_1)$ et $z_2 \in \operatorname{ext}(C_2)$. Il s'agit de montrer que $(z_1, z_2) \in \operatorname{ext}(C_1 \times C_2)$. Supposons que l'on ait $(z_1, z_2) = (1-t)(x_1, x_2) + t(y_1, y_2)$ avec $t \in]0,1[$ et $(x_1, x_2),\ (y_1, y_2) \in C_1 \times C_2$. Il suffit maintenant de montrer que $(x_1, x_2) = (y_1, y_2) = (z_1, z_2)$ ou, par symétrie, que $x_1 = y_1 = z_1$. Or $z_1 \in \operatorname{ext}C_1$ vérifie $z_1 = (1-t)x_1 + ty_1$ avec avec $t \in]0,1[$ et $x_1,\ y_1 \in C_1$; donc $x_1 = y_1 = z_1$.

2) [$\Rightarrow$] Soit $x \in \operatorname{ext} C$. Supposons que l'on ait $Ax = (1-t)Ax_0 + tAx_1$ avec $t \in]0,1[$ et $x_0,\ x_1 \in C$. Il s'agit de montrer que $Ax_0 = Ax_1$. Par l'injectivité de A, on a $x = (1-t)x_0 + tx_1$ et donc $x_0 = x_1$ parce que $x \in \operatorname{ext} C$.

[$\Leftarrow$] Supposons que $Ax \in \operatorname{ext}(A(C))$. Par l'injectivité de A, on a $x \in C$. Supposons donc que l'on ait $x = (1-t)x_0 + tx_1$ avec $t \in]0,1[$ et $x_0,\ x_1 \in C$. Il s'agit de montrer que $x_0 = x_1$. On a $Ax = (1-t)Ax_0 + tAx_1$ et donc $Ax_0 = Ax_1$ parce que $Ax \in \operatorname{ext}(A(C))$. On en déduit que $x_0 = x_1$ par l'injectivité de A.

L'implication [$\Rightarrow$] équivaut à $A(\operatorname{ext}(C)) \subseteq \operatorname{ext}(A(C))$, même sans l'injectivité de A. L'implication [$\Leftarrow$] équivaut à $A(\operatorname{ext}(C)) \supseteq \operatorname{ext}(A(C))$, avec l'injectivité de A (l'implication serait plus forte que l'inclusion, si A n'était pas injective).

2.11. On note $\alpha = \min\{\langle c, x\rangle : x \in C\}$. Soient x_0 et $x_1 \in C$ tels que $x = (x_0 + x_1)/2 \in E$. Alors $\langle c, x_i\rangle \geqslant \alpha$ et $\langle c, x_0\rangle + \langle c, x_1\rangle = 2\langle c, x\rangle = 2\alpha$. Donc $\langle c, x_i\rangle = \alpha$ et les $x_i \in E$.

Le singleton $\{0\}$ de l'ensemble $\mathbb{R}^2_+ \cup \{x \in \mathbb{R}^2 : x_{(2)} \geqslant x^2_{(1)}\}$ est une face, mais n'est pas exposé.

2.12. 1) L'implication $(i) \Rightarrow (ii)$ est claire. Réciproquement, il suffit de monter que, quel que soit $z \in \mathbb{E}$, $x_t := (1-t)x + tz \in C$ pour un $t < 0$. Comme on peut écrire $x_t = x + (-t)(x - z)$, c'est le cas par (ii) avec $y = x - z$.

2) Soient $x_0 \in C_1 \cap C_2^{\diamond}$, qui est non vide par hypothèse, et $x_1 \in (C_1 \cap \operatorname{aff} C_2)^{\diamond}$, qui est aussi non vide puisque $C_1 \cap \operatorname{aff} C_2$ contient x_0 (on utilise ensuite le point 1 de la proposition 2.16). Si $x_0 = x_1$, le résultat est démontré, car ce point appartient alors à $(C_1 \cap \operatorname{aff} C_2)^{\diamond} \cap C_2^{\diamond}$. Sinon, pour $t > 0$ assez petit, $x_t = (1-t)x_0 + tx_1$ appartient à $(C_1 \cap \operatorname{aff} C_2)^{\diamond} \cap C_2^{\diamond}$ qui est donc non vide. En effet, $x_t \in]x_0, x_1] \subseteq (C_1 \cap \operatorname{aff} C_2)^{\diamond}$, car $x_0 \in C_1 \cap \operatorname{aff} C_2$ et $x_1 \in (C_1 \cap \operatorname{aff} C_2)^{\diamond}$ (lemme 2.14). D'autre part, $x_t \in C_2^{\diamond}$ pour $t > 0$ petit, car $x_0 \in C_2^{\diamond}$ et $x_1 \in \operatorname{aff} C_2$.

3) Il suffit de prendre l'intérieur relatif des deux membres de $\overline{C} = \mathbb{E}$ et d'utiliser le point 3 de la propoistion 2.16, pour voir que $C^{\diamond} = \mathbb{E}$ et donc $C = \mathbb{E}$, certainement.

2.15. En utilisant les points 1 et 3 de la proposition 2.17 avec $A : \mathbb{E} \times \mathbb{E} \to \mathbb{E} : (x_1, x_2) \mapsto x_1 - x_2$ et $C = C_1 \times C_2$, on obtient $(C_1 - C_2)^{\diamond} = A(C_x \times C_2)^{\diamond} = A((C_1 \times C_2)^{\diamond}) = A(C_1^{\diamond} \times C_2^{\diamond}) = C_1^{\diamond} - C_2^{\diamond}$. Dès lors $0 \in (C_1 - C_2)^{\diamond}$ si, et seulement si, $0 \in C_1^{\diamond} - C_2^{\diamond}$ si, et seulement si, $C_1^{\diamond} \cap C_2^{\diamond} \neq \varnothing$.

2.13. 1) Comme $P^{\diamond} \subseteq P$, on a $\operatorname{aff} P^{\diamond} \subseteq \operatorname{aff} P$. Inversement, soit $x \in \operatorname{aff} P$ et supposons que $P^{\diamond} \neq \varnothing$. Soit $x_0 \in P^{\diamond}$, si bien qu'il existe un $r > 0$ tel que $B(x_0, r) \cap \operatorname{aff} P \subseteq P$. Comme $x \in \operatorname{aff} P$, $x_t = (1-t)x_0 + tx \in P^{\diamond}$ pour $t > 0$ petit. Alors $x = (1 - \frac{1}{t})x_0 + \frac{1}{t}x_t \in \operatorname{aff}\{x_0, x_t\} \subseteq \operatorname{aff} P^{\diamond}$.

2) Comme aff conserve l'inclusion et que A est un sous-espace affine, on a $\operatorname{aff} C^{\diamond} \subseteq A \subseteq \operatorname{aff} C$. Mais $\operatorname{aff} C^{\diamond} = \operatorname{aff} C$ (point 1 de la proposition 2.16), donc $A = \operatorname{aff} C$. Ensuite $A \subseteq C \subseteq \operatorname{aff} C$ conduit à $A = C$.

3) On a clairement $\overline{P^{\diamond}} \subseteq \overline{P}$ (car $P^{\diamond} \subseteq P$) et $P^{\diamond} \subseteq (\overline{P})^{\diamond}$ (car $P \subseteq \overline{P}$ et $\operatorname{aff} P = \operatorname{aff} \overline{P}$).

On suppose à présent que P est convexe.

Soit $x \in \overline{P}$ et montrons que $x \in \overline{P^{\diamond}}$. On peut choisir $x_0 \in P^{\diamond}$ qui est non vide (proposition 2.16). D'après le lemme 2.14, $[x_0, x[\subseteq P^{\diamond}$ et donc x est limite de points de $P^{\diamond}$ (ceux du segment $[x_0, x[$). Dès lors $x \in \overline{P^{\diamond}}$.

Soit $x \in (\overline{P})^{\diamond}$ et montrons que $x \in P^{\diamond}$. On peut choisir $x_0 \in P^{\diamond}$ qui est non vide (proposition 2.16). Comme $x_0 \in \operatorname{aff} \overline{P}$ et $x \in (\overline{P})^{\diamond}$, le point $x_t = (1-t)x_0 + tx \in (\overline{P})^{\diamond} \subseteq \overline{P}$ pour un $t > 1$. D'après le lemme 2.14, $[x_0, x_t[\subseteq P^{\diamond}$ et donc $x \in P^{\diamond}$.

4) Cela résulte des points précédents. En ce qui concerne la frontière relative, on a : $\overline{C^{\diamond}} \setminus C^{\diamond} = \overline{C} \setminus C^{\diamond}$ et $\overline{C} \setminus (\overline{C})^{\diamond} = \overline{C} \setminus C^{\diamond}$.

5)

6) Si $x \notin F(x)^{\diamond}$, on peut séparer les deux convexes non vides disjoints $\{x\}$ et $F(x)^{\diamond}$ dans $\operatorname{aff} F(x)$ (théorème 2.38) : il existe $\xi \in \mathbb{E}$, non nul et parallèle à $\operatorname{aff} F(x)$, tel que pour tout $y \in F(x)^{\diamond}$, on a $\langle \xi, x\rangle \leqslant \langle \xi, y\rangle$. On a l'inégalité stricte pour au moins un $y \in F(x)^{\diamond}$, car dans le cas contraire $\langle \xi, \cdot\rangle$ serait constante sur $F(x)^{\diamond}$: pour $y \in F(x)^{\diamond}$ et $t > 0$ assez petit, on aurait $\langle \xi, y + t\xi\rangle = \langle \xi, y\rangle$ (on utilise le fait que $\xi \in \operatorname{aff} F(x)$), ce qui impliquerait la nullité de ξ. Alors $F' := \{z \in C : \langle \xi, z\rangle \leqslant \langle \xi, x\rangle\}$ serait une

face de C (exposée, voir l'exercice 2.11) qui contient x mais n'est pas incluse dans $F(x)$. La minimalité de la face $F(x)$ serait contredite.

2.14. Montrons que $\overline{K}$ est un cône. Si K est vide, il n'y a rien à démontrer. Soient $x \in \overline{K}$ et $\alpha > 0$. On veut montrer que $\alpha x \in \overline{K}$, ce qui sera le cas si, pour un $\varepsilon > 0$ arbitraire donné, $K \cap B(\alpha x, \varepsilon) \neq \varnothing$. Mais, $K \cap B(x, \varepsilon/\alpha)$ contient un point x_ε. Alors $\alpha x_\varepsilon \in K \cap B(\alpha x, \varepsilon)$.

Montrons que $K^\ominus$ est un cône. Soient $x \in K^\ominus$ et $\alpha > 0$. On souhaite montrer que $\alpha x \in K^\ominus$, ce qui sera le cas si l'on trouve un $\varepsilon > 0$ tel que $B(\alpha x, \varepsilon) \cap \operatorname{vect} K \subseteq K$. Comme $x \in K^\ominus$, on peut trouver un $\varepsilon > 0$ tel que $B(x, \varepsilon/\alpha) \cap \operatorname{vect} K \subseteq K$, d'où l'on déduit que $B(\alpha x, \varepsilon) \cap \operatorname{vect} K \subseteq K$.

2.16. 1) Soient x, $y \in C$ tels que $(x+y)/2 \in F_0$. Il faut montrer que x et $y \in F_0$. On observe que $(x+y)/2 \in F$, qui est une face de C, donc x et $y \in F$. Comme F_0 est une face de F, que x, $y \in F$ et que $(x+y)/2 \in F_0$, on a x, $y \in F_0$.

2) C'est un cas particulier du point 1, dans lequel F_0 est réduit à un point.

3) Il suffit de montrer que $\operatorname{ext}(C) \neq \varnothing \Rightarrow \operatorname{ext}(F) \neq \varnothing$ (la réciproque découle du point 2). Soient $\hat{x} \in \operatorname{ext}(C)$ et $x \in F$. On démarre une démonstration par récurrence. Si x est un point extrême de F, le résultat est démontré. Sinon, on peut trouver une direction non nulle $d \in \mathbb{E}$ telle que $x + td \in F$ pour $|t|$ petit. Par ailleurs, soit $x + td$, soit $x - td$ n'est plus dans F pour t grand. En effet, dans le cas contraire, en passant à la limite pour $t \to \infty$ dans $(1-1/t)\hat{x} + (1/t)(x \pm td) \in C$, on trouverait que $\hat{x} \pm d \in C$ (fermé), ce qui contredirait l'extrémalité de $\hat{x}$. On peut donc trouver un point $x' = x + td$ qui est sur la frontière relative de F. Alors, la face F' engendrée par x' est strictement incluse dans F (exercice 2.13, point 6). On poursuit la récurrence qui conduit nécessairement à un point extrême d'une face F'' de C (les faces déterminées ainsi sont de plus en plus petites et si une face est réduite à un point, celui-ci est un point extrême). Comme un point extrême de F'' est un point extrême de F (point 2), le résultat est démontré.

2.17. Commençons par la boule unité pour la norme ℓ_1. Un dessin en dimension 2 suggère que ses points extrêmes sont les points vérifiant $\exists j$ tel que $x_j = \pm 1$ et $x_i = 0$ pour $i \neq j$. Ce sont clairement des points extrêmes car si un tel point s'écrit $x = (y+z)/2$, avec y et z dans la boule unité, on a nécessairement $x_j = y_j = z_j$ et donc $x = y = z$. D'autre part, il n'y a pas d'autres points extrêmes. En effet, si x dans la boule unité vérifie pour un indice j, $x_j \neq 0$ et $x_j \neq \pm 1$, il existe un autre indice, disons k, tel que $x_k \neq 0$ et $x_k \neq \pm 1$. Alors, pour $t > 0$ suffisamment petit, $y = x + (\operatorname{sgn} x_j)te_j - (\operatorname{sgn} x_k)te_k$ et $y = x - (\operatorname{sgn} x_j)te_j + (\operatorname{sgn} x_k)te_k$ sont dans la boule unité (on a noté e_i le i-ième vecteur de la base canonique de $\mathbb{R}^n$), sont distincts de x et vérifient $x = (y+z)/2$. Donc x n'est pas un point extrême.

Par un raisonnement analogue, on montre que les points extrêmes de la boule unité pour la norme ℓ_∞ sont les points vérifiant $x_i = \pm 1$ pour tout i.

2.18. Soit $\mathbb{P}^n$ l'ensemble des matrices de permutation d'ordre n. Il est clair que $\mathbb{P}^n \subseteq \operatorname{ext}(\mathbb{G}^n)$: il suffit d'examiner les matrices élément par élément et d'utiliser le fait que 0 et 1 sont des points extrêmes de $[0,1]$.

Inversement, soit $G \in \operatorname{ext}(\mathbb{G}^n)$. Si $G_{i_1 j_1} \in]0,1[$, un autre élément de la ligne i_1 est dans $]0,1[$, disons $G_{i_1 j_2}$ (car la somme des éléments de la ligne i_1 vaut 1). Dès lors, un autre élément de la colonne j_2 est dans $]0,1[$, disons $G_{i_2 j_2}$ (car la somme des éléments de la colonne j_2 vaut 1). Pour la même raison, on trouve ensuite $G_{i_2 j_3} \in]0,1[$. Si $j_3 = j_1$ on s'arrête. Sinon, on poursuit jusqu'à ce qu'un indice de ligne ou de colonne

se retrouve. On peut supposer qu'il s'agit de i_1 ou j_1 (si ce n'est pas le cas, on élimine tous les autres éléments précédent celui dont on a retrouvé un indice de ligne ou de colonne). On a donc formé une chaîne :

$$G_{i_1j_1} \quad G_{i_1j_2} \quad G_{i_2j_2} \quad G_{i_2j_3} \quad \cdots \quad G_{i_mj_{m'}},$$

avec $i_m = i_1$ ou $j_{m'} = j_1$ (et pas les deux). On peut même supposer que $i_m \neq i_1$ et $j_{m'} = j_1$ (alors $m' = m+1$), car dans le cas contraire, on peut éliminer $G_{i_1j_1}$ de la liste et la parcourir « en sens inverse » : $G_{i_1j_2}, G_{i_mj_{m'}}, G_{i_{m'}j_{m'}}, \ldots, G_{i_2j_3}, G_{i_2j_2}$. On retrouve ainsi la situation

$$G_{i_1j_1} \quad G_{i_1j_2} \quad G_{i_2j_2} \quad G_{i_2j_3} \quad \cdots \quad G_{i_mj_m} \quad G_{i_mj_1}.$$

Chaque ligne [resp. colonne] de G contient soit 2 éléments de cette liste, soit aucun. On peut donc modifier la matrice en ajoutant et retranchant alternativement une petite quantité $\varepsilon > 0$, tout en conservant la propriété d'appartenir à $\mathbb{G}^n$. De manière plus précise, on choisit $\varepsilon > 0$ tel que $\varepsilon \leqslant \min(G_{i_1j_1}, \ldots, G_{i_mj_1})$ et $\varepsilon \leqslant \min(1 - G_{i_1j_1}, \ldots, 1 - G_{i_mj_1})$. Ensuite, on construit les matrices G^- et G^+ par $G^-_{ij} = G^+_{ij} = G_{ij}$, pour tous les couples d'indices (i,j) ne correspondants pas aux éléments de G listés ci-dessus. Pour ceux-ci, on prend

$$G^-_{i_1j_1} = G_{i_1j_1} - \varepsilon, \quad G^-_{i_1j_2} = G_{i_1j_2} + \varepsilon, \quad G^-_{i_2j_2} = G_{i_2j_2} - \varepsilon, \quad G^-_{i_2j_3} = G_{i_2j_4} + \varepsilon, \quad \ldots$$

et

$$G^+_{i_1j_1} = G_{i_1j_1} + \varepsilon, \quad G^+_{i_1j_2} = G_{i_1j_2} - \varepsilon, \quad G^+_{i_2j_2} = G_{i_2j_2} + \varepsilon, \quad G^+_{i_2j_3} = G_{i_2j_4} - \varepsilon, \quad \ldots$$

Il est clair que G^- et $G^+ \in \mathbb{G}^n$ et sont différentes de G. Cette dernière matrice s'écrit $G = \frac{1}{2}(G^- + G^+)$, si bien que G n'est pas un point extrême de $\mathbb{G}^n$. Nous avons montré que les points extrêmes de $\mathbb{G}^n$ n'ont que des 0 ou des 1 et même exactement un unique 1 par ligne et par colonne. Les points extrêmes sont donc les matrices de permutation.

2.19. 1) Si K est convexe, $x, y \in K$, $\alpha, \beta \in \mathbb{R}_+$ et $\alpha + \beta \neq 0$, alors on a $\alpha x + \beta y = (\alpha+\beta)[\alpha/(\alpha+\beta)x + \beta/(\alpha+\beta)y] \in \mathbb{R}_+K = K$. La réciproque est évidente.

2) Observons d'abord que $\alpha > 0$ et $x \in K^\bullet$ impliquent que $\alpha x \in K^\bullet$ (si $B(x,r) \cap \operatorname{aff} K \subseteq K$, on a $B(\alpha x, \alpha r) \cap \operatorname{aff} K \subseteq K$). Comme $\alpha/(\alpha+\beta)x + \beta/(\alpha+\beta)y \in K^\bullet$ (lemme 2.14), ce qui précède montre que $\alpha x + \beta y = (\alpha+\beta)[\alpha/(\alpha+\beta)x + \beta/(\alpha+\beta)y] \in K^\bullet$.

3) Clairement, $K \subseteq \operatorname{vect}\{x_1, \ldots, x_p\}$ et ce dernier ensemble est un sous-espace affine, donc $\operatorname{aff} K \subseteq \operatorname{vect}\{x_1, \ldots, x_p\}$. Inversement si $x = \sum_{i=1}^p \alpha_i x_i$, avec des $\alpha_i \in \mathbb{R}$, on a $x = y_0 + y_1 - y_2$, où $y_0 = 0 \in K$, $y_1 = \sum_{i=1}^p \alpha_i^+ x_i \in K$ et $y_2 = \sum_{i=1}^p \alpha_i^- x_i \in K$. Comme la somme des coefficients des y_i vaut 1, $x \in \operatorname{aff} K$ (proposition 2.1).

Montrons que $K^\bullet \subseteq \{\sum_{i=1}^p \alpha_i x_i : \alpha_i > 0\}$. Soit $x = \sum_{i=1}^p \alpha_i x_i \in K^\bullet$, avec des $\alpha_i \geqslant 0$, et $y = \sum_{i=1}^p x_i \in K$. Alors il existe un $t > 1$ tel que $(1-t)y + tx \in K$, c'est-à-dire $(1-t)y + tx = \sum_{i=1}^p \alpha'_i x_i$, avec des $\alpha'_i \geqslant 0$. On en déduit que $x = \sum_{i=1}^p \frac{1}{t}(t-1+\alpha'_i)x_i$ avec $\frac{1}{t}(t-1+\alpha'_i) > 0$ pour tout i.

Réciproquement, montrons que $x = \sum_{i=1}^p \alpha_i x_i$, avec des $\alpha_i > 0$, est dans $K^\bullet$. Soit $y = \sum_{i=1}^p \beta_i x_i \in K$, avec des $\beta_i \geqslant 0$. D'après le lemme 2.14, il suffit de montrer qu'il existe $t > 1$ tel que $(1-t)y + tx \in K$. On voit qu'il suffit de prendre $t/(t-1) \geqslant \max_i(\beta_i/\alpha_i)$.

2.20. 1) D'abord, il est clair qu'un ensemble de la forme $F = \{x \in P : (Bx - b)_I = 0\}$ est une face de P. En effet, F est convexe. D'autre part, si x_0 et $x_1 \in P$ sont tels que

$x = (x_0 + x_1)/2 \in F$, alors $0 = (Bx - b)_I = \frac{1}{2}(Bx_0 - b)_I + \frac{1}{2}(Bx_1 - b)_I$. Comme $(Bx_i - b)_I \leqslant 0$, on en déduit que $(Bx_i - b)_I = 0$, c'est-à-dire $x_i \in F$.

Montrons la réciproque. Soit F une face (non vide) de P. Pour définir I, on prend $x_0 \in F^\oplus$ (non vide) et ensuite $I = \{i : (Bx_0 - b)_i = 0\}$. Montrons que $F = \{x \in P : (Bx - b)_I = 0\}$. Soit $x \in F$. Alors $x \in P$. Comme $x_0 \in F^\oplus$, $(1-t)x_0 + tx \in F \subseteq P$ pour $|t|$ petit, ce qui implique que $0 \geqslant (1-t)(Bx_0 - b)_I + t(Bx - b)_I = t(Bx - b)_I$. Donc $(Bx - b)_I = 0$. Inversement, montrons qu'un $x \in P$ vérifiant $(Bx - b)_I = 0$ est dans F. Il suffit de montrer que, pour $t < 0$ proche de zéro, $x_t := (1-t)x_0 + tx$ est dans P. D'une part $(Bx_t - b)_I = 0$. D'autre part, $(Bx_t - b)_{I^c} = (1-t)(Bx_0 - b)_{I^c} + t(Bx - b)_{I^c} \leqslant 0$, pour $t < 0$ assez proche de zéro, car $(Bx_0 - b)_{I^c} < 0$.

2) Soit $F = \{x \in P : (Bx - b)_I = 0\}$ une face de P, supposée non vide. Il faut trouver $c \in \mathbb{R}^n$ tel que $F = \arg\min\{c^\mathsf{T} x : x \in P\}$. Les conditions d'optimalité de ce problème contiennent $A^\mathsf{T} y - B^\mathsf{T} s = c$, $s \geqslant 0$ et $s^\mathsf{T}(Bx - b) = 0$, ce qui donne la forme que doit prendre c). On définit $s \in \mathbb{R}^m$ par $s_i = 1$ si $i \in I$ et $s_i = 0$ sinon (complémentarité stricte pour la variable duale s associée aux contraintes $Bx \leqslant b$), y arbitraire et $c = A^\mathsf{T} y - B^\mathsf{T} s$. On pose $\alpha := a^\mathsf{T} y - b^\mathsf{T} s$ (c'est la valeur optimale duale). Pour tout $x \in P$, on a $c^\mathsf{T} x \geqslant \alpha$ (on multiplie $Ax = a$ par y et $Bx \leqslant b$ par $-s$). Pour $x \in F$, on a $c^\mathsf{T} x = \alpha$. Donc $F \subseteq \arg\min\{c^\mathsf{T} x : x \in P\}$. Il reste à montrer que si $x \in \arg\min\{c^\mathsf{T} x' : x' \in P\}$, alors $x \in F$. Dans ce cas, on a $c^\mathsf{T} x' \geqslant c^\mathsf{T} x$, pour tout $x' \in P$, ce qui s'écrit aussi $s^\mathsf{T}(Bx' - Bx) \leqslant 0$, pour tout $x' \in P$. En prenant $x' \in F$, on a $s^\mathsf{T}(b - Bx) \leqslant 0$. En utilisant $Bx \leqslant b$ dans cette inégalité, on obtient $(Bx - b)_I = 0$.

3) Soit $x \in P$. La matrice $(A^\mathsf{T} \;\; B_I^\mathsf{T})^\mathsf{T}$ n'est pas injective $\iff$ il existe $d \in \mathbb{R}^n$ non nul tel que $Ad = 0$ et $B_I d = 0 \iff$ il existe $d \neq 0$ tel que $x + \alpha d \in P$ pour $|\alpha|$ petit $\iff$ x n'est pas un sommet de P.

4) Soit $x \in P$. Alors $d \in P^\infty$ ssi $x + td \in P$ pour tout $t \geqslant 0$, ce qui revient à dire que $Ad = 0$ et $Bd \leqslant 0$.

5) Soit $\mathcal{A} := \{x \in \mathbb{R}^n : Ax = a,\ B_{I^c} x = b_{I^c}\}$. Comme $\mathcal{A}$ est affine et contient P, on a $\operatorname{aff} P \subseteq \mathcal{A}$. Inversement, si $I = \varnothing$, $\mathcal{A} = P \subseteq \operatorname{aff} P$ et l'identité $\operatorname{aff} P = \mathcal{A}$ est démontrée. Si $I \neq \varnothing$, l'ensemble $P_I := \{x \in P : B_I x < b_I\}$ est non vide car, par définition de I, pour tout $i \in I$, on peut trouver un $x^i \in P$ tel que $B_i x^i < b_i$ et alors $\sum_{i \in I} x^i/|I| \in P_I$. Ensuite, soit $x \in \mathcal{A}$ et montrons que $x \in \operatorname{aff} P$, ce qui suffit. On prend $y \in P_I \neq \varnothing$. Alors, pour $t > 0$ assez petit, $x_t := tx + (1-t)y \in P$, donc $x = \frac{1}{t} x_t + (1-\frac{1}{t})y$ avec $x_t, y \in P$. Donc $x \in \operatorname{aff} P$.

Montrons à présent que $P^\oplus = P_I$. [$\supseteq$] Si $x \in P_I$, on voit qu'il existe un voisinage V de x tel que tout point de $\operatorname{aff} P \cap V$ est encore dans P, ce qui implique que $x \in P^\oplus$. [$\supseteq$] Inversement, soient $x \in P^\oplus$ et $y \in P_I$ (non vide et dans P). Alors, pour $t > 1$ proche de 1, on a $(1-t)y + tx \in P$, donc $(1-t)B_I y + tB_I x \leqslant b_I$. On en déduit que $B_I x \leqslant b_I + ((t-1)/t)[By - b]_I < b_I$. Donc $x \in P_I$.

6) Si $P_1 = \{x_1 \in \mathbb{E}_1 : A_1 x_1 \leqslant b_1\}$ et $P_2 = \{x_2 \in \mathbb{E}_2 : A_2 x_2 \leqslant b_2\}$ sont deux polyèdres convexes, $P_1 \times P_2 = \{(x_1, x_2) \in \mathbb{E}_1 \times \mathbb{E}_2 : A_1 x_1 \leqslant b_1,\ A_2 x_2 \leqslant b_2\}$ est un polyèdre convexe de $\mathbb{E}_1 \times \mathbb{E}_2$. Comme $\sigma : \mathbb{E} \times \mathbb{E} \to \mathbb{E} : (x_1, x_2) \mapsto x_1 + x_2$ est linéaire, $\sigma(P_1 \times P_2) = P_1 + P_2$ est un polyèdre convexe de $\mathbb{E}$ (proposition 2.18) et celui-ci s'écrit $P_1 + P_2 = \{x_1 + x_2 \in \mathbb{E} : A_1 x_1 \leqslant b_1,\ A_2 x_2 \leqslant b_2\}$.

7) Soient en effet $T : \mathbb{R}^n \to \mathbb{R}^m$ une application linéaire et $P = \{y \in \mathbb{R}^m : By \leqslant b\}$ un polyèdre convexe de $\mathbb{R}^m$. On voit immédiatement que $T^{-1}(P) = \{x \in \mathbb{R}^n : BTx \leqslant b\}$, qui est bien un polyèdre convexe de $\mathbb{R}^n$.

8) Soient $\mathbb{E}$ et $\mathbb{F}$ deux espaces vectoriels, et $P \subseteq \mathbb{E}$ et $Q \subseteq \mathbb{F}$ deux polyèdres convexes. Donc P est de la forme $\{x \in \mathbb{E} : Ax \leqslant b\}$ et Q est de la forme $\{y \in \mathbb{F} : \tilde{A}y \leqslant \tilde{b}\}$. Alors $(x, y) \in P \times Q$ si, et seulement si, $(x, y) \in \mathbb{E} \times \mathbb{F}$ avec

$$\begin{pmatrix} A & 0 \\ 0 & \tilde{A} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \leqslant \begin{pmatrix} b \\ \tilde{b} \end{pmatrix},$$

ce qui montre que $P \times Q$ est un polyèdre convexe de $\mathbb{E} \times \mathbb{F}$.

2.21. 1) Soit $\hat{x}$ un sommet de P. Alors $\{\hat{x}\}$ est une face exposée de P (exercice 2.11) : il existe $c \in \mathbb{E}$ tel que $c^\mathsf{T}\hat{x} < c^\mathsf{T}x$ pour tout $x \in P$, $x \neq \hat{x}$. On a $\hat{x} = \sum_i \alpha_i x_i + \sum_j \beta_j y_j$ avec $\alpha := (\alpha_1, \ldots, \alpha_p) \in \Delta$ et les $\beta_j \geqslant 0$. Dans l'égalité $c^\mathsf{T}\hat{x} = \sum_i \alpha_i c^\mathsf{T}x_i + \sum_j \beta_j c^\mathsf{T}y_j$, les $\beta_j c^\mathsf{T}y_j$ sont positifs (car $\hat{x} + y_j \in P$, donc $c^\mathsf{T}(\hat{x} + y_j) \geqslant c^\mathsf{T}\hat{x}$) et $\sum_i \alpha_i c^\mathsf{T}x_i \geqslant c^\mathsf{T}\hat{x}$ (car $\sum_i \alpha_i c^\mathsf{T}x_i \in P$); il faut donc que tous les $\beta_j c^\mathsf{T}y_j$ soient nuls pour que l'égalité puisse avoir lieu. Donc $c^\mathsf{T}\hat{x} = \sum_i \alpha_i c^\mathsf{T}x_i$. Comme $c^\mathsf{T}x_i > c^\mathsf{T}\hat{x}$ si $x_i \neq \hat{x}$, il faut que tous les $\alpha_i = 0$ si $x_i \neq \hat{x}$; donc $\hat{x}$ est un des x_i.

On n'a pas pas égalité si $P := \mathrm{co}\{0, 1, 2\} \subseteq \mathbb{R} : 1 \notin \mathrm{ext}(P)$.

2) Évidemment, les $y_j \in P^\infty$ et donc $\mathrm{cone}\{y_1, \ldots, y_q\} \subseteq P^\infty$ puisque P^∞ est un cône convexe (proposition 2.7). Inversement, si $d \in P^\infty$, $x + td \in P$ pour tout $t \geqslant 0$ (x est arbitraire dans P). On peut donc écrire $x + td = \sum_i \alpha_i^t x_i + \sum_j \beta_j^t y_j$, avec $(\alpha_1^t, \ldots, \alpha_p^t) \in \Delta$ et des $\beta_j^t \geqslant 0$. Lorsque $t \to \infty$, $\alpha_1^t/t \to 0$ et $(x + td)/t \to d$, donc $\sum_j (\beta_j^t/t) y_j$ converge, et sa limite ne peut être qu'un élément de $\mathrm{cone}\{y_1, \ldots, y_q\}$ (c'est un fermé d'après la remarque 2.19). Donc $d \in \mathrm{cone}\{y_1, \ldots, y_q\}$.

3) On peut écrire P comme l'image du convexe $\Delta_p \times \mathbb{R}^q_+$ par l'application linéaire $T : (\alpha, \beta) \in \mathbb{R}^p \times \mathbb{R}^q \mapsto \sum_i \alpha_i x_i + \sum_j \beta_j y_j \in \mathbb{R}^n$. Alors (point 3 de la proposition 2.17), $P^\ominus = T((\Delta \times \mathbb{R}^q_+)^\ominus)$. On en déduit le résultat.

4) Montrons que $K := \mathrm{cone}\{x_1, \ldots, x_p, y_1, \ldots, y_q\} \subseteq \mathrm{cone}\, P$. Les $x_i \in \mathrm{cone}\, P$, car les $x_i \in P$. Par ailleurs, pour tout $t > 0$, $x_1 + ty_j \in \mathrm{cone}\, P$ car $x_1 + ty_j \in P$; donc pour tout $t > 0$, $(x_1 + ty_j)/t \in \mathrm{cone}\, P$ car $\mathrm{cone}\, P$ est un cône; donc les $y_j \in \mathrm{cone}\, P$; on en déduit que $K \subseteq \mathrm{cone}\, P$. Réciproquement, comme $P \subseteq K$, on a $\mathrm{cone}\, P \subseteq K$.

5) Soient $P = \mathrm{co}\{x_i : i \in I\} + \mathrm{cone}\{y_j : j \in J\}$ et $\tilde{P} = \mathrm{co}\{\tilde{x}_k : k \in K\} + \mathrm{cone}\{\tilde{y}_l : l \in L\}$ deux polyèdres convexes (I, J, K et L étant des ensembles d'indices finis). D'après le point 2 de la proposition 2.5, $P + \tilde{P} = \mathrm{co}\{x_i + \tilde{x}_k : i \in I,\ k \in K\} + \mathrm{cone}\{y_j : j \in J\} + \mathrm{cone}\{\tilde{y}_l : l \in L\}$, qui est bien un polyèdre convexe.

6) Soient en effet $T : \mathbb{E} \to \mathbb{F}$ une application linéaire et $P = \mathrm{co}\{x_1, \ldots, x_p\} + \mathrm{cone}\{y_1, \ldots, y_q\}$ un polyèdre de $\mathbb{E}$. On voit immédiatement que $T(P) = \mathrm{co}\{Tx_1, \ldots, Tx_p\} + \mathrm{cone}\{Ty_1, \ldots, Ty_q\}$, qui est bien un polyèdre convexe.

2.22. On note $P_1' := \{x_1 \in \mathbb{R}^{n_1} : \text{il existe un } x_2 \in \mathbb{R}^{n_2} \text{ tel que } A_1x_1 + A_2x_2 \leqslant b\}$ la projection de P sur $\mathbb{R}^{n_1}$. Il est clair que P_1' est un polyèdre convexe (point 6 de l'exercice 2.21). Montrons que $P_1' = P_1$. [$\subseteq$] Cette inclusion s'obtient facilement en multipliqnat chaque membre de l'inégalité $A_1x_1 + A_2x_2 \leqslant b$ par un $y \geqslant 0$ dans le noyau de A_2^T. [$\supseteq$] Si $x_1 \in P_1$, on a

$$\begin{aligned} b - A_1x_1 &\in (\mathbb{R}^n_+ \cap \mathcal{N}(A_2^\mathsf{T}))^+ \\ &= \{z : (A_2 \ -A_2 \ I)^\mathsf{T} z \geqslant 0\}^+ \\ &= \{A_2x_2 + s : s \geqslant 0\}, \end{aligned}$$

si bien qu'il existe un x_2 tel que $A_1x_1 + A_2x_2 \leqslant b$.

2.23. Observons que $A(K) = \cup_{r\geqslant 0} A\{(x_1,x_2,x_3) : x_1^2+x_2^2 \leqslant r^2,\, x_3 = r\}$, si bien que $A(K)$ est une réunion de disques de $\mathbb{R}^2$, de centre $(0,r)$ et de rayon r. On en déduit que $A(K) = \{0\} \cup \{y \in \mathbb{R}^2 : y_2 > 0\}$.

2.24. Soient $K \subseteq \mathbb{R}^3$ et $A : \mathbb{R}^3 \to \mathbb{R}^2$ le cône convexe fermé et l'application linéaire de l'exercice 2.23. On a $K + \mathcal{N}(A) = A^{-1}(A(K)) = A^{-1}(\{y \in \mathbb{R}^2 : y_2 > 0\} \cup \{0\}) = \{x \in \mathbb{R}^3 : x_2 + x_3 > 0\} \cup \mathcal{N}(A)$, qui n'est pas fermé.

2.27. [$(i) \Rightarrow (ii)$] D'après l'exercice 2.26, $x = P_K(z)$ si (1) $\langle x - z, x\rangle = 0$ ou encore $\langle y, x\rangle = 0$, ce qui est supposé dans (i) et si (2) pour tout $u \in K$: $\langle x - z, u\rangle \geqslant 0$ ou encore $\langle y, u\rangle \leqslant 0$, ce qui est bien vérifié car $y \in K^-$. De même, $y = P_{K^-}(z)$ si (1) $\langle y - z, y\rangle = 0$ ou encore $\langle x, y\rangle = 0$, ce qui est supposé dans (i) et si (2) pour tout $u \in K^-$: $\langle y - z, u\rangle \geqslant 0$ ou encore $\langle x, u\rangle \leqslant 0$, ce qui est bien vérifié car $x \in K$.

[$(ii) \Rightarrow (i)$] D'après l'exercice 2.26, $x = P_K(z)$ vérifie $\langle x - z, x\rangle = 0$ et $\langle x - z, u\rangle \geqslant 0$ pour tout $u \in K$. Si on pose $g := z - x$, on a $\langle g, x\rangle = 0$ et $\langle g, u\rangle \leqslant 0$ pour tout $u \in K$. On en déduit que (1) $g \in K^-$, (2) $0 = -\langle x, g\rangle = \langle g - z, g\rangle$ et finalement (3) $\forall v \in K^-$: $\langle g - z, v\rangle = -\langle x, v\rangle \geqslant 0$, car $x \in K$. D'après l'exercice 2.26, ceci implique que $g = P_{K^-}(z)$, qui n'est autre que y. On a donc bien $z = x + y$, avec $x \in K$ et $y \in K^-$. De plus $\langle x - z, x\rangle = 0$ montre que $\langle x, y\rangle = 0$.

L'implication $\Rightarrow$ dans (2.52) se déduit aisément de l'implication $(ii) \Rightarrow (i)$. L'implication $\Rightarrow$ dans (2.52) se déduit aisément de l'implication $(i) \Rightarrow (ii)$, puisque l'on peut écrire $z = 0 + z$ avec $0 \in K$ et $z \in K^-$, donc $P_K(z) = 0$.

2.28. Observons que $A_p := V\Lambda^+V^\mathsf{T} \in \mathcal{S}_+^n$. Comme $\mathcal{S}_+^n$ est un cône convexe fermé non vide, il suffit de vérifier que les conditions (2.51) données dans l'exercice 2.26 sont vérifiées. On note $\Lambda^- := \Lambda^+ - \Lambda$, qui est une matrice diagonale aux éléments positifs. D'une part, on a $\langle A_p - A, A_p\rangle = \langle V\Lambda^-V^\mathsf{T}, V\Lambda^+V^\mathsf{T}\rangle = \mathrm{tr}(V\Lambda^-\Lambda^+V^\mathsf{T}) = 0$, car $\Lambda^-\Lambda^+ = 0$. D'autre part, quel que soit $B \in \mathcal{S}_+^n$, on a $\langle A_p - A, B\rangle = \langle V\Lambda^-V^\mathsf{T}, B\rangle \geqslant 0$, car $V\Lambda^-V^\mathsf{T} \succcurlyeq 0$ et $B \succcurlyeq 0$.

2.29. Soient $\mathbb{E}_1$ et $\mathbb{E}_2$ les sous-espaces vectoriels parallèles aux sous-espaces affines. Pour $i = 1, 2$, on peut trouver des éléments $a_i \in A_i$ tels que $A_i = a_i + \mathbb{E}_i$. Par définition de $\bar{x}$, on a $\langle x - \bar{x}, a_1 + x_1 + a_2 + x_2 - \bar{x}\rangle = 0$ pour tout $x_i \in \mathbb{E}_i$. Mais $\bar{x} \in A$ est de la forme $a_1 + \bar{x}_1 + a_2 + \bar{x}_2$ avec des $\bar{x}_i \in \mathbb{E}_i$. Dès lors, pour tout $x_i \in \mathbb{E}_i$, $\langle x - \bar{x}, (x_1 - \bar{x}_1) + (x_2 - \bar{x}_2)\rangle = 0$. On en déduit que $(x - \bar{x}) \perp \mathbb{E}_i$.

2.30. Soient $x \in \mathbb{E}$, $\bar{x}_\mathcal{A} := P_\mathcal{A}x$ et $\bar{x}_C := P_C(\bar{x}_\mathcal{A})$. Il suffit de montrer que $\bar{x}_C = P_C(x)$ ou encore que

$$\forall y \in C : \quad \langle y - \bar{x}_C, \bar{x}_C - x\rangle \geqslant 0. \tag{D.11}$$

Soit donc $y \in C$. Comme $\bar{x}_C$ est la projection de $\bar{x}_\mathcal{A}$ sur C on a

$$\langle y - \bar{x}_C, \bar{x}_C - \bar{x}_\mathcal{A}\rangle \geqslant 0. \tag{D.12}$$

Par ailleurs, la projection $\bar{x}_\mathcal{A}$ vérifie $\langle z - \bar{x}_\mathcal{A}, \bar{x}_\mathcal{A} - x\rangle = 0$, $\forall z \in \mathcal{A}$. En prenant $z = y - \bar{x}_C + \bar{x}_\mathcal{A} \in \mathcal{A}$ (car y, $\bar{x}_C \in C$ et $\bar{x}_\mathcal{A}$ sont tous dans $\mathcal{A}$ et la somme des coefficients vaut 1) dans cette relation, que l'on ajoute à (D.12), on obtient (D.11).

2.31. L'ensemble $C := P_1 - P_2$ étant un convexe fermé non vide d'un espace euclidien, il existe un unique point $\bar{x} \in C$ minimisant la distance de 0 à C (proposition 2.25). Comme $x \in C$ si, et seulement si, $x = x_1 - x_2$ avec des $x_i \in P_i$, $\bar{x}$ vérifie

$$\|\bar{x}\| = \inf_{x\in C} \|x\| = \inf_{\substack{x_1\in P_2\\ x_2\in P_2}} \|x_1 - x_2\|.$$

En particulier $\bar{x} = \bar{x}_1 - \bar{x}_2$ et ces $\bar{x}_i$ (non nécessairement uniques) sont ceux cherchés.

2.32. Montrons d'abord que $\operatorname{aff} P = \operatorname{aff}(\operatorname{co} P)$. L'inclusion « $\subseteq$ » vient de $P \subseteq \operatorname{co} P$. L'inclusion inverse vient de la proposition 2.3 et de ce qu'une combinaison affine de combinaisons convexes est une combinaison affine.

Soit à présent P une partie relativement ouverte de $\mathbb{E}$ et montrons que $P \subseteq (\operatorname{co} P)^{\diamond}$. Soit $z \in P$. Comme P est relativement ouvert, il existe $r > 0$ tel que $(B(z,r) \cap \operatorname{aff} P) \subseteq P \subseteq \operatorname{co} P$; dès lors $(B(z,r) \cap \operatorname{aff}(\operatorname{co} P)) \subseteq \operatorname{co} P$, ce qui signifie que $z \in (\operatorname{co} P)^{\diamond}$.

Du résultat précédent, on déduit que $\operatorname{co} P \subseteq (\operatorname{co} P)^{\diamond}$ (car $(\operatorname{co} P)^{\diamond}$ est convexe, proposition 2.16). Comme l'inclusion inverse est claire, il s'en ensuit que $\operatorname{co} P = (\operatorname{co} P)^{\diamond}$.

2.33. Comme $X \subseteq \overline{\operatorname{co}}\, X$, on a $\inf\{\langle c,x\rangle : x \in X\} \geqslant \inf\{\langle c,x\rangle : x \in \overline{\operatorname{co}}\, X\}$. Si l'inégalité est stricte, on peut trouver $x_0 \in \overline{\operatorname{co}}\, X$ tel que $\langle c,x_0\rangle < \alpha := \inf\{\langle c,x\rangle : x \in X\}$, ce qui signifie que $H := \{x \in \mathbb{E} : \langle c,x\rangle \geqslant \alpha\}$ est un demi-espace fermé qui contient X et que $x_0 \notin H$. Ceci est en contradiction avec le fait que $x_0 \in \overline{\operatorname{co}}\, X$ (car $\overline{\operatorname{co}}\, X$ est l'intersection de tous les demi-espaces fermés qui contiennent X, proposition 2.40). On a démontré (2.53).

Par ailleurs, si $\bar{x} \in \arg\min\{\langle c,x\rangle : x \in X\}$, alors $H := \{x \in \mathbb{E} : \langle c,x\rangle \geqslant \langle c,\bar{x}\rangle\}$ est un demi-espace fermé qui contient X, donc qui contient $\overline{\operatorname{co}}\, X$, ce qui implique que $\bar{x} \in \arg\min\{\langle c,x\rangle : x \in \overline{\operatorname{co}}\, X\}$. On a montré que $S_1 := \arg\min\{\langle c,x\rangle : x \in X\}$ est contenu dans $S_2 := \arg\min\{\langle c,x\rangle : x \in \overline{\operatorname{co}}\, X\}$. Comme S_2 est l'intersection des convexes fermés $\overline{\operatorname{co}}\, X$ et $\{x \in \mathbb{E} : \langle c,x\rangle = \langle c,\bar{x}\rangle\}$, c'est un convexe fermé. Dès lors $\overline{\operatorname{co}}\, S_1 \subseteq S_2$. Ceci démontre (2.54).

Si X est compact et si $\bar{x} \in S_2$, alors $\bar{x} \in \overline{\operatorname{co}}\, X = \operatorname{co} X$, qui est un ensemble compact (corollaire 2.4). Donc $\bar{x}$ peut s'écrire comme une combinaison convexe d'un nombre fini de points $\bar{x}_i$ de X : $\bar{x} = \sum_i \alpha_i \bar{x}_i$ avec $\alpha \in \Delta$ (simplexe unité). Comme le coût est linéaire, $\langle c,\bar{x}\rangle = \sum_i \alpha_i \langle c,\bar{x}_i\rangle$, ce qui montre que $\langle c,\bar{x}_i\rangle = \langle c,\bar{x}\rangle$ (ou $\bar{x}_i \in S_1$) pour tout i tel que $\alpha_i > 0$. Dès lors $\bar{x} \in \operatorname{co} S_1 \subseteq \overline{\operatorname{co}}\, S_1$.

On n'a pas nécessairement l'égalité en (2.54) si X n'est pas compact. Par exemple, si $X = \{x \in \mathbb{R}^2 : x_1 x_2 \geqslant 1\} \cup \{(0,0)\}$ et $c = e^2$, on a $\overline{\operatorname{co}}\, X = \mathbb{R}^2_+$, $S_1 = \{(0,0)\}$ et $S_2 = \mathbb{R}_+ e^1$.

2.34. 1) Comme $C - x$ est convexe (somme de deux convexes), l'expression (2.45) de $\mathrm{T}^a_x C = \mathbb{R}_+(C-x)$ montre que cet ensemble est convexe (point 1 de l'exercice 2.1) et qu'il contient 0 (car $x \in C$). Le cône tangent $\mathrm{T}_x C$ est un cône convexe comme adhérence d'un cône convexe (exercice 2.14 et point 2 de la proposition 2.16).

2) On a $\operatorname{aff} \mathrm{T}_x C = \operatorname{aff}(\operatorname{adh} \mathrm{T}^a_x C) = \operatorname{aff} \mathrm{T}^a_x C$ [$\mathrm{T}^a_x C$ est convexe et point 2 de la proposition 2.16] $= \operatorname{aff}(C-x)$ [$0 \in C - x$ et point 5 de l'exercice 2.2] $= (\operatorname{aff} C) - x$ [point 3 de l'exercice 2.2].

3) [$(\mathrm{T}_x C)^{\diamond} = (\mathrm{T}^a_x C)^{\diamond}$] En effet, on a $(\mathrm{T}_x C)^{\diamond} = (\operatorname{adh} \mathrm{T}^a_x C)^{\diamond}$ [(2.46)] $= (\mathrm{T}^a_x C)^{\diamond}$ [point 3 de l'exercice 2.13].

[$(\mathrm{T}_x C)^{\diamond} \subseteq R_{++}(C^{\diamond} - x)$] On note $\bar{B}$ la boule unité fermée de $\mathbb{E}$. Soit $d \in (\mathrm{T}_x C)^{\diamond}$; il s'agit de montrer que $x + \mathbb{R}_{++} d$ rencontre $C^{\diamond}$. Si ce n'est pas le cas, on peut séparer ces deux convexes [théorème 2.38] : il existe un $\xi \in \mathbb{E} \setminus \{0\}$ tel que $\langle \xi, x + \alpha d\rangle \leqslant \langle \xi, x'\rangle$ pour tout $\alpha > 0$ et pour tout $x' \in C^{\diamond}$ et donc aussi pour tout $x' \in C$, car $\operatorname{adh}(C^{\diamond}) = \operatorname{adh} C \supseteq C$. On peut supposer que $\xi \in \operatorname{vect}(C - x) = \operatorname{vect} \mathrm{T}_x C$. Par ailleurs, comme $d \in (\mathrm{T}_x C)^{\diamond}$, il existe un $\varepsilon > 0$ tel que $d + \varepsilon(\bar{B} \cap \operatorname{vect} \mathrm{T}_x C) \subseteq (\mathrm{T}_x C)^{\diamond}$. On introduit $d' := d - \varepsilon \xi / \|\xi\|$, qui est donc dans $(\mathrm{T}_x C)^{\diamond} \subseteq \mathrm{T}^a_x C$ [première étape]. Il existe donc un $x' \in C$ de la forme $x' = x + \alpha d'$ avec $\alpha > 0$. Alors l'inégalité de séparation

donne pour le $\alpha > 0$ et le $x' \in C$ trouvés : $\langle \xi, x + \alpha d\rangle \leqslant \langle \xi, x'\rangle = \langle \xi, x + \alpha d'\rangle = \langle \xi, x + \alpha d - \alpha\varepsilon\xi/\|\xi\|\rangle = \langle \xi, x + \alpha d\rangle - \alpha\varepsilon\|\xi\|$, ce qui contredit $\xi \neq 0$.

$[(\mathrm{T}_x^a C)^\ominus \supseteq R_{++}(C^\ominus - x)]$ Comme $(\mathrm{T}_x^a C)^\ominus$ est un cône (exercice 2.14), il suffit de montrer que $(\mathrm{T}_x^a C)^\ominus \supseteq C^\ominus - x$. Soit $y \in C^\ominus$. Pour montrer que $y - x \in (\mathrm{T}_x^a C)^\ominus$, il suffit que pour une direction $d \in \mathrm{T}_x^a C$ arbitraire on peut trouver un $t > 1$ tel que $(1-t)d + t(y-x) \in \mathrm{T}_x^a C$ (lemme 2.14). En fait, puisque $x + d \in \operatorname{aff} C$ [point 2] et $y \in C^\ominus$, il existe un $t > 1$ tel que $(1-t)(x+d)+ty \in C$. Dès lors $[(1-t)(x+d)+ty]-x \in \mathrm{T}_x^a C$ ou encore $(1-t)d + t(y-x) \in \mathrm{T}_x^a C$.

4) On a les équivalences

$$\begin{aligned} \nu \in \mathrm{N}_x C &\iff \langle \nu, d\rangle \leqslant 0, \quad \forall\, d \in C - x \\ &\iff \langle \nu, d\rangle \leqslant 0, \quad \forall\, d \in \mathbb{R}_+(C - x) \\ &\iff \langle \nu, d\rangle \leqslant 0, \quad \forall\, d \in \mathrm{T}_x\, C \qquad [(2.46)] \\ &\iff \nu \in (\mathrm{T}_x\, C)^-. \end{aligned}$$

Dès lors, $\mathrm{N}_x C = (\mathrm{T}_x\, C)^-$. On en déduit que $(\mathrm{N}_x C)^- = (\mathrm{T}_x\, C)^{--} = \mathrm{T}_x\, C$, car $\mathrm{T}_x\, C$ est un cône convexe fermé [point 1 et point 5 de la proposition 2.43].

2.35. 1) Clair en testant les éléments de $(\mathbb{R}^n_+)^+$ sur les vecteurs de base de $\mathbb{R}^n$.

2) Soient $K := \{(x,z) \in \mathbb{R}^n \times \mathbb{R} : \|x\|_2 \leqslant z\}$ et $(d,e) \in K^+$. En testant (d,e) sur $(-d, \|d\|_2) \in K$, on obtient $\|d\|_2 \leqslant e$, si bien que $(d,e) \in K$; on a montré que $K^+ \subseteq K$. Inversement, soient (x,z) et $(d,e) \in K$. Par Cauchy-Schwarz, $-x^\mathsf{T} d \leqslant \|x\|_2\, \|d\|_2 \leqslant ze$. Donc $(x,z)^\mathsf{T}(d,e) \geqslant 0$, ce qui implique que $K \subseteq K^+$.

3) Soit $D \in (\mathcal{S}^n_+)^+$. Pour tout $v \in \mathbb{R}^n$, $vv^\mathsf{T} \in \mathcal{S}^n_+$, si bien que $0 \leqslant \langle D, vv^\mathsf{T}\rangle = v^\mathsf{T} D v$. Comme v est arbitraire, on en déduit que $D \in \mathcal{S}^n_+$; donc $(\mathcal{S}^n_+)^+ \subseteq \mathcal{S}^n_+$. Inversement, si $D \in \mathcal{S}^n_+$ et $A \in \mathcal{S}^n_+$, on a nécessairement $A = \sum_{i=1}^n \lambda_i v_i v_i^\mathsf{T}$ (factorisation spectrale) avec des $\lambda_i \geqslant 0$. Dès lors, $\langle D, A\rangle = \sum_{i=1}^n \lambda_i v_i^\mathsf{T} D v_i \geqslant 0$, car $D \in \mathcal{S}^n_+$. On en déduit que $\mathcal{S}^n_+ \subseteq (\mathcal{S}^n_+)^+$.

4) Soit $d \in (\mathbb{R}^n_\leqslant)^+$. En prenant $x \in \mathbb{R}^n_\leqslant$ tel que $x_1 = \cdots = x_j < 0$ et $x_{j+1} = \cdots = x_n = 0$ $(j < n)$ on trouve que $\sum_{i=1}^j d_i \leqslant 0$. En prenant $x \in \mathbb{R}^n_\leqslant$ tel que $x_1 = \cdots = x_n$ quelconque dans $\mathbb{R}$ on trouve que $\sum_{i=1}^n d_i = 0$. Inversement soient $x \in \mathbb{R}^n_\leqslant$ et $d \in \mathbb{R}^n$ vérifiant $\sum_{i=1}^j d_i \leqslant 0$, pour $j = 1, \ldots, n-1$, et $\sum_{i=1}^n d_i = 0$. En utilisant successivement $(\sum_{i=1}^j d_i)x_j \geqslant (\sum_{i=1}^j d_i)x_{j+1}$, pour $j = 1, \ldots, n-1$, puis $\sum_{i=1}^n d_i = 0$, on trouve que $d^\mathsf{T} x \geqslant 0$.

5) On peut écrire $\mathcal{C}^n = \cap_{x \geqslant 0}\{M \in \mathcal{S}^n : x^\mathsf{T} M x \geqslant 0\}$, si bien que $\mathcal{C}^n$ est une intersection de demi-espaces fermés ; donc un convexe fermé. C'est aussi clairement un cône non vide.

L'ensemble $\mathcal{C}^{n+}$ est clairement un cône non vide. Par ailleurs $\mathcal{C}^{n+} = \mathrm{co}\{xx^\mathsf{T} : x \in \mathbb{R}^n_+\}$ (d'une part $BB^\mathsf{T} = \sum_i b_i b_i^\mathsf{T} = \sum_i \frac{1}{m}(\sqrt{m}b_i)(\sqrt{m}b_i)^\mathsf{T}$ où m est le nombre de colonnes de B et les $b_i \geqslant 0$ sont les colonnes de B ; d'autre part pour $\alpha \in \Delta$, $\sum_i \alpha_i x_i x_i^\mathsf{T} = BB^\mathsf{T}$, où la i-ième colonne de B est $\sqrt{\alpha_i}x_i \geqslant 0$), ce qui montre que $\mathcal{C}^{n+}$ est convexe. Montrons maintenant que, dans la représentation d'une matrice $M \in \mathcal{C}^{n+}$ par $M = BB^\mathsf{T}$, on peut supposer que B est de type $n \times [n(n+1)/2+1]$: comme on peut prendre des colonnes de B nulles, il suffit de montrer qu'il faut *au plus* $n(n+1)/2+1$ vecteurs x_i pour représenter M par $\sum_i x_i x_i^\mathsf{T}$, ce qui découle en fait du théorème 2.3 de Carathéodory et du fait que la dimension de $\mathcal{S}^n$ est $n(n+1)/2$. Il reste à montrer que $\mathcal{C}^{n+}$ est fermé : supposons qu'une suite $\{M_k\} \subseteq \mathcal{C}^{n+}$ converge vers M et montrons que $M \in \mathcal{C}^{n+}$; on a $M_k = B_k B_k^\mathsf{T}$ avec des $B_k \geqslant 0$; d'abord, $\{B_k\}$ est bonée, car si $x_k \in \arg\max\{\|B_k^\mathsf{T} x\|_2 : \|x\|_2 = 1\}$, on a $\|B_k\|_2^2 = \|B_k^\mathsf{T}\|_2^2 =$

$x_k^{\mathsf{T}} B_k B_k^{\mathsf{T}} x_k \leqslant \sup\{x^{\mathsf{T}} M_k x : \|x\|_2 = 1\} = \|M_k\|_2$ [car $M_k \in \mathcal{S}_+^n$] ; en extrayant une sous-suite convergente des $B_k \to B$, on voit que la même sous-suite de M_k converge vers $BB^{\mathsf{T}} = M$; comme $B \geqslant 0$, $M \in \mathcal{C}^{n+}$.

Montrons maintenant que $\mathcal{C}^n \subseteq (\mathcal{C}^{n+})^+$. Si $M \in \mathcal{C}^n$ et $BB^{\mathsf{T}} \in \mathcal{C}^{n+}$ (avec $B \geqslant 0$ dont les colonnes sont notées b_i), on a en effet $\langle M, BB^{\mathsf{T}}\rangle = \operatorname{tr} B^{\mathsf{T}} M B = \sum_i b_i^{\mathsf{T}} M b_i \geqslant 0$, car les $b_i \geqslant 0$. Inversement, $(\mathcal{C}^{n+})^+ \subseteq \mathcal{C}^n$, car si $M \in (\mathcal{C}^{n+})^+$, alors pour tout $xx^{\mathsf{T}} \in \mathcal{C}^{n+}$ (avec $x \in \mathbb{R}_+^n$), on a $0 \leqslant \langle M, xx^{\mathsf{T}}\rangle = x^{\mathsf{T}} M x$, c'est-à-dire que $M \in \mathcal{C}^n$. On a donc montré que $(\mathcal{C}^{n+})^+ = \mathcal{C}^n$. En prenant le dual de chaque membre et en utilisant le fait que $\mathcal{C}^{n+}$ est un cône convexe fermé, on voit que $(\mathcal{C}^n)^+ = \mathcal{C}^{n+}$.

Les inclusions dans $\mathcal{C}^{n+} \subseteq \mathcal{S}_+^n \subseteq \mathcal{C}^n$ sont claires.

2.36. 1) Si K_1 et K_2 sont des cônes non vides, alors $0 \in \operatorname{adh}(K_1) \cap \operatorname{adh}(K_2)$, si bien que le point 6 de la proposition 2.43 conduit au résultat.

2) Si les K_i sont des cônes convexes fermés non vide, on a $K_i^{+\,+} = K_i$ (point 5 de la proposition 2.43). En remplaçant K_i par K_i^+ dans l'identité du point 1, on trouve $(K_1^+ + K^+)^+ = K_1 \cap K_2$ et donc $(K_1 \cap K_2)^+ = (K_1^+ + K_2^+)^{++} = \operatorname{adh}(K_1^+ + K_2^+)$ (point 4 de la proposition 2.43).

2.37. [(2.56) $\Rightarrow$] Si $d \in \operatorname{intr} P^+$, il existe un $\varepsilon > 0$ tel que, pour tout $u \in B \cap \operatorname{aff} P^+$ (B désigne la boule unité de $\mathbb{E}$), on a $d + \varepsilon u \in P^+$. Alors pour tout $x \in P \setminus \{0\}$ et tout $u \in B \cap \operatorname{aff} P^+$: $\langle d + \varepsilon u, x\rangle \geqslant 0$. En prenant $u = -P_{\operatorname{aff} P^+} x / \|P_{\operatorname{aff} P^+} x\| \in B \cap \operatorname{aff} P^+$, on obtient pour tout $x \in P \setminus \{0\}$: $\langle d, x\rangle \geqslant \varepsilon \langle P_{\operatorname{aff} P^+} x, x\rangle / \|P_{\operatorname{aff} P^+} x\| = \varepsilon \|P_{\operatorname{aff} P^+} x\|$ (évidemment, si $x = 0$, l'inégalité a aussi lieu).

[(2.56) $\Leftarrow$] Clairement, $d \in P^+$. Il suffit donc de montrer que si $u \in B \cap \operatorname{aff} P^+$, alors $d + \varepsilon u \in P^+$. Pour $x \in P$, on a certainement que $\langle u, x\rangle = \langle P_{\operatorname{aff} P^+} u, x\rangle = \langle u, P_{\operatorname{aff} P^+} x\rangle \geqslant -\|P_{\operatorname{aff} P^+} x\|$ et donc, en tenant compte de l'affirmation de droite dans (2.56), que $\langle d + \varepsilon u, x\rangle \geqslant \varepsilon \|P_{\operatorname{aff} P^+} x\| + \varepsilon \langle u, x\rangle \geqslant 0$, qui est l'inégalité recherchée.

[(2.55) $\Rightarrow$] Se déduit de « (2.56) $\Rightarrow$ », car si $d \in \operatorname{int} P^+$, alors P^+ est d'intérieur non vide et donc $\operatorname{aff} P^+ = \mathbb{E}$.

[(2.55) $\Leftarrow$] Pour tout $u \in B$ et tout $x \in P$ on a certainement que $\langle u, x\rangle \geqslant -\|x\|$ et donc, en tenant compte de l'affirmation de droite dans (2.55), que $\langle d + \varepsilon u, x\rangle \geqslant \varepsilon \|x\| + \varepsilon \langle u, x\rangle \geqslant 0$. Donc $d + \varepsilon B \subseteq P^+$, si bien que d est intérieur à P^+.

2.38. 1) On considère l'application linéaire $B : (x, y) \in \mathbb{E} \times \mathbb{F} \mapsto Ax + y \in \mathbb{F}$ et l'on munit $\mathbb{E} \times \mathbb{F}$ du produit scalaire $((x, y), (x', y')) \mapsto \langle x, x'\rangle + \langle y, y'\rangle$. Alors, l'adjointe de B est l'application $B^* : z \in \mathbb{F} \mapsto (A^* z, z) \in \mathbb{E} \times \mathbb{F}$. Selon le lemme de Farkas, $\overline{B(K \times L)} = \{y \in \mathbb{F} : B^*(y) \in (K \times L)^+\}^+$. Il reste à observer que $B(K \times L) = A(K) + L$, que $(K \times L)^+ = K^+ \times L^+$, et que $B^*(y) \in (K \times L)^+$ s'écrit $A^+ y \in K^+$ et $y \in L^+$.

2) $\{Ax : x \geqslant x_0\} = Ax_0 + \{A(x - x_0) : x - x_0 \geqslant 0\} = Ax_0 + \{y : A^{\mathsf{T}} y \geqslant 0\}^+$.

2.39. 1) (i) revient à dire que $b \in \{Ax : x \geqslant 0\}$. D'après le lemme de Farkas, ce dernier ensemble est $\{y : A^{\mathsf{T}} y \geqslant 0\}^+$, si bien que (i) revient à dire que $\forall y$ tel que $A^{\mathsf{T}} y \geqslant 0$ on a $b^{\mathsf{T}} y \geqslant 0$, ce qui est l'affirmation opposée de (ii).

2) Par homogénéité, le point (i) est équivalent à l'existence de x tel que $Ax = 0$, $Bx \leqslant 0$ et $Cx \leqslant -e$ et donc à l'existence de x, $y \geqslant 0$ et $z \geqslant 0$ tels que $Ax = 0$, $Bx + y = 0$, $Cx + z = -e$, ce qui s'exprime matriciellement par

$$\begin{pmatrix} 0 \\ 0 \\ -e \end{pmatrix} \in \begin{pmatrix} A & 0 & 0 \\ B & I & 0 \\ C & 0 & I \end{pmatrix} (\mathbb{R}^n \times \mathbb{R}_+^{m_B} \times \mathbb{R}_+^{m_C}).$$

Par le lemme de Farkas, cette dernière condition équivaut à dire que le produit scalaire $(0,0,-e)^\mathsf{T}(\alpha,\beta,\gamma) = -e^\mathsf{T}\gamma$ est positif pour tout (α,β,γ) vérifiant

$$\begin{pmatrix} A^\mathsf{T} & B^\mathsf{T} & C^\mathsf{T} \\ 0 & I & 0 \\ 0 & 0 & I \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} \in (\mathbb{R}^n \times \mathbb{R}^{m_B}_+ \times \mathbb{R}^{m_C}_+)^+ = \{0\} \times \mathbb{R}^{m_B}_+ \times \mathbb{R}^{m_C}_+,$$

c'est-à-dire vérifiant $A^\mathsf{T}\alpha + B^\mathsf{T}\beta + C^\mathsf{T}\gamma = 0$, $\beta \geqslant 0$ et $\gamma \geqslant 0$. On a donc démontré que (i) est équivalent au fait que $A^\mathsf{T}\alpha + B^\mathsf{T}\beta + C^\mathsf{T}\gamma = 0$, $\beta \geqslant 0$ et $\gamma \geqslant 0$ impliquent que $\gamma = 0$, ce qui est bien l'oppposé de (ii).

3) La démonstration suit celle de l'alternative de Motzkin homogène. Le point (i) est équivalent à l'existence de x et de $\varepsilon > 0$ tels que $Ax = a$, $Bx \leqslant b$ et $Cx \leqslant c - \varepsilon e$ et donc à l'existence de x, $\varepsilon > 0$, $y \geqslant 0$ et $z \geqslant 0$ tels que $Ax = a$, $Bx + y = b$, $Cx + z = c - \varepsilon e$, ce qui s'exprime matriciellement par

$$\exists\, \varepsilon > 0 : \quad \begin{pmatrix} a \\ b \\ c - \varepsilon e \end{pmatrix} \in \begin{pmatrix} A & 0 & 0 \\ B & I & 0 \\ C & 0 & I \end{pmatrix} (\mathbb{R}^n \times \mathbb{R}^{m_B}_+ \times \mathbb{R}^{m_C}_+).$$

Par le lemme de Farkas, cette dernière condition équivaut à affirmer l'existence de $\varepsilon > 0$ tel que le produit scalaire $(a,b,c-\varepsilon e)^\mathsf{T}(\alpha,\beta,\gamma)$ est positif pour tout (α,β,γ) vérifiant

$$\begin{pmatrix} A^\mathsf{T} & B^\mathsf{T} & C^\mathsf{T} \\ 0 & I & 0 \\ 0 & 0 & I \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} \in (\mathbb{R}^n \times \mathbb{R}^{m_B}_+ \times \mathbb{R}^{m_C}_+)^+ = \{0\} \times \mathbb{R}^{m_B}_+ \times \mathbb{R}^{m_C}_+,$$

c'est-à-dire qui est dans

$$\Omega := \{(\alpha,\beta,\gamma) : A^\mathsf{T}\alpha + B^\mathsf{T}\beta + C^\mathsf{T}\gamma = 0,\ \beta \geqslant 0,\ \gamma \geqslant 0\}.$$

On a donc démontré que (i) est équivalent à

$$\exists\, \varepsilon > 0 : \quad (\alpha,\beta,\gamma) \in \Omega \quad \Longrightarrow \quad a^\mathsf{T}\alpha + b^\mathsf{T}\beta + c^\mathsf{T}\gamma \geqslant \varepsilon\, e^\mathsf{T}\gamma. \tag{D.13}$$

Il reste à montrer que (D.13) est équivalent à l'opposé de (ii).

- [(D.13) $\Rightarrow$ non (ii)] Supposons que l'on ait $(\alpha,\beta,\gamma,\gamma_0) \in \mathbb{R}^{m_A} \times \mathbb{R}^{m_B}_+ \times \mathbb{R}^{m_C}_+ \times \mathbb{R}_+$ tel que $A^\mathsf{T}\alpha + B^\mathsf{T}\beta + C^\mathsf{T}\gamma = 0$ et $a^\mathsf{T}\alpha + b^\mathsf{T}\beta + c^\mathsf{T}\gamma + \gamma_0 = 0$. Alors $(\alpha,\beta,\gamma) \in \Omega$ et par (D.13) on obtient $\gamma_0 + \varepsilon\, e^\mathsf{T}\gamma \leqslant 0$, d'où on déduit que $\gamma_0 = 0$ et $\gamma = 0$, c'est-à-dire non (ii).
- [(D.13) $\Leftarrow$ non (ii)] Si $(\alpha,\beta,\gamma) \in \Omega$, on a $-\gamma_0 := a^\mathsf{T}\alpha + b^\mathsf{T}\beta + c^\mathsf{T}\gamma \geqslant 0$ (sinon, $\gamma_0 > 0$ et donc (ii) serait vérifié). Mais il faut montrer un peu plus pour avoir (D.13), en particulier lorsque $\gamma \neq 0$. Mais lorsque $(\alpha,\beta,\gamma) \in \Omega$ et $\gamma \neq 0$, on a $-\gamma_0 := a^\mathsf{T}\alpha + b^\mathsf{T}\beta + c^\mathsf{T}\gamma > 0$ (sinon, $\gamma_0 \geqslant 0$, $(\gamma_0,\gamma) \neq 0$ et donc (ii) serait vérifié). Alors par compacité de $\Omega \cap \bar{B}_1$, $\varepsilon := \min(1, \inf\{a^\mathsf{T}\alpha + b^\mathsf{T}\beta + c^\mathsf{T}\gamma : (\alpha,\beta,\gamma) \in \Omega$ et $\|\gamma\|_1 = 1\})$ est strictement positif. On en déduit (D.13).

4) En prenant $x + te$ avec $t > 0$ petit et $e := (1\ \cdots\ 1)^\mathsf{T} \in \mathbb{R}^n$, on voit que (i) revient à dire qu'il existe un $x > 0$ tel que $Ax > 0$ ou encore, qu'il existe un x tel que $(A^\mathsf{T}\ I)^\mathsf{T}x > 0$. Par l'alternative de Gordon, l'affirmation « non (i) » est alors équivalente au fait qu'il existe $(y,z) \geqslant 0$ non nul tel que $A^\mathsf{T}y + z = 0$ ou encore qu'il existe $y \geqslant 0$ non nul tel que $A^\mathsf{T}y \leqslant 0$.

5.a) L'existence d'un x tel que $Ax \leqslant b$ revient à montrer que

$$\begin{aligned} b &\in \{Ax_1 - Ax_2 + s : (x_1, x_2, s) \geqslant 0\} \\ &= \left\{ y : \begin{pmatrix} A^\mathsf{T} \\ -A^\mathsf{T} \\ I \end{pmatrix} y \geqslant 0 \right\}^+ \\ &= \{y : A^\mathsf{T} y = 0, y \geqslant 0\}^+. \end{aligned}$$

5.b) L'existence d'un x tel que $a \leqslant Ax \leqslant b$ revient à montrer que

$$\begin{aligned} (a, b) &\in \{(Ax_1 - Ax_2 - s_1, Ax_1 - Ax_2 + s_2) : (x_1, x_2, s_1, s_2) \geqslant 0\} \\ &= \left\{ (y, z) : \begin{pmatrix} A^\mathsf{T} & A^\mathsf{T} \\ -A^\mathsf{T} & -A^\mathsf{T} \\ -I & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix} \geqslant 0 \right\}^+ \\ &= \{(y, z) : A^\mathsf{T}(y + z) = 0, y \leqslant 0 \leqslant z\}^+. \end{aligned}$$

On obtient le résultat par la substitution $y \to -y$.

2.40. Si $Ax \leqslant b$ n'a pas de solution, il existe $y \geqslant 0$ tel que $A^\mathsf{T} y = 0$ et $\beta := b^\mathsf{T} y < 0$ (lemme de Farkas, voir aussi le théorème de l'alternative à l'exercice 2.39). On note $C = (A\ b)$ et $c = (0, \beta)$. Alors y appartient au polyèdre convexe $\{y : C^\mathsf{T} y = c \text{ et } y \geqslant 0\}$, qui est donc non vide. Soit $\hat{y}$ un sommet de ce polyèdre (corollaire 2.23) et $I \subseteq \{1, \ldots, m\}$ l'ensemble des indices de ses composantes non nulles. On a $|I| \leqslant n+1$, qui est le nombre de lignes de C^T (proposition 2.21). Évidemment $\hat{y}_I$ vérifie $\hat{y}_I \geqslant 0$, $A_I^\mathsf{T} \hat{y}_I = 0$ et $b^\mathsf{T} \hat{y}_I = \beta < 0$. Donc, par le lemme de Farkas, $A_I x \leqslant b_I$ n'a pas de solution.

2.41. xxx

2.42. On doit montrer que $\forall\, \varepsilon > 0$, $\exists\, \delta > 0$ tel que $\|x' - x\| < \delta$ et $d' \in \mathrm{N}_C(x') \cap S$ impliquent que $\mathrm{dist}(d', \mathrm{N}_C(x) \cap S) < \varepsilon$. Montrons cela par l'absurde. Si ce n'est pas le cas, il existerait un $\varepsilon > 0$ et des suites $\{x_k\} \to x$ et des directions $d_k \in \mathrm{N}_C(x_k) \cap S$ telles que $\mathrm{dist}(d_k, \mathrm{N}_C(x) \cap S) \geqslant \varepsilon$. En extrayant une sous-suite au besoin, on peut upooser que $d_k \to d \in S$. Puisque $d_k \in \mathrm{N}_C(x_k)$, on a pour tout $x' \in C$: $\langle x' - x_k, d_k \rangle \leqslant 0$. En passant à la limite, il vient que, pour tout $x' \in C$, on a $\langle x' - x, d \rangle \leqslant 0$, ce qui implique que $d \in \mathrm{N}_C(x) \cap S$. Par ailleurs, en prenant la limite dans $\mathrm{dist}(d_k, \mathrm{N}_C(x) \cap S) \geqslant \varepsilon$, on trouve $\mathrm{dist}(d, \mathrm{N}_C(x) \cap S) \geqslant \varepsilon$, ce qui est en contradiction avec $d \in \mathrm{N}_C(x) \cap S$.

2.43. D'après (2.46), il suffit de montrer que $\mathrm{T}_x P \subseteq \mathrm{T}_x^a P$. Soit $d \in \mathrm{T}_x P$. Par définition, il existe $d_k \to d$ et $t_k \downarrow 0$ tels que $x_k := x + t_k d_k \in P$. Soit $\bar{t}_k := \max\{t : A(x + t d_k) \leqslant b,\ t \leqslant 1\}$. Clairement $\bar{t}_k > 0$. Montrons que $\bar{t}_k$ reste uniformément positif. Dans le cas contraire, il existe une sous-suite d'indices k telle que $\bar{t}_k \downarrow 0$ et un indice i tel que $[A(x + \bar{t}_k d_k)]_i = b_i$ et $[A(x + 2\bar{t}_k d_k)]_i > b_i$ (comme le nombre d'indices est fini, on peut prendre i indépendant de k). En passant à la limite dans la dernière égalité, on voit que $(Ax)_i = b_i$ et donc que $(Ad_k)_i = 0$, ce qui contredit la dernière inégalité. Puisque $\bar{t}_k$ est dans un compact de $]0, 1]$, on peut extraire une sous-suite d'indices k tels que $\bar{t}_k \to \bar{t} > 0$. En passant à la limite dans $A(x + \bar{t}_k d_k) \leqslant b$, on trouve $A(x + \bar{t} d) \leqslant b$, si bien que $d \in \mathrm{T}_x^a P$.

Montrons à présent que $\mathrm{T}_x P = \{d \in \mathbb{R}^n : (Ad)_I \leqslant 0\}$. Si $d \in \mathrm{T}_x P$, il existe $t > 0$ tel que $A(x + td) \leqslant b$. Pour les $i \in I$, on obtient $(Ad)_i \leqslant 0$. Inversement, soit $d \in \mathbb{R}^n$ tel que $(Ad)_I \leqslant 0$. Pour $i \in I$, on a $(A(x + td) - b)_i = t(Ad)_i \leqslant 0$ pour tout $t > 0$. Pour $i \notin I$, on a $(Ax - b)_i < 0$, et donc pour $t > 0$ petit, on a aussi $(A(x + td) - b)_i \leqslant 0$.

Le cône normal s'obtient par le lemme de Farkas : $\mathrm{N}_x P = -(\mathrm{T}_x P)^+ = -\{-A_I^\mathsf{T} y : y \geqslant 0\} = \mathrm{cone}\{A_i^\mathsf{T} : i \in I\}$.

2.44. Soit $D \in \mathrm{T}_S\, \mathcal{S}_+^n$. Alors il existe une suite $\{S_k\} \subseteq \mathcal{S}_+^n$, une suite $\{t_k\} \downarrow 0$ telles que $(S_k - S)/t_k \to D$. Pour $v \in \mathcal{N}(S)$, on a alors $v^\mathsf{T}(S_k - S)v/t_k = v^\mathsf{T} S_k v/t_k \geqslant 0$, qui à la limite donne $v^\mathsf{T} D v \geqslant 0$.

Pour l'inclusion réciproque, on introduit

$$\begin{array}{rcl} T & := & \{D \in \mathcal{S}^n : v^\mathsf{T} D v \geqslant 0, \text{ pour tout } v \in \mathcal{N}(S)\} \\ T_0 & := & \{D \in \mathcal{S}^n : v^\mathsf{T} D v > 0, \text{ pour tout } v \in \mathcal{N}(S) \setminus \{0\}\}. \end{array}$$

Observons d'abord que $T_0 \subseteq \mathrm{T}_S\, \mathcal{S}_+^n$: en effet, si $D \in T_0$, $D + (1/t)S \succ 0$ pour $t > 0$ assez petit (lemme B.3 de Finsler) et donc $S + tD \succ 0$ pour $t > 0$ assez petit, ce qui implique que $D \in \mathrm{T}_S\, \mathcal{S}_+^n$. Par ailleurs, comme $\mathrm{T}_S\, \mathcal{S}_+^n$ est fermé, $\mathrm{adh}\, T_0 \subseteq \mathrm{T}_S\, \mathcal{S}_+^n$; de plus $T \subseteq \mathrm{adh}\, T_0$ car si $D \in T$, $D_k := D + t_k I \in T_0$ lorsque $t_k > 0$ et $D_k \to D$ lorsque $t_k \downarrow 0$; donc $T \subseteq \mathrm{T}_S\, \mathcal{S}_+^n$.

$[N_S \mathcal{S}_+^n \subseteq \{N \in \mathcal{S}_-^n : \langle S, N\rangle = 0\}]$ Soit $N \in N_S \mathcal{S}_+^n$. Alors, $\langle X - S, N\rangle \leqslant 0$ pour tout $X \in \mathcal{S}_+^n$. En prenant $X = 0$ et $X = 2S$, on obtient $\langle S, N\rangle = 0$. Alors, on a aussi $\langle X, N\rangle \leqslant 0$ pour tout $X \in \mathcal{S}_+^n$; donc $N \in \mathcal{S}_-^n$ (point 1 du lemme 20.1).

$[N_S \mathcal{S}_+^n \supseteq \{N \in \mathcal{S}_-^n : \langle S, N\rangle = 0\}]$ Soit $N \preccurlyeq 0$ telle que $\langle S, N\rangle = 0$. Alors, pour tout $X \in \mathcal{S}_+^n$, $\langle X - S, N\rangle \leqslant 0$, donc $N \in N_S \mathcal{S}_+^n$.

Fonctions convexes

3.1. 1) $[\subseteq]$ Clairement $\mathrm{epi}\, f \subseteq (\mathrm{dom}\, f) \times \mathbb{R}$. L'inclusion s'obtient alors en prenant l'enveloppe affine des deux membres et en utilisant le point 8 de l'exercice 2.2.

$[\supseteq]$ Soit $(x, \alpha) := (\sum_i t_i x_i, \alpha) \in (\mathrm{aff}(\mathrm{dom}\, f)) \times \mathbb{R}$, donc avec un nombre fini de $t_i \in \mathbb{R}$ tels que $\sum_i t_i = 1$, des $x_i \in \mathrm{dom}\, f$ et $\alpha \in \mathbb{R}$. Il faut montrer que $(x, \alpha) \in \mathrm{aff}(\mathrm{epi}\, f)$. Pour tout i, posons $s_i := f(x_i) + 1 - \alpha$. Alors, $\alpha = s_i f(x_i) + (1 - s_i)(f(x_i) + 1)$ et on a

$$\begin{array}{rcl} (x, \alpha) & = & \sum_i t_i (x_i, \alpha) \\ & = & \sum_i t_i \big(s_i x_i + (1 - s_i)x_i, s_i f(x_i) + (1 - s_i)(f(x_i) + 1)\big) \\ & = & \sum_i t_i s_i (x_i, f(x_i)) + \sum_i t_i (1 - s_i)(x_i, f(x_i) + 1), \end{array}$$

qui est bien dans $\mathrm{aff}(\mathrm{epi}\, f)$ puisque $(x_i, f(x_i))$ et $(x_i, f(x_i) + 1)$ sont dans $\mathrm{epi}\, f$ et que $\sum_i t_i s_i + \sum_i t_i (1 - s_i) = \sum_i t_i = 1$.

2) $[\Rightarrow]$ Soient (x, α) et $(y, \beta) \in \mathrm{epi}_s\, f$ et $t \in]0, 1[$. On a $f((1-t)x + ty) \leqslant (1-t)f(x) + tf(y)$ (par convexité de f) $< (1-t)\alpha + t\beta$ (car $(1-t) > 0$, $t > 0$, $f(x) < \alpha$ et $f(y) < \beta$). Donc $(1-t)(x, \alpha) + t(y, \beta) \in \mathrm{epi}_s\, f$.

$[\Leftarrow]$ Soient (x, α) et $(y, \beta) \in \mathrm{epi}\, f$ et $\varepsilon > 0$. Alors $(x, \alpha + \varepsilon)$ et $(y, \beta + \varepsilon) \in \mathrm{epi}_s\, f$ et donc $(1-t)(x, \alpha + \varepsilon) + t(y, \beta + \varepsilon) \in \mathrm{epi}_s\, f$. Ceci se traduit par $f((1-t)x + ty) < (1-t)\alpha + t\beta + \varepsilon$. Comme $\varepsilon > 0$ est arbitraire, on a $f((1-t)x + ty) \leqslant (1-t)\alpha + t\beta$ et donc $(1-t)(x, \alpha) + t(y, \beta) \in \mathrm{epi}\, f$.

3) Si on restreint f à son domaine, on voit qu'il s'agit de montrer que $(\mathrm{epi}\, f)^\circ = \{(x, \alpha) : x \in (\mathrm{dom}\, f)^\circ,\ \alpha > f(x)\}$. L'inclusion $\subseteq$ étant claire, montrons donc l'inclusion inverse. Soient $\bar{x} \in (\mathrm{dom}\, f)^\circ$ et $\bar{\alpha} > f(\bar{x})$. On peut alors trouver des points $x_1, \ldots, x_n \in \mathrm{dom}\, f$ tels que $P := \mathrm{co}\{x_1, \ldots, x_n\}$ contient $\bar{x}$ dans son intérieur. Soit $\beta := \max\{f(x_i) : 1 \leqslant i \leqslant n\}$. Alors $\Omega := P^\circ \times]\beta, +\infty[$ est un ouvert contenu dans

epi f [en effet, si $(x',\beta')\in\Omega$, on a $x'=\sum_{i=1}^n t_i x_i$, avec $(t_1,\ldots,t_n)\in\Delta_n$, et donc $f(x')\leqslant\sum_{i=1}^n t_i f(x_i)\leqslant\beta\leqslant\beta'$]. En particulier, $(\bar x,\beta+1)\in(\operatorname{epi} f)^\circ$. Alors, comme il existe un $t:=(\beta+1-f(\bar x))/(\beta+1-\bar\alpha)>1$ tel que $(1-t)(\bar x,\beta+1)+t(\bar x,\bar\alpha)\in\operatorname{epi} f$, on en déduit que $(\bar x,\bar\alpha)\in(\operatorname{epi} f)^\circ$. [517 ; lemme 7.3]

3.2. [$\Rightarrow$] Supposons que $t>1$ (on s'y prend de la même manière si $t<0$). Si $x_t:=(1-t)x+ty\notin\operatorname{dom} f$ l'inégalité est démontrée. Dans le cas contraire, x et $x_t\in\operatorname{dom} f$, $s=1/t\in[0,1]$ et $y=(1-s)x+sx_t$, si bien que par convexité de f :

$$f(y)\leqslant(1-s)f(x)+sf(x_t)=(1-\tfrac1t)f(x)+\tfrac1t f((1-t)x+ty).$$

Après multiplication par $t>0$, on obtient l'inégalité de (3.73).

[$\Leftarrow$] Soient x, $y\in\operatorname{dom} f$ et $s\in\,]0,1[$. Il suffit de montrer que

$$f((1-s)x+sy)\leqslant(1-s)f(x)+sf(y). \tag{D.14}$$

Posons $x_s:=(1-s)x+sy$. Pour $t:=1/s>1$, on a $y=(1-t)x+tx_s$. Alors (3.73) donne

$$f(y)\geqslant(1-t)f(x)+tf(x_s)=(1-\tfrac1s)f(x)+\tfrac1s f((1-s)x+sy).$$

Après multiplication par $s>0$, on obtient (D.14).

3.3. Supposons qu'il existe $\hat x\in\mathbb{E}$ tel que $f(\hat x)=-\infty$. Donc $\hat x\in\operatorname{dom} f$ et il existe $t>0$ tel que $x_t:=x+t(x-\hat x)\in\operatorname{dom} f$. On peut donc trouver un $r\in\mathbb{R}$ tel que $(x_t,r)\in\operatorname{epi} f$. D'autre part, $(\hat x,\hat r)\in\operatorname{epi} f$, pour tout $\hat r\in\mathbb{R}$. Par convexité de l'épigraphe, en utilisant le fait que $x=\frac{1}{1+t}x_t+\frac{t}{1+t}\hat x$, on voit que $(x,\frac{1}{1+t}r+\frac{t}{1+t}\hat r)\in\operatorname{epi} f$, pour tout $\hat r\in\mathbb{R}$. Ceci implique que $f(x)=-\infty$, ce qui contredit l'hypothèse.

3.4. 1) La condition est clairement suffisante (proposition 3.2). Le fait que la condition est nécessaire peut se montrer par récurrence sur m. Elle est vraie pour $m=2$. Soit alors $t=\sum_{i=2}^m t_i$. Si $t=0$ la condition est triviale. Traitons le cas où $t\in\,]0,1]$. Par convexité de $\operatorname{dom} f$, $y:=\sum_{i=1}^m(t_i/t)x_i\in\operatorname{dom} f$. D'autre part, comme $t_1+t=1$, $f(\sum_{i=1}^m t_i x_i)=f(t_1x_1+ty)\leqslant t_1f(x_1)+tf(y)$ [par convexité de f] $\leqslant\sum_{i=1}^m t_i f(x_i)$ [par l'hypothèse de récurrence].

2) La condition est nécessaire. Soient F et μ choisis tels que $\mu(\Omega)=1$ et $x:=\int_\Omega F\,\mathrm{d}\mu\in\operatorname{dom} f$. Cas 1. Si f est sous-différentiable en x et $g\in\partial f(x)$, on a pour tout $z\in\Omega$: $f(F(z))\geqslant f(x)+\langle g,F(z)-x\rangle$. En intégrant cette inégalité on trouve (3.75) (l'intégrale du second terme est nulle). Cas 2. Si f n'est pas sous-différentiable en x, alors x est sur la frontière relative de $C_0:=\operatorname{dom} f$ (proposition 3.62). On peut supposer que $\{z\in\operatorname{supp}\mu : F(z)\notin C_0\}$ est de mesure nulle sinon $\int_\Omega(f\circ F)\,\mathrm{d}\mu=+\infty$ et il n'y a rien à démontrer. Soit alors $\nu\in\mathrm{N}_x(C_0)\cap\operatorname{aff} C_0$ qui est non vide (proposition 2.29). Montrons que, pour presque tout $z\in\Omega$, $F(z)$ est dans le convexe $C_1:=(x+\nu^\perp)\cap(\partial C_0)$, dont la dimension est strictement plus petite que celle de C_0. Comme on a supposé que $F(z)\in C_0$, on a nécessairement $\langle\nu,F(z)\rangle\leqslant\langle\nu,x\rangle$. Dès lors si la propriété que l'on veut démontrer n'était pas vérifiée, l'ensemble $\Omega_1:=\{z\in\Omega:\langle\nu,F(z)\rangle<\langle\nu,x\rangle\}$ serait de mesure non nulle et

$$\begin{aligned}
\langle\nu,x\rangle &= \int_\Omega\langle\nu,F(z)\rangle\,\mathrm{d}\mu(z)\quad\text{[linéarité de l'intégrale]}\\
&= \int_{\Omega_1}\underbrace{\langle\nu,F(z)\rangle}_{<\langle\nu,x\rangle}\mathrm{d}\mu(z)+\int_{\Omega\setminus\Omega_1}\underbrace{\langle\nu,F(z)\rangle}_{=\langle\nu,x\rangle}\mathrm{d}\mu(z)\\
&< \langle\nu,x\rangle\mu(\Omega_1)+\langle\nu,x\rangle\mu(\Omega\setminus\Omega_1)\quad\text{[235 ; page 109]}\\
&= \langle\nu,x\rangle,
\end{aligned}$$

qui est la contradiction attendue. Si la restriction f_1 de f à C_1 est sous-différentiable en x, le résultat se démontre comme dans le Cas 1. Sinon, x est sur la frontière relative de C_1 et, par le même argument que dans le Cas 2, on en déduit que, pour presque tout $z \in \operatorname{supp}\mu$, $F(z)$ est inclus dans une partie convexe C_2 de la frontière relative de C_1 de dimension srtrictement plus petite que C_1. La récurrence ne peut se poursuivre indéfiniment car on finit par montrer que, pour presque tout $z \in \operatorname{supp}\mu$, $F(z) = x$, auquel cas les deux membres de (3.75) valent tous deux $f(x)$ et le résultat est démontré.

La condition est suffisante. Soient x et $y \in \operatorname{dom} f$ et $t \in [0,1]$. On prend $\Omega = \mathbb{E}$, $\mathcal{A}$ l'ensemble des boréliens de $\mathbb{E}$, $\mu = (1-t)\delta_x + t\delta_y$ (δ_x est la mesure de Dirac en x) qui est une mesure positive sur $\mathbb{E}$ vérifiant $\mu(\mathbb{E}) = (1-t)\delta_x(\mathbb{E}) + t\delta_y(\mathbb{E}) = (1-t) + t = 1$, et $F = \mathrm{I}_{\mathbb{E}}$ qui vérifie $\int_\Omega \mathrm{I}_{\mathbb{E}}\,\mathrm{d}\mu = (1-t)\int_\Omega \mathrm{I}_{\mathbb{E}}\,\mathrm{d}\delta_x + t\int_\Omega \mathrm{I}_{\mathbb{E}}\,\mathrm{d}\delta_y = (1-t)x + ty \in \operatorname{dom} f$. Alors, en utilisant (3.75), $f((1-t)x+ty) = f(\int_\Omega \mathrm{I}_{\mathbb{E}}\,\mathrm{d}\mu) \leqslant \int_\Omega (f\circ \mathrm{I}_{\mathbb{E}})\,\mathrm{d}\mu = (1-t)f(x) + tf(y)$, ce qui montre la convexité de f.

3) Soit $x_t := (1-t)x_0 + tx_1$. Le point x_1 est combinaison convexe de x_t et x_0 : $x_1 = (1/t)x_t + ((t-1)/t)x_0$. Dès lors $f(x_1) \leqslant (1/t)f(x_t) + ((t-1)/t)f(x_0)$, qui est le résultat attendu.

4) Comme le logarithme est croissant, cela revient à montrer que $\frac{1}{n}(\sum_{i=1}^n \log a_i) \leqslant \log(\frac{1}{n}\sum_{i=1}^n a_i)$, ce qui se déduit de la concavité du logarithme.

5) En appliquant l'inégalité de convexité $\varphi(a) \geqslant \varphi(b^{p'/p}) + \varphi'(b^{p'/p})(a - b^{p'/p})$ à la fonction convexe $a \in \mathbb{R}_+ \mapsto \varphi(a) := \frac{1}{p}a^p$, on obtient $\frac{1}{p}a^p \geqslant \frac{1}{p}b^{p'} + b(a - b^{p'/p})$. Après réarrangement, on obtient l'inégalité souhaitée. Du fait de la stricte convexité de φ, on aura égalité ssi $a = b^{p'/p}$, c'est-à-dire ssi $a^p = b^{p'}$.

3.5. Soit $x \in \mathbb{E}$ tel que $a(x) \in \operatorname{dom} f$ et $d \in \mathbb{E}$. Alors $x \in \operatorname{dom}(f\circ a)$ et donc $(f\circ a)^\infty(d) = \lim_{t\to\infty}(f\circ a)(x+td)/t = \lim_{t\to\infty} f((a(x)+tAd)/t = f^\infty(Ad)$.

3.6. Démontrons d'abord l'identité sur les épigraphes strictes. Soit $(x,\alpha) \in \operatorname{epi}_s(f \uplus g)$. Alors $\alpha > (f \uplus g)(x) = \inf\{f(y) + g(z) : x = y + z\}$ et donc il existe y et z tels que $\alpha > f(y)+g(z)$. On en déduit qu'il existe $\beta > f(y)$ et $\gamma > g(z)$ tels que $\alpha = \beta+\gamma$ (par exemple, on pose $\varepsilon := \alpha - f(y) - g(z) > 0$ et on prend $\beta = f(y)+\varepsilon/2$ et $\gamma = g(z)+\varepsilon/2$). Alors $(x,\alpha) = (y,\beta) + (z,\gamma)$ montre que $(x,\alpha) \in \operatorname{epi}_s f + \operatorname{epi}_s g$. Réciproquement, si $(x,\alpha) \in \operatorname{epi}_s f + \operatorname{epi}_s g$, on a $(x,\alpha) = (y,\beta) + (z,\gamma)$, avec $(y,\beta) \in \operatorname{epi}_s f$ et $(z,\gamma) \in \operatorname{epi}_s g$. Donc $\alpha = \beta + \gamma > f(y) + g(z) \geqslant \inf\{f(y') + g(z') : x = y' + z'\} = (f \uplus g)(x)$, c'est-à-dire $(x,\alpha) \in \operatorname{epi}_s(f \uplus g)$.

Les points 1–4 de la proposition 3.36 sont des conséquences immédiates de l'identité (3.26). Démontrons le point 5. Le domaine de $f \uplus g$ n'est pas vide si ceux de f et g ne le sont pas (par le point 1). D'autre part, si $x \mapsto \ell(x) + \alpha$ est une minorante affine commune à f et g, on a $f(y) + g(x-y) \geqslant \ell(x) + 2\alpha$, si bien que $\ell(\cdot) + 2\alpha$ est une minorante affine de $f \uplus g$.

Les propriétés (1) et (2) se déduisent directement de la définition (3.25), tandis que pour (3) on utilise (3.26).

3.7. La fonction A ne peut pas être convexe, car elle a des ensembles de sous-niveau qui ne sont pas convexes, alors que ceux d'une fonction convexe sont convexes : si f est convexe, si x, $y \in N_\nu(f) := \{z : f(z) \leqslant \nu\}$ et si $t \in [0,1]$, alors $(1-t)x + ty \in N_\nu(f)$ puisque

$$f((1-t)x + ty) \leqslant (1-t)f(x) + tf(y) \leqslant (1-t)\nu + t\nu = \nu.$$

La fonction B peut être convexe. Les ensembles de sous-niveau de la norme ℓ_1 ont cette allure et une norme est convexe (exercice 3.31).

La fonction C ne peut pas être convexe, car l'« écart » entre les courbes 2 et 1 est plus petit que celui entre les courbes 3 et 2. De manière plus précise, on peut supposer que l'origine est au centre des cercles (translation de l'origine si nécessaire) et prendre des points x_1, x_2 et x_3 le long d'un même rayon et vérifiant $f(x_i) = i$ pour tout i. On a $x_2 = (1-t)x_1 + tx_3$ avec $t < \frac{1}{2}$. Si f était convexe, on aurait $2 = f(x_2) \leqslant (1-t)f(x_1) + tf(x_3) = 1 + 2t < 2$, ce qui est absurde.

3.8. 1)

2)

3) Soit K le cône asymptotique commun aux sous-ensembles de niveau non vide de f.

Si $d \in K$, quels que soient $x \in \mathbb{R}^n$ et $t > 0$, on a $f(x+td) - f(x) \leqslant 0$ et donc $t(g+Hx)^\mathsf{T} d + (t^2/2) d^\mathsf{T} H d \leqslant 0$. En divisant cette inégalité par t^2 et en prenant la limite quand $t \to \infty$, on obtient $d^\mathsf{T} H d \leqslant 0$; alors la semi-définie positivité de H implique que $d \in \mathcal{N}(H)$. En réinjectant $Hd = 0$ dans l'inégalité, on obtient $g^\mathsf{T} d \leqslant 0$.

Inversement, si $d \in \mathcal{N}(H)$ vérifie $g^\mathsf{T} d \leqslant 0$, quels que soient $x \in \mathbb{R}^n$ et $t \geqslant 0$, on a : $f(x+td) - f(x) = t(g+Hx)^\mathsf{T} d + (t^2/2) d^\mathsf{T} H d = t g^\mathsf{T} d \leqslant 0$, si bien que $d \in K$.

3.9. On a

$$\begin{aligned}(f \boxplus g)(x) &= \inf_y \langle Ax, x\rangle + \langle B(x-y), x-y\rangle \\ &= \inf_y \langle (A+B)y, y\rangle - 2\langle Bx, y\rangle + \langle Ax, x\rangle \\ &= \begin{cases} \langle Ay_x, x\rangle & \text{si } A+B \succcurlyeq 0 \text{ et } Bx \in \mathcal{R}(A+B) \\ -\infty & \text{sinon,} \end{cases}\end{aligned}$$

où y_x est une solution arbitraire de $(A+B)y = Bx$. Comme celle-ci peut être prise comme une fonction linéaire de x, $f \boxplus g$ est quadratique. Si A et B sont inversibles et si $A+B \succ 0$, on trouve que $(f \boxplus g)(x) = \langle (A^{-1}+B^{-1})^{-1} x, x\rangle$.

3.10. On peut supposer que $I(x) = I := [1\,{:}\,m]$ car, par continuité, les f_i avec $i \notin I$ n'affectent pas $f_{\max}$ dans le voisinage de x. D'une part, le fait que pour tout $i \in I$ on ait $f(x+td) \geqslant f_i(x+td)$ et $f(x) = f_i(x)$, permet d'écrire $\liminf_{t \to 0^+} [f(x+td) - f(x)]/t \geqslant \max_i f_i'(x;d)$. Il reste à montrer que $\limsup_{t \downarrow 0} [f(x+td) - f(x)]/t \leqslant \max_i f_i'(x;d)$. Si ce n'est pas le cas, il existe une suite $\{t_k\} \downarrow 0$ et $\varepsilon > 0$ tels que $[f(x+t_k d) - f(x)]/t_k \geqslant \max_i f_i'(x;d) + \varepsilon$. Comme I est fini, on peut trouver une sous-suite des indices précédents et un indice $j \in I$ tels que $f(x+t_k d) = f_j(x+t_k d)$. On déduirait alors, en passant à la limite dans l'inégalité précédente que $f_j'(x;d) \geqslant \max_i f_i'(x;d) + \varepsilon$, une contradiction. [78 ; 2000; proposition 2.3.2]

3.11. xxx

3.12. 1) Soit $x \in X$. Comme $F(X)$ est ouvert et $0 \notin F(X)$, on peut trouver $\bar{y} \in F(X)$ tel que $\|\bar{y}\| < \|F(x)\|$. On a $\bar{y} = F(\bar{x})$, avec un $\bar{x} \in X$. Donc x n'est pas solution. Ce point étant arbitraire, le problème ne peut pas avoir de solution.

2) Si $\|x_k\| \not\to \infty$, il existe une sous-suite de $\{x_k\}$ convergeant vers un point $\bar{x} \in \mathbb{R}^n$. On ne peut pas avoir $\bar{x} \in X$, car $\bar{x}$ serait solution du problème ($x \mapsto \|F(x)\|$ est continue et $\{x_k\}$ est minimisante), donc $\liminf \mathrm{d}_{X^c}(x_k) = 0$.

3.13. Soit $F_* = F(x_*)$ la face de C engendrée par x_*. Si f n'était pas constante sur F_*, il existerait un $x_0 \in F_*$ avec $f(x_0) > f(x_*)$. Alors, avec $x_t := (1-t)x_0 + tx_*$ et $t > 1$, on

aurait par la concavité de f : $f(x_t) \leqslant (1-t)f(x_0)+tf(x_*) < f(x_*)$, ce qui montrerait que $x_t \notin F_*$. Donc x_* serait sur la frontière relative de F_*, ce qui contredirait le fait que F_* est la face engendrée par x_* (exercice 2.13, point 6).

Pour la seconde partie, il suffit de se rappeler (exercice 2.16, point 3) que si C est fermé et a un point extrême, la face F_* en a également un. Ce dernier point extrême est alors solution du probleme.

3.14. 1) Le point $x_* \in X$ est Pareto optimal ssi il n'existe pas de $x \in X$ tel que $f(x) < f(x_*)$, c'est-à-dire ssi il n'existe pas de $x \in X$ et de $p \in \mathbb{R}^m_{++}$ tels que $f(x) = f(x_*)-p$, c'est-à-dire ssi $(f(x_*) - \mathbb{R}^m_{++}) \cap f(X) = \varnothing$.

2) Pour déterminer l'image par f des points Pareto optimaux, on utilise la caractérisation du numéro 1 (représentée à gauche à la figure D.1). En ce qui concerne

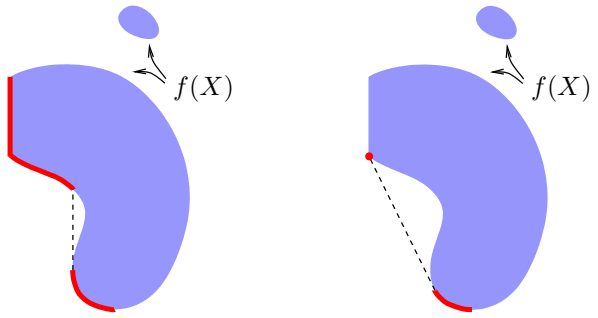


Fig. D.1. Image par f des points Pareto optimaux (g) et moyennement optimaux (d)

les points moyennement optimaux, on observe que $r^\mathsf{T} f$ est constant dans l'espace image le long des droites $\{(f_1, f_2) : r_1 f_1 + r_2 f_2 = \gamma \text{ (une constante)}\}$. L'image par f des points $x \in X$ rendant $r^\mathsf{T} f(x)$ constant sont donc les intersections de ces droites avec $f(X)$. L'image par f des points $x \in X$ minimisant $r^\mathsf{T} f(\cdot)$ est donc, si elle existe, l'intersection avec $f(X)$ de la droite $r^\mathsf{T} f = \gamma$ avec la plus petite constante γ (représentée à droite à la figure D.1).

3) Si x_* n'est pas Pareto optimal, il existe un $x \in X$ tel que $f(x) < f(x_*)$. Alors, avec $r \geqslant 0$ non nul, on a $\sum_i r_i(f_i(x) - f_i(x_*)) < 0$ et donc $r^\mathsf{T} f(x_*) > r^\mathsf{T} f(x)$, ce qui contredit le fait que x_* minimise $r^\mathsf{T} f$.

4) Observons d'abord que $(f(x_*)-\mathbb{R}^m_{++})\cap(f(X)+\mathbb{R}^m_+) = \varnothing$, car si on pouvait trouver un $x \in X$ et un $p \geqslant 0$ tels que $f(x) + p \in f(x_*) - \mathbb{R}^m_{++}$, alors $f(x) - f(x_*) < 0$, ce qui contredirait la Pareto optimalité de x_*.

On sépare les deux convexes disjoints $f(x_*) - \mathbb{R}^m_{++}$ et $f(X) + \mathbb{R}^m_+$ dans $\mathbb{R}^m$: il existe $r \in \mathbb{R}^m$ non nul tel que pour tout $p \in \mathbb{R}^m_{++}$ et pour tout $x \in X$ on a $r^\mathsf{T}(f(x_*) - p) \leqslant r^\mathsf{T} f(x)$ (on a choisit l'élément nul dans $\mathbb{R}^m_+$). On prend p de la forme tp, avec $p \in \mathbb{R}^m_{++}$. Lorsque $t \to 0$, on trouve que x_* minimise $r^\mathsf{T} f$. Lorsque $t \to +\infty$, on trouve que $r^\mathsf{T} p \geqslant 0$ quel que soit $p \in \mathbb{R}^m_{++}$, donc $r \geqslant 0$.

5) Soient $f(x) + p$ et $f(x') + p' \in f(X) + \mathbb{R}^m_+$, avec x, $x' \in X$ et p, $p' \in \mathbb{R}^m_+$. Par la convexité des f_i, on a l'inégalité vectorielle $(1-t)f(x) + tf(x') \geqslant f((1-t)x + tx')$. On en déduit $(1-t)(f(x) + p) + t(f(x') + p') \geqslant f((1-t)x + tx')$, donc $(1-t)(f(x) + p) + t(f(x') + p') \in f(X) + \mathbb{R}^m_+$.

$f(\mathbb{R})$ n'est pas convexe dans $\mathbb{R}^2$ si $f_1(x) = x$ et $f_2(x) = \frac{1}{2}x^2$.

3.15. 1) Soient $f = f_1 + f_2$ la somme de deux fonctions convexes polyédriques. On les écrit sous une forme duale : $f_i = \max_{1\leqslant j\leqslant k_i} \mathcal{A}_{ij} + \mathcal{I}_{P_i}$, où les $\mathcal{A}_{ij}(x) = \langle \xi_{ij}, x\rangle + t_{ij}$ définissent des fonctions affines et les $P_i = \operatorname{dom} f_i$ sont des polyèdres convexes. On a clairement $\mathcal{I}_{P_1} + \mathcal{I}_{P_1} = \mathcal{I}_{P_1\cap P_2}$ et

$$\begin{aligned} f(x) &= \max_j \left(\langle \xi_{1j}, x\rangle + t_{1j}\right) + \mathcal{I}_{P_1} + \max_k \left(\langle \xi_{2k}, x\rangle + t_{2k}\right) + \mathcal{I}_{P_2} \\ &= \max_{j,k} \left(\langle \xi_{1j} + \xi_{2k}, x\rangle + t_{1j} + t_{2k}\right) + \mathcal{I}_{P_1\cap P_2}. \end{aligned}$$

Comme $P_1 \cap P_2$ est polyédrique, f est bien une fonction convexe polyédrique.

2) D'après l'exercice 2.21, $\operatorname{epi} f_1 + \operatorname{epi} f_2$ est un polyèdre convexe (éventuellement vide), donc un fermé. Alors la formule (3.27) implique que $\operatorname{epi} f_1 \uplus f_2 = \operatorname{epi} f_1 + \operatorname{epi} f_2$, si bien que $f_1 \uplus f_2$ est polyédrique. Si $x \in \operatorname{dom}(f_1 \uplus f_2)$, l'infimum dans la formule (3.27) est atteint : $(f \uplus g)(x) = \min\{\alpha : (x,\alpha) \in \operatorname{epi} f + \operatorname{epi} g\}$, ce qui implique que $(f \uplus g)(x)$ s'écrit $f(y) + g(z)$ avec $y + z = x$. Donc l'infimum dans la définition de $(f_1 \uplus f_2)(x)$ est également atteint.

3) Si on utilise la représentation primale (3.4) de la fonction convexe polyédrique f, on obtient :

$$\begin{aligned} f^*(x^*) &= \sup_{x\in\mathbb{E}} \left(\langle x^*, x\rangle - f(x)\right) \\ &= \sup_{\substack{t_i\geqslant 0,\ s_j\geqslant 0 \\ \sum_i t_i = 1}} \left(\sum_{1\leqslant i\leqslant p} t_i(\langle x^*, x_i\rangle - \alpha_i) + \sum_{1\leqslant j\leqslant q} s_j(\langle x^*, y_j\rangle - \beta_j) \right) \\ &= \sup_{\substack{t_i\geqslant 0 \\ \sum_i t_i = 1}} \left(\sum_{1\leqslant i\leqslant p} t_i(\langle x^*, x_i\rangle - \alpha_i) \right) + \sup_{s_j\geqslant 0} \left(\sum_{1\leqslant j\leqslant q} s_j(\langle x^*, y_j\rangle - \beta_j) \right) \\ &= \begin{cases} \max_{1\leqslant i\leqslant p} \left(\langle x^*, x_i\rangle - \alpha_i\right) & \text{si } \langle x^*, y_j\rangle - \beta_j \leqslant 0 \text{ pour tout } j = 1,\dots,q \\ +\infty & \text{sinon.} \end{cases} \end{aligned}$$

C'est la représentation duale (3.5) d'une fonction convexe polyédrique.

3.16. 1)

2)

3)

4) Il est clair que $f_{\inf}$ est propre avec une minorante affine. Pour $x^* \in \mathbb{E}$, on a $f^*_{\inf}(x^*) = \sup_{x\in\mathbb{E}}(\langle x^*, x\rangle - f_{\inf}(x)) = \sup_{x\in\mathbb{E}} \sup_{i\in I}(\langle x^*, x\rangle - f_i(x)) = \sup_{i\in I} \sup_{x\in\mathbb{E}}(\langle x^*, x\rangle - f_i(x)) = \sup_{i\in I} f_i^*(x^*)$.

5) On sait que $f_{\sup} \in \Gamma_0(\mathbb{E})$. Soit $x_0 \in \operatorname{dom} f_{\sup}$; donc $x_0 \in \operatorname{dom} f_i$ pour tout $i \in I$. Comme $f_i \in \Gamma_0(\mathbb{E})$, on peut définir $g_i = f_i^*$ qui est dans $\Gamma_0(\mathbb{E})$. De plus $g_i(x^*) \geqslant \langle x^*, x_0\rangle - f_i(x_0)$, si bien que $x^* \mapsto \langle x^*, x_0\rangle - f_i(x_0)$ est une minorante affine commune à tous les g_i. On peut donc calculer la conjuguée de l'enveloppe inférieure des g_i : $(\inf_{i\in I} g_i)^* = \sup_{i\in I} g_i^* = \sup_{i\in I} f_i$. En conjuguant cette relation, on obtient l'identité recherchée.

3.17. 1) Par le corollaire 3.47, $f^{**} \leqslant f$. Par ailleurs, $f^*(0) = -\inf f$, si bien que, quel que soit $x \in \mathbb{E}$: $f^{**}(x) \geqslant -f^*(0) = \inf f$.

2) Si $\arg\min f = \varnothing$, il n'y a rien à démontrer. Si f prend la valeur $-\infty$, alors $f^* \equiv +\infty$, $f^{**} =\equiv -\infty$ et l'inclusion a trivialement lieu. Si $f \equiv +\infty$, alors $f^* \equiv -\infty$, $f^{**} =\equiv +\infty$ et l'inclusion a trivialement lieu.

On suppose désormais que f est propre et que $\arg\min f \neq \varnothing$. Alors f a une minorante affine (la fonction constante $x \mapsto \inf f$), si bien que $f^{**} \in \mathrm{C}\overline{\mathrm{onv}}(\mathbb{E})$ (proposition 3.45). Comme cela implique que $\arg\min f^{**}$ est un convexe fermé, il suffit de montrer que $\arg\min f \subseteq \arg\min f^{**}$. Soit $\bar{x}$ est un minimum de f. On a $f^{**}(\bar{x}) \leqslant f(\bar{x})$ [corollaire 3.47] $= \inf f = \inf f^{**}$ [point 1]. Donc $\bar{x} \in \arg\min f^{**}$.

L'inclusion sera stricte pour $f : x \in \mathbb{R} \mapsto \sqrt{|x|}$, puisqu'alors $f^* = \mathcal{I}_{\{0\}}$ et $f^{**} \equiv 0$.

3) Si l'égalité $\overline{\mathrm{co}}(\arg\min f) = \arg\min f^{**}$ n'a pas lieu, grâce à l'inclusion du point 2, on peut trouver un point $\bar{x}_0 \in \arg\min f^{**}$ qui n'est pas dans le convexe fermé non vide $\overline{\mathrm{co}}(\arg\min f)$ et, par conséquent (point 3 du corollaire 2.37), un couple $(x_0^*, \alpha) \in \mathbb{E} \times \mathbb{R}$ tel que $\sup\{\langle x_0^*, \bar{x}\rangle : \bar{x} \in \arg\min f\} < \alpha < \langle x_0^*, \bar{x}_0\rangle$. Pour tout $t > 0$, on introduit l'application affine $a_t : x \in \mathbb{E} \mapsto a_t(x) = t(\langle x_0^*, x\rangle - \alpha) + \inf f$. Comme $f^{**}(\bar{x}_0) = \inf f^{**} = \inf f < a_t(\bar{x}_0)$, a_t n'est pas une minorante affine de f (car celles-ci sont des minorantes affines de f^{**}, point 1 de la proposition 3.46). Donc, il existe un point $x_t \in \mathbb{E}$ tel que $f(x_t) < a(x_t)$, qui doit donc être dans le demi-espace fermé $H := \{x \in \mathbb{R} : \langle x_0^*, x\rangle \geqslant \alpha\}$. La suite $\{x_t\}_{t\downarrow 0}$ n'est pas bornée, car dans le cas contraire on pourrait en extraire une sous-suite convergente, disons vers un point $x_0 \in H$, qui vérifierait $f(x_0) \leqslant \liminf f(x_t)$ [f est fermée] $\leqslant \lim a(x_t) = \inf f$; donc on aurait $x_0 \in \arg\min f$, alors que $H \cap \arg\min f = \varnothing$ (une contradiction). On extrait ensuite une sous-suite convergente de $\{u_t\}$, où $u_t := (x_t - \bar{x}_0)/\|x_t - \bar{x}_0\|$, qui converge vers u. Montrons que $u \in (\arg\min f^{**})^\infty$, ce qui aboutira à une contradiction car $u \neq 0$ alors que le cône asymptotique $(\arg\min f^{**})^\infty$ est réduit à $\{0\}$ parce que $\arg\min f^{**}$ est supposé borné (corollaire 2.8). Il s'agit de montrer que $\bar{x}_0 + su \in \arg\min f^{**}$, quel que soit $s > 0$. Considérons les points $y_t := (1 - s/\|x_t - \bar{x}_0\|)\bar{x}_0 + (s/\|x_t - \bar{x}_0\|)x_t = \bar{x}_0 + su_t$, qui convergent vers $\bar{x}_0 + su$. Par convexité de f^{**}, on a $f^{**}(y_t) \leqslant (1 - s/\|x_t - \bar{x}_0\|)f^{**}(\bar{x}_0) + (s/\|x_t - \bar{x}_0\|)f^{**}(x_t)$. Mais $f^{**}(\bar{x}_0) = \inf f < a_t(\bar{x}_0)$ et $f^{**}(x_t) \leqslant f(x_t) < a_t(x_t)$, si bien que $\inf f \leqslant f^{**}(y_t) < a_t(y_t) \to \inf f$. À la limite on trouve donc $f^{**}(\bar{x}_0 + su) = \inf f$ ou $\bar{x}_0 + su \in \arg\min f^{**}$.

3.18. 1) On note $x_t := (1-t)x_0 + tx$. Comme f est s.c.i. et $f^{**} \leqslant f$, on a $f^{**}(x) \leqslant \liminf_{t\uparrow 1} f^{**}(x_t) \leqslant \liminf_{t\uparrow 1} f(x_t)$. Il reste donc à montrer que $f^{**}(x) \geqslant \limsup_{t\uparrow 1} f(x_t)$. Soient $\alpha := f^{**}(x)$ et $\alpha_0 > f(x_0)$. On a $(x, \alpha) \in \operatorname{epi} f^{**} = \operatorname{adh}\operatorname{epi} f$ et $(x_0, \alpha_0) \in (\operatorname{epi} f)^\diamond$ (exercice 3.1). Dès lors, pour $t \in [0, 1[$, $(1-t)(x_0, \alpha_0) + t(x, \alpha) \in (\operatorname{epi} f)^\diamond \subseteq \operatorname{epi}_s f$, ce qui implique $f((1-t)x_0 + tx) < (1-t)\alpha_0 + t\alpha$. On en déduit $\limsup_{t\uparrow 1} f(x_t) \leqslant \alpha = f^{**}(x)$. [517 ; lemme 7.5]

2) Comme $\operatorname{epi} f \subseteq \operatorname{epi} f^{**}$ (corollaire 3.47), on a certainement $\operatorname{dom} f \subseteq \operatorname{dom} f^{**}$ et donc $\operatorname{aff}\operatorname{dom} f \subseteq \operatorname{aff}\operatorname{dom} f^{**}$. Remarquons que

$$x_0 \in (\operatorname{dom} f)^\diamond \text{ et } x \in \operatorname{dom} f^{**} \quad \Longrightarrow \quad [x_0, x[\in (\operatorname{dom} f)^\diamond. \tag{D.15}$$

Ceci est dû au fait que $[x_0, x[\subseteq \operatorname{dom} f$ (d'après le point 1) et donc que $[x_0, x[\subseteq (\operatorname{dom} f)^\diamond$ (par le critère d'intériorité relative du lemme 2.14). On déduit de (D.15) que $\operatorname{dom} f^{**} \subseteq \operatorname{aff}(\operatorname{dom} f)^\diamond \subseteq \operatorname{aff}\operatorname{dom} f$. On a donc démontré que $\operatorname{aff}\operatorname{dom} f = \operatorname{aff}\operatorname{dom} f^{**}$. Mais alors, par (2.12), $(\operatorname{dom} f)^\diamond \subseteq (\operatorname{dom} f^{**})^\diamond$. Réciproquement, si $x_1 \in (\operatorname{dom} f^{**})^\diamond$, alors pour $t > 1$ proche de 1, $x = (1-t)x_0 + tx_1 \in \operatorname{dom} f^{**}$ (car $x_0 \in \operatorname{aff}\operatorname{dom} f^{**}$) et donc $x_1 \in [x_0, x[\subseteq (\operatorname{dom} f)^\diamond$ (par (D.15)). On a donc démontré que $(\operatorname{dom} f)^\diamond = (\operatorname{dom} f^{**})^\diamond$. On déduit de cette dernière identité et du point 4 de l'exercice 2.13 que $\operatorname{adh}\operatorname{dom} f = \operatorname{adh}(\operatorname{dom} f)^\diamond = \operatorname{adh}(\operatorname{dom} f^{**})^\diamond = \operatorname{adh}\operatorname{dom} f^{**}$.

3) On définit $f := f_1 + f_2$. Par hypothèse, le domaine de f est non vide; on prend $x_0 \in (\operatorname{dom} f)^\circ$. D'après le numéro 1, on a pour tout $x \in \mathbb{E}$ et $x_t := (1-t)x_0 + tx$:

$$f^{**}(x) = \lim_{t\uparrow 1} f(x_t) = \lim_{t\uparrow 1} f_1(x_t) + \lim_{t\uparrow 1} f_2(x_t).$$

Si les $f_i \in \overline{\mathrm{Conv}}(\mathbb{E})$, $f_i(x) \leqslant \liminf_{t\uparrow 1} f_i(x_t) \leqslant \limsup_{t\uparrow 1} f_i(x_t) \leqslant \limsup_{t\uparrow 1}(1-t) f_i(x_0) + t f_i(x) = f_i(x)$. Donc $f_i(x_t) \to f(x)$ et $f^{**}(x) = f(x)$, si bien que $f = f^{**} \in \overline{\mathrm{Conv}}(\mathbb{E})$. Si les $(\operatorname{dom} f_i)^\circ$ ont un point en commun, alors $(\operatorname{dom} f)^\circ = (\operatorname{dom} f_2)^\circ \cap (\operatorname{dom} f_2)^\circ$ (proposition 2.17 (3)). Alors, comme $x_0 \in (\operatorname{dom} f_i)^\circ$, $f_i(x_t) \to f_i^{**}(x)$ par le point 1. On en déduit que $f^{**} = f_1^{**} + f_2^{**}$.

3.19. Observons d'abord que $\delta_x = f'(x;\cdot)$ a son domaine non vide ($\delta_x(0) = 0$) et a une minorante affine ($\langle x^*, \cdot\rangle$, où x^* est un sous-gradient arbitraire en x). On peut donc calculer δ_x^*, ce que l'on fait en observant que $\delta_x(d) = \inf_{t>0} \frac{1}{t}(f(x+td) - f(x))$:

$$\begin{aligned}
\delta_x^*(x^*) &= \sup_{d\in\mathbb{E}} \Big[\langle x^*, d\rangle - \inf_{t>0} \frac{1}{t}(f(x+td) - f(x))\Big] \\
&= \sup_{t>0} \frac{1}{t}\Big[\sup_{d\in\mathbb{E}}\Big(\langle x^*, td\rangle - f(x+td)\Big) + f(x)\Big] \\
&= \sup_{t>0} \frac{1}{t}\Big[f^*(x^*) + f(x) - \langle x^*, x\rangle\Big] \\
&= \mathcal{I}_{\partial f(x)}(x^*).
\end{aligned}$$

On en déduit que $\delta_x^{**} = \mathcal{I}^*_{\partial f(x)} = \sigma_{\partial f(x)}$.

3.20. 1) Si $\bar{B}_\varepsilon \subseteq \partial f(\bar{x})$, la formule du max (3.52) donne pour $h \in \mathbb{E}$:

$$f'(\bar{x};h) = \max_{g\in\partial f(\bar{x})} \langle g, h\rangle \geqslant \varepsilon\,\|h\|,$$

en prenant $g = \varepsilon\, h/\|h\| \in \partial f(\bar{x})$. Supposons inversement que $f'(\bar{x};h) \geqslant \varepsilon\|h\|$ pour tout $h \in \mathbb{E}$. Si g est tel que $\|g\| \leqslant \varepsilon$, on a pour tout $h \in \mathbb{E}$: $\langle g, h\rangle \leqslant \varepsilon\|h\| \leqslant f'(\bar{x};h)$. Donc $g \in \partial f(\bar{x})$.

2) Pour tout $g \in \partial f(x)$, on a $\pm f'(x;h) = f'(x;\pm h)$ (par l'hypothèse de différentiabilité) $\geqslant \langle g, \pm h\rangle$ (car $g \in \partial f(x)) = \pm\langle g, h\rangle$. Donc $f'(x;h) = \langle g, h\rangle$, pour tout $g \in \partial f(x)$. On en déduit que $\partial f(x) \subseteq f'(x;h)h/\|h\|^2 + h^\perp$.

3.21. A : S n'est pas convexe et ne peut donc pas être $x + \partial f(x)$.

B : On peut avoir $S = x + \partial f(x)$. Dans ce cas, $x \in x + \partial f(x)$ ou encore $0 \in \partial f(x)$, si bien que x est un minimiseur de f. On a même $0 \in \operatorname{int}\partial f(x)$; dès lors, quel que soit $y \in \mathbb{R}^2$ différent de x, on peut trouver un $\varepsilon > 0$ tel que $\varepsilon(y-x) \in \partial f(x)$; on en déduit que

$$f(y) \geqslant f(x) + \varepsilon(y-x)^{\mathsf{T}}(y-x) > f(x),$$

si bien que x est l'unique minimiseur de f.

C : Étant donné la forme de $\mathcal{C}$, f croît en x lorsque l'on se dirige vers le bas, donc $x + \partial f(x)$ doit être en dessous de $\mathcal{C}$, si bien que $S \neq x + \partial f(x)$. De manière plus précise, supposons que $S = x + \partial f(x)$. Il suffit de montrer que cela conduit à une contradiction. Pour $s \in \partial f(x)$, on peut trouver un $\alpha \in]0,1[$ tel que $x' := x + \alpha s \in \mathcal{C}$. Alors, pour $\varepsilon > 0$ assez petit, $x'' := x + (\alpha + \varepsilon)s$ s'écrit comme une combinaison convexe de deux points de $\mathcal{C}$, donc $f(x'') \leqslant f(x')$. Par ailleurs, par convexité, pour un $t \in]0,1[$, on a $f(x') \leqslant (1-t)f(x'') + tf(x) \leqslant (1-t)f(x') + tf(x)$, si bien que

$f(x') \leqslant f(x)$. Cette inégalité est en contradiction avec $f(x') \geqslant f(x) + \langle s, x' - x\rangle = f(x) + \alpha\|s\|^2 > f(x)$.

D : S peut être $x + \partial f(x)$. Si $S = x + \partial f(x)$, x ne minimise pas f car $x \notin x + \partial f(x)$.

3.23. Soit $(x_1^*, x_2^*) \in \partial f(x_1, x_2)$. Alors, pour tout $(y_1, y_2) \in \mathbb{E}_1 \times \mathbb{E}_2$, on a

$$f(y_1, y_2) \geqslant f(x_1, x_2) + \langle (x_1^*, x_2^*), (y_1, y_2) - (x_1, x_2)\rangle.$$

En prenant $y_2 = x_2$, on a pour tout $y_1 \in \mathbb{E}_1$:

$$f(y_1, x_2) \geqslant f(x_1, x_2) + \langle (x_1^*, x_2^*), (y_1 - x_1, 0)\rangle$$

ou

$$f^1_{x_2}(y_1) \geqslant f^1_{x_2}(x_1) + \langle x_1^*, y_1 - x_1)\rangle.$$

Ceci implique que $x_1^* \in \partial f^1_{x_2}(x_1)$. Pour la même raison, $x_2^* \in \partial f^2_{x_1}(x_2)$. L'inclusion est démontrée.

On n'a pas égalité pour la norme ℓ_2, $(x_1, x_2) \in \mathbb{R}^2 \to (x_1^2 + x_2^2)^{1/2} \in \mathbb{R}$ dont le sous-différentiel en 0 est la boule unité pour cette norme (voir le point 4 de l'exercice 3.30), qui n'est pas un produit d'intervalles.

3.24. [$(i) \Rightarrow (ii)$] Soient $d \in \mathbb{E}$ et $t > 0$. On prend $x = \bar{x} + td$ dans (i), ce qui donne $[f(\bar{x} + td) - f(\bar{x})]/t \geqslant \alpha\|d\|$. En prenant la limite pour $t \downarrow 0$, on trouve (ii).

[$(ii) \Rightarrow (iii)$] Soit $u \in \bar{B}$. Alors pour tout $d \in \mathbb{E}$, on a $f'(\bar{x}; d) \geqslant \alpha\|d\|$ [(ii)] $\geqslant \langle \alpha u, d\rangle$ [Cauchy-Schwarz]. On en déduit que $\alpha u \in \partial f(\bar{x})$ (propriété (i) de la proposition 3.55).

[$(iii) \Rightarrow (i)$] Soit $x \in \mathbb{E}$. Si $x = \bar{x}$, l'inégalité est trivialement vérifiée. Dans le cas contraire, $u = (x - \bar{x})/\|x - \bar{x}\| \in \bar{B}$. Alors $\alpha u \in \partial f(\bar{x})$ par (iii). Par la propriété (ii) de la proposition 3.55, on en déduit que $f(x) \geqslant f(\bar{x}) + \alpha\langle u, x - \bar{x}\rangle = f(\bar{x}) + \alpha\|x - \bar{x}\|$.

3.26. [$\Rightarrow$] Soit en effet $x^* \in \partial f(x)$, supposé non vide. Alors, pour tout $y \in \mathbb{E}$, $f(y) \geqslant f(x) + \langle x^*, y - x\rangle \geqslant f(x) - L\|y - x\|$, où $L := \|x^*\|$.

[$\Leftarrow$] On considère les ensembles convexes

$$C_1 = \operatorname{epi} f, \qquad \text{et} \qquad C_2 = \{(y, \alpha) \in \mathbb{E} \times \mathbb{R} : \alpha < f(x) - L\|y - x\|\}.$$

Par hypothèse, ce sont deux convexes disjoints. On peut donc trouver $(\xi, t) \in \mathbb{E} \times \mathbb{R}$ et $\gamma \in \mathbb{R}$ tels que

$$\forall (y_1, \alpha_1) \in C_1, \quad \forall (y_2, \alpha_2) \in C_2 : \quad \langle \xi, y_1\rangle - t\alpha_1 \leqslant \gamma < \langle \xi, y_2\rangle - t\alpha_2.$$

L'inégalité stricte est justifiée par le fait que C_2 est ouvert (le résultat de séparation de la proposition 2.38 affirme l'existence de $(\xi, t) \neq 0$ tel que $\gamma \leqslant \langle \xi, y_2\rangle - t\alpha_2$ pour tout $(y_2, \alpha_2) \in C_2$; mais si l'on avait l'égalité pour $(y_2, \alpha_2) \in C_2$, on aurait pour $(y_2 - \varepsilon\xi, \alpha_2 + \varepsilon t)$ qui est encore dans C_2 pour $\varepsilon > 0$ petit : $\gamma \leqslant \langle \xi, y_2 - \varepsilon\xi\rangle - t(\alpha_2 + \varepsilon t) = \gamma - \varepsilon(\|\xi\|^2 + t^2)$, ce qui contredirait $(\xi, t) \neq 0$). On exploite ces inégalités comme suit.

- En prenant $(x, f(x)) \in C_1$, on obtient $\langle \xi, x\rangle - tf(x) \leqslant \gamma$. En prenant $(x, f(x) - \varepsilon) \in C_2$, avec $\varepsilon > 0$ arbitraire, on obtient $\gamma \leqslant \langle \xi, x\rangle - tf(x)$. Donc $\gamma = \langle \xi, x\rangle - tf(x)$.
- Alors, en prenant $(x, f(x) - 1) \in C_2$, on obtient maintenant $t > 0$.

En introduisant $x^* := \xi/t$, la première inégalité devient

$$\forall (y_1, \alpha_1) \in C_1 : \quad f(x) + \langle x^*, y_1 - x\rangle \leqslant \alpha_1.$$

Soit $y \in \mathbb{E}$ quelconque. En prenant $(y, f(y)) \in C_1$, on obtient $f(x) + \langle x^*, y - x\rangle \leqslant f(y)$, qui implique que x^* est un sous-gradient de f en x.

3.27. On introduit l'application affine $a : \mathbb{E}_0 \to \mathbb{E} : x \mapsto x_0 + x$, dont l'opérateur associé est l'injection canonique $i : \mathbb{E}_0 \to \mathbb{E}$. Alors $f|_{x_0+\mathbb{E}_0} = f \circ i$. Comme l'adjointe $i^* = P_{\mathbb{E}_0}$, le théorème 3.72 donne (3.58) et les conditions d'égalité qui suivent.

3.28. 1) $\mathcal{I}_P$ est propre car $P \neq \varnothing$. D'autre part, $\operatorname{epi}\mathcal{I}_P = P \times \mathbb{R}_+$ et, comme un produit d'ensembles, est convexe (resp. convexe fermé) ssi chacun d'eux est convexe (resp. convexe fermé). On en déduit que $P \times \mathbb{R}_+$ est convexe (resp. convexe fermé) ssi P est convexe (resp. convexe fermé).

2) On a $\mathcal{I}'_X(x;d) = \lim_{t\downarrow 0} \mathcal{I}(x+td)/t$, qui est nul si $x + td \in C$ pour tout $t > 0$ petit (c'est-à-dire si $d \in \mathrm{T}^a_C(x)$) et vaut $+\infty$ autrement.

3) Quel que soit $x^* \in \mathbb{E}$, on a $\mathcal{I}^*_P(x^*) = \sup_{x\in\mathbb{F}} \langle x^*, x\rangle - \mathcal{I}_P(x) = \sup_{x\in P} \langle x^*, x\rangle = \sigma_P(x^*)$. Le supremum $\sup_{x\in K} \langle x^*, x\rangle$ est nul si $x^* \in K^-$ (car alors $\langle x^*, x\rangle \leqslant 0$ pour tout $x \in K$ et la valeur nulle est obtenue à la limite en $t \downarrow 0$ en prenant $x = tx_0$ avec $x_0 \in K$) et vaut $+\infty$ sinon (on prend $x = tx_0$ avec $t \to \infty$ et x_0 tel que $\langle x^*, x_0\rangle > 0$).

4) Puisque C est convexe non vide, $\mathcal{I}_C \in \operatorname{Conv}(\mathbb{E})$ par le point 1. D'après le point (ii) de la proposition 3.55, $x^* \in \partial\mathcal{I}_C(x)$ ssi $0 \geqslant \langle x^*, y - x\rangle$, pour tout $y \in C$. Cela revient à dire que $x^* \in \mathrm{N}_C(x)$.

3.29. 1) D'après le point 2 de l'exercice 3.28, $\sigma_C = \mathcal{I}^*_C$. Alors $\sigma^*_C = \mathcal{I}^{**}_C$. Comme C est un convexe fermé non vide, $\mathcal{I}_C \in \overline{\operatorname{Conv}}(\mathbb{E})$ (proposition 3.6) et se confond donc avec sa biconjuguée (point 3 de la proposition 3.46). Donc $\sigma^*_C = \mathcal{I}_C$.

2) Cela résulte des équivalences suivantes :

$$
\begin{aligned}
& x \in \partial\sigma_C(x^*) \\
\iff\ & \sigma_C(x^*) + \sigma^*_C(x) = \langle x^*, x\rangle && \text{[proposition 3.55 } (v)\text{]} \\
\iff\ & \sigma_C(x^*) + \mathcal{I}_C(x) = \langle x^*, x\rangle && \text{[point 1]} \\
\iff\ & x \in C \text{ et } \sup\{\langle x^*, x'\rangle : x' \in C\} = \langle x^*, x\rangle && \text{[définition de } \sigma_C\text{]}.
\end{aligned}
$$

3.30. 1) Soient $x, y \in \mathbb{E}$ et $t \in [0,1]$. Grâce à l'inégalité triangulaire, on a $\|(1-t)x + ty\| \leqslant \|(1-t)x\| + \|ty\| = (1-t)\|x\| + t\|y\|$, ce qui montre la convexité de la norme. La fonction est propre car elle ne prend que des valeurs finies. Elle est fermée, parce que continue. Elle est nminorée par la fonction nulle.

2) Si $\|x^*\|_{\mathrm{D}} \leqslant 1$, on a $0 \leqslant f^*(x^*) \leqslant \sup_x(\|x^*\|_{\mathrm{D}} - 1)\|x\| \leqslant 0$; donc $f^*(x^*) = 0$. Si $\|x^*\|_{\mathrm{D}} > 1$, on peut trouver $x_1 \in \mathbb{R}^n$ tel que $\|x_1\| = 1$ et $\langle x^*, x_1\rangle > 1$. Alors, quel que soit $t > 0$, $f^*(x^*) \geqslant \langle x^*, tx_1\rangle - \|tx_1\| = t(\langle x^*, x_1\rangle - 1)$; donc $f^*(x^*) = +\infty$.

3) On a $\|x\|_{\mathrm{DD}} = \sup\{\langle x, x^*\rangle : \|x^*\|_{\mathrm{D}} \leqslant 1\} = \sup\{\langle x, x^*\rangle - \mathcal{I}_{\bar{B}_{\mathrm{D}}}(x^*) : x^* \in \mathbb{E}\} = \sup\{\langle x, x^*\rangle - f^*(x^*) : x^* \in \mathbb{E}\} = f^{**}(x) = f(x) = \|x\|$, car $f \in \overline{\operatorname{Conv}}(\mathbb{E})$. L'identité (3.81) se déduit alors de $\|x\| = \sup\{\langle x, x^*\rangle : x^* \in \bar{B}_{\mathrm{D}}\}$ et du fait que $\bar{B}_{\mathrm{D}}$ est un compact ($\mathbb{E}$ est de dimension finie).

4) En utilisant le point 2 : $x^* \in \partial f(x) \iff f(x) + f^*(x^*) = \langle x^*, x\rangle \iff \|x^*\|_{\mathrm{D}} \leqslant 1$ et $\langle x^*, x\rangle = \|x\|$. Enfin : $x^* \in \arg\max\{\langle y, x\rangle : \|y\|_{\mathrm{D}} \leqslant 1\}$ ssi $\|x^*\|_{\mathrm{D}} \leqslant 1$ et $\langle x^*, x\rangle = \|x^*\|_{\mathrm{DD}} = \|x\|$ (point 3), ce qui veut dire que $x^* \in \partial f(x)$.

Cas particuliers.

- $\partial f(0) = \bar{B}_D$ par la deuxième égalité de (3.82).
- Par (3.82) : $\|x\| = \langle x^*, x\rangle \leqslant \|x^*\|_{\mathrm{D}}\,\|x\|$, donc $\|x^*\|_{\mathrm{D}} \geqslant 1$ si $x \neq 0$.
- Si $x^* \in \partial(\|\cdot\|_1)(x)$, alors $x^* \in \bar{B}_\infty$ et $\sum_{i\in[1\,:\,n]} x^*_i x_i = \sum_{i\in[1\,:\,n]} |x_i|$ par (3.82). Cette dernière identité s'écrit aussi $\sum_{i\in[1\,:\,n]}(x^*_i \operatorname{sgn}(x_i) - 1)|x_i| = 0$. Comme

$x_i^* \operatorname{sgn}(x_i) \leqslant 1$, il vient que $x_i^* = \operatorname{sgn}(x_i)$, pour tout i tel que $x_i \neq 0$. L'inclusion réciproque est claire.
La norme $\|\cdot\|_p$ est différentiable en $x \neq 0$ et son gradient est donné par (3.83b). Si $x^* \in \partial(\|\cdot\|_\infty)(x)$, alors $x^* \in \bar{B}_1$ et $\sum_{i\in[1:n]} x_i^* x_i = \max_{i\in[1:n]} |x_i|$. On en déduit que

$$\max_{i\in[1:n]} |x_i| = \sum_{i\in[1:n]} x_i^* x_i \leqslant \sum_{i\in[1:n]} |x_i^*|\,|x_i| \leqslant \max_{i\in[1:n]} |x_i|, \tag{D.16}$$

car $x^* \in \bar{B}_1$. On a donc égalité partout dans (D.16). Ensuite, on déduit de $\sum_{i\in[1:n]} |x_i^*|\,|x_i| = \|x\|_\infty \neq 0$ et $|x^*| \in \Delta_n$ (le simplexe unité de $\mathbb{R}^n$) que $x_i^* = 0$ si $|x_i| < \|x\|_\infty$ et $|x_i^*| = 1$ si $|x_i| = \|x\|_\infty$ (c'est-à-dire $i \in I$). Alors la première égalité dans (D.16) montre que $\operatorname{sgn}(x_i^*) = \operatorname{sgn}(x_i)$ pour $i \in I$ et que $\sum_{i\in I} |x_i^*| = 1$. On a montré que $x^* \in \operatorname{co}\{\operatorname{sgn}(x_i)e^i : i \in I\}$. L'inclusion réciproque est claire.

5)

– [$\geqslant$] Soit $x^* \in \bar{B}_D$ tel que $\langle x^*, x\rangle = \|x\|$ (un tel x^* existe par (3.81)). Il suffit de montrer que $(\|\cdot\|)'(x;h) \geqslant \langle x^*, h\rangle$. On a pour $t > 0$:

$$\begin{aligned} \|x+th\| &\geqslant \langle x^*, x+th\rangle \qquad [\|x^*\|_D \leqslant 1 \text{ et (A.9)}] \\ &= \|x\| + t\langle x^*, h\rangle \qquad [\langle x^*, x\rangle = \|x\|]. \end{aligned}$$

Alors, $(\|\cdot\|)'(x;h) = \lim_{t\downarrow 0}[\|x+th\| - \|x\|]/t \geqslant \langle x^*, h\rangle$.
[$\leqslant$] Soit $t > 0$. On peut trouver un $x_t^* \in \bar{B}_D$ tel que $\langle x_t^*, x+th\rangle = \|x+th\|$. Alors, $\langle x_t^*, x\rangle \leqslant \|x\|$ et donc

$$\frac{\|x+th\| - \|x\|}{t} \leqslant \frac{\langle x_t^*, x+th\rangle - \langle x_t^*, x\rangle}{t} = \langle x_t^*, h\rangle.$$

On note $\bar{x}_t^*$ le projeté orthogonal de x_t^* sur le convexe fermé non vide $C := \{x^* \in \bar{B}_D : \langle x^*, x\rangle = \|x\|\}$. Avec l'inégalité précédente, on voit que l'on a certainement

$$\frac{\|x+th\| - \|x\|}{t} \leqslant \langle \bar{x}_t^*, h\rangle + \langle x_t^* - \bar{x}_t^*, h\rangle \leqslant \max_{x^*\in C} \langle \bar{x}^*, h\rangle + \langle x_t^* - \bar{x}_t^*, h\rangle.$$

Pour conclure, il reste à montrer que $x_t^* - \bar{x}_t^* \to 0$ lorsque $t \downarrow 0$. Or $\{x_t^*\}_{t\downarrow 0} \subseteq \bar{B}_D$ est une suite bornée, dont on peut extraire une sous-suite convergente, disons vers $\bar{x}^*$. Comme ce $\bar{x}^* \in C$, la définition du projeté implique que l'on a $\|x_t^* - \bar{x}_t^*\| \leqslant \|x_t^* - \bar{x}^*\| \to 0$.

–

– [(3.85a)]
[(3.85b)]
[(3.85c)] Si x^* est tel que $\|x^*\|_1 \leqslant 1$ et $(x^*)^\mathsf{T} x = \|x\|_\infty$, on a nécessairement $\|x^*\|_1 = 1$ et $\sum_i |x_i^*|(\operatorname{sgn} x_i^* x_i)|x_i| = \|x\|_\infty$, si bien que $\|x\|_\infty$ est combinaison convexe de quantités $(\operatorname{sgn} x_i^* x_i)|x_i|$ qui sont $\leqslant \|x\|_\infty$. Cela implique que $x_i^* = 0$ si $|x_i| < \|x\|_\infty$ et $\operatorname{sgn} x_i^* = \operatorname{sgn} x_i$ si $|x_i| = \|x\|_\infty$. Si l'on note $I(x) := \{i \in [1:n] : |x_i| = \|x\|_\infty\}$ et $\Delta_{|I(x)|} := \{\alpha \in \mathbb{R}_+^{|I(x)|} : \sum_i \alpha_i = 1\}$, l'on déduit alors de (3.84) que

$$(\|\cdot\|_\infty)'(x;h) = \max_{\alpha\in\Delta_{|I(x)|}} \sum_{i\in I(x)} \alpha_i(\operatorname{sgn} x_i)h_i = \max_{i\in I(x)} (\operatorname{sgn} x_i)h_i.$$

3.31. 1) La convexité de f vient du fait qu'elle est la composition d'une fonction convexe (la norme, voir le point 1 de l'exercice 3.30) et de la fonction convexe croissante $r \in \mathbb{R}_+ \mapsto r^p$ ($p \geqslant 1$) (proposition 3.33). La fonction est propre car elle ne prend que

des valeurs finies. Elle est fermée, parce que continue (composition de deux fonctions continues).

2) En utilisant l'inégalité de Cauchy-Schwarz généralisée (A.9), $\langle s, x\rangle \leqslant \|s\|_{\mathrm{D}}\,\|x\|$, on obtient

$$f^*(s) := \sup_{x\in\mathbb{E}} \left(\langle s, x\rangle - \frac{1}{p}\|x\|^p\right) \leqslant \sup_{x\in\mathbb{E}} \left(\|s\|_{\mathrm{D}}\,\|x\| - \frac{1}{p}\|x\|^p\right) = \frac{1}{p'}\|s\|_{\mathrm{D}}^{p'},$$

car le maximum de $t \mapsto \|s\|_{\mathrm{D}} t - t^p/p$ est obtenu pour $t = \|s\|_{\mathrm{D}}^{1/(p-1)}$. Pour l'inégalité inverse, on prend $x = \|s\|_{\mathrm{D}}^{1/(p-1)} x_s$ comme argument du supremum dans la définition de $f^*(s)$, où $x_s \in \arg\max\{\langle s, x\rangle : \|x\| = 1\} \neq \varnothing$. Donc $\langle s, x_s\rangle = \|s\|_{\mathrm{D}}$ et $\|x_s\| = 1$, si bien que l'on obtient

$$f^*(s) \geqslant \|s\|_{\mathrm{D}}^{1/(p-1)} \langle s, x_s\rangle - \frac{1}{p}\|s\|_{\mathrm{D}}^{p/(p-1)} = \frac{1}{p'}\|s\|_{\mathrm{D}}^{p'}.$$

3) Par l'une des définitions du sous-différentiel (point (v) de la proposition 3.55), on a

$$s \in \partial f(x) \quad\Longleftrightarrow\quad f(x) + f^*(s) = \langle s, x\rangle \quad\Longleftrightarrow\quad \frac{1}{p}\|x\|^p + \frac{1}{p'}\|s\|_{\mathrm{D}}^{p'} = \langle s, x\rangle.$$

Donc si $s \in \partial f(x)$, on a $\|x\|^p/p + \|s\|_{\mathrm{D}}^{p'}/p' \leqslant \|s\|_{\mathrm{D}}\,\|x\|$. Par le point 5 de l'exercice 3.4, on en déduit que l'on a égalité et donc que $\|x\|^p = \|s\|_{\mathrm{D}}^{p'} = \langle s, x\rangle$. Réciproquement, si $s \in \mathbb{E}$ vérifie $\|x\|^p = \|s\|_{\mathrm{D}}^{p'} = \langle s, x\rangle$, alors $\|x\|^p/p + \|s\|_{\mathrm{D}}^{p'}/p' = \langle s, x\rangle$ et la chaîne d'équivalences ci-dessus montre que $s \in \partial f(x)$.

3.32. La conjuguée de f s'écrit en $x^* \in \mathbb{R}^n$: $f^*(x^*) = \sup\{(x^*)^{\mathsf{T}} x - \|x\|_0 : x \in \bar{B}_1\}$. En prenant $x = 0 \in \bar{B}1$, on obtient $f^*(x^*) \geqslant 0$. Ensuite, en utilisant l'inégalité de Hölder, on obtient

$$f^*(x^*) \leqslant \sup_{x\in B_1} \|x^*\|_\infty \|x\|_1 - \|x\|_0 \leqslant (\|x^*\|_\infty - 1)^+,$$

car l'argument du supremum vaut 0 si $x = 0$ et est majoré par $\|x^*\|_\infty - 1$ si $x \in B_1 \setminus \{0\}$. En utilisant les minoration et majoration de $f^*(x^*)$ précédentes, on voit que $f^*(x^*) = 0$ si $\|x^*\|_\infty \leqslant 1$. Supposons maintenant que $\|x^*\|_\infty > 1$. Si $\|x^*\|_\infty = |x_i^*|$, on obtient en prenant $x = \operatorname{sgn}(x_i^*) e^i \in B_1$ dans la définition de la conjuguée que $f^*(x^*) \geqslant \|x^*\|_\infty - 1$; donc l'égalité a lieu grâce à l'inégalité exposée ci-dessus. On a démontré (3.86).

Venons-en à (3.87). Par définition, on a

$$f^{**}(x) = \sup_{x^*\in\mathbb{R}^n} (x^*)^{\mathsf{T}} x - f^*(x^*) = \sup_{x^*\in\mathbb{R}^n} (x^*)^{\mathsf{T}} x - (\|x^*\|_\infty - 1)^+.$$

Supposons dans un premier temps que $\|x\|_1 \leqslant 1$. Considérons successivement les supremums en $x^* \in \mathbb{R}^n$ avec les contraintes additionnelles $\|x^*\|_\infty \leqslant 1$ et $\|x^*\|_\infty \geqslant 1$:

$$\sup_{\|x^*\|_\infty\leqslant 1} (x^*)^{\mathsf{T}} x - f^*(x^*) = \sup_{\|x^*\|_\infty\leqslant 1} (x^*)^{\mathsf{T}} x = \|x\|_1$$

et

$$\begin{aligned}
&\sup_{\|x^*\|_\infty \geqslant 1} (x^*)^\mathsf{T} x - f^*(x^*) \\
&= \sup_{\|x^*\|_\infty \geqslant 1} (x^*)^\mathsf{T} x - (\|x^*\|_\infty - 1) \\
&\leqslant \sup_{\|x^*\|_\infty \geqslant 1} \|x^*\|_\infty \underbrace{(\|x\|_1 - 1)}_{\leqslant 0} + 1 \qquad \text{[Hölder]} \\
&= \|x\|_1 \qquad \text{[en prenant } \|x^*\|_\infty = 1\text{, le plus petit possible]}.
\end{aligned}$$

Dès lors $f^{**}(x) = \|x\|_1$ si $\|x\|_1 \leqslant 1$.

Supposons maintenant que $\|x\|_1 > 1$. Comme la norme ℓ_1 est la norme duale de la norme ℓ_∞, on peut trouver un x_1^* tel que $\|x_1^*\|_\infty \leqslant 1$ et $(x_1^*)^\mathsf{T} x > 1$. En prenant x^* de la forme tx_1^* avec $t > 0$, on voit que

$$f^{**}(x) \geqslant \sup_{t \geqslant 1/\|x_1^*\|_\infty} t(x_1^*)^\mathsf{T} x - t\|x_1^*\|_\infty + 1 \geqslant \sup_{t \geqslant 1/\|x_1^*\|_\infty} t\underbrace{[(x_1^*)^\mathsf{T} x - \|x_1^*\|_\infty]}_{>0} + 1 = +\infty.$$

3.33. 1) La fonction f est à valeur dans $\mathbb{R}$, donc elle est propre. Elle est aussi l'enveloppe supérieure des fonctions linéaires $x \mapsto x_i$, donc elle est convexe et fermée (proposition 3.34).

2) On a $f^*(x^*) = \sup_{x \in \mathbb{R}^n}((x^*)^\mathsf{T} x - \max_i x_i)$. Clairement $f^*(x^*) \geqslant 0$ (en prenant $x = 0$). Par ailleurs, si $x^* \geqslant 0$ et $\sum_{i=1}^n x_i^* = 1$, on a $f^*(x^*) \leqslant \sup_x([\sum_j x_j^* - 1] \max_i x_i) = 0$. Si $x^* \not\geqslant 0$, il existe un indice j tel que $x_j^* < 0$. Alors, en prenant $x = -te^j$ (avec $t > 0$) dans la définition de $f^*(x^*)$, on trouve que $f^*(x^*) \geqslant -tx_j^* \to +\infty$ lorsque $t \to +\infty$. Si $\sum_i x_i^* \neq 1$, en prenant $x = te$ (e est le vecteur dont toutes les composantes valent 1) dans la définition de $f^*(x^*)$, on trouve que $f^*(x^*) \geqslant (\sum_i x_i^* - 1)t \to +\infty$ lors que $t \to -\infty$ ou $+\infty$ (selon le signe de $\sum_i x_i^* - 1$).

3) On a $x^* \in \partial f(x)$, ssi $f^*(x^*) + f(x) = (x^*)^\mathsf{T} x$, ssi $x^* \geqslant 0$, $\sum_{i=1}^n x_i^* = 1$ et $\max_i x_i = (x^*)^\mathsf{T} x$ ($\max_i x_i$ est donc une combinaison convexe des x_i), ssi $x^* \geqslant 0$, $\sum_{i=1}^n x_i^* = 1$ et $x_{I^c}^* = 0$, ssi $x^* \in \operatorname{co}\{e^i : i \in I\}$.

4) f est différentiable en x ssi $\partial f(x)$ est un singleton, c'est-à-dire ssi I est un singleton. Dans ce cas, $\partial f(x) = \{\nabla f(x)\}$, qui n'est autre que e^i, où i est l'unique élément de I.

3.34. 1) On sait que $\lambda_{\max}$ est l'enveloppe supérieure des fonctions linéaires $A \mapsto v^\mathsf{T} A v$, pour v sur la sphère unité de $\mathbb{R}^n$; donc $\lambda_{\max}$ est convexe (proposition 3.34).

2) On a $\lambda_{\max}^*(A^*) = \sup_{A \in \mathcal{S}^n}(\langle A^*, A\rangle - \lambda_{\max}(A))$. En prenant $A = 0$, on voit que $\lambda_{\max}^*(A^*) \geqslant 0$. Si $A^* \in \mathcal{S}_+^n$ et $\operatorname{tr} A^* = 1$, A^* est de la forme $\sum_i \lambda_i v_i v_i^\mathsf{T}$, où les valeurs propres $\lambda_i \geqslant 0$, $\sum_i \lambda_i = 1$ et les vecteurs propres sont pris unitaires : $\|v_i\|_2 = 1$. Alors $\operatorname{tr} A^* A = \sum_i \lambda_i v_i^\mathsf{T} A v_i \leqslant \lambda_{\max}(A)$; dans ce cas $\lambda_{\max}^*(A^*) = 0$. Si $\operatorname{tr} A^* \neq 1$, en prenant $A = \alpha I$ dans la définition de $\lambda_{\max}^*(A^*)$, on trouve que $\lambda_{\max}^*(A^*) \geqslant \sup_{\alpha \in \mathbb{R}} \alpha(\operatorname{tr} A^* - 1) = +\infty$. Si $A^* \notin \mathcal{S}_+^n$, A^* a une valeur propre $\lambda < 0$ et un vecteur propre associé v de norme 1 ; en prenant $A = \alpha v v^\mathsf{T}$, on trouve que $\lambda_{\max}^*(A^*) \geqslant \sup_{\alpha \in \mathbb{R}}(\alpha\lambda - \alpha^+) = \sup_{\alpha < 0} \alpha\lambda = +\infty$.

3) D'après le point 2, $A^* \in \partial\lambda_{\max}(A) \iff \lambda_{\max}(A) + \lambda_{\max}^*(A^*) = \langle A^*, A\rangle \iff A^* \in \mathcal{S}_+^n$, $\operatorname{tr} A^* = 1$ et $\langle A^*, A\rangle = \lambda_{\max}(A)$. On a donc montré

$$\partial\lambda_{\max}(A) = \{A^* \in \mathcal{S}_+^n : \operatorname{tr} A^* = 1,\ \operatorname{tr} A^* A = \lambda_{\max}(A)\}.$$

Soit $A^* \in \partial\lambda_{\max}(A)$. Avec la factorisation spectrale $A^* = \sum_{i=1}^n \lambda_i v_i v_i^\mathsf{T}$, on doit avoir $(\lambda_1, \dots, \lambda_n) \in \Delta$ et $\operatorname{tr} A^* A = \sum_{i=1}^n \lambda_i v_i^\mathsf{T} A v_i = \lambda_{\max}(A)$. Ceci implique que $\lambda_i = 0$ si v_i n'est pas dans l'espace propre de A associé à $\lambda_{\max}(A)$. Dès lors $\partial\lambda_{\max}(A) \subseteq$

$\mathrm{co}\{vv^\mathsf{T} : \|v\|_2 = 1,\ Av = \lambda_{\max}(A)v\}$. Réciproquement, si $A^* = \sum_{i=1}^m \lambda_i v_i v_i^\mathsf{T}$ avec $(\lambda_1, \dots, \lambda_m) \in \Delta$ et les v_i vecteurs propres unitaires maximaux de A, on a $A^* \in \mathcal{S}_+^n$, $\operatorname{tr} A^* = 1$, $\operatorname{tr} A^* A = \sum_{i=1}^m \lambda_i v_i^\mathsf{T} A v_i = \lambda_{\max}(A)$. Donc $A^* \in \partial\lambda_{\max}(A)$.

4.a) Si $\lambda_{\max}(A)$ est simple, $\partial\lambda_{\max}(A)$ est le singleton $\{vv^\mathsf{T}\}$, où $\pm v$ sont les uniques vecteurs propres unitaires maximaux, ce qui implique que $\lambda_{\max}(\cdot)$ est différentiable et que $\nabla\lambda_{\max}(A) = vv^\mathsf{T}$ (proposition 3.68).

Pour utiliser le théorème des fonctions implicites (théorème C.14), on introduit l'application $F : \mathbb{R}^n \times \mathbb{R} \times \mathcal{S}^n \to \mathbb{R}^n \times \mathbb{R}$ définie par

$$F(v, \lambda, A) = \begin{pmatrix} Av - \lambda v \\ \|v\|_2^2 - 1 \end{pmatrix}.$$

Les couples propres unitaires (v, λ) de A sont ceux vérifiant $F(v, \lambda, A) = 0$ et cette équation définit implicitement la dépendance de (v, λ) par rapport à A. Pour que (v, λ) puisse s'exprimer comme une fonction de A dans le voisinage d'un triplet (v_1, λ_1, A_1) vérifiant $F(v_1, \lambda_1, A_1) = 0$, il suffit de montrer l'inversibilité de $F'_{v,\lambda}(v_1, \lambda_1, A_1)$. On suppose que λ_1 est la plus grande valeur propre de A_1 et qu'elle est simple. Si $F'_{v,\lambda}(v_1, \lambda_1, A_1) \cdot (w, \mu) = 0$, on a

$$\begin{aligned} A_1 w - \lambda_1 w - \mu v_1 &= 0 \\ v_1^\mathsf{T} w &= 0. \end{aligned}$$

En multipliant scalairement la première équation par $v_1 \neq 0$, en utilisant $A_1 v_1 = \lambda_1 v_1$ et la seconde équation, on trouve $\mu = 0$. En multipliant scalairement la première équation par w et en utilisant la seconde équation, on trouve $w^\mathsf{T} A_1 w = \lambda_1 \|w\|_2^2$. Comme λ_1 est la valeur propre maximale et qu'elle est simple, on a nécessairement $w /\!\!/ v_1$; la seconde équation donne alors $w = 0$. Le théorème des fonctions implicites (théorème C.14) assure l'existence d'une fonction implicite $A \mapsto (v(A), \lambda(A))$ telle que $F(v(A), \lambda(A), A) = 0$ pour tout matrice $A \in \mathcal{S}^n$ voisine de A_1. Si on note $\dot{v}_1 := v'(A_1) \cdot B$ et $\dot{\lambda}_1 := \lambda'(A_1) \cdot B$, on a $F'_v(v_1, \lambda_1, A_1) \cdot \dot{v}_1 + F'_\lambda(v_1, \lambda_1, A_1) \cdot \dot{\lambda}_1 + F'_A(v_1, \lambda_1, A_1) \cdot B = 0$ ou encore

$$\begin{aligned} A_1 \dot{v}_1 - \lambda_1 \dot{v}_1 - \dot{\lambda}_1 v_1 + B v_1 &= 0 \\ v_1^\mathsf{T} \dot{v}_1 &= 0. \end{aligned}$$

En multipliant la première équation par v_1, on trouve $\dot{\lambda}_1 = v_1^\mathsf{T} B v_1 = \operatorname{tr} v_1 v_1^\mathsf{T} B = \langle v_1 v_1^\mathsf{T}, B \rangle$. Donc $\nabla\lambda_1(A_1) = v_1 v_1^\mathsf{T}$.

4.b) Si $A = 0$, on a $\partial\lambda_{\max}(0) = \{\sum_i \lambda_i v_i v_i^\mathsf{T} : \lambda \in \Delta_n \text{ et } \|v_i\|_2 = 1 \text{ pour tout } i\}$. Avec $\Lambda = \mathrm{Diag}(\lambda_1, \dots, \lambda_n)$ et $V = (v_1 \cdots v_n)$, on a $\partial\lambda_{\max}(0) = \{V\Lambda V^\mathsf{T} : \Lambda \text{ diagonale } \succcurlyeq 0, \operatorname{tr}\Lambda = 1 \text{ et } V \text{ orthogonale}\} = \{S \in \mathcal{S}_+^n : \operatorname{tr} S = 1\}$.

4.c) En utilisant la formule du max et le sous-différentiel de $\lambda_{\max}$, on trouve $\lambda'_{\max}(A; D) = \sup\{\langle vv^\mathsf{T}, D\rangle : \|v\|_2 = 1,\ Av = \lambda_{\max}(A)v\}$. Évidemment $\langle vv^\mathsf{T}, D\rangle = v^\mathsf{T} D v$. Par ailleurs, si V est une matrice dont les colonnes forment une base orthonormale de l'espace propre associé à $\lambda_{\max}(A)$, un vecteur unitaire v de ce sous-espace vectoriel s'écrit $v = Vu$, avec $\|u\|_2 = 1$. On voit que $\lambda'_{\max}(A; D) = \lambda_{\max}(V^\mathsf{T} D V)$.

3.35. 1) En prenant la décomposition spectrale (B.17) de $A = V\Lambda V^\mathsf{H}$, on voit que pour tout $x \in \mathbb{C}^n$ tel que $\|x\| = 1$, on a avec $y := V^\mathsf{H} x$:

$$x^\mathsf{H} A x = y^\mathsf{H} \Lambda y = \sum_i \bar{y}_i \lambda_i y_i = \sum_i \lambda_i |y_i|^2 \qquad \text{et} \qquad \sum_i |y_i|^2 = \|y\|^2 = 1.$$

Donc $x^{\mathsf{H}}Ax$ est une combinaison convexe des valeurs propres (réelles) λ_i. Celle-ci est maximale si $y_i = 0$ lorsque $\lambda_i < \lambda_{\max}(A)$, c'est-à-dire si x est orthogonal aux vecteurs propres associés aux valeurs propres différentes de la valeur propre maximale ou encore si x est dans l'espace propre associé à la valeur prepre maximale. En particulier, comme dans le cas réel, on a (B.18), où le maximum est obtenu en prenant pour vecteur x un vecteur propre unitaire associé à la valeur propre maximale.

Par (B.17), $\lambda_{\max}$ est l'enveloppe supérieure des fonctions linéaires $A \mapsto x^{\mathsf{H}}Ax$, pour x sur la sphère unité de $\mathbb{R}^n$; donc $\lambda_{\max}$ est convexe (proposition 3.34).

2) On a $\lambda^*_{\max}(A^*) = \sup_{A\in\mathcal{H}^n}(\langle A^*, A\rangle - \lambda_{\max}(A))$. En prenant $A = 0$, on voit que $\lambda^*_{\max}(A^*) \geqslant 0$. Si $A^* \succcurlyeq 0$ et $\operatorname{tr} A^* = 1$, on voit par (B.17) que A^* est de la forme $\sum_i \lambda_i v_i v_i^{\mathsf{H}}$, où les valeurs propres $\lambda_i \geqslant 0$, $\sum_i \lambda_i = 1$ et les vecteurs propres sont pris unitaires : $\|v_i\| = 1$. Alors $\operatorname{tr} A^*A = \sum_i \lambda_i v_i^{\mathsf{H}} A v_i \leqslant \lambda_{\max}(A)$; dans ce cas $\lambda^*_{\max}(A^*) = 0$. Si $\operatorname{tr} A^* \neq 1$, en prenant $A = \alpha I$ dans la définition de $\lambda^*_{\max}(A^*)$, on trouve que $\lambda^*_{\max}(A^*) \geqslant \sup_{\alpha\in\mathbb{R}} \alpha(\operatorname{tr} A^* - 1) = +\infty$. Si $A^* \not\succcurlyeq 0$, A^* a une valeur propre $\lambda < 0$ et un vecteur propre associé v de norme 1 ; en prenant $A = \alpha vv^{\mathsf{H}}$, on trouve que $\lambda^*_{\max}(A^*) \geqslant \sup_{\alpha\in\mathbb{R}}(\alpha\lambda - \alpha^+) = \sup_{\alpha<0} \alpha\lambda = +\infty$.

3) D'après le point 2, $A^* \in \partial\lambda_{\max}(A) \iff \lambda_{\max}(A) + \lambda^*_{\max}(A^*) = \langle A^*, A\rangle \iff A^* \succcurlyeq 0$, $\operatorname{tr} A^* = 1$ et $\langle A^*, A\rangle = \lambda_{\max}(A)$, si bien que

$$\partial\lambda_{\max}(A) = \{A^* \succcurlyeq 0 : \operatorname{tr} A^* = 1,\ \operatorname{tr} A^*A = \lambda_{\max}(A)\}.$$

On peut en déduire l'expression plus expressive (3.89) du sous-différentiel. Soit $A^* \in \partial\lambda_{\max}(A)$. Avec la factorisation spectrale (B.17) de $A^* = \sum_{i=1}^n \lambda_i v_i v_i^{\mathsf{H}}$, on doit avoir $(\lambda_1, \dots, \lambda_n) \in \Delta$ et $\operatorname{tr} A^*A = \sum_{i=1}^n \lambda_i v_i^{\mathsf{H}} A v_i = \lambda_{\max}(A)$. Ceci implique que $\lambda_i = 0$ si v_i n'est pas dans l'espace propre de A associé à $\lambda_{\max}(A)$. Dès lors $\partial\lambda_{\max}(A) \subseteq \operatorname{co}\{vv^{\mathsf{H}} : \|v\| = 1,\ Av = \lambda_{\max}(A)v\}$. Réciproquement, si $A^* = \sum_{i=1}^m \lambda_i v_i v_i^{\mathsf{H}}$ avec $(\lambda_1, \dots, \lambda_m) \in \Delta$ et les v_i vecteurs propres unitaires maximaux de A, on a $A^* \succcurlyeq 0$, $\operatorname{tr} A^* = 1$, $\operatorname{tr} A^*A = \sum_{i=1}^m \lambda_i v_i^{\mathsf{H}} A v_i = \lambda_{\max}(A)$. Donc $A^* \in \partial\lambda_{\max}(A)$.

La compacité du sous-différentiel provient du fait que $\lambda_{\max}$ est à valeurs finies (proposition 3.64).

3.36. 1) La fonction $(x, y) \in \mathbb{E}\times\mathbb{E} \mapsto \|x-y\|$ est convexe, comme pré-composition de la norme (qui est convexe) par l'application affine $(x, y) \mapsto x-y$. Il en est donc de même de $(x, y) \mapsto d(x, y) + \mathcal{I}_{\mathbb{E}\times C}(x, y)$, car $\mathbb{E}\times C$ est un convexe de $\mathbb{E}\times\mathbb{E}$ et la somme de deux fonctions convexes est convexe. Comme fonction marginale de cette dernière fonction, d_C est convexe (proposition 3.35).

On peut donner une démonstration directe (qui n'est autre celle qui démontre qu'une fonction marginale d'une fonction convexe est convexe). Soient x, $y \in \mathbb{E}$, $t \in [0, 1]$ et $\varepsilon > 0$. Alors on peut trouver z_x et $z_y \in C$ tels que $\|x-z_x\| \leqslant d_C(x) + \varepsilon$ et $\|y-z_y\| \leqslant d_C(y)+\varepsilon$. Alors $d_C((1-t)x+ty) \leqslant \|(1-t)x+ty-(1-t)z_x-tz_y\|$ [convexité de C] $\leqslant (1-t)\|x-z_x\| + t\|y-z_y\|$ [convexité de la norme] $\leqslant (1-t)d_C(x) + td_C(y) + \varepsilon$. Comme $\varepsilon > 0$ est arbitraire, on a $d_C((1-t)x+ty) \leqslant (1-t)d_C(x)+td_C(y)$, c'est-à-dire la convexité de d_C.

2) Montrons en premier lieu que $\frac{1}{t}\operatorname{P}_C(x+tp) \to \operatorname{P}_{C^\infty}(p)$, quand $t \to \infty$. Observons d'abord que $\frac{1}{t}\operatorname{P}_C(x+tp)$ est borné quand $t \to \infty$. En effet

$$\frac{1}{t}\|\operatorname{P}_C(x+tp)\| \leqslant \underbrace{\frac{1}{t}\|\operatorname{P}_C(x+tp) - \operatorname{P}_C(x)\|}_{\leqslant \frac{1}{t}\|(x+tp)-x\| = \|p\|} + \underbrace{\frac{1}{t}\|\operatorname{P}_C(x)\|}_{\to 0},$$

où l'on a utilisé la propriété de contraction (2.29) de l'opérateur de projection. Soit à présent un point limite z la fonction $t \in \mathbb{R}_{++} \mapsto z_t := \frac{1}{t}\operatorname{P}_C(x+tp)$ lorsque $t \to \infty$.

Il est suffisant de montrer que $z = \mathrm{P}_{C^\infty}(p)$. Comme C^∞ est un cône convexe fermé, il suffit de montrer que

$$\langle z - p, z\rangle \leqslant 0 \qquad \text{et} \qquad \forall\, q \in C^\infty : \quad \langle z - p, q\rangle \geqslant 0.$$

Soit $x_0 \in C$. La première inégalité provient de la limite dans

$$\begin{aligned}\langle z_t - p, z_t\rangle \quad &= \quad \underbrace{\langle \mathrm{P}_C(x+tp) - (x+tp), \mathrm{P}_C(x+tp) - x_0\rangle}_{\leqslant 0}/t^2 \\ &\quad + \underbrace{\langle \mathrm{P}_C(x+tp) - (x+tp), x_0\rangle/t^2}_{\to 0} + \underbrace{\langle x, \mathrm{P}_C(x+tp)\rangle/t^2}_{\to 0},\end{aligned}$$

où la négatilité du premier terme découle de la caractérisation (2.28a) de la projection, tandis que la convergence vers zéro des deux autres termes se déduit de $\|\mathrm{P}_C(x+tp)\| \leqslant \|x_0\| + \|x + tp - x_0\|$, venant de l'inégalité (2.29) de contraction $\|\mathrm{P}_C(x+tp) - x_0\| \leqslant \|x + tp - x_0\|$, si bien que $\|\mathrm{P}_C(x+tp)\|/t^2 \to 0$. Soit $q \in C^\infty$; la seconde inegalité provient de la limite dans

$$\langle z_t - p, q\rangle = \underbrace{\langle \mathrm{P}_C(x+tp) - (x+tp), [\mathrm{P}_C(x+tp) + tq] - [\mathrm{P}_C(x+tp)]\rangle}_{\geqslant 0}/t^2 + \underbrace{\langle x, q\rangle/t}_{\to 0},$$

où la positivité du premier terme provient de la caractérisation (2.28a) de la projection et le fait que $\mathrm{P}_C(x+tp) + tq \in C$, parce que $q \in C^\infty$.

Venons-en maintenant au calcul de d_C^∞. On sait que $\mathrm{d}_C \in \mathrm{C}\overline{\mathrm{onv}}(\mathbb{F})$. Soit x un point arbitraire de $\operatorname{dom} \mathrm{d}_C = \mathbb{F}$. Alors

$$\begin{aligned}\mathrm{d}_C^\infty(p) \quad &= \quad \lim_{t\to\infty} \frac{1}{t}\, \mathrm{d}_C(x+tp) \qquad \text{[définition de la fonction asymptotique]} \\ &= \quad \lim_{t\to\infty} \frac{1}{t}\, \|\mathrm{P}_C(x+tp) - (x+tp)\| \qquad \text{[définition de } \mathrm{d}_C\text{]} \\ &= \quad \|\mathrm{P}_{C^\infty}(p) - p\| \qquad \text{[première partie]} \\ &= \quad \|\mathrm{P}_{(C^\infty)^-}(p)\| \qquad \text{[décomposition de Moreau].}\end{aligned}$$

3) Soit $x^* \in \mathbb{E}$. On a successivement $d_C^*(x^*) = \sup_{x\in\mathbb{E}} \sup_{y\in C}(\langle x^*, x\rangle - \|x - y\|) = \sup_{y\in C}(\sup_{x\in\mathbb{E}}(\langle x^*, x\rangle - \|x\|) + \langle x^*, y\rangle) = \mathcal{I}_{\bar{B}_D}(x^*) + \sigma_C(x^*)$.

4) On sait que $x^* \in \partial d_C(x)$ ssi $d_C^*(x^*) + d_C(x) = \langle x^*, x\rangle$ c'est-à-dire ssi $x^* \in \bar{B}_D$ et $\sigma_C(x^*) + \|x - \bar{x}\| = \langle x^*, x\rangle$. Cette dernière condition s'écrit encore

$$\sup_{y\in C} \langle x^*, y - \bar{x}\rangle = \langle x^*, x - \bar{x}\rangle - \|x - \bar{x}\|.$$

Supposons d'abord que $x^* \in \partial d_C(x)$. Par l'inégalité de Cauchy-Schwarz généralisée et $x^* \in \bar{B}_D$, on trouve que $\langle x^*, x - \bar{x}\rangle \leqslant \|x - \bar{x}\|$, si bien que l'identité ci-dessus implique que $\langle x^*, y - \bar{x}\rangle \leqslant 0$ pour tout $y \in C$, c'est-à-dire que $x^* \in \mathrm{N}_C(\bar{x})$. Alors $\sup_{y\in C} \langle x^*, y - \bar{x}\rangle = 0$ (en prenant $y = \bar{x}$) et $\langle x^*, x - \bar{x}\rangle = \|x - \bar{x}\|$.

Réciproquement, soit $x^* \in \bar{B}_D \cap \mathrm{N}_C(\bar{x})$ tel que $\langle x^*, x - \bar{x}\rangle = \|x - \bar{x}\|$. Alors $\sup_{y\in C} \langle x^*, y - \bar{x}\rangle = 0$ et on a l'identité ci-dessus, donc $x^* \in \partial d_C(x)$.

5) Dorénavant $\|\cdot\|_{\mathrm{D}} = \|\cdot\|$. Soit $x^* \in \partial d_C(x)$. • Si $x \notin C$, alors $x \neq \bar{x}$. De plus $\langle x^*, x - \bar{x}\rangle = \|x - \bar{x}\|$ et $x^* \in \bar{B}$ impliquent que $\|x^*\| = 1$ et $x^* \parallel (x - \bar{x})$ (Cauchy-Schwarz), c'est-à-dire $x^* = (x - \bar{x})/\|x - \bar{x}\|$. Le sous-différentiel étant réduit à ce seul point x^*, d_C est différentiable (voir la figure D.2, à gauche). • Si $x \in C^\circ$, alors $\bar{x} = x$,

$\mathrm{N}_C(\bar{x}) = \{0\}$, si bien que $\partial d_C(x) = \{0\}$. Donc d_C est différentiable en x avec un gradient nul. • Si $x \in C \setminus C^\circ$, alors $\bar{x} = x$ et donc $\partial d_C(x) = \bar{B} \cap \mathrm{N}_C(x)$.

6) On prend par exemple $P := \{x \in \mathbb{R}^2 \colon x_2 \geqslant x_1\} \cup \{x \in \mathbb{R}^2 \colon x_2 \leqslant -x_1\}$, qui n'est pas convexe, et $x = (1,0) \notin P$ (voir la figure D.2, à droite). Pour $t \in [0,1]$, on a

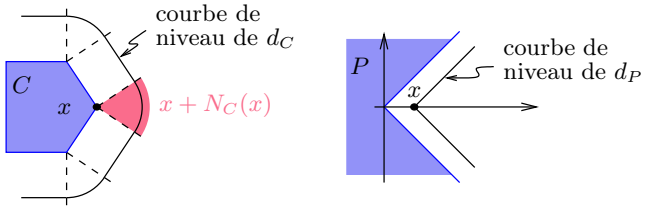


Fig. D.2. Différentiabilité de d_C et non-différentiabilité de d_P

$d_P(x+te^2) = d_P(x-te^2) = (1-t)/\sqrt{2}$, si bien que $d'_P(x;e^2) = d'_P(x;-e^2) = -1/\sqrt{2}$. Dès lors d_P n'est pas différentiable en x.

3.37. 1) La fonction e_C est convexe comme composition de la fonction convexe d_C (exercice 3.36) et de la fonction convexe croissante $t \mapsto (t^+)^2$.

2) En utilisant (2.28a) pour $\mathrm{P}_C(x)$ avec $\mathrm{P}_C(x+h)$ comme point-test dans C, on obtient l'inégalité de gauche

$$0 \leqslant \langle \mathrm{P}_C(x) - x, \mathrm{P}_C(x+h) - \mathrm{P}_C(x) \rangle \tag{D.17}$$

Pour l'inégalité de droite, on décompose le facteur de gauche du produit scalaire comme suit, de manière à faire apparaître h :

$$\mathrm{P}_C(x) - x = \big[\mathrm{P}_C(x) - \mathrm{P}_C(x+h)\big] + \big[\mathrm{P}_C(x+h) - (x+h)\big] + \big[h\big].$$

Le produit scalaire de ces trois termes avec $\mathrm{P}_C(x+h) - \mathrm{P}_C(x)$ donne

$$\begin{aligned}
&\langle \mathrm{P}_C(x) - \mathrm{P}_C(x+h), \mathrm{P}_C(x+h) - \mathrm{P}_C(x) \rangle \leqslant 0, \\
&\langle \mathrm{P}_C(x+h) - (x+h), \mathrm{P}_C(x+h) - \mathrm{P}_C(x) \rangle \leqslant 0 \quad [(2.28a) \text{ pour } \mathrm{P}_C(x+h)], \\
&\langle h, \mathrm{P}_C(x+h) - \mathrm{P}_C(x) \rangle \leqslant \|h\| \, \|\mathrm{P}_C(x+h) - \mathrm{P}_C(x)\| \leqslant \|h\|^2 \quad [(A.7), (2.29)].
\end{aligned}$$

On en déduit l'inégalité de droite en sommant.

3) On a

$$\|(x+h) - \mathrm{P}_C(x+h)\|^2 = \|x - \mathrm{P}_C(x+h)\|^2 + 2\langle x - \mathrm{P}_C(x+h), h \rangle + \|h\|^2. \tag{D.18}$$

Le premier terme à droite dans (D.18) peut s'estimer comme suit

$$\begin{aligned}
&\|x - \mathrm{P}_C(x+h)\|^2 \\
&\quad = \|x - \mathrm{P}_C(x)\|^2 + 2\langle x - \mathrm{P}_C(x), \mathrm{P}_C(x) - \mathrm{P}_C(x+h) \rangle + \|\mathrm{P}_C(x) - \mathrm{P}_C(x+h)\|^2 \\
&\quad = \|x - \mathrm{P}_C(x)\|^2 + O(\|h\|^2) \qquad [\text{point 2 et } \|\mathrm{P}_C(x) - \mathrm{P}_C(x+h)\| \leqslant \|h\|].
\end{aligned}$$

Le second terme à droite dans (D.18) peut s'estimer comme suit

$$\begin{aligned}
&2\langle x - \mathrm{P}_C(x+h), h \rangle \\
&\quad = 2\langle x - \mathrm{P}_C(x), h \rangle + 2\langle \mathrm{P}_C(x) - \mathrm{P}_C(x+h), h \rangle \\
&\quad = 2\langle x - \mathrm{P}_C(x), h \rangle + O(\|h\|^2) \qquad [\|\mathrm{P}_C(x) - \mathrm{P}_C(x+h)\| \leqslant \|h\|].
\end{aligned}$$

Dès lors, (D.18) devient

$$\|(x+h)-\mathrm{P}_C(x+h)\|^2=\|x-\mathrm{P}_C(x)\|^2+2\langle x-\mathrm{P}_C(x),h\rangle+O(\|h\|^2).$$

On en déduit la différentiabilité de e_C et son gradient.

4) D'après le point 3, $\mathrm{P}_C=\nabla(q-e_C)$, où $q(x)=\frac{1}{2}\|x\|^2$.

Remarque: on trouvera chez Moreau [441 ; proposition 7.d] une démonstration indirecte (via l'application prox), mais plus courte de ce résultat.

3.38. Soient $x_1\in A^{\ominus}$, $x_2\in A$ et $x^*\in\partial f(x_1)$. Il suffit de montrer que $x^*\in\partial f(x_2)$. (A) Montrons que $\langle x^*,x_1\rangle=\langle x^*,x_2\rangle$. Comme $x^*\in\partial f(x_1)$, on a $\langle x^*,x_1\rangle-f(x_1)=f^*(x^*)\geqslant\langle x^*,x_2\rangle-f(x_2)$; donc $\langle x^*,x_1\rangle\geqslant\langle x^*,x_2\rangle$ puisque $f(x_1)=f(x_2)$. Ensuite, comme $x_1\in A^{\ominus}$ et $x_2\in A$, on peut trouver un $t>1$ tel que $tx_1+(1-t)x_2\in A$. Ce dernier point peut jouer le rôle de x_2 ci-dessus, si bien que $\langle x^*,x_1\rangle\geqslant\langle x^*,tx_1+(1-t)x_2\rangle$ ou encore $\langle x^*,x_1\rangle\leqslant\langle x^*,x_2\rangle$. (B) On conclut en observant que $\langle x^*,x_1\rangle-f(x_1)=f^*(x^*)$, $\langle x^*,x_1\rangle=\langle x^*,x_2\rangle$ et $f(x_1)=f(x_2)$ impliquent que $\langle x^*,x_2\rangle-f(x_2)=f^*(x^*)$, c'est-à-dire $x^*\in\partial f(x_2)$.

3.39. 1) D'après (i), il existe $r>0$ et $\varepsilon>0$ tels que pour tout x avec $\|x\|\geqslant r$, on a $f(x)\geqslant\varepsilon\|x\|$. On en déduit que si $x\in N_\nu(f)$, alors $\|x\|\leqslant\max(r,\nu/\varepsilon)$. Donc $N_\nu(f)$ est borné. La réciproque est fausse pour $x\in\mathbb{R}\mapsto|x|^{1/2}$.

2) Supposons à présent que $f\in\mathrm{Conv}(\mathbb{E})$.

[$(ii)\Rightarrow(i)$] Si (i) n'a pas lieu, alors

$$\forall k\geqslant 1,\quad \exists\, x_k\in\mathbb{E}:\quad \frac{\|x_k\|}{k}\to\infty\quad\text{et}\quad\frac{f(x_k)}{\|x_k\|}\leqslant\frac{1}{k}.$$

Soit $x\in\operatorname{dom}f$. On considère le point $y_k=x+(k/\|x_k\|)(x_k-x)$, qui est dans le domaine de f pour k grand (car alors $k/\|x_k\|\in[0,1]$ et x et $x_k\in\operatorname{dom}f$). On a $\|y_k-x\|\geqslant k-(k/\|x_k\|)\|x\|\to\infty$; donc $\{y_k\}$ n'est pas bornée. D'autre part, par convexité de f, on a $f(y_k)\leqslant(1-k/\|x_k\|)f(x)+(k/\|x_k\|)f(x_k)\leqslant\nu:=f(x)+2$, pour k grand. Donc l'ensemble de sous-niveau $N_\nu(f)$ n'est pas borné.

[$(i)\Leftrightarrow(iii)$] La condition (i) est équivalente au fait qu'il existe $r>0$ et $\varepsilon>0$ tels que $\|x\|\geqslant r$ implique $f(x)\geqslant\varepsilon\|x\|$. D'autre part, comme élément de $\mathrm{Conv}(\mathbb{E})$, f a une minorante affine (proposition 3.7) et est donc bornée inférieurement sur B_r. On en déduit que, lorsque $f\in\mathrm{Conv}(\mathbb{E})$, (i) est équivalente à

$$\exists\,\varepsilon>0,\ \exists\,\alpha>0,\ \forall x\in\mathbb{E}:\quad f(x)\geqslant\varepsilon\|x\|-\alpha. \tag{D.19}$$

Comme $f\in\mathrm{Conv}(\mathbb{E})$, on peut prendre sa conjuguée (proposition 3.7). Si (D.19) a lieu, on a

$$f^*(x^*)=\sup_{x\in\mathbb{E}}\Big(\langle x^*,x\rangle-f(x)\Big)\leqslant\sup_{x\in\mathbb{E}}\Big(\langle x^*,x\rangle-\varepsilon\|x\|\Big)+\alpha=\mathcal{I}_{\varepsilon\bar{B}_{\mathrm{D}}}(x^*)+\alpha,$$

ce qui montre que f^* est continue en zéro (lemme 3.13). Inversement, si f^* est continue en zéro, il existe $\varepsilon>0$ et $\alpha>0$ tels que $f^*\leqslant\mathcal{I}_{\varepsilon\bar{B}_{\mathrm{D}}}+\alpha$. En prenant la conjuguée de f^*, on a alors

$$f^{**}(x)=\sup_{x^*\in\mathbb{E}}\Big(\langle x^*,x\rangle-f^*(x^*)\Big)\geqslant\sup_{\|x^*\|_D\leqslant\varepsilon}\langle x^*,x\rangle-\alpha=\varepsilon\|x\|-\alpha.$$

Comme $f(x)\geqslant f^{**}(x)$, on a (D.19) et donc (i).

3) [(ii), $(iii) \Rightarrow (iv)$] Il s'agit de montrer qu'il existe un voisinage V de $0 \in \mathbb{E}$ tel que, pour tout $x_0^* \in V$, il existe un x vérifiant $x_0^* \in \partial f(x)$. Cela revient au même de dire que le problème $\inf\{f(x) - \langle x_0^*, x\rangle : x \in \mathbb{E}\}$ a une solution lorsque x_0^* est voisin de 0. Ce sera le cas si les ensembles de sous-niveau de $f_0 : x \mapsto f(x) - \langle x_0^*, x\rangle$ sont compacts, ce qui revient à dire que f_0^* est continue en 0 (équivalence $(ii) \Leftrightarrow (iii)$). Mais $f_0^*(x^*) = f^*(x^* + x_0^*)$, si bien qu'il suffit de montrer que f^* est continue en x_0^*. Or f^* est continue en 0 (les ensembles de sous-niveau de f sont compacts par (ii)) et donc f^* est continue dans un voisinage de 0 (lemme 3.13). On en déduit que f^* est continue en x_0^*, si ce point est voisin de 0.

[$(iv) \Rightarrow (iii)$] Supposons que les ensembles de sous-niveau de f ne soient pas bornés. Il existe alors une direction d non nulle dans leur cône asymptotique. D'après (iv), $g := \varepsilon d$ est un sous-gradient de f si $\varepsilon > 0$ est assez petit : il existe $x \in \mathbb{E}$ tel que $g \in \partial f(x)$. Par convexité de f, $f(x+d) \geqslant f(x) + \langle g, d\rangle = f(x) + \varepsilon\|d\|^2$. D'autre part, comme $d \in \mathrm{Niv}(f, f(x))$, on a $f(x+d) \leqslant f(x)$. On en déduirait que $d = 0$, une contradiction.

3.40. Dans le cas contraire, il existerait un $r > 0$ et des suites $\{x_k\}$ et $\{x_k^*\}$ telles que $x_k^* \in \partial f(x_k)$, $x_k^* \to 0$ et $\mathrm{d}_{\mathcal{S}}(x_k) > r$. On note $\mathcal{S}_r := \mathcal{S} + \bar{B}_r$, qui est un convexe fermé non vide. On peut donc considérer la projection $\bar{x}_k := P_{\mathcal{S}}(x_k)$ de x_k sur $\mathcal{S}$ et introduire $x_k' := \bar{x}_k + \alpha_k(x_k - \bar{x}_k)$, où $\alpha_k := r/\|x_k - \bar{x}_k\| < 1$. Alors $x_k' \in \partial\mathcal{S}_r$ (en effet $x_k' \in \mathcal{S}_r$ car $\|x_k' - \bar{x}_k\| = r$ et, d'autre part, pour tout $\varepsilon > 0$, $x_k^\varepsilon := \bar{x}_k + (\alpha_k+\varepsilon)(x_k - \bar{x}_k) \notin \mathcal{S}_r$ puisque $P_{\mathcal{S}}(x_k^\varepsilon) = \bar{x}_k$ implique que $\mathrm{d}_{\mathcal{S}}(x_k^\varepsilon) = \|x_k^\varepsilon - \bar{x}_k\| = (\alpha_k+\varepsilon)\|x_k - \bar{x}_k\| > r$). Par convexité de f et du fait que $x_k^* \in \partial f(x_k)$, on a

$$\frac{f(x_k') - f(\bar{x}_k)}{\alpha_k} \leqslant f(x_k) - f(\bar{x}_k) \leqslant \langle x_k^*, x_k - \bar{x}_k\rangle.$$

En extrayant une sous-suite au besoin, on peut supposer que $x_k' \to x'$ ($\partial\mathcal{S}_r$ est un compact). On divise alors les inégalités ci-dessus par $\|x_k - \bar{x}_k\|$ et on passe à la limite : on trouve que $f(x') \leqslant f(\bar{x})$. On aurait donc une solution $x' \in \partial\mathcal{S}_r$, ce qui est absurde puisque x' est à une distance r de $\mathcal{S}$.

Le résultat n'est plus vrai si $\mathcal{S}$ n'est pas borné. Comme contre-exemple, on peut considérer la fonction $f \in \overline{\mathrm{Conv}}(\mathbb{R}^2)$, valant x_1^2/x_2 si $x_2 \geqslant 1$ et $+\infty$ ailleurs. Les points $(1, x_2)$ avec $x_2 \to \infty$ sont à une distance 1 de $\mathcal{S} = \{(0, x_2) : x_2 \geqslant 1\}$, alors que le gradient de f en ces points tend vers zéro.

3.41. [$(i) \Leftrightarrow (ii)$] C'est clair !

[$(ii) \Rightarrow (iii)$] Pour obtenir une minoration locale de f, on travaille sur une majoration locale de f^*. On note $P_{\mathcal{S}}$ le projecteur orthogonal sur $\mathcal{S}$, qui est un convexe fermé non vide. On introduit $V_1 := \bar{B}_{r/L}$, avec $r > 0$ assez petit pour que $V_1 \subseteq V \cap V_0$, où V est le voisinage donné par (ii) et V_0 est le voisinage donné par l'exercice 3.39-(iv). Comme les points de $\mathcal{S}$ minimise f, on a certainement pour $x \in \mathcal{S} + \bar{B}_r$ et $\alpha = 0$:

$$f(x) \geqslant \inf f + \alpha\|x - P_{\mathcal{S}}(x)\|^2. \tag{D.20}$$

Nous allons montrer que cette inégalité reste vraie sur $\mathcal{S} + \bar{B}_r$ pour un $\alpha > 0$, ce qui démontrera (iii). Soit $x^* \in V_1$. Comme $x^* \in V_0$, c'est un sous-gradient : il existe $x \in \mathbb{E}$ tel que $x^* \in \partial f(x)$. D'autre part, $x^* \in V$ et $\|x^*\| \leqslant r/L$ impliquent par (ii) que $x \in \mathcal{S} + \bar{B}_r$. En utilisant successivement (D.20) et l'inégalité de Cauchy-Schwarz, on trouve

$$\begin{aligned} f^*(x^*) &= \langle x^*, x\rangle - f(x) \\ &\leqslant \langle x^*, x - P_{\mathcal{S}}(x)\rangle + \langle x^*, P_{\mathcal{S}}(x)\rangle - \inf f - \alpha\|x - P_{\mathcal{S}}(x)\|^2 \qquad \text{(D.21)} \\ &\leqslant \|x^*\|\,\|x - P_{\mathcal{S}}(x)\| - \alpha\|x - P_{\mathcal{S}}(x)\|^2 + \langle x^*, P_{\mathcal{S}}(x)\rangle - \inf f. \end{aligned}$$

Les deux premiers termes du dernier membre forment une fonction quadratique $t \mapsto \|x^*\|t - \alpha t^2$, qui est croissante sur $[0, \|x^*\|/(2\alpha)]$. Tant que $\alpha \leqslant 1/(2L)$ (on ne prétend pas pouvoir prendre un α plus grand), $\|x - P_{\mathcal{S}}(x)\| \leqslant L\|x^*\| \leqslant \|x^*\|/(2\alpha)$. On peut donc conclure

$$f^*(x^*) \leqslant L(1-\alpha L)\|x^*\|^2 + \sup_{\bar{x}\in\mathcal{S}} \langle x^*, \bar{x}\rangle - \inf f, \quad \forall x^* \in V_1.$$

Venons-en à la minoration de f. Comme $f \in \mathrm{C}\overline{\mathrm{onv}}(\mathbb{E})$, $f = f^{**}$. On en déduit pour $x \in \mathcal{S} + \bar{B}_r$:

$$\begin{aligned} f(x) &= f^{**}(x) \\ &\geqslant \sup_{x^*\in V_1} \left(\langle x^*, x\rangle - f^*(x^*)\right) \\ &\geqslant \inf f + \sup_{x^*\in V_1} \left(\langle x^*, x\rangle - \sup_{\bar{x}\in\mathcal{S}} \langle x^*, \bar{x}\rangle - L(1-\alpha L)\|x^*\|^2\right) \\ &\geqslant \inf f - \inf_{x^*\in V_1} \sup_{\bar{x}\in\mathcal{S}} \left(\langle x^*, \bar{x} - x\rangle + L(1-\alpha L)\|x^*\|^2\right), \end{aligned}$$

où on a utilisé la majoration locale de f^* obtenue ci-dessus. On peut inverser l'inf et le sup, car V_1 et $\mathcal{S}$ sont des convexes compacts non vides et la fonction entre parenthèses est convexe-concave (elle a donc un point-selle, théorème **??**) :

$$f(x) \geqslant \inf f - \sup_{\bar{x}\in\mathcal{S}} \inf_{x^*\in V_1} \left(\langle x^*, \bar{x} - x\rangle + L(1-\alpha L)\|x^*\|^2\right). \qquad \text{(D.22)}$$

L'infimum avec contrainte est atteint en $x^* = (x-\bar{x})/(2L(1-\alpha L)+\sigma)$, avec un $\sigma \geqslant 0$ pour que la contrainte $x^* \in V_1$ soit satisfaite, si bien que

$$f(x) \geqslant \inf f + \frac{L(1-\alpha L)+\sigma}{[2L(1-\alpha L)+\sigma]^2}\,(\mathrm{d}_{\mathcal{S}}(x)^)2.$$

Ceci prouve (*iii*).

Lorsque $\mathcal{S}$ est un singleton, on peut aller un peu plus loin. Dans ce cas, $\mathcal{S}$ est le singleton $\{\bar{x}\}$ et (D.22) devient

$$f(x) \geqslant \inf f - \inf_{x^*\in V_1} \left(\langle x^*, \bar{x} - x\rangle + L(1-\alpha L)\|x^*\|^2\right).$$

L'infimum sans contrainte ci-dessus est atteint en $x^* = (x-\bar{x})/(2L(1-\alpha L))$, qui est dans $\bar{B}_{r/L} = V_1$ lorsque $\alpha \leqslant 1/(2L)$. Dès lors

$$f(x) \geqslant f(\bar{x}) + \frac{1}{4L(1-\alpha L)}\,\|x - \bar{x}\|^2.$$

Nous venons de démontrer que si l'inégalité (D.20) est vérifiée sur $\bar{B}(\bar{x}, r)$ avec $\alpha \in [0, 1/(2L)]$, elle l'est aussi avec la constante α valant $\varphi(\alpha)$, où

$$\varphi : \alpha \in [0, 1/(2L)] \mapsto \varphi(\alpha) = \frac{1}{4L(1-\alpha L)}.$$

L'application φ est croissante et $\varphi(\alpha) \geqslant \alpha$. Dès lors (D.20) est aussi vérifié sur $\bar{B}(\bar{x}, r)$ pour α égal au point fixe de φ, à savoir $\alpha = 1/(2L)$.

[$(iii) \Rightarrow (ii)$] Soient $\alpha > 0$ et $r > 0$ donnés en (iii) et V donné par l'exercice 3.40. Supposons que $x^* \in \partial f(x) \cap V$ avec $x \notin \mathcal{S}$ et montrons que $\alpha\, \mathrm{d}_{\mathcal{S}}(x) \leqslant \|x^*\|$ (si $x \in \mathcal{S}$, il n'y a rien à démontrer). On note $\bar{x}$ la projection de x sur le convexe fermé non vide $\mathcal{S}$, si bien que $\mathrm{d}_{\mathcal{S}}(x) = \|x - \bar{x}\|$. On a $\|x - \bar{x}\| \leqslant r$ (exercice 3.40) et on peut utiliser l'inégalité du point (iii). En utilisant aussi la convexité de f et l'inégalité de Cauchy-Schwarz, on trouve

$$f(\bar{x}) \geqslant f(x) + \langle x^*, \bar{x} - x\rangle \geqslant f(\bar{x}) + \alpha\|x - \bar{x}\|^2 - \|x^*\|\,\|x - \bar{x}\|.$$

On en déduit que $\|x^*\| \geqslant \alpha\|x - \bar{x}\|$, c'est-à-dire (ii) avec $L = 1/\alpha$. On ne peut pas prendre L plus petit dans le cas de la valeur absolue $f(x) = |x|$: sur $[-1, 1]$ on a $f(x) \geqslant f(\bar{x}) + |x - \bar{x}|^2$ et on n'a pas mieux que $|x - \bar{x}| \leqslant |x^*|$ sur le même intervalle.

3.42. Soient $0 < r_1 < r_2$. Par optimalité de $x_p(r_1)$ et $x_p(r_2)$, on a

$$\begin{aligned} f(x_p(r_1)) + \frac{1}{2r_1}\|x_p(r_1) - x\|^2 &\leqslant f(x_p(r_2)) + \frac{1}{2r_1}\|x_p(r_2) - x\|^2 \\ f(x_p(r_2)) + \frac{1}{2r_2}\|x_p(r_2) - x\|^2 &\leqslant f(x_p(r_1)) + \frac{1}{2r_2}\|x_p(r_1) - x\|^2. \end{aligned}$$

En sommant ces inégalités on trouve $\|x_p(r_1) - x\| \leqslant \|x_p(r_2) - x\|$, c'est-à-dire (i). En utilisant ce résultat dans la seconde inégalité ci-dessus, on trouve (ii). Si $\bar{x} \in \arg\min f$, supposé non vide, on a

$$f(\bar{x}) + \frac{1}{2r}\|x_p(r) - x\|^2 \leqslant f(x_p(r)) + \frac{1}{2r}\|x_p(r) - x\|^2 \leqslant f(\bar{x}) + \frac{1}{2r}\|\bar{x} - x\|^2. \quad \text{(D.23)}$$

On en déduit que $\|x_p(r) - x\| \leqslant \|\bar{x} - x\|$ et donc que $\{x_p(r)\}_{r\uparrow\infty}$ est bornée. Soit $\hat{x}$ un point d'adhérence de cette suite ($\mathbb{E}$ est de dimension finie). En passant à la limite dans la dernière inégalité et dans (D.23), on trouve que $\hat{x}$ est la projection de x sur $\arg\min f$. Comme celle-ci est unique ($\arg\min f$ est convexe), toute la suite $\{x_p(r)\}_{r\uparrow\infty}$ converge vers cette projection.

3.43. [(i)] $\tilde{x} = P_{f,M}(x)$ ssi 0 est dans le sous-différentiel de $y \mapsto f(y) + \frac{1}{2}\|y - x\|_M^2$, c'est-à-dire ssi il existe $\tilde{g} \in \partial f(\tilde{x})$ tel que $0 = \tilde{g} + M(x_p - x)$.

[(ii)] Comme $x - x_p = M^{-1}g_p$, il reste à montrer que $g_p = P_{f^*,M^{-1}}(Mx)$. Mais $g_k \in \partial f(x_p)$ implique que $x_p \in \partial f^*(g_p)$. Alors la relation $g_p = Mx - Mx_p$ montre que $g_p = P_{f^*,M^{-1}}(Mx)$.

[(iii)] En retranchant les deux relations $x_p = x - M^{-1}g_p^x$ et $y_p = y - M^{-1}g_p^y$, on trouve

$$y - x = (y_p - x_p) + M^{-1}(g_p^y - g_p^x). \quad \text{(D.24)}$$

En la multipliant par M, en prenant le produit scalaire avec $y_p - x_p$ et en utilisant la monotonie du sous-différentiel, on trouve le résultat.

[(iv)] La relation s'obtient en multipliant à gauche (D.24) par $M^{1/2}$, en l'élevant au carré et en utilisant la monotonie du sous-différentiel.

Conditions d'optimalité

4.1. 1) Soit $d \in \mathrm{T}_x(\cap_{i\in I} X_i)$. Alors, il existe une suite $\{x_k\} \subseteq (\cap_{i\in I} X_i)$ et une suite $t_k \downarrow 0$ telles que $\frac{x_k - x}{t_k} \to d$. Pour tout $i \in I$, $\{x_k\} \subseteq X_i$ et donc $d \in \mathrm{T}_x X_i$. L'inclusion inverse n'a pas nécessairement lieu, car si $X_1 = \{x \in \mathbb{R} : x > 0\}$ et $X_2 = -X_1$, on a $X_1 \cap X_2 = \varnothing$, alors qu'en $x = 0$, $\mathrm{T}_x X_1 = \{d \in \mathbb{R} : d \geqslant 0\}$, $\mathrm{T}_x X_2 = -\mathrm{T}_x X_1$ et donc $(\mathrm{T}_x X_1) \cap (\mathrm{T}_x X_2) = \{0\}$.

2) Si $d \in \cup_{i\in I}(\mathrm{T}_x\, X_i)$, il existe un indice $i \in I$ tel que $d \in \mathrm{T}_x\, X_i$. Il existe alors une suite $\{x_k\} \subseteq X_i$ et une suite $t_k \downarrow 0$ telles que $\frac{x_k - x}{t_k} \to d$. Comme $\{x_k\} \subseteq \cup_{i\in I} X_i$, $d \in \mathrm{T}_x(\cup_{i\in I} X_i)$. L'inclusion inverse n'a pas nécessairement lieu. Par exemple si $X_i = \{\frac{1}{i}\} \subseteq \mathbb{R}$, on a en $x = 0$, $\mathrm{T}_x(\cup_{i\in I} X_i) = \mathbb{R}_+$, alors que pour tout i, $\mathrm{T}_x\, X_i = \varnothing$. Cependant, si I est fini et si $d \in \mathrm{T}_x(\cup_{i\in I} X_i)$, il existe une suite $\{x_k\} \subseteq (\cup_{i\in I} X_i)$ et une suite $t_k \downarrow 0$ telles que $\frac{x_k - x}{t_k} \to d$. Mais pour au moins un indice $i \in I$, on a une sous-suite de $\{x_k\}$ incluse dans X_i. Donc $d \in \mathrm{T}_x\, X_i$.

3) Si $(p, q) \in \mathrm{T}_{(x,y)}(X \times Y)$, il existe une suite $\{(x_k, y_k)\} \subseteq X \times Y$ et une suite $\{t_k\} \downarrow 0$ telles que $((x_k, y_k) - (x, y))/t_k \to (p, q)$. Alors $((x_k, y_k) - (x, y))/t_k \to (p, q)$ et $((x_k, y_k) - (x, y))/t_k \to (p, q)$, ce qui montre que $(p, q) \in (\mathrm{T}_x\, X) \times (\mathrm{T}_y\, Y)$. Si X et Y sont des parties convexes, on a

$$\begin{aligned}
\mathrm{T}_{(x,y)}(X \times Y) &= \overline{\mathbb{R}_+[(X - x) \times (Y - y)]} \\
&= \overline{[\mathbb{R}_+(X - x)] \times [\mathbb{R}_+(Y - y)]} \\
&= \overline{\mathbb{R}_+(X - x)} \times \overline{\mathbb{R}_+(Y - y)} \\
&= (\mathrm{T}_x\, X) \times (\mathrm{T}_y\, Y)
\end{aligned}$$

Pour la seconde égalité ci-dessus, on utilise le fait que, si C et D sont des convexes contenant zéro, alors $\mathbb{R}_+(C \times D) = (\mathbb{R}_+ C) \times (\mathbb{R}_+ D)$. L'inclusion $\subseteq$ est claire. Pour l'inclusion réciproque, on écrit $t_x x + t_y y$, avec t_1 et $t_y \in \mathbb{R}_+$, comme

$$(t_x x, t_y y) = \max(t_x, t_y) \left(\frac{t_x}{\max(t_x, t_y)} x, \frac{t_y}{\max(t_x, t_y)} y \right) \in \mathbb{R}_+(C \times D),$$

puisque, par exemple, $t_x / \max(t_x, t_y) x = (1 - t_x / \max(t_x, t_y))0 + t_x / \max(t_x, t_y) x \in C$ lorsque $0 \in C$ et C est convexe.

4.2. 2) Notons $\mathrm{T}_x\, C$ et $\mathrm{T}'_x\, C$ les définitions

Par définition, une direction tangente est limite de directions d telles que $x + \mathbb{R}_{++} d$ rencontre C (en des points de plus en plus proches de x). Donc $\mathrm{T}_x\, C \subseteq \operatorname{adh} \mathrm{T}^a_x C$. Inversement $\mathrm{T}^a_x C \subseteq \mathrm{T}_x\, C$ par convexité de C (et $x \in C$) et $\operatorname{adh} \mathrm{T}^a_x C \subseteq \mathrm{T}_x\, C$ parce que $\mathrm{T}_x\, C$ est fermé.

4) On a $\operatorname{aff} \mathrm{T}_x\, C = \operatorname{aff}(\operatorname{adh} \mathrm{T}^a_x C)$ [point 2] $= \operatorname{aff} \mathrm{T}^a_x C$ [$\mathrm{T}^a_x C$ est convexe et point 2 de la proposition 2.16] $= \operatorname{aff}(C - x)$ [$0 \in C - x$ et point 3 de l'exercice 2.2] $= (\operatorname{aff} C) - x$ [point 2 de l'exercice 2.2].

5) [$(\mathrm{T}_x\, C)^\ominus = (\mathrm{T}^a_x C)^\ominus$] En effet, on a $(\mathrm{T}_x\, C)^\ominus = (\overline{\mathrm{T}^a_x C})^\ominus$ [point 2] $= (\mathrm{T}^a_x C)^\ominus$ [point 3 de l'exercice 2.13].

[$(\mathrm{T}_x\, C)^\ominus \subseteq \mathrm{T}^a_x C^\ominus$] On note B la boule unité fermée de $\mathbb{E}$. Soit $d \in (\mathrm{T}_x\, C)^\ominus$; il s'agit de montrer que $x + \mathbb{R}_{++} d$ rencontre $C^\ominus$. Si ce n'est pas le cas, on peut séparer ces deux convexes [théorème 2.38] : il existe un $\xi \in \mathbb{E} \setminus \{0\}$ tel que $\langle \xi, x + \alpha d \rangle \leqslant \langle \xi, x' \rangle$ pour tout $\alpha > 0$ et pour tout $x' \in C^\ominus$ et donc aussi pour tout $x' \in C$, car $\operatorname{adh}(C^\ominus) = \operatorname{adh} C \supseteq C$. Par ailleurs, comme $d \in (\mathrm{T}_x\, C)^\ominus$, il existe un $\varepsilon > 0$ tel que $d + \varepsilon(B \cap \operatorname{vect} \mathrm{T}_x\, C) \subseteq (\mathrm{T}_x\, C)^\ominus$. On introduit $d' := d - \varepsilon\xi / \|\xi\|$, qui est donc dans $(\mathrm{T}_x\, C)^\ominus \subseteq \mathrm{T}^a_x C$ [première étape]. Il existe donc un $x' \in C$ de la forme $x' = x + \alpha d'$ avec $\alpha > 0$. Alors l'inégalité de séparation donne pour le $\alpha > 0$ et le $x' \in C$ trouvés : $\langle \xi, x + \alpha d \rangle \leqslant \langle \xi, x' \rangle = \langle \xi, x + \alpha d' \rangle = \langle \xi, x + \alpha d - \alpha\varepsilon\xi / \|\xi\| \rangle = \langle \xi, x + \alpha d \rangle - \alpha\varepsilon \|\xi\|$, ce qui contredit $\xi \neq 0$.

[$\mathrm{T}^a_x C^\ominus \subseteq (\mathrm{T}^a_x C)^\ominus$] Comme $\mathrm{T}^a_x C^\ominus = \mathbb{R}_{++}(C^\ominus - x)$ [point 1] et que $(\mathrm{T}^a_x C)^\ominus$ est un cône, il suffit de montrer que $C^\ominus - x \subseteq (\mathrm{T}^a_x C)^\ominus$. Soit $y \in C^\ominus$. Pour montrer que $y - x \in (\mathrm{T}^a_x C)^\ominus$, il suffit que pour une direction $d \in \mathrm{T}^a_x C$ arbitraire on puisse trouver

un $t>1$ tel que $(1-t)d+t(y-x)\in \mathrm{T}^a_x C$ [lemme 2.14]. En fait, puisque $x+d\in \operatorname{aff} C$ [point 4] et $y\in C^\infty$, il existe un $t>1$ tel que $(1-t)(x+d)+ty\in C$. Dès lors $[(1-t)(x+d)+ty]-x\in \mathrm{T}^a_x C$ c'est-à-dire $(1-t)d+t(y-x)\in \mathrm{T}^a_x C$.

6) D'après ce qui précède, $\{p:\langle p,y-x\rangle\leqslant 0,\ \forall y\in C\}=(\mathrm{T}^a_x C)^-=(\operatorname{adh}\mathrm{T}^a_x C)^-$ [point 3 de la proposition 2.43] $=(\mathrm{T}_x C)^-$.

6) On a $\mathrm{T}_x C=(\mathrm{T}_x C)^{--}$ [$\mathrm{T}_x C$ est un cône convexe fermé et point 5 de la proposition 2.43] $=\mathrm{N}_x C^-$ [(4.8)].

4.3. On notera toujours $y:=a(x)$.

1) On a

$$\begin{aligned}
\mathrm{T}_y\, a(C) &= \overline{\mathbb{R}_{++}(a(C)-a(x))} && [(2.46)\text{ et } y=a(x)]\\
&= \overline{\mathbb{R}_{++}A(C-x))} && [\text{définition de } a \text{ et linéarité}]\\
&= \overline{A(\mathbb{R}_{++}(C-x))} && [\text{linéarité}]\\
&= \overline{A(\overline{\mathbb{R}_{++}(C-x)})} && [\text{convexité de } \mathbb{R}_{++}(C-x) \text{ et } (2.18)]\\
&= \overline{A(\mathrm{T}_x C)} && [(2.46)].
\end{aligned}$$

Remarque. On ne peut pas enlever l'adhérence à droite, par exemple, lorsque $C:=\{x\in\mathbb{R}^3: x_1^2+x_2^2\leqslant x_3^2,\ x_3\geqslant 0\}$, $b=0$, $A: x\in\mathbb{R}^3\to\mathbb{R}^2: x\mapsto(x_1,x_2+x_3)$ et $x=0\in\mathbb{R}^3$, car alors $\mathrm{T}_x C=C$ et $A(\mathrm{T}_x C)=A(C)=\{0\}\cup\{y\in\mathbb{R}^2: y_2>0\}$, qui n'est pas fermé (exemple emprunté à l'exercice 2.23).

2) Il n'est pas difficile de montrer que

$$a^{-1}(C)-x=A^{-1}([C\cap\mathcal{R}(a)]-y). \tag{D.25}$$

Montrons maintenant que l'ensemble convexe $\mathcal{C}:=\mathbb{R}_{++}([C\cap\mathcal{R}(a)]-y)$ vérifie

$$A^{-1}(\mathcal{C}^\infty)\neq\varnothing. \tag{D.26}$$

En effet, comme $[C\cap\mathcal{R}(a)]-y=(C-y)\cap\mathcal{R}(A)$, on peut écrire $\mathcal{C}=[\mathbb{R}_{++}(C-y)]\cap\mathcal{R}(A)$; ensuite, comme $\mathcal{C}$ contient 0, $\mathcal{C}^\infty$ est non vide ; finalement, le fait que $\mathcal{C}$ est contenu dans $\mathcal{R}(A)$ implique que $\mathcal{C}^\infty\subseteq\mathcal{R}(A)$, si bien que (D.26) a lieu. L'intérêt de (D.26) est d'impliquer par le point 4 de la proposition 2.17 :

$$\overline{A^{-1}(\mathcal{C})}=A^{-1}(\overline{\mathcal{C}}). \tag{D.27}$$

On peut maintenant conclure :

$$\begin{aligned}
\mathrm{T}_x(a^{-1}(C)) &= \overline{\mathbb{R}_{++}(a^{-1}(C)-x)} && [(2.46)]\\
&= \overline{\mathbb{R}_{++}(A^{-1}([C\cap\mathcal{R}(a)]-y))} && [(D.25)] \qquad (D.28)\\
&= \overline{A^{-1}(\mathbb{R}_{++}([C\cap\mathcal{R}(a)]-y))} && [\mathbb{R}_{++}(A^{-1}(P))=A^{-1}(\mathbb{R}_{++}(P))]\\
&= A^{-1}(\overline{\mathbb{R}_{++}([C\cap\mathcal{R}(a)]-y)}) && [(D.27)]\\
&= A^{-1}(\mathrm{T}_y[C\cap\mathcal{R}(a)]) && [(2.46)].
\end{aligned}$$

3) Pour tout $x'\in C$ et $v\in\mathbb{F}$, on a

$$\langle a(x')-a(x),v\rangle=\langle A(x'-x),v\rangle=\langle x'-x,A^*v\rangle.$$

Si $v\in\mathrm{N}_{a(x)}a(C)$, le membre de gauche est négatif et la négativité du membre de droite implique que $A^*v\in\mathrm{N}_x C$ ou $v\in(A^*)^{-1}\mathrm{N}_x C$; nous avons montré que

$\mathrm{N}_{a(x)}a(C) \subseteq (A^*)^{-1}(\mathrm{N}_x C)$. L'inclusion réciproque est obtenue en inversant les implications sprécédentes.

4) [$\subseteq$] Soit $v \in \mathrm{N}_x(a^{-1}(C))$ et $x' \in a^{-1}(C)$. Alors $\langle x'-x, v\rangle \leqslant 0$. Pour tout $t \in \mathbb{R}$ et tout $h \in \mathcal{N}(A)$, on a $x'+th \in a^{-1}(C)$, si bien que $\langle x'+th-x, v\rangle \leqslant 0$. En divisant par $t \neq 0$ et en faisant tendre $t \to \pm\infty$, on obtient $\langle h, v\rangle = 0$. Donc $v \in \mathcal{N}(A)^\perp = \mathcal{R}(A^*)$ et $v = A^*u$ pour un certain $u \in \mathbb{F}$. Il s'agit maintenant de montrer que $u \in \mathrm{N}_y(C \cap \mathcal{R}(a))$. Pour tout $y' \in C \cap \mathcal{R}(a)$, on peut trouver un $x' \in a^{-1}(y')$, si bien que $\langle y'-y, u\rangle = \langle a(x')-a(x), u\rangle = \langle A(x'-x), u\rangle = \langle x'-x, v\rangle \leqslant 0$ (car $x' \in a^{-1}(C)$). Ceci montre que $u \in \mathrm{N}_y(C \cap \mathcal{R}(a))$. [$\supseteq$] Soient $v \in \mathrm{N}_y(C \cap \mathcal{R}(a))$ et $x' \in a^{-1}(C)$. Posons $y' := a(x')$. Alors $\langle x'-x, A^*v\rangle = \langle A(x'-x), v\rangle = \langle a(x')-a(x), v\rangle = \langle y'-y, v\rangle \leqslant 0$, car $y' \in C \cap \mathcal{R}(a)$. Ceci montre que $A^*v \in \mathrm{N}_x(a^{-1}(C))$.

Remarque. Il peut paraître curieux que $A^*(\mathrm{N}_{a(x)}[C \cap \mathcal{R}(a)])$ soit fermé, parce que l'image par une application linéaire d'un cône convexe fermé n'est pas nécessairement fermée (exercice 2.23). En guise de certification, voici une autre preuve de ce fait. On sait que l'image par une application linéaire d'un convexe fermé est fermée si l'intersection de son cône asymptotique avec le noyau de l'application linéaire est un sous-espace vectoriel (point 6 de la proposition 2.9). Dans le cas présent, $\mathrm{N}_{a(x)}[C \cap \mathcal{R}(a)]$ est un cône convexe fermé, qui est donc égal à son cône asymptotique (point 1 de la proposition 2.9). Par ailleurs, ce cône contient $\mathcal{R}(A)^\perp$, si bien que son intersection avec $\mathcal{N}(A^*) = \mathcal{R}(A)^\perp$ est le sous-espace vectoriel $\mathcal{R}(A)^\perp$. Donc $A^*(\mathrm{N}_{a(x)}[C \cap \mathcal{R}(a)])$ est fermé.

4.4. On déduit facilement le résultat en utilisant le développement de Taylor du théorème C.19. Voici un contre-exemple qui montre que la réciproque est fausse. On considère la fonction $f : \mathbb{R}^2 \to \mathbb{R}$ définie par $f(x) = \frac{1}{2}x_1^2 e^{-x_2^2}$. Elle atteint un minimum en 0, mais sa hessienne $\nabla^2 f(x_1, 0) = \mathrm{diag}(1, -x_1^2)$ n'est pas semi-définie positive dans un voisinage de 0.

4.5. Il suffit de constater que les conditions énoncées impliquent que x_* est un minimum local strict du problème $\min(-f(x))$.

4.6. 1) Comme le critère est convexe ($A \succ 0$), le problème a une solution ssi son équation d'optimalité $Ax + b = 0$ en a une, ce qui est clairement le cas si $A \succ 0$. L'unicité vient de la stricte convexité de f (car $A \succ 0$).

2) [(a) $\Rightarrow$ (c)] Clair !

[(c) $\Rightarrow$ (b)] Soit C est une constante minorant f:

$$\forall\, x \in \mathbb{R}^n : \quad \frac{1}{2}\, x^\mathsf{T} A x + b^\mathsf{T} x \geqslant C. \tag{D.29}$$

En prenant $x = tv$ dans (D.29), en divisant par t^2 et en faisant tendre $t \to \infty$, on obtient $v^\mathsf{T} A v \geqslant 0$; comme v est arbitraire dans $\mathbb{R}^n$, on en déduit que $A \succcurlyeq 0$. En prenant $x = tv$ dans (D.29), avec $v \in \mathcal{N}(A)$, en utilisant $Av = 0$, en divisant par t et en faisant tendre $t \to \infty$, on obtient $b^\mathsf{T} v = 0$; comme v est arbitraire dans $\mathcal{N}(A)$, on en déduit que $b \in \mathcal{N}(A)^\perp = \mathcal{R}(A)$.

[(b) $\Rightarrow$ (a)] Le critère étant convexe ($A \succcurlyeq 0$), le problème a une solution ssi son équation d'optimalité $Ax + b = 0$ en a une, ce qui est clairement le cas puisque $b \in \mathcal{R}(A)$.

Lorsque les conditions équivalentes (a), (b) ou (c) ont lieu, $\bar{x}$ est une solution du problème de minimisation de f ssi $A\bar{x}+b = 0$. On voit alors clairement que l'ensemble des solutions est de la forme $x_0 + \mathcal{N}(A)$, où x_0 est une solution particulière. Par

ailleurs le coût optimal s'écrit $\frac{1}{2}\bar{x}^\mathsf{T}A\bar{x}+b^\mathsf{T}\bar{x} = -\frac{1}{2}\bar{x}^\mathsf{T}A\bar{x} = -\frac{1}{2}\bar{x}^\mathsf{T}AA^\dagger A\bar{x} = -\frac{1}{2}b^\mathsf{T}A^\dagger b$, où l'on a utilisé (B.12).

3) Si $A \not\succcurlyeq 0$ ou $b \notin \mathcal{R}(A)$, la valeur optimale du problème quadratique dans le membre de gauche vaut $-\infty$ (implication $(c) \Rightarrow (b)$ du point 2). Il en est de même dans le membre de droite, car l'ensemble admissible du problème de minimisation est alors vide. En effet, si $A \not\succcurlyeq 0$ et $\alpha \geqslant 0$, $\alpha A \not\succcurlyeq 0$, si bien que $\alpha A \succcurlyeq bb^\mathsf{T}$ ne peut avoir lieu. Par ailleurs si $b \notin \mathcal{R}(A)$, on peut décomposer $b = b_0 + b_1$, avec $b_0 \in \mathcal{R}(A)^\perp = \mathcal{N}(A)$ non nul et $b_1 \in \mathcal{R}(A)$; alors $b_0^\mathsf{T}(\alpha A - bb^\mathsf{T})b_0 = -\|b_0\|^4 < 0$, si bien que $\alpha A \succcurlyeq bb^\mathsf{T}$ ne peut avoir lieu.

On suppose à présent que $A \succcurlyeq 0$ et $b \in \mathcal{R}(A)$. Soit $\bar{x}$ une solution du problème quadratique à gauche (implication $(b) \Rightarrow (a)$ du point 2). Deux cas sont encore à considérer. *Premier cas* : $b = 0$. Alors les deux membres de l'identité prennent la valeur nulle. *Deuxième cas* : $b \neq 0$, alors $A \neq 0$ (sinon $b \notin \mathcal{R}(A)$) et la valeur optimale du problème quadratique est $\frac{1}{2}(b^\mathsf{T}\bar{x}) < 0$. Quant à la valeur du membre de droite de l'identité, elle s'écrit $-\frac{1}{2}\sup\{(b^\mathsf{T}x)^2/(x^\mathsf{T}Ax) : x^\mathsf{T}Ax \neq 0\} = \frac{1}{2}\inf\{-(b^\mathsf{T}x)^2 : x^\mathsf{T}Ax = 1\}$. Ce dernier problème a une solution x qui doit vérifier $\lambda Ax = (b^\mathsf{T}x)b$ pour une certain multiplicateur $\lambda \in \mathbb{R}$. En multipliant scalairement cette dernière condition par x et en utilisant $x^\mathsf{T}Ax = 1$, on trouve que $\lambda = (b^\mathsf{T}x)^2$, qui est non nul (car c'est l'opposé de la valeur optimale et celle-ci est < 0, car $\bar{x}^\mathsf{T}A\bar{x} > 0$ et $b^\mathsf{T}\bar{x} < 0$). L'équation d'optimalité précédente devient $(b^\mathsf{T}x)Ax = b$, ce qui montre que $-(b^\mathsf{T}x)x$ est une solution du problème quadratique. Dès lors, la valeur optimale du problème quadratique s'écrit $\frac{1}{2}(b^\mathsf{T}\bar{x}) = -\frac{1}{2}(b^\mathsf{T}x)^2$, qui est bien la valeur du membre de droite de l'identité.

4.7. On peut monter l'existence de solution d'au moins deux manières. *Première méthode*. On observe que le critère est continu et qu'il tend vers l'infini à l'infini (il est minoré par $\frac{1}{2}\|M\|_F^2 - \lambda_{\max}(M)\|v\|_2^2 + \frac{1}{2}\|v\|_2^4$, qui tend vers l'infini lorsque $\|v\| \to \infty$). *Deuxième méthode*. On récrit le problème comme suit : $\inf_X\{\|M - X\|_F^2 : X \in \mathcal{S}_+^n, \operatorname{rg}(X) \leqslant 1\}$. Ce problème a une solution car son ensemble admissible est fermé (le rang est s.c.i., exercice B.4) non vide (la matrice nulle est admissible) et son critère est continu et tend vers l'infini à l'infini.

Si on note $f(v) = \frac{1}{2}\|M - vv^\mathsf{T}\|_F^2$ le critère à minimiser, on a pour tout $h \in \mathbb{R}^n$: $f'(v)\cdot h = -2\,v^\mathsf{T}Mh + 2(v^\mathsf{T}h)\,\|v\|^2$. La condition d'optimalité du premier ordre s'écrit $f'(v)\cdot h = 0$ pour tout $h \in \mathbb{R}^n$, c'est-à-dire $Mv = \|v\|^2 v$. Donc une solution v est soit le vecteur nul, soit un vecteur propre de M de valeur propre $\lambda = \|v\|^2$.

La condition du second ordre s'écrit $f''(v)\cdot h^2 \geqslant 0$ pour tout $h \in \mathbb{R}^n$. Or on a $f''(v)\cdot h^2 = -2\,h^\mathsf{T}Mh + 2\,\|v\|^2\,\|h\|^2 + 4(v^\mathsf{T}h)^2$. Dès lors, on doit avoir $h^\mathsf{T}Mh \leqslant \|v\|^2\|h\|^2 = \lambda\|h\|^2$ pour tout $h \perp v$ et (par la CN1) $h^\mathsf{T}Mh = \|v\|^2\|h\|^2$ pour $h \parallel v$. Si $M \preccurlyeq 0$, v doit être nul (dans le cas contraire M aurait un vecteur propre de valeur propre $\|v\|^2 > 0$) ; sinon $v \neq 0$ et v est un vecteur propre de M de valeur propre maximale $\lambda_{\max}$ et de norme $\lambda^{1/2}$.

On a donc montré qu'une solution v du problème vérifie

$$v = \begin{cases} 0 & \text{si } M \preccurlyeq 0 \\ \left[\begin{array}{l}\text{un vecteur propre de valeur propre}\\ \text{maximale } \lambda_{\max} \text{ et de norme } \lambda_{\max}^{1/2}\end{array}\right] & \text{sinon.} \end{cases} \tag{D.30}$$

Montrons la réciproque : tout vecteur v vérifiant (D.30) est solution du problème. Soient $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n$ les valeurs propres de M et v_i des vecteurs propres orthonormaux associés. Soit $\overline{v}$ une solution du problème (qui vérifie donc (D.30)).

Sans perte de généralité, on peut supposer que $v_1 \parallel \overline{v}$ si $\lambda_1 > 0$ (grâce à (D.30)). On a la factorisation spectrale $M = \sum_i \lambda_i v_i v_i^\mathsf{T}$, si bien que le coût optimal vaut

$$\mathrm{val}(P) = \frac{1}{2} \left\| \sum_{i=1}^{n} \lambda_i v_i v_i^\mathsf{T} - \overline{v}\overline{v}^\mathsf{T} \right\|_F^2 = \begin{cases} \frac{1}{2} \sum_{i=1}^{n} \lambda_i^2 & \text{si } M \preccurlyeq 0 \\ \frac{1}{2} \sum_{i=2}^{n} \lambda_i^2 & \text{sinon.} \end{cases}$$

Ce coût optimal est obtenu par tout vecteur v vérifiant (D.30).

4.8. Il faut montrer que x_* est un minimum local strict du problème $\min\{-f(x) : c(x) = 0\}$. Mais ce dernier est équivalent au problème $\min\{-f(x) : -c(x) = 0\}$ dont le lagrangien $\tilde{\ell}$ est l'opposé de celui ℓ du problème $\min\{f(x) : c(x) = 0\}$. On en déduit que $\nabla\tilde{\ell}(x_*, \lambda_*) = 0$ et que $d^\mathsf{T}\nabla^2_{xx}\tilde{\ell}(x_*, \lambda_*)d > 0$ pour tout d non nul dans $\mathcal{N}(c'(x_*))$. Le théorème 4.24 permet de conclure.

4.9. On montre facilement que, si $\alpha \leqslant 1/2$, le problème (4.132) a un unique minimum $\bar{x} = 0$, et que, si $\alpha > 1/2$, (4.132) un unique maximum local strict en $\bar{x} = 0$ et deux minima globaux stricts en $\bar{x} = (\pm(\sqrt{\alpha - 1/2})/\alpha, 1 - 1/(2\alpha))$.

Par ailleurs, l'unique solution de (4.133) est $\bar{x}_2 = 1 - 1/(2\alpha)$. Si $\alpha \geqslant 1/2$, on retrouve bien la solution $\bar{x}_2$ trouvée ci-dessus. Si $\alpha < 1/2$, alors $\bar{x}_2 < 0$ qui ne vérifie pas la contrainte de (4.132) et n'est donc pas solution de (4.132).

Première explication (difficilement généralisable). Il y a dans (4.132), une contrainte implicite $x_2 \geqslant 0$ qui n'est plus nécessairement satisfaite après élimination de x_1^2, puisque l'on a minimisé le critère résultant sur $\mathbb{R}$. Si on ajoutait la contrainte $x_2 \geqslant 0$ au problème (4.133), on trouverait alors $\bar{x}_2 = 0$, c'est-à-dire la solution trouvée ci-dessus. Mais la notion de *contrainte implicite* n'est pas bien définie.

Deuxième explication (plus convaincante). Puisque l'on a écrit x_1 comme une fonction de x_2, on peut voir l'opération réalisée comme une paramétrisation de la variété des contraintes X par $x_2 \in \mathbb{R} \mapsto (\pm\sqrt{x_2/\alpha}, x2) \in X$. Mais cette paramétrisation est illicite : elle a deux branches et n'est pas définie pour des $x_2 < 0$. D'ailleurs, il n'y a pas de problème si l'on paramétrise X par $x_1 \in \mathbb{R} \mapsto (x_1, \alpha x_1^2) \in X$. Ceci conduit au problème quartique

$$\min_{x_1} x_1^2 + \left(\alpha x_1^2 - 1\right)^2,$$

dont les points stationnaires sont, d'une part, $x_1 = 0$, qui est un minimum global strict si $\alpha \leqslant 1/2$ (le critère est alors convexe) et un maximum local strict si $\alpha > 1/2$, et, d'autre part, $x_1 = \pm(\sqrt{\alpha - 1/2})/\alpha$, qui requiert d'avoir $\alpha \geqslant 1/2$ et qui sont des minima globaux stricts.

4.10. Il faut déterminer les multiplicateurs optimaux λ associés à la contrainte en chaque point stationnaire (ceux-ci existent bien car l'unique contrainte c est qualifiée en tout point admissible x parce que son gradient $\nabla c(x) = x$ est non nul). Si l'on prend comme lagrangien $(x, \lambda) \in \mathbb{R}^2 \times \mathbb{R} \mapsto \ell(x, \lambda) = \frac{1}{3}(x_1^3 + x_2^3) + \frac{\lambda}{2}(x_1^2 + x_2^2 - 1)$, les conditions de Lagrange s'écrivent

$$\begin{cases} x_1(x_1 + \lambda) = 0 \\ x_2(x_2 + \lambda) = 0 \\ x_1^2 + x_2^2 = 1. \end{cases}$$

Les multiplicateurs associés λ sont donc

$$\begin{aligned} x = \pm(0, 1) \quad &\longrightarrow \quad \lambda = \mp 1, \\ x = \pm(1, 0) \quad &\longrightarrow \quad \lambda = \mp 1, \\ x = \pm(1/\sqrt{2}, 1/\sqrt{2}) \quad &\longrightarrow \quad \lambda = \mp 1/\sqrt{2}. \end{aligned}$$

La hessienne du lagrangien s'écrit

$$L \equiv \nabla^2_{xx}\ell(x,\lambda) = \begin{pmatrix} 2x_1+\lambda & 0 \\ 0 & 2x_2+\lambda \end{pmatrix}$$

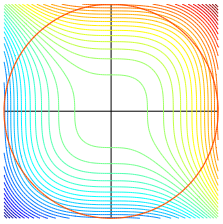
et, par la qualification de la contrainte, l'espace tangent à la contrainte est

$$T_x X := \{d : x_1 d_1 + x_2 d_2 = 0\}.$$

Considérons chaque point stationnaire.

- Point $x = \pm(0,1)$. L'espace tangent est $T_x X := \{d : d_2 = 0\}$ et $d^\mathsf{T} L d = \lambda d_1^2 = \mp d_1^2$, si bien que
  - $x = (0,1)$ est un maximum local strict (critère valant 1/3),
  - $x = (0,-1)$ est un minimum local strict (critère valant -1/3),
- Point $x = \pm(1,0)$. L'espace tangent est $T_x X := \{d : d_1 = 0\}$ et $d^\mathsf{T} L d = \lambda d_2^2 = \mp d_2^2$, si bien que
  - $x = (1,0)$ est un maximum local strict (critère valant 1/3),
  - $x = (-1,0)$ est un minimum local strict (critère valant -1/3),
- Point $x = \pm(1/\sqrt{2}, 1/\sqrt{2})$. L'espace tangent est $T_x X := \{d : d_1 + d_2 = 0\}$ et $d^\mathsf{T} L d = -\lambda \|d\|_2^2 = \pm(1/\sqrt{2})\|d\|_2^2$, si bien que
  - $x = (1/\sqrt{2}, 1/\sqrt{2})$ est un minimum local strict (critère valant $1/(3\sqrt{2})$),
  - $x = -(1/\sqrt{2}, 1/\sqrt{2})$ est un maximum local strict (critère valant $-1/(3\sqrt{2})$),

Le dernier cas peut paraître contre-intuitif si l'on ne regarde que les valeurs du critère aux deux points stationnaires du cas (pas si l'on compare avec les valeurs du critère aux autres points stationnaires), car la valeur du critère est plus grande au minimum local qu'au maximum local. En réalité les courbes de niveau du critère sont comme à la figure ci-contre et la théorie ne nous a pas trahi. Un minimum global est donc atteint en $x = (0,-1)$ et $x = (-1,0)$.



4.11. 1) Considérons le cas où $p = 1$. Le problème est alors équivalent à $\min\{c^\mathsf{T} x : \sum_{i=1}^n v_i = 1,\ -v \leqslant x \leqslant v\}$. En utilisant le lagrangien $\ell(x,v,y,s^+,s^-) = c^\mathsf{T} x + y\left(\sum_{i=1}^n v_i - 1\right) + \sum_{i=1}^n s_i^+(x_i - v_i) - \sum_{i=1}^n s_i^-(x_i + v_i)$, on obtient les conditions d'optimalité suivantes : (a) $c + s^+ - s^- = 0$, (b) $ye - s^+ - s^- = 0$, (c) $\sum_{i=1}^n v_i = 1$, (d) $-v \leqslant x \leqslant v$, (e) $s^+ \geqslant 0$, $s^- \geqslant 0$ et (f) $s_i^+(x_i - v_i) = 0$, $s_i^-(x_i + v_i) = 0$, $\forall i$, où $e = (1,\ldots,1) \in \mathbb{R}^n$. Les équations (a) et (b) sont équivalentes à

$$s^+ = \frac{1}{2}(ye - c) \qquad \text{et} \qquad s^- = \frac{1}{2}(ye + c). \tag{D.31}$$

Par (e), on voit que $y \geqslant \|c\|_\infty$. Si $y > \|c\|_\infty$, $s^+ > 0$, $s^- > 0$ et on aurait par (f), pour tout i : $x_i = v_i = -v_i$, donc $x_i = v_i = 0$, ce qui contredirait (c). On a donc démontré que $y = \|c\|_\infty$. On distinguent les trois cas suivants. (i) Si $|c_i| < \|c\|_\infty$, alors $s_i^+ > 0$, $s_i^- > 0$ puis $x_i = v_i = -v_i$, donc $x_i = v_i = 0$. (ii) Si $c_i = \|c\|_\infty$, alors $s_i^+ = 0$, $s_i^- > 0$ donc $x_i = -v_i \leqslant 0$. (iii) Si $c_i = -\|c\|_\infty$, alors $s_i^+ > 0$, $s_i^- = 0$ donc $x_i = v_i \geqslant 0$. Si on note $I := \{i : |c_i| = \|c\|_\infty\}$, les solutions x vérifient

$$x_{I^c} = 0, \quad \operatorname{sgn}(x_i) = -\operatorname{sgn}(c_i) \quad \text{et} \quad \|x\|_1 = 1.$$

Réciproquement les x vérifiant ces conditions vérifient (c)-(f), avec s^+ et s^- donnés par (D.31) et $v = |x|$.

Venons-en à l'inégalité de Hölder. Quels que soient c et $x \in \mathbb{R}^n$ de norme $\|x\|_1 \leqslant 1$, on a donc $c^\mathsf{T} x \geqslant \sum_{i\in I} \|c\|_\infty \operatorname{sgn}(c_i) x_i = -\|c\|_\infty$. On en déduit l'inégalité de Hölder : quels que soient c et $x \in \mathbb{R}^n$ on a $|c^\mathsf{T} x| \leqslant \|c\|_\infty \|x\|_1$.

2) Considérons le cas où $1 < p < \infty$. Le lagrangien du problème s'écrit (on prend la contrainte équivalente $\|x\|_p^p/p \leqslant 1/p$ pour des questions de différentiabilité et pour éviter le facteur p après différentiation) : $\ell(x,\lambda) = c^\mathsf{T} x + (\lambda/p)(\sum_{i=1}^n |x_i|^p - 1)$. Les conditions du premier ordre s'écrivent (a) $c_i + \lambda x_i |x_i|^{p-2} = 0$, (b) $\|x\|_p \leqslant 1$, (c) $\lambda \geqslant 0$, (d) $\lambda(\|x\|_p - 1) = 0$. Nécessairement $\lambda \neq 0$ (puisque $c \neq 0$), donc $\|x\|_p = 1$ (quatrième équation), $\lambda = \|c\|_{p'}$ où $1/p + 1/p' = 1$ (première et troisième équation) et finalement

$$x_i = -\operatorname{sgn}(c_i) \left(\frac{|c_i|}{\|c\|_{p'}} \right)^{\frac{p'}{p}}.$$

Dans le cas où $p = 2$, on obtient simplement $x = -c/\|c\|_2$.

Quels que soient c et $x \in \mathbb{R}^n$ de norme $\|x\|_p \leqslant 1$, on a donc

$$c^\mathsf{T} x \geqslant -\sum_{i=1}^n c_i \operatorname{sgn}(c_i) \left(\frac{|c_i|}{\|c\|_{p'}} \right)^{\frac{p'}{p}} = -\left(\frac{1}{\|c\|_{p'}} \right)^{\frac{p'}{p}} \sum_{i=1}^n |c_i|^{\frac{p'}{p}+1} = -\frac{\|c\|_{p'}^{p'}}{\|c\|_{p'}^{p'/p}} = -\|c\|_{p'},$$

car $p' - p'/p = 1$. On en déduit l'inégalité de Hölder : quels que soient c et $x \in \mathbb{R}^n$ on a $|c^\mathsf{T} x| \leqslant \|c\|_{p'} \|x\|_p$. L'unicité de la solution du problème implique que l'on a égalité dans l'inégalité de Hölder ssi x est parallèle au vecteur de composante $\operatorname{sgn}(c_i)|c_i|^{p'/p}$ (qui est c si $p = 2$).

3) Considérons le cas où $p = \infty$. Il s'agit de résoudre $\min_x \{\sum_i c_i x_i : -1 \leqslant x_i \leqslant 1, \forall i\}$. Il s'agit des n problèmes indépendants : $\min_{x_i} \{c_i x_i : -1 \leqslant x_i \leqslant 1\}$, dont les solutions vérifient $x_i = -1$ (si $c_i > 0$), $x_i = 1$ (si $c_i < 0$), $-1 \leqslant x_i \leqslant 1$ (si $c_i = 0$).

Quels que soient c et $x \in \mathbb{R}^n$ de norme $\|x\|_\infty \leqslant 1$, on a donc $c^\mathsf{T} x \geqslant -\sum_{i=1}^n \operatorname{sgn} c_i c_i = -\|c\|_1$. On en déduit l'inégalité de Hölder : quels que soient c et $x \in \mathbb{R}^n$, on a $|c^\mathsf{T} x| \leqslant \|c\|_1 \|x\|_\infty$.

4.12. 1) Sur $\mathbb{E} \setminus \{0\}$, on a $q(x) = (x/\|x\|)^\mathsf{T} A (x/\|x\|)$ et $x/\|x\|$ est de norme 1, si bien que (4.135) et (4.136) ont les mêmes valeurs optimales. Par ailleurs, si x est solution de (4.135), alors $x/\|x\|$ est solution de (4.136) ; inversement, si x est solution de (4.136), tx est solution de (4.135), quel que soit $t \neq 0$. Enfin, (4.136) a une solution car on y minimise une fonction continue ($x^\mathsf{T} A x$ est un polynôme) sur la sphère unité qui est compacte en dimension finie (théorème 1.2).

2) Tout est évident lorsqu'on prend comme lagrangien du problème (4.136) la fonction $\ell : (x,\lambda) \in \mathbb{E} \times \mathbb{R} \mapsto \ell(x,\lambda) = \langle Ax, x\rangle - \lambda(\|x\|^2 - 1)$, puisqu'alors les conditions de Lagrange (théorème 4.17) s'écrivent $A\bar{x} = \bar{\lambda}\bar{x}$ et $\|\bar{x}\|^2 = 1$, ce qui caractérise bien un couple propre unitaire de A. Par ailleurs, la valeur critique vaut bien $\langle A\bar{x}, \bar{x}\rangle = \bar{\lambda}\|\bar{x}\|^2 = \bar{\lambda}$.

3) Comme la contrainte de (4.136) est qualifiée en tout point admissible, si $\bar{x}$ est solution du problème, il vérifie les conditions de Lagrange pour un certain multiplicateur $\bar{\lambda}$. Par le point 2, $\bar{x}$ est vecteur propre unitaire de A de valeur prorpe $\bar{\lambda}$. Si (x,λ) est un couple propre de A, avec x unitaire, x est admissible pour (4.136), si bien que $\langle A\bar{x}, \bar{x}\rangle \leqslant \langle Ax, x\rangle$ par optimalité de $\bar{x}$, ce qui s'écrit $\bar{\lambda} \leqslant \lambda$; dès lors $\bar{\lambda}$ est la plus petite valeur prore. Réciproquement, si $\bar{x}$ est un vecteur propre unitaire de valeur propre minimale $\lambda_{\min}(A)$, $\bar{x}$ est admissible pour (4.136) et le critère y prend la valeur $\langle A\bar{x}, \bar{x}\rangle = \lambda_{\min}(A)$, donc $\bar{x}$ est solution de (4.136). Enfin, (4.137) est la

conséquence du fait que les valeurs optimales de (4.136) et (4.135) sont identiques et valent $\lambda_{\min}(A)$.

4) [$\Rightarrow$] On raisonne par l'absurde en supposant qu'il y a deux vecteurs propres unitaires orthogonaux, disons $\bar{x}$ et $\bar{x}'$, correspondant à la plus petite valeur prore $\bar{\lambda}$. On contredira les CS2 en montrant qu'il existe un $d \in \mathrm{T}_{\bar{x}} X \setminus \{0\}$ tel que $\langle \nabla^2_{xx} \ell(\bar{x}, \bar{\lambda}) d, d\rangle = 0$. Ici, $\mathrm{T}_{\bar{x}} X = \{d : \langle \bar{x}, d\rangle = 0\}$ (par la qualification de la contrainte) et $\nabla^2_{xx} \ell(\bar{x}, \bar{\lambda}) = 2(A - \bar{\lambda} I)$. On voit qu'il suffit de prendre $d = \bar{x}'$.

[$\Leftarrow$] On suppose à présent que la plus petite valeur propre $\bar{\lambda}$ de A est simple. Soit $d \in \mathrm{T}_{\bar{x}} X \setminus \{0\}$, c'est-à-dire $\langle \bar{x}, d\rangle = 0$ et $d \neq 0$. On cherche à montrer que $\langle \nabla^2_{xx} \ell(\bar{x}, \bar{\lambda}) d, d\rangle > 0$. On peut décomposer d sur une base orthonormale de vecteurs propres $\{\bar{x}_i\}$ de A, incluant $\bar{x} = \bar{x}_1$ et de valeurs propres associées $\bar{\lambda}_i$: $d = \sum_{i>1} \langle d, \bar{x}_i\rangle \bar{x}_i$ (car $\langle d, \bar{x}_1\rangle = 0$). Alors $\langle \nabla^2_{xx} \ell(\bar{x}, \bar{\lambda}) d, d\rangle = 2\langle (A - \bar{\lambda} I) d, d\rangle = 2 \sum_{i>1} \langle d, \bar{x}_i\rangle^2 (\bar{\lambda}_i - \bar{\lambda}) > 0$, car $\bar{\lambda}_i > \bar{\lambda}$ pour tout $i > 1$ et, d étant non nul, un des $\langle d, \bar{x}_i\rangle \neq 0$.

5) La matrice A étant auto-adjointe, $\mathbb{E}$ a un système orthonormal $\{\bar{x}_i\}_i$, formé de vecteurs propres $\bar{x}_i$ avec les valeurs propres correspondantes $\bar{\lambda}_i$. On peut donc écrire tout vecteur x sous la forme $x = \sum_i \alpha_i \bar{x}_i$. Alors $\|A\|^2 = \sup\{\|Ax\|^2 : \|x\|^2 = 1\} = \sup\{\sum_i \alpha_i^2 \bar{\lambda}_i^2 : \sum_i \alpha_i^2 = 1\} = \max_i \bar{\lambda}_i^2$. Lorsque $A \succcurlyeq 0$, toutes les valeurs propres $\bar{\lambda}_i$ sont $\geqslant 0$, d'où le résultat. Lorsque $A \succ 0$, les valeurs propres de A sont les inverses de celles de A^{-1}, si bien que $\lambda_{\min}(A) = 1/\lambda_{\max}(A^{-1}) = 1/\|A^{-1}\|$.

4.13. 1) Le problème (4.138) consiste à maximiser une fonction continue sur un compact (fermé borné) de $E \times F$; il a donc une solution.

Si l'on note $c(x, y) = (\frac{1}{2}(\|x\|^2 - 1), \frac{1}{2}(\|y\|^2 - 1))$ la contrainte, on a pour $(h, k) \in \mathbb{E} \times \mathbb{F}$: $c'(x, y) \cdot (h, k) = (\langle x, h\rangle, \langle y, k\rangle)$. Si (x, y) est admissible, $c'(x, y)$ est surjective, car on peut atteindre $(\alpha, \beta) \in \mathbb{R}^2$ par $c'(x, y)$ en prenant $(h, k) = (\alpha x, \beta y)$. Donc la contrainte est qualifiée.

En utilisant le lagrangien $\ell(x, y, \sigma, \tau) = -\langle Ax, y\rangle + \frac{\sigma}{2}(\|x\|^2 - 1) + \frac{\tau}{2}(\|y\|^2 - 1)$, les conditions de Lagrange s'écrivent

$$A^* y = \sigma x, \quad Ax = \tau y, \quad \|x\|^2 = 1 \quad \text{et} \quad \|y\|^2 = 1.$$

En multipliant la première équation par x, la seconde par y et en tenant compte des contraintes, on voit que $\sigma = \tau$. Donc, avec $z = (x, y)$, on a $\hat{A} z = \sigma z$, ce qui montre que z est un vecteur propre de $\hat{A}$, de valeur propre σ. En ce point stationnaire, le critère vaut $\langle Ax, y\rangle = \tau \|y\|^2 = \sigma$.

2) Si $\bar{\sigma} = 0$, alors $A = 0$, $\hat{A} = 0$ et le résultat est clair. Supposons à présent que $\bar{\sigma} > 0$ et que σ soit une valeur propre de $\hat{A}$ de vecteur propre (x, y). Il suffit de montrer que $\sigma \leqslant \bar{\sigma}$. On peut donc ne considérer que le cas où $\sigma > 0$ (sinon le résultat est clair). Observons que $x \neq 0$ (sinon $Ax = \sigma y$ impliquerait que $y = 0$, ce qui contredirait le fait que (x, y) est un vecteur propre; de même $y \neq 0$) et que $\|x\| = \|y\|$ (en multipliant scalairement $Ax = \sigma y$ par y, on a $\sigma \|y\|^2 = \langle Ax, y\rangle = \langle x, A^* y\rangle = \sigma \|x\|^2$). Comme $(x/\|x\|, y/\|y\|)$ est admissible pour (4.138), on a $\bar{\sigma} \geqslant \langle Ax/\|x\|, y/\|y\|\rangle = \sigma \|y\|/\|x\| = \sigma$.

4.14. Le premier problème a ses contraintes qualifiées en sa solution $\bar{x}$ (elles sont affines). Il existe donc des vecteurs $y \in \mathbb{R}^m$ et $s \in \mathbb{R}^n$ tels que

$$\nabla f(\bar{x}) + A^{\mathsf{T}} y + s = 0, \quad \bar{x}_I \geqslant 0, \quad s_I \leqslant 0, \quad s_I^{\mathsf{T}} \bar{x}_I = 0, \quad \bar{x}_{I^c} = 0.$$

On définit $J := I \cup \{i \in I^c : s_i \leqslant 0\}$. Alors

$$\nabla f(\bar{x}) + A^\mathsf{T} y + s = 0, \quad \bar{x}_J \geqslant 0, \quad s_J \leqslant 0, \quad s_J^\mathsf{T} \bar{x}_J = 0,$$
$$\bar{x}_{J^c} \leqslant 0, \quad s_{J^c} \geqslant 0, \quad s_{J^c}^\mathsf{T} \bar{x}_{J^c} = 0.$$

Par convexité, ces conditions suffisent à montrer que $\bar{x}$ est aussi solution du second problème.

4.15. Les contraintes sont qualifiées en $x = 0$, car on a $\mathrm{T}_x X = \mathrm{T}'_x X = \{d \in \mathbb{R}^2 : d_2 = 0\}$. Cependant, les conditions de qualification de Mangasarian-Fromovitz n'ont pas lieu puisque $\nabla c_1(0) + \nabla c_2(0) = (0 \ -1)^\mathsf{T} + (0 \ 1)^\mathsf{T} = 0$.

4.16. Dans le cas contraire, il existerait une suite $\{x_k\} \subseteq X$ convergeant vers x et une suite $\{\alpha_k\} \subseteq \mathbb{R}^m \setminus \{0\}$ tels que l'on ait $\sum_{i \in E \cup I} (\alpha_k)_i \nabla c_i(x_k) = 0$, $(\alpha_k)_I \geqslant 0$ et $(\alpha_k)_I^\mathsf{T} c_I(x_k) = 0$. Comme $\alpha_k \neq 0$, par homogénéité des conditions précédentes, on peut supposer que $\|\alpha_k\| = 1$. On pourrait alors extraire une sous-suite telle que $\alpha_k \to \alpha \neq 0$. À la limite dans les conditions précédentes, on trouverait $\sum_{i \in E \cup I} \alpha_i \nabla c_i(x) = 0$, $\alpha_I \geqslant 0$ et $\alpha_I^\mathsf{T} c_I(x) = 0$, ce qui montrerait que (QC-MF) ne serait pas vérifié en x.

4.17. On cherche un vecteur d petit; il est donc naturel de considérer le problème $\min\{\frac{1}{2}\|d\|_2^2 : A_E d = v_E, \ A_J d \leqslant v_J\}$. Grâce à (4.39), ce problème est réalisable et son *unique* solution vérifie (les contraintes sont qualifiées, soit par (QC-A) ou (QC-MF)) : $\exists \lambda \in \mathbb{R}^{m_E + m_J}$ tel que $d + A_E^\mathsf{T} \lambda_E + A_J^\mathsf{T} \lambda_J = 0$, $A_E d = v_E$, $A_J d \leqslant v_J$, $\lambda_J \geqslant 0$ et $\lambda_J^\mathsf{T}(A_J d - v_J) = 0$. En multipliant l'équation d'optimalité par d et en utilisant la complémentarité, on trouve $\|d\|_2^2 = -\lambda_E^\mathsf{T} v_E - \lambda_J^\mathsf{T} v_J \leqslant \|\lambda\|_2 \|v\|_2$. Il reste à montrer qu'il existe une constante C telle que $\|\lambda\|_2 \leqslant C\|d\|_2$.

Ceci ne dépend pas du d particulier que l'on vient de calculer. Montrons donc qu'il existe une constant C telle que si $h \in \mathbb{R}^n$ et $\lambda \in \mathbb{R}^m$ vérifient $h + A_E^\mathsf{T} \lambda_E + A_J^\mathsf{T} \lambda_J = 0$ et $\lambda_J \geqslant 0$, alors $\|\lambda\| \leqslant C\|h\|$ (pour des normes arbitraires). S'il n'en était pas ainsi, on pourrait trouver deux suites de vecteurs $\{h_k\}$ et $\{\lambda_k\}$ tels que $h_k + A_E^\mathsf{T}(\lambda_k)_E + A_J^\mathsf{T}(\lambda_k)_J = 0$, $(\lambda_k)_J \geqslant 0$ et $\|\lambda_k\| > k\|h_k\|$. Donc $\|\lambda_k\| \neq 0$ et on peut extraire une sous-suite convergente de $\lambda_k / \|\lambda_k\| \to \nu$. On passe ensuite à la limite lorsque $k \to \infty$ dans les relations précédentes après les avoir divisées par $\|\lambda_k\|$ pour trouver: $A_E^\mathsf{T} \nu_E + A_J^\mathsf{T} \nu_J = 0$, $\nu_J \geqslant 0$ et $\nu \neq 0$. Ceci serait en contradiction avec (4.39).

4.18. Le point (ii) s'écrit $\nabla f(x_*)^\mathsf{T} v \geqslant 0$ pour tout $v \in \mathrm{T}'_{x_*} X$, ce qui est équivalent à $\nabla f(x_*) \in (\mathrm{T}'_{x_*} X)^+$, encore équivalent à (i) (c'est la démonstration du théorème 4.32). Évidemment, (iii) implique (ii), puisque l'on déduit de (iii) que $\nabla f(x_*)^\mathsf{T} v \geqslant 0$ pour tout $v \in \mathrm{T}_{x_*} X = \mathrm{T}'_{x_*} X$ (par qualification des contraintes en x_*). Il reste à montrer que $(ii) \Rightarrow (iii)$. Comme $\mathrm{T}'_{x_*} X$ est une convexe fermé mon vide, il existe un $v_* \in \mathrm{T}'_{x_*} X$ tel que $\mathrm{d}_{\mathrm{T}'_{x_*} X}(v) = \|v - v_*\|_2$. Alors

$$\begin{aligned} \nabla f(x_*)^\mathsf{T} v + \alpha \, \mathrm{d}_{\mathrm{T}_{x_*} X}(v) &= \nabla f(x_*)^\mathsf{T} v_* + \nabla f(x_*)^\mathsf{T}(v - v_*) + \alpha \|v - v_*\|_2 \\ &\geqslant \nabla f(x_*)^\mathsf{T} v_* + (\alpha - \|\nabla f(x_*)\|_2)\|v - v_*\|_2 \\ &\geqslant 0. \end{aligned}$$

4.19. Si (QC-MF) n'a pas lieu, on a l'identité énoncée avec $\lambda_0 = 0$. Dans le cas contraire, il existe $\hat{d} \in \mathbb{R}^n$ tel que $c'_E(x) \cdot \hat{d} = 0$ et $c'_{I^0(x)}(x) \cdot \hat{d} < 0$ (proposition 4.42). Alors $d \in \mathrm{T}_x X$ implique que $\nabla f(x)^\mathsf{T} d \geqslant 0$ (on prend $\epsilon > 0$ et $d_\epsilon = d + \epsilon \hat{d}$ qui vérifie $c'_E(x) \cdot d_\epsilon = 0$ et $c'_{I^0(x)}(x) \cdot d_\epsilon < 0$; par hypothèse, $\nabla f(x)^\mathsf{T} d_\epsilon = \nabla f(x)^\mathsf{T} d + \epsilon \nabla f(x)^\mathsf{T} \hat{d} \geqslant 0$; on utilise alors le fait que $\epsilon > 0$ est arbitraire). Autrement dit $\nabla f(x) \in (\mathrm{T}_x X)^+$. On en déduit (4.33).

Réciproquement, supposons qu'il existe $d \in \mathbb{R}^n$ tel que $f'(x) \cdot d < 0$, $c'_E(x) \cdot d = 0$ et $c'_{I^0(x)}(x) \cdot d < 0$. On déduit des conditions de Fritz-John que $\lambda_0 = 0$ et que $\lambda_I = 0$; donc $\sum_{i \in E} \lambda_i \nabla c_i(x) = 0$ et $\lambda_E \neq 0$, ce qui contredit la surjectivité de $c'_E(x)$.

4.20. Si x_* n'est pas un point stationnaire isolé, il existe une suite de points stationnaires $\{x_k\} \subseteq X$ telle que $x_k \to x_*$, $x_k \neq x_*$. Comme $\{x_k\} \subseteq X$, en extrayant une sous-suite au besoin, on a $d_k := (x_k - x_*)/\|x_k - x_*\| \to d$, avec $d \in \mathrm{T}_{x_*} X \setminus \{0\}$. Montrons que $d \in C_* \setminus \{0\}$ ou encore que $f'(x_*) \cdot d = 0$. Pour cela, on utilise la stationnarité de x_k : il existe λ_k tel que

$$\nabla f(x_k) + \sum_{i \in E \cup I} (\lambda_k)_i \nabla c_i(x_k) = 0, \quad (\lambda_k)_I \geqslant 0 \quad \text{et} \quad (\lambda_k)_I^{\mathsf{T}} c_I(x_k) = 0.$$

Le suite $\{\lambda_k\}$ est bornée (sinon en divisant ces conditions par $\|\lambda_k\| \neq 0$ et en passant à la limite, on montrerait que (QC-MF) n'a pas lieu) et on peut encore extraire une sous-suite d'indices tels que $\lambda_k \to \lambda_*$. Par passage à la limite dans les conditions ci-dessus, on voit que $\lambda_* \in \Lambda_*$. D'autre part, pour $i \in I^0_*$, on a $0 = (\lambda_k)_i c_i(x_k) = (\lambda_k)_i c'_i(x_*) \cdot (x_k - x_*) + o(\|x_k - x_*\|)$. Par passage à la limite, après avoir divisé par $\|x_k - x_*\|$, on obtient : $(\lambda_*)_i c'_i(x_*) \cdot d = 0$, $\forall i \in I^0_*$. Dès lors : $f'(x_*) \cdot d = -\sum_{i \in E \cup I} (\lambda_*)_i c'_i(x_*) \cdot d = 0$.

Il reste à exhiber une contradiction, ce qui ne semble pas pouvoir s'obtenir par un simple développement du gradient du lagrangien. On s'y prend comme suit. On considère la fonction

$$t \in [0,1] \mapsto \varphi_k(t) := \nabla_x \ell(x_t, \lambda_t)^{\mathsf{T}} (x_k - x_*) - (\lambda_k - \lambda_*)^{\mathsf{T}} c(x_t),$$

où $(x_t, \lambda_t) := (1-t)(x_*, \lambda_*) + t(x_k, \lambda_k)$. Dans ce qui suit, on suppose k suffisamment grand. On a alors $\varphi_k(0) = -(\lambda_k - \lambda_*)^{\mathsf{T}} c(x_*) = -\lambda_k^{\mathsf{T}} c(x_*) = 0$ et $\varphi_k(1) = -(\lambda_k - \lambda_*)^{\mathsf{T}} c(x_k) = \lambda_*^{\mathsf{T}} c(x_k) \leqslant 0$. Comme φ_k est continue sur $[0,1]$ et dérivable sur $]0,1[$, le théorème de Rolle implique l'existence d'un $t_k \in]0,1[$ tel que

$$0 \geqslant \varphi'_k(t_k) = (x_k - x_*)^{\mathsf{T}} L(x_{t_k}, \lambda_{t_k})(x_k - x_*)$$

On divise par $\|x_k - x_*\|^2$ et on passe à la limite pour obtenir $d^{\mathsf{T}} L_* d \leqslant 0$, en contradiction avec (4.63), puisque $d \in C_* \setminus \{0\}$.

4.21. 1) Le cône tangent à $\tilde{X}_{\lambda_*}$ en x_* est inclus dans son cône linéarisant qui n'est autre que C_* par la proposition 4.29 et la formule (4.45c). Puis on observe que (QC-A) ou (QC-IL) assure que l'on a égalité entre ces deux cônes.

2) Soit $d \in C_*$. Par le point 1, il existe alors des suites $\{x_k\} \subseteq \tilde{X}_{\lambda_*}$ et $\{t_k\} \downarrow 0$ telles que $(x_k - x_*)/t_k \to d$. En développant au deuxième ordre $\ell(\cdot, \lambda_*)$ en x_k autour de x_* et en utilisant la nullité du gradient du lagrangien en (x_*, λ_*), on a

$$\ell(x_k, \lambda_*) = \ell(x_*, \lambda_*) + \frac{1}{2} \langle L_*(x_k - x_*), x_k - x_* \rangle + o(\|x_k - x_*\|^2).$$

Clairement $\ell(x_*, \lambda_*) = f(x_*)$ et $\ell(x_k, \lambda_*) = f(x_k)$ $[x_k \in \tilde{X}_{\lambda_*}] \geqslant f(x_*)$ pour k assez grand [optimalité de x_*, $\{x_k\} \subseteq \tilde{X}_{\lambda_*} \subseteq X_{EI}$ et $x_k \to x_*$]. Dès lors, $0 \leqslant \langle L_*(x_k - x_*), x_k - x_* \rangle + o(\|x_k - x_*\|^2)$. En divisant cette inégalité par t_k^2 et en passant à la limite, on trouve que $\langle L_* d, d \rangle \geqslant 0$.

3) On peut raisonner par l'absurde en disant que si l'égalité du point 1 avait lieu pour un certain $\lambda_* \in \Lambda_*$, alors la propriété du point 2 serait aussi vraie (par le raisonnement tenu au point 2) : on aurait alors des conditions d'optimalité du second

ordre semi-fortes. Or on sait que ce n'est pas le cas dans cet exemple ($\forall\, \lambda_* \in \Lambda_*$, $\exists\, d \in C_*$, tel que $\langle L_* d, d\rangle < 0$). Donc l'égalité du point 1 ne peut avoir lieu.

On peut aussi le montrer directement. Comme au moins un multiplicateur est non nul (car leur somme vaut 1) et que $\tilde{X}_{\lambda_*}$ décroît lorsqu'il y a plus de multiplicateurs non nuls, il suffit de montrer que $\mathrm{T}_{x_*} \tilde{X}_{\lambda_*} \neq C_*$ lorsqu'il y a un unique multiplicateur non nul. Supposons que seul $(\lambda_*)_1 \neq 0$ (on raisonne de la même manière pour les autres cas), si bien que $\lambda_* = (1,0,0)$. Alors $L_* = \mathrm{Diag}(2,-2,0)$, ce qui suggère de montrer que $d := (0,1,0)$ n'est pas dans $\mathrm{T}_{x_*} \tilde{X}_{\lambda_*}$ (alors qu'il est dans C_*, car $d_3 = 0$). S'il l'était, il existerait des suites $\{x_k\} \subseteq \tilde{X}_{\lambda_*}$ et $\{t_k\} \downarrow 0$ telles que $x_k/t_k \equiv (x_k - x_*)/t_k \to d$. La première contrainte est active dans $\tilde{X}_{\lambda_*}$ et s'écrit donc $x_3 = (x_1 + x_2)(x_1 - x_2)$, soit en x_k : $(x_k)_3 = ((x_k)_1 + (x_k)_2)((x_k)_1 - (x_k)_2)$. En divisant chaque membre par t_k^2 et en passant à la limite, on obtient la contradiction $0 = -1$.

4.22. [(i)] C'est évident à partir du point (iii) de la proposition 4.68.

[(ii)] Soit $y \in \mathbb{F}$. Alors $p := c'(x)d - ty \in \mathrm{T}_{c(x)}\, K$ pour $t > 0$ assez petit, si bien que $y = c'(x)(d/t) - p/t \in c'(x)\mathbb{E} - \mathrm{T}_{c(x)}\, K$, ce qui montre que la condition (iii) de la proposition 4.68 a lieu.

4.23. 1) Soit $d \in \mathrm{T}_x\, X_{QG}$. Alors, il existe des suites $\{x_k\} \subseteq X_{QG}$ et $\{t_k\} \downarrow 0$ telles que $(x_k - x)/t_k \to d$. Comme x_k et $x \in Q$, on a certainement $d \in \mathrm{T}_x\, Q$. Ensuite, par la différentiabilité de c, il vient $c(x_k) = c(x) + c'(x)(x_k - x) + o(\|x_k - x\|)$. En divisant cette identité par t_k et en prenant la limite en k, on obtient $(c(x_k) - c(x))/t_k \to c'(x)d$. Comme $c(x_k)$ et $c(x) \in G$, on en déduit que $c'(x)d \in \mathrm{T}_{c(x)}\, G$.

2) Soit $K := (\mathrm{T}_{x_*}\, Q)^+ \times (\mathrm{T}_{c(x_*)}\, G)^+$. Les ensembles $(\mathrm{T}_{x_*}\, Q)^+$ et $(\mathrm{T}_{c(x_*)}\, G)^+$ étant des cônes convexes fermés contenant zéro, le dual de K s'écrit $K^+ = (\mathrm{T}_{x_*}\, Q) \times (\mathrm{T}_{c(x_*)}\, G)$ (point 8 de la proposition 2.43). Ensuite,

$$
\begin{array}{rcll}
\nabla f(x_*) & \in & (\mathrm{T}_{x_*}\, X_{QG})^+ & \text{[Peano-Kantorovitch (4.12)]} \\
& = & (\mathrm{T}'_{x_*}\, X_{QG})^+ & \text{[qualification (4.140)]} \\
& = & \{d \in \mathrm{T}_{x_*}\, Q : c'(x_*)d \in \mathrm{T}_{c(x_*)}\, G\}^+ & \text{[(4.139)]} \\
& = & \{d \in \mathbb{E} : \tilde{c}'(x_*)d \in K^+\}^+ & \text{[avec } \tilde{c} \text{ définie par (D.1)]} \\
& = & \overline{\tilde{c}'(x_*)^*(K)} & \text{[lemme de Farkas]} \\
& = & \overline{(\mathrm{T}_{x_*}\, Q)^+ + c'(x_*)^* \left[(\mathrm{T}_{c(x_*)}\, G)^+\right]} & \\
& = & (\mathrm{T}_{x_*}\, Q)^+ + c'(x_*)^* \left[(\mathrm{T}_{c(x_*)}\, G)^+\right] & \text{[qualification (4.140)].}
\end{array}
$$

Ceci signifie que l'on peut trouver $-\lambda_* \in (\mathrm{T}_{c(x_*)}\, G)^+ = -\mathrm{N}_{c(x_*)} G$ tel que $\nabla f(x_*) + c'(x_*)^* \lambda_* \in (\mathrm{T}_{x_*}\, Q)^+$; c'est bien (4.141).

3) Commençons par montrer que l'équivalence (4.142) $\Leftrightarrow$ (D.2).

- [(4.142) $\Rightarrow$ (D.2)] On montre (D.2) par sa forme équivalente (voir la proposition 4.68):

$$\tilde{c}'(x_0)\mathbb{E} - \mathrm{T}^a_{\tilde{c}(x_0)}\, \tilde{G} = \mathbb{E} \times \mathbb{F}. \tag{D.32}$$

Soit $(u,v) \in \mathbb{E} \times \mathbb{F}$. Par (4.142), on peut trouver un $t \geqslant 0$ tel que

$$v - c'(x_0)u \in t\Big(c(x_0) + c'(x_0)(Q - x_0) - G\Big)$$

ou

$$v \in c'(x_0)[t(Q - x_0) + u] - t(G - c(x_0)).$$

Ceci implique qu'il existe un $d \in t(Q - x_0) + u$ tel que $v \in c'(x_0)d - t(G - c(x_0))$, ce qui conduit à

$$\begin{pmatrix} u \\ v \end{pmatrix} \in \begin{pmatrix} d \\ c'(x_0)d \end{pmatrix} - t \begin{pmatrix} Q - x_0 \\ G - c(x_0) \end{pmatrix} = \tilde{c}'(x_0)d - t(\tilde{G} - \tilde{c}(x_0)) \subseteq \tilde{c}'(x_0)\mathbb{E} - \mathrm{T}^a_{\tilde{c}(x_0)}\,\tilde{G}.$$

Donc (D.32) a bien lieu.

- [(4.142) ⇐ (D.2)] Pour obtenir (4.142), il suffit de montrer que 0 est absorbant dans $c(x_0) + c'(x_0)(Q - x_0) - G$. Soit donc $v \in \mathbb{F}$; il s'agit de trouver un $t \geqslant 0$ tel que $v \in t[c(x_0) + c'(x_0)(Q - x_0) - G]$. Nous utilisons à nouveau le fait que la condition (D.2) est équivalente à (D.32). Par conséquent, $(0, v) \in \mathbb{E} \times \mathbb{F}$ est un élément de l'ensemble du membre de gauche de (D.32): il existe $d \in \mathbb{E}$ et $t \geqslant 0$ tels que $(0, v) \in \tilde{c}'(x_0)d - t(\tilde{G} - \tilde{c}(x_0))$ ou

$$\begin{pmatrix} 0 \\ v \end{pmatrix} \in \begin{pmatrix} d \\ c'(x_0)d \end{pmatrix} - t \begin{pmatrix} Q - x_0 \\ G - c(x_0) \end{pmatrix}.$$

  Dès lors, $v \in c'(x_0)[t(Q - x_0)] - t(G - c(x_0))$ ou $v \in t[c(x_0) + c'(x_0)(Q - x_0) - G]$, comme souhaité.

Supposons à présent que (4.142) ait lieu. Alors (D.2) a également lieu, par l'équivalence démontrée précédemment, et donc, $\tilde{c}$ est qualifiée en x_0 pour représenter $\tilde{c}^{-1}(\tilde{G}) = X_{QG}$ par la proposition 4.69. Ceci signifie que

$$\begin{cases} \mathrm{T}_{x_0} X_{QG} = \{d \in \mathbb{E} : \tilde{c}'(x_0)d \in \mathrm{T}_{\tilde{c}(x_0)}\,\tilde{G}\} \\ \tilde{c}'(x_0)^*[(\mathrm{T}_{\tilde{c}(x_0)}\,\tilde{G})^+] \text{ est fermé.} \end{cases} \tag{D.33}$$

Comme

$$\tilde{c}'(x_0)d = \begin{pmatrix} d \\ c'(x_0)d \end{pmatrix} \qquad \text{et} \qquad \mathrm{T}_{\tilde{c}(x_0)}\,\tilde{G} = \mathrm{T}_{x_0}\,Q \times \mathrm{T}_{c(x_0)}\,G,$$

la première propriété dans (D.33) est la première propriété dans (4.140). Par ailleurs, on a successivement (la seconde égalité vient du point 8 de la proposition 2.43)

$$\begin{aligned} \left(\mathrm{T}_{\tilde{c}(x_0)}\,\tilde{G}\right)^+ &= \left(\mathrm{T}_{x_0}\,Q \times \mathrm{T}_{c(x_0)}\,G\right)^+ = (\mathrm{T}_{x_0}\,Q)^+ \times \left(\mathrm{T}_{c(x_0)}\,G\right)^+, \\ \tilde{c}'(x_0)^* \left[\left(\mathrm{T}_{\tilde{c}(x_0)}\,\tilde{G}\right)^+\right] &= (\mathrm{T}_{x_0}\,Q)^+ + c'(x_0)^* \left[\left(\mathrm{T}_{c(x_0)}\,G\right)^+\right]. \end{aligned}$$

Dès lors, la seconde propriété dans (D.33) est la seconde propriété dans (4.140).

4) Par le théorème 4.58, on sait que (D.2), donc (4.142), est équivalent à l'existence d'un $\mu \geqslant 0$ tel que pour tout $(x, y) \in \mathbb{E} \times (\mathbb{E} \times \mathbb{F})$ proche de $(x_0, 0)$, on a

$$\operatorname{dist}(x, \tilde{c}^{-1}(y + \tilde{G})) \leqslant \mu \operatorname{dist}(\tilde{c}(x), y + \tilde{G}).$$

En posant $y = (q, g)$, on obtient $\tilde{c}^{-1}(y + \tilde{G}) = \tilde{c}^{-1}((q + Q) \times (g + G)) = \{x \in q + Q : c(x) \in g + G\} = X^{qg}_{QG}$ et $\operatorname{dist}(\tilde{c}(x), y + \tilde{G}) = \operatorname{dist}((x, c(x)), (q + Q) \times (g + G)) = \operatorname{dist}(x, q + Q) + \operatorname{dist}(c(x), g + G)$, pour une définition appropriée de la distance sur l'espae produit $\mathbb{E} \times \mathbb{F}$. L'énégalité précédente s'écrit donc comme en (4.143).

4.24. 1) La fonction valeur v est convexe grâce à la convexité du problème de gauche (proposition 4.75). En sommant les contraintes du problème perturbé, on trouve $p_1 + p_2 \leqslant -2x_1^2 \leqslant 0$, si bien que lorsque $p_1 + p_2 > 0$, les contraintes sont incompatibles et $v(p) = +\infty$. Si $p_1 + p_2 = 0$, on déduit de ce qui précèse que $x_1 = 0$; alors $x_2 = p_1$ est seul à vérifier les contraintes ; on a aussi $v(p) = p_1$. Si $p_1 + p_2 \leqslant -\alpha^2/2$, on vérifie aisément que le problème perturbé a pour unique solution primale $(-\alpha/2, p_1 + \alpha^2/4)$ et donc pour valeur optimale $v(p) = p_1 - \alpha^2/4$. Enfin si $p_1 + p_2 \in]-\alpha^2/2, 0[$, l'unique

solution primale est à une intersection des paraboles, soit en $(-\operatorname{sgn}(\alpha)\sqrt{|p_1+p_2|/2}, (p_1-p_2)/2)$, ce qui donne $v(p) = (p_1-p_2)/2 - |\alpha|\sqrt{|p_1+p_2|/2}$.

2) Si $\alpha = 0$, v est linéaire sur son domaine $\{p : p_1+p_2 \leqslant 0\}$; donc cette fonction linéaire $p \mapsto p_1$ est dans $\partial v(0)$ et v est sous-différentiable en zéro. À l'inverse si $\alpha \neq 0$, l'opposée de la racine carrée rend v non différentiable en zéro ; en effet, quel que soit $p \in (\operatorname{dom} v)^\ominus = \{p : p_1+p_2 < 0\}$, la dérivé directionnelle $v'(0;p) = (p_1-p_2)/2 - |\alpha|\sqrt{|p_1+p_2|/2}(\lim_{t\downarrow 0}\sqrt{t}/t) = -\infty$, ce qui implique que $\partial v(0) = \varnothing$ (proposition 3.62).

3) Soit $p^* \in \partial v(0)$; alors, pour tout $p \in \operatorname{dom} v = \{p \in \mathbb{R}^2 : p_1+p_2 \leqslant 0\}$, on a $p_1 = v(0) + (p^*)^\mathsf{T} p = p_1^* p_1 + p_2^* p_2$; en prenant $p_1 = -p_2 = \pm 1$, on trouve que $p_1^* = p_2^* + 1$; en prenant $p_1 = -1$ et $p_2 = 0$, on trouve que $p_1^* \geqslant 1$. Inversément, si $1 \leqslant p_1^* = p_2^* + 1$, on a $p_2^* \geqslant 0$ et donc, pour tout $p \in \operatorname{dom} v$, on trouve $p_1^* p_1 + p_2^* p_2 = p_2^*(p_1+p_2) + p_1 \leqslant p_1$. Un multiplicateur $\lambda_* \in \mathbb{R}^2$ est optimal en la solution $x_* = 0$ s'il vérifie $\nabla f(0) + (\lambda_*)_1 \nabla c_1(0) + (\lambda_*)_2 \nabla c_2(0) = 0$, $0 \leqslant \lambda_* \perp c(0) \leqslant 0$, ce qui se récrit $1 - (\lambda_*)_1 + (\lambda_*)_2 = 0$ et $(\lambda_*)_1 \geqslant 0$; on retrouve donc bien les éléments de $\partial v(0)$.

4.25. En utilisant la proposition 1.13, on voit que le problème peut se récrire comme suit

$$\begin{cases} \min_{(x,v)\in\mathbb{E}\times\mathbb{R}} v \\ \max_{i\in[1:m]} f_i(x) \leqslant v \end{cases} \qquad \text{ou} \qquad \begin{cases} \min_{(x,v)\in\mathbb{E}\times\mathbb{R}} v \\ f_i(x) \leqslant v, \quad \forall\, i \in [1:m]. \end{cases}$$

Soit $\bar{x}$ une solution du problème original et $\overline{v} = f(\bar{x})$. Alors $(\bar{x}, \overline{v})$ est solution de ce dernier problème. De plus, (QC-MF) est vérifiée en ce point, car si

$$\sum_{i\in[1:m]} \alpha_i \begin{pmatrix} \nabla f_i(\bar{x}) \\ -1 \end{pmatrix} = 0 \qquad \text{et} \qquad \alpha \geqslant 0,$$

on a $\sum_{i\in[1:m]} \alpha_i = 0$ et $\alpha \geqslant 0$, si bien que $\alpha = 0$. On introduit le lagrangien ℓ du problème, qui en $(x, v, \lambda) \in \mathbb{E}\times\mathbb{R}\times\mathbb{R}$ prend la valeur $\ell(x,v,\lambda) = v + \sum_{i\in[1:m]} \lambda_i(f_i(x) - v)$. Par les conditions d'optimalité de KKT (4.33), il existe alors un multiplicateur optimal $\bar{\lambda} \in \mathbb{R}^m$ tels que

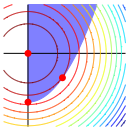
$$\begin{cases} \sum_{i\in[1:m]} \bar{\lambda}_i \nabla f_i(\bar{x}) = 0 \\ \sum_{i\in[1:m]} \bar{\lambda}_i = 1 \\ 0 \leqslant \bar{\lambda} \perp (f(\bar{x}) - \overline{v}e) \leqslant 0. \end{cases}$$

Si $i \notin I$, alors $f_i(\bar{x}) < \overline{v}$ et donc $\bar{\lambda}_i = 0$ par complémentarité. On a donc bien obtenu les conditions énoncées avec $\alpha_I = \bar{\lambda}_I$.

4.26. Le lagrangien $\ell : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ du problème est défini en (x, λ) par $\ell(x,\lambda) = -\frac{1}{2}(x_1^2 + x_2^2) + \lambda_1(x_1^2 - x_2 - 1) + \lambda_2(-x_1)$. Son gradient et sa hessienne s'écrivent

$$\nabla_x \ell(x,\lambda) = \begin{pmatrix} -x_1 + 2\lambda_1 x_1 - \lambda_2 \\ -x_2 - \lambda_1 \end{pmatrix} \quad \text{et} \quad L \equiv \nabla^2_{xx}\ell(x,\lambda) = \begin{pmatrix} 2\lambda_1 - 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

On trouve donc comme points stationnaires et unique multiplicateur associé :



$$x = 0 \quad \text{et} \quad \lambda = 0, \tag{D.34a}$$

$$x = (0,-1) \quad \text{et} \quad \lambda = (1,0), \tag{D.34b}$$

$$x = (\sqrt{2}/2, -1/2) \quad \text{et} \quad \lambda = (1/2, 0). \tag{D.34c}$$

On se trouve donc dans le cas plus simple où il n'y a qu'un unique multiplciateur associé à chaque point stationnaire. Le cône critique C dépend des contraintes actives. Examinons chaque point stationnaire.

1) Pour le point-stationnaire (D.34a), on a $C = \{d \in \mathbb{R}^2 : d_1 \geqslant 0\}$ et $L = \mathrm{Diag}(-1,-1)$. Pour tout $d \in C\backslash\{0\}$, on a $d^{\mathsf{T}}Ld = -\|d\|^2$, qui n'est pas $\geqslant 0$. Dès lors, $x = 0$ *n'est pas un minimum local*. Le fait que $d^{\mathsf{T}}Ld < 0$ pour tout $d \in C\backslash\{0\}$ ne permet pas de conclure que l'on a un maximum local strict. Pour cela, on doit examiner le problème qui consiste à maximiser la même fonction sous les mêmes contraintes. Le lagrangien de ce dernier problème s'écrit $\tilde{\ell}(x,\lambda) = \frac{1}{2}(x_1^2 + x_2^2) + \lambda_1(x_1^2 - x_2 - 1) + \lambda_2(-x_1)$, si bien que

$$\nabla_x \tilde{\ell}(x,\lambda) = \begin{pmatrix} x_1 + 2\lambda_1 x_1 - \lambda_2 \\ x_2 - \lambda_1 \end{pmatrix} \quad \text{et} \quad \tilde{L} \equiv \nabla^2_{xx}\tilde{\ell}(x,\lambda) = \begin{pmatrix} 2\lambda_1 + 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

On voit que $(x,\lambda) = (0,0)$ est l'unique point stationnaire de ce problème de maximisation. Le cône critique associé est toujours $\tilde{C} = \{d \in \mathbb{R}^2 : d_1 \geqslant 0\}$ et on a $d^{\mathsf{T}}\tilde{L}d = \|d\|^2 > 0$ pour tout $d \in \tilde{C}\backslash\{0\}$. On peut donc en conclure que $x = 0$ *est un maximum local strict* (c'est en fait un maximum global).

2) Pour le point-stationnaire (D.34b), on a $C = \{d \in \mathbb{R}^2 : d_1 \geqslant 0,\ d_2 = 0\}$ et $L = \mathrm{Diag}(1,-1)$. Alors, $d^{\mathsf{T}}Ld = d_1^2 > 0$ pour tout $d \in C\backslash\{0\}$. Dès lors, $x = (0,-1)$ *est un minimum local strict* (mais pas un minimum global, car le problème est non borné).

3) Pour le point-stationnaire (D.34c), on a $C = \{d \in \mathbb{R}^2 : d_2 = \sqrt{2}d_1\}$ et $L = \mathrm{Diag}(0,-1)$. Alors, $d^{\mathsf{T}}Ld = -d_2^2$, qui n'est pas $\geqslant 0$ pour tout $d \in C\backslash\{0\}$. Dès lors $x = (\sqrt{2}/2, -1/2)$ *n'est pas un minimum local*. Comme ce n'est pas un point stationnaire du problème de maximisation, on ne peut en dire plus.

Prolégomènes à l'algorithmique

5.2. [Proposition 5.4] Si $\{x_k\}$ converge q-linéairement, il existe un réel $\tau \in [0,1[$ et un indice k_1 tels que pour tout $k \geqslant k_1$ on ait $\|x_{k+1} - x_*\| \leqslant \tau\|x_k - x_*\|$. En divisant par $\|x_*\| \neq 0$ et appliquant $-\log_{10}$ aux deux membres de l'inégalité, on obtient $\sigma_{k+1} \geqslant \sigma_k + \sigma$. La réciproque s'obtient effectuant l'opération inverse.

[Proposition 5.6] Si $\{x_k\}$ converge q-superlinéairement, $\|x_{k+1} - x_*\|/\|x_k - x_*\| = (\|x_{k+1} - x_*\|/\|x_*\|)/(\|x_k - x_*\|/\|x_*\|) \to 0$. En appliquant $-\log_{10}$, on obtient $\sigma_{k+1} - \sigma_k \to \infty$. La réciproque s'obtient effectuant l'opération inverse.

[Proposition 5.11] Si $\{x_k\}$ converge q-quadratiquement, $\|x_{k+1} - x_*\| \leqslant C_0\,\|x_k - x_*\|^2$. En divisant par $\|x_*\| \neq 0$ et en appliquant $-\log_{10}$, on obtient $\sigma_{k+1} \geqslant 2\sigma_k + C$. La réciproque s'obtient effectuant l'opération inverse. Ensuite en divisant les deux membres de $\sigma_{k+1} \geqslant 2\sigma_k + C$ par $\sigma_k > 0$, qui tend vers $+\infty$, et en prenant la limite inférieure lorsque $k \to \infty$, on obtient le dernier résultat.

5.3. [Proposition 5.7] Si $\{x_k\}$ converge q-superlinéairement, $\|x_{k+1} - x_*\|/\|x_k - x_*\| \to 0$. Par (5.2) et l'équivalence de norme on en déduit que $\|F(x_{k+1})\|/\|F(x_k)\| \to 0$. En prenant le logarithme, on obtient le résultat. La réciproque s'obtient effectuant l'opération inverse.

[Proposition 5.13] Si $\{x_k\}$ converge q-quadratiquement, $\|x_{k+1} - x_*\| = O(\|x_k - x_*\|^2)$. Par (5.2) et l'équivalence de norme on en déduit que $\|F(x_{k+1})\| = O(\|F(x_k)\|^2)$. En prenant le logarithme, on obtient le résultat. La réciproque s'obtient effectuant l'opération inverse. Ensuite en divisant les deux membres de $\log \|F(x_{k+1}\| \leqslant$

$2\log\|F(x_k)\| + C$ par $\log\|F(x_k)\|$, qui tend vers $-\infty$, et en prenant la limite inférieure lorsque $k \to \infty$, on obtient le dernier résultat.

5.4. 1) Si $\{x_k\}$ converge quadratiquement vers x_*, elle converge superlinéairement et donc $s_k \sim (x_k - x_*)$ (lemme 5.8). Alors l'inégalité de la définition (5.10) conduit à l'existence d'une constante $C' \geqslant 0$ telle que $\|s_{k+1}\| \leqslant C'\|s_k\|^2$, c'est-à-dire à la convergence quadratique de $\{s_k\}$ vers zéro.

2) Si $\{x_k\}$ est telle que la suite des déplacements $\{s_k\}$ converge quadratiquement vers zéro, celle-ci converge superlinéairement. Alors par le point 2 de la proposition 5.9, la suite $\{x_k\}$ est de Cauchy et converge superlinéairement vers une limite x_*. Par le lemme 5.8, on a alors $s_k \sim (x_k - x_*)$ et de la convergence quadratique de $\{s_k\}$, on déduit celle de $\{x_k\}$.

5.5. 1) En développant $\|x_{k+1} - x\|^2 \leqslant \|x_k - x\|^2$, on voit que $x \in C(\{x_k\})$ ssi $\langle x_k - x_{k+1}, x_k + x_{k+1} - 3x\rangle \geqslant 0$. L'affinité en x de cette dernière expression permet d'écrire $C(\{x_k\})$ comme une intersection de demi-espaces fermés; cet ensemble est donc un convexe fermé.

2) Soit $x \in C(\{x_k\})$, qui est non vide. Par récurrence, on a $\|x_k - x\| \leqslant \|x_0 - x\|$, dont on déduit la bornitude de $\{x_k\}$.

3) Comme $\bar{x} \in C(\{x_k\})$, on a pour tout k : $\|x_{k+1} - \bar{x}\| \leqslant \|x_k - \bar{x}\|$. Alors la suite $\{\|x_k - \bar{x}\|\}$ converge et ce ne peut être que vers zéro car, $\bar{x}$ étant un point d'adhérence de $\{x_k\}$, une sous-suite de $\{\|x_k - \bar{x}\|\}$ converge vers zéro.

4) C'est clair : pour tout $x \in C(\{x_k\})$, $\|x_{k+1} - x\| \leqslant \|x_k - x\|$ et donc $x_{k+1} \in \bar{B}(x, \|x_k - x\|)$.

5.6. On écrit f comme un lagrangien en introduisant un état adjoint $p : [0,T] \to \mathbb{R}^n$ que l'on définit plus loin. On a $f(u) = \int_0^T [J(y(t,u),u(t),t) + p(t)^\mathsf{T}(\dot{y}(t) - \varphi(y(t,u),u(t),t))]\,\mathrm{d}t + J^T(y(T,u))$, lorsque y est solution de l'équation d'état (5.24). En intégrant par parties, on trouve $f(u) = \int_0^T [J(y(t,u),u(t),t) - \dot{p}(t)^\mathsf{T} y(t,u) - p(t)^\mathsf{T}\varphi(y(t,u),u(t),t)]\,\mathrm{d}t + p(T)^\mathsf{T} y(T,u) - p(0)^\mathsf{T} y_0 + J^T(y(T,u))$. En dérivant $f'(u)\cdot v = \int_0^T [J'_y(y(t),u(t),t)\cdot(y'(t,u)\cdot v) + J'_u(y(t),u(t),t)\cdot v(t) - \dot{p}(t)^\mathsf{T}(y'(t,u)\cdot v) - p(t)^\mathsf{T}(\varphi'_y(y(t),u(t),t)\cdot(y'(t,u)\cdot v) + \varphi'_u(y(t),u(t),t)\cdot v(t))]\,\mathrm{d}t + p(T)^\mathsf{T}(y'(T,u)\cdot v) + (J^T)'(y(T))\cdot(y'(T,u)\cdot v)$. Ceci conduit à l'état adjoint défini dans l'énoncé, ainsi qu'à la valeur de $f'(u)\cdot v$ donnée.

5.8. Considérons d'abord le code linéaire tangent. On se rappelle que $\dot{x} = -A^{-1}\dot{A}A^{-1}b + A^{-1}\dot{b}$. Dès lors, après avoir calculé $x := A^{-1}b$, on obtient $\dot{x}$ par $\dot{x} := A^{-1}(\dot{b} - \dot{A}x)$. Intéressons-nous à présent au code adjoint. On se place dans le cadre de l'exercice 5.7 en notant $\varphi(A,b) := A^{-1}b$ l'instruction à valeurs vectorielles qui définit la solution du système linéaire. Si on introduit $x := A^{-1}b$ et $z := A^{-\mathsf{T}}\overline{x}$, on a $\overline{x}^\mathsf{T}\varphi'(A,b)\cdot(H,d) = \overline{x}^\mathsf{T}(-A^{-1}HA^{-1}b + A^{-1}d) = z^\mathsf{T}(-Hx + d)$. Alors, d'après (5.34), $\overline{A}_{ij} := \overline{A}_{ij} + \overline{x}^\mathsf{T}\varphi(A,b)\cdot(E^{ij},0) = \overline{A}_{ij} - z_i x_j$. Vectoriellement, cela s'écrit $\overline{A} := \overline{A} - zx^\mathsf{T}$. De même $\overline{b}_i := \overline{b}_i + \overline{x}^\mathsf{T}\varphi(A,b)\cdot(0,e^i) = \overline{b}_i + z_i$ ou vectoriellement $\overline{b} := \overline{b} + z$. Comme la valeur précédente de la variable x qui est modifiée par l'instruction n'intervient pas dans ce calcul (elle est « écrasée » par l'instruction), on doit annuler $\overline{x}$ à la fin de code adjoint (c'est ce qu'exprime la dernière instruction de (5.34)). On remarque enfin, que l'on peut stocker z dans $\overline{x}$.

Si le code original s'écrit $b := A^{-1}b$, le code tangent devient

$$b := A^{-1}b, \quad \dot{b} := A^{-1}(\dot{b} - \dot{A}b)$$

et le code adjoint devient

$$b := A^{-1}b, \quad \overline{b} := A^{-\mathsf{T}}\overline{x}, \quad \overline{A} := \overline{A} - \overline{b}b^{\mathsf{T}}.$$

Ces instructions peuvent s'obtenir avec le même raisonnement que ci-dessus. On peut aussi considérer que le code original est formé des deux instructions $x := A^{-1}b$ et $b := x$, dont les codes tangent et adjoint s'écrivent

$$\begin{array}{c} x := A^{-1}b, \quad \dot{x} := A^{-1}(\dot{b} - \dot{A}x), \quad \dot{b} := \dot{x} \\ x := A^{-1}b, \quad \overline{x} := A^{-\mathsf{T}}\overline{x}, \quad \overline{b} := \overline{b} + \overline{x}, \quad \overline{A} := \overline{A} - \overline{x}x^{\mathsf{T}}, \quad \overline{x} := 0. \end{array}$$

En observant qu'à la première utilisation $\overline{x} = 0$ (x n'existe pas dans le code). En simplifiant on retrouve les codes proposés ci-dessus.

5.9. Le code direct s'écrit

$$\begin{aligned} f &:= \frac{1}{2}\,x^{\mathsf{T}}Ax + b^{\mathsf{T}}x \\ \dot{f} &:= (Ax+b)^{\mathsf{T}}\dot{x} + \dot{b}^{\mathsf{T}}x + \frac{1}{2}\,x^{\mathsf{T}}\dot{A}x. \end{aligned}$$

Le code adjoint s'écrit

$$\begin{aligned} f &:= \frac{1}{2}\,x^{\mathsf{T}}Ax + b^{\mathsf{T}}x \\ \overline{x} &:= \overline{x} + \overline{f}(Ax+b) \\ \overline{b} &:= \overline{b} + \overline{f}x \\ \overline{A} &:= \overline{A} + \frac{1}{2}\,\overline{f}\,xx^{\mathsf{T}} \\ \overline{f} &:= 0. \end{aligned}$$

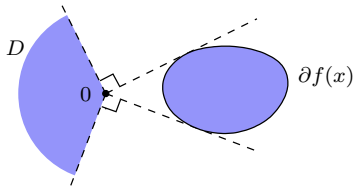
Méthodes à directions de descente

6.1. 1) Les courbes de niveau étant des ellipses, la fonction est soit strictement convexe, soit strictement concave. La direction du gradient en y montre qu'elle est strictement concave (faire une coupe le long de la droite $y+\mathbb{R}\nabla f(y)$); la courbe de niveau encercle une « colline » de f. C'est donc dans le dessin A que d est de descente.

2) On élève l'inégalité au carré et on développe : $\|\nabla f(x) + \alpha d\|^2 = \|\nabla f(x)\|^2 + 2\alpha\langle\nabla f(x), d\rangle + \alpha^2\|d\|^2 \leqslant \|\nabla f(x)\|^2$. On en déduit que $\langle\nabla f(x), d\rangle \leqslant -(\alpha/2)\|d\|^2 < 0$, car $\alpha > 0$ et $d \neq 0$. Donc d est une direction de descente de f en x.

6.2. 1) Il suffit d'écrire l'inégalité de convexité en x et y : $f(x) + f'(x)\cdot(y-x) \leqslant f(y) < f(x)$. On en déduit que $f'(x)\cdot(y-x) < 0$, ce qui montre que $y-x$ est une direction de descente de f en x.

2.a) En utilisant la formule du max, on a $D := \{d \in \mathbb{R}^2 : g^{\mathsf{T}}d \leqslant 0,\ \forall g \in \partial f(x)\} = \partial f(x)^-$ (le cône dual négatif de $\partial f(x)$). On en déduit que l'ensemble D est le cône représenté ci-dessous.



2.b) En prenant $f : \mathbb{R}^2 \to \mathbb{R}$ définie en $x \in \mathbb{R}^2$ par $f(x) = \max(x_1 + \alpha x_2, x_1 - \alpha x_2)$, avec $\alpha > 0$, on trouve $\partial f(0) = \mathrm{co}\{(1,\alpha),(1,-\alpha)\}$. Par la formule du max : $f'(0;d) = d_1 + \alpha|d_2|$, si bien que $D = \{d \in \mathbb{R}^2 : \alpha|d_2| \leqslant -d_1\}$. Donc si $\alpha > 1$, $d = (-1,\alpha)$ est dans $-\partial f(0)$ mais pas dans D.

2.c) Par définition de la projection, on a pour tout $g \in \partial f(x)$: $\langle g - p, p - 0\rangle \geqslant 0$ ou $\langle g, -p\rangle \leqslant -\|p\|^2$. Par la formule du max, on a $f'(x;-p) \leqslant -\|p\|^2$.

3) Soit $\nu(\cdot) := \|\cdot\|$. La fonction $f = \nu \circ F$ a des dérivées directionnelles, car il s'agit de la composition de deux fonctions ayant des dérivées directionnelles avec la seconde ν lipschitzienne. Si on note $r(x) := F(x) + F'(x)\cdot d$, on a $f'(x;d) = \nu'(F(x);F'(x)\cdot d) = \nu'(F(x);r(x) - F(x)) = \lim_{t\downarrow 0} \frac{1}{t}(\|F(x) + t[r(x) - F(x)]\| - \|F(x)\|)$. Par convexité de la norme, on a $\|F(x) + t[r(x) - F(x)]\| \leqslant t\|r(x)\| + (1-t)\|F(x)\|$, dès que $0 \leqslant t \leqslant 1$, si bien que $f'(x;d) \leqslant \|r(x)\| - \|F(x)\| < 0$.

6.3. On considère l'algorithme suivant, qui s'inspire de l'algorithme de Fletcher-Lemaréchal déterminant un pas de Wolfe (l'algorithme 6.6).

Algorithme D.1 (détermination d'un pas de Goldstein)

1. Soient $\underline{\alpha} := 0$, $\overline{\alpha} := +\infty$, $\tau_i \in]0, \frac{1}{2}[$ et $\tau_e > 1$;
   On se donne un premier pas $\alpha > 0$;
2. Répéter :
   2.1. Si (6.10a) n'est pas vérifiée avec $\alpha_k = \alpha$;
   2.2. Alors $\overline{\alpha} = \alpha$ et on choisit un nouveau pas α dans l'intervalle

$$[(1-\tau_i)\underline{\alpha} + \tau_i\overline{\alpha}, \tau_i\underline{\alpha} + (1-\tau_i)\overline{\alpha}]; \tag{D.35}$$

   2.3. Sinon
      2.3.1. Si (6.10b) est vérifiée avec $\alpha_k = \alpha$,
      2.3.2. Alors on sort avec $\alpha_k = \alpha$;
      2.3.3. Sinon
         2.3.3.1. $\underline{\alpha} = \alpha$;
         2.3.3.2. Si $\overline{\alpha} = +\infty$;
         2.3.3.3. Alors choisir un nouveau pas $\alpha \in [\tau_e\underline{\alpha}, \infty[$;
         2.3.3.4. Sinon choisir un nouveau pas α dans l'intervalle (D.35).

Les conditions de terminaison finie de l'algorithme D.1 sont semblables à celles de l'algorithme de Fletcher-Lemaréchal pour la détermination d'un pas de Wolfe (proposition 6.7), à savoir : si d_k est une direction de descente de f en x_k, si $h_k : \mathbb{R}_+ \to \mathbb{R}$ définie par (6.8) est *continue*, dérivable en zéro et bornée inférieurement et si $0 < \omega_1 < \omega_1' < 1$, alors l'algorithme ci-dessus trouve un pas $\alpha_k > 0$ vérifiant les conditions (6.10) en un nombre fini d'étapes. Démontrons cette affirmation.

On raisonne par l'absurde en supposant que la boucle « répéter » (instruction 2) est parcourue indéfiniment. Pour la même raison que dans l'algorithme 6.6 ((6.10a) est vérifiée pour des pas $\underline{\alpha} \to +\infty$), l'instruction 2.3.3.3 n'est exécutée qu'un nombre fini de fois et donc $\overline{\alpha} < +\infty$ après un nombre fini d'étapes. Ensuite les nouveaux pas-candidats α sont choisis dans l'intervalle $]\underline{\alpha}, \overline{\alpha}[$ avec $\underline{\alpha} \to \hat{\alpha}$ et $\overline{\alpha} \to \hat{\alpha}$ pour un certain $\hat{\alpha} > 0$ (on ne peut avoir $\hat{\alpha} = 0$ car (6.10a) est vérifiée pour α petit). En passant à la limite dans $h_k(\overline{\alpha}) > h_k(0) + \omega_1\underline{\alpha}h_k'(0)$, on trouve $h_k(\hat{\alpha}) \geqslant h_k(0) + \omega_1\hat{\alpha}h_k'(0)$. En passant à la limite dans $h_k(\underline{\alpha}) < h_k(0) + \omega_1'\underline{\alpha}h_k'(0)$, on trouve $h_k(\hat{\alpha}) \leqslant h_k(0) +$

$\omega_1'\hat{\alpha}h_k'(0)$. Comme $\hat{\alpha} > 0$, $\omega_1 < \omega_1'$ et $h_k'(0) < 0$, les deux inégalités vérifiées par $\hat{\alpha}$ sont contradictoires.

6.4. Par hypothèse, on dispose d'une direction $d \in \mathbb{E}$ telle que $f'(x) \cdot d < 0$. Soit $\nabla f(x)$ le gradient de f en x pour un produit scalaire $\langle \cdot, \cdot \rangle$ équipant $\mathbb{E}$. On considère l'application $\sigma : (u, v) \mapsto \langle Su, v \rangle$, où

$$S := I - \frac{\nabla f(x) \otimes \nabla f(x)}{\langle \nabla f(x), d \rangle} - \frac{d \otimes d}{\|d\|^2},$$

où I est l'identité sur $\mathbb{E}$ et $u \otimes v : \mathbb{E} \to \mathbb{E}$ est l'opérateur défini en $z \in \mathbb{E}$ par $(u \otimes v)z = \langle v, z \rangle u$ (on reconnaît la formule de BFGS (11.9) de mise à jour de l'identité). Comme S est suto-adjoint et défini positif pour le produit scalaire $\langle \cdot, \cdot \rangle$, l'application bilinéaire σ est un produit scalaire sur $\mathbb{E}$. Par ailleurs, on a $Sd = -\nabla f(x)$, si bien que le gradient de f pour le produit scalaire σ, qui vaut $S^{-1}\nabla f(x)$, n'est autre que $-d$.

Premiers algorithmes

7.1. [(i)] Soit $g \in \partial f(x)$. Pour tout $\bar{g} \in \partial f(\bar{x})$, on a par monotonie du sous-différentiel : $\langle \bar{g}, x - \bar{x} \rangle \leqslant \langle g, x - \bar{x} \rangle$. En prenant $\bar{g} = \varepsilon(x - \bar{x})/\|x - \bar{x}\| \in \partial f(\bar{x})$, on a $\varepsilon\|x - \bar{x}\| \leqslant \langle g, x - \bar{x} \rangle \leqslant \|g\| \, \|x - \bar{x}\|$. Dès lors $x \neq \bar{x} \Longrightarrow \|g\| \geqslant \varepsilon$.

[(ii)] Si $\|x - \bar{x}\| \leqslant \varepsilon$, l'hypothèse $\bar{B}_\varepsilon \subseteq \partial f(\bar{x})$ implique qu'il existe $\bar{g} \in \partial f(\bar{x})$ tel que $x - \bar{x} = \bar{g}$ ou $\bar{x} = x - \bar{g}$. Par l'équivalence (3.64), $x_p = \bar{x}$.

[(iii)] Pour $x \in \bar{B}(\bar{x}, \varepsilon)$, $x_p = \bar{x}$ ne dépend plus de x et donc $\tilde{f}(x) = f(\bar{x}) + \frac{1}{2}\|\bar{x} - x\|^2$ qui est bien quadratique en x.

Optimisation quadratique

8.1. (i) Montrons d'abord que $E_p \subseteq K_p$ par récurrence. L'inclusion est claire pour $p = 1$. Supposons donc qu'elle est vérifiée pour $p - 1$ (avec $p \geqslant 2$) et montrons là pour p. Comme $E_{p-1} \subseteq K_{p-1} \subseteq K_p$, il suffit de montrer que $d_p \in K_p$. Mais $d_p = -g_p + \beta_p d_{p-1}$ et $d_{p-1} \in E_{p-1} \subseteq K_p$; il suffit donc de montrer que $g_p \in K_p$. On utilise pour cela le fait que $g_p = g_{p-1} + \alpha_{p-1}Ad_{p-1}$, que $g_{p-1} \in E_{p-1} \subseteq K_p$ et que $Ad_{p-1} \in AK_{p-1} \subseteq K_p$.

L'inclusion réciproque $K_p \subseteq E_p$ se montre aussi par récurrence. Elle est claire pour $p = 1$. Supposons donc qu'elle est vérifiée pour $p - 1$ (avec $p \geqslant 2$) et montrons là pour p. Par récurrence, il suffit de montrer que $A^{p-1}g_1 \in E_p$. Comme $A^{p-2}g_1 \in E_{p-1} = \operatorname{vect}\{d_1, \ldots, d_{p-1}\}$ et que pour $i = 1, \ldots, p-2$ on a par récurrence $Ad_i \in AE_i \subseteq AK_i \subseteq K_{i+1} \subseteq E_{i+1} \subseteq E_p$, il reste à montrer que $Ad_{p-1} \in E_p$. Pour voir cela, on écrit $\alpha_{p-1}Ad_{p-1} = g_p - g_{p-1} \in E_p$.

8.2. Comme $d_i^{\mathsf{T}} g_j = 0$ si $j > i$, on a pour $1 \leqslant i < k$

$$d_i^{\mathsf{T}} d_k = \beta_k d_i^{\mathsf{T}} d_{k-1} = \beta_k \beta_{k-1} \cdots \beta_{i+1} \|d_i\|^2 > 0.$$

On en déduit

$$\|x_{k+1} - x_1\|_2^2 = \|x_k - x_1\|_2^2 + 2\alpha_k d_k^{\mathsf{T}}(x_k - x_1) + \alpha_k^2 \|d_k\|_2^2 > \|x_k - x_1\|_2^2,$$

car $d_k^{\mathsf{T}}(x_k - x_1) = \sum_{i=1}^{k-1} \alpha_i d_k^{\mathsf{T}} d_i > 0$. De manière similaire

$$\|x_{k+1} - x_*\|_2^2 = \|x_k - x_*\|_2^2 + 2\alpha_k d_k^{\mathsf{T}}(x_k - x_*) + \alpha_k^2 \|d_k\|_2^2 < \|x_k - x_*\|_2^2,$$

car $d_k^{\mathsf{T}}(x_k - x_*) = -\sum_{i=k}^{p-1} \alpha_i d_k^{\mathsf{T}} d_i < 0$ et $d_k^{\mathsf{T}}(x_k - x_*) + \alpha_k \|d_k\|_2^2 = -\sum_{i=k+1}^{p-1} \alpha_i d_k^{\mathsf{T}} d_i < 0$ (s'il faut p itérations pour converger).

(ii) On montre les trois relations suivantes : si $k \geqslant 2$ et $g_k \neq 0$

$$\begin{aligned} \Big(d_1^{\mathsf{T}}(x_k - x_1)\Big)\|d_k\|_2^2 &> \Big(d_1^{\mathsf{T}} d_k\Big)\Big(d_k^{\mathsf{T}}(x_k - x_1)\Big) \\ \Big(d_1^{\mathsf{T}}(x_k - x_1)\Big)\Big(d_k^{\mathsf{T}}(x_k - x_1)\Big) &\geqslant \Big(d_1^{\mathsf{T}} d_k\Big)\|x_k - x_1\|_2^2 \qquad \text{(D.36)} \\ \Big(d_1^{\mathsf{T}}(x_k - x_1)\Big)\|d_k\|_2 &> \Big(d_1^{\mathsf{T}} d_k\Big)\|x_k - x_1\|_2. \qquad \text{(D.37)} \end{aligned}$$

Les deux premières s'obtiennent par récurrence en y substituant $d_k = -g_k + \beta_k d_{k-1}$, $\|d_k\|_2^2 = \|g_k\|_2^2 + \beta_k^2 \|d_{k-1}\|_2^2$ et $x_k = x_{k-1} + \alpha_{k-1} d_{k-1}$. En utilisant les deux premières relations on trouve $(d_1^{\mathsf{T}}(x_k - x_1))^2 \|d_k\|_2^2 > (d_1^{\mathsf{T}}(x_k - x_1))(d_1^{\mathsf{T}} d_k)(d_k^{\mathsf{T}}(x_k - x_1)) \geqslant (d_1^{\mathsf{T}} d_k)^2 \|x_k - x_1\|_2^2$, qui montre que la troisième relation a lieu.

Pour montrer la croissance stricte de l'angle entre $x_2 - x_1$ et $x_k - x_1$, on montre la décroissance stricte de son cosinus multiplié par $\|x_2 - x_1\|/\alpha_1$. Il s'agit de voir si

$$\frac{\Big(d_1^{\mathsf{T}}(x_{k+1} - x_1)\Big)^2}{\|x_{k+1} - x_1\|_2^2} < \frac{\Big(d_1^{\mathsf{T}}(x_k - x_1)\Big)^2}{\|x_k - x_1\|_2^2}.$$

En y substituant $x_k = x_{k-1} + \alpha_{k-1} d_{k-1}$, cette relation sera vérifiée ssi

$$\frac{\Big(d_1^{\mathsf{T}}(x_k - x_1)\Big)^2}{\|x_k - x_1\|_2^2} > \frac{2\Big(d_1^{\mathsf{T}}(x_k - x_1)\Big)\Big(d_1^{\mathsf{T}} d_k\Big) + \alpha_k \Big(d_1^{\mathsf{T}} d_k\Big)^2}{2\Big(d_k^{\mathsf{T}}(x_k - x_1)\Big) + \alpha_k \|d_k\|_2^2}.$$

Cette dernière inégalité se vérifie grâce aux relations (D.36) et (D.37).

8.3. Considérons l'algorithme des directions conjuguées dans lequel $\tilde{d}_k = -Pg_k$. D'après (8.10), on a

$$d_k = \begin{cases} -Pg_1 & \text{si } k = 1 \\ -Pg_k + \sum_{i=1}^{k-1} \frac{g_k^{\mathsf{T}} P A d_i}{d_i^{\mathsf{T}} A d_i} d_i & \text{si } k \geqslant 2. \end{cases}$$

Le procédé de Gram-Schmidt donne la première égalité de (ii). Pour la seconde, on procède comme à l'exercice 8.1.

Ensuite, on a (iii) parce que g_k est orthogonal à $\mathrm{vect}\{d_1, \ldots, d_k\}$ (proposition 8.5). En utilisant $As_i = y_i$ et les propriétés d'orthogonalité de (iii), on obtient pour $k \geqslant 2$

$$d_k = -Pg_k + \frac{g_k^{\mathsf{T}} P g_k}{d_{k-1}^{\mathsf{T}} y_{k-1}} d_{k-1}.$$

Il reste à observer que $d_{k-1}^{\mathsf{T}} y_{k-1} = -d_{k-1}^{\mathsf{T}} g_{k-1} = g_{k-1}^{\mathsf{T}} P g_{k-1}$ pour constater qu'il s'agit bien de l'algorithme du GC préconditionné par P (formule (8.16)).

8.4. Avec les notations introduites au point 8.4 de la section D.1, on montre que, pour tout $k \geqslant 1$, on a $Z d_{k-1}^r = d_{k-1}$ et $x_k = x_1 + Z u_k$. Les identités ont lieu pour $k = 1$, car $d_0^r = 0$, $d_0 = 0$ et $u_1 = 0$. Pour $k \geqslant 2$, on procède par récurrence : en supposant que les relations ont lieu jusqu'à l'indice k, on montre qu'elles ont lieu pour l'indice $k+1$. Observons d'abord que $\nabla f^r(u) = Z^{\mathsf{T}} \nabla f(x_1 + Zu)$. D'autre part, le β de conjugaison du GC réduit s'écrit $\beta_k^r = \|\nabla f^r(u_k)\|^2 / \|\nabla f^r(u_{k-1})\|^2$. Comme $\|\nabla f^r(u_k)\|^2 = \|Z^{\mathsf{T}} \nabla f(x_k)\|^2 = \nabla f(x_k)^{\mathsf{T}} P \nabla f(x_k)$, on en déduit que β_k^r est le β_k de conjugaison de l'algorithme du GC préconditionné par P. Dès lors $Z d_k^r =$

$-Z\nabla f^r(u_k) + \beta_k^r Z d_{k-1}^r = -P\nabla f(x_k) + \beta_k d_{k-1} = d_k$. Ensuite $x_1 + Zu_{k+1} = x_1 + Zu_k + \alpha_k^r Z d_k^r = x_k + \alpha_k^r d_k$, si bien qu'il reste à montrer que α_k^r est aussi le pas optimal α_k du GC préconditionné. Or on a bien $\alpha_k^r = -(\nabla f^r(u_k)^\mathsf{T} d_k^r)/((d_k^r)^\mathsf{T} Z^\mathsf{T} A Z(d_k^r)) = -(\nabla f(x_k)^\mathsf{T} d_k)/(d_k^\mathsf{T} A d_k) = \alpha_k$.

8.5. D'après l'exercice 8.4, le GC projeté dans l'espace des $\tilde{x}$ a pour direction de recherche

$$\tilde{d}_k = -\tilde{P}\tilde{g}_k + \frac{\tilde{g}_k^\mathsf{T}\tilde{P}\tilde{g}_k}{\tilde{g}_{k-1}^\mathsf{T}\tilde{P}\tilde{g}_{k-1}}\,\tilde{d}_{k-1},$$

où $\tilde{P} = \tilde{Z}\tilde{Z}^\mathsf{T}$ est le projecteur sur le sous-espace affine $\tilde{\mathcal{A}} := \{Lx : x \in \mathcal{A}\}$. La matrice orthogonale Z associée au sous-espace affine $\mathcal{A}$ étant donnée, on peut prendre $\tilde{Z} = LZ(Z^\mathsf{T}L^\mathsf{T}LZ)^{-1/2}$ comme matrice orthogonale associée à $\tilde{\mathcal{A}}$. Le retour à l'espace des x donne

$$d_k = L^{-1}\tilde{d}_k = -Pg_k + \frac{g_k^\mathsf{T} P g_k}{g_{k-1}^\mathsf{T} P g_{k-1}}\,d_{k-1},$$

où $P = Z(Z^\mathsf{T}L^\mathsf{T}LZ)^{-1}Z^\mathsf{T}$.

Algorithme de Newton

10.1. xxx

10.2. Comme F est différentiable, φ a des dérivées directionnelles et on a $\varphi'(x;d) = (\|\cdot\|)'(F(x);F'(x)d) = (\|\cdot\|)'(F(x);-F(x)) = -\|F(x)\|$. Il s'agit donc de voir si $\|F(x+d)\| \leqslant (1-\omega)\|F(x)\|$ pour x proche de x_*. La régularité de F et $F(x)+F'(x)d = 0$ assurent l'existence de constantes $C_i > 0$ telles que $\|F(x+d)\| \leqslant C_1\|d\|^2 \leqslant C_2\|F(x)\|^2$. On en déduit que $\|F(x+d)\| - (1-\omega)\|F(x)\| \leqslant [C_2\|F(x)\| - (1-\omega)]\|F(x)\| \leqslant 0$, pour x proche de x_*.

Si $\varphi(x) = \|F(x)\|^p$, on a $\varphi'(x;d) = -p\|F(x)\|^p$ et il s'agit de montrer que $\|F(x+d)\|^p \leqslant (1-p\omega)\|F(x)\|^p$ ou encore $\|F(x+d)\| \leqslant (1-p\omega)^{1/p}\|F(x)\|$. Comme ci-dessus, cette inégalité est vérifiée pour x voisin de x_*.

10.3. 1) Soit $\xi := g \circ x$. Pour $t \in I$, on a $\xi'(t) = H(x(t))x'(t) = -g(x(t)) = -\xi(t)$. On en déduit que $\xi(t) = e^{-t}\xi(0)$. Si $\xi(0) = g(x_0) \neq 0$, on a $\xi(t) \neq 0$ pour tout $t \geqslant 0$ et donc $\xi(t)/\|\xi(t)\| = \xi(0)/\|\xi(0)\|$ est constant.

2)

3) Par le point 1, on a $e^t x'(t) = -e^t[H(x(t))^{-1}g(x(t))] = -H(x(t))^{-1}g(x_0)$. À la limite, on obtient le résultat.

4) Cela résulte de la stricte convexité de f sur $\mathcal{N}_0$. Supposons, en effet, que $g(x_1) = g(x_2)$ avec $x_1, x_2 \in \mathcal{N}_0$. On considère l'application $x \mapsto \phi(x) = f(x) - g(x_1)^\mathsf{T}x$, qui est telle que $\nabla\phi(x_1) = \nabla\phi(x_2) = 0$. Comme ϕ est strictement convexe, on a $x_1 = x_2$.

5) Si $x(t) \notin \overline{\Omega}$, on a $\tilde{g}(x(t)) = g(x(t))$ [car $f = \tilde{f}$ sur l'ouvert $\overline{\Omega}^c$] $= e^{-t}g(x_0)$ [par le point 1] $= e^{-t}\tilde{g}(x_0)$ [car $f = \tilde{f}$ sur l'ouvert $\overline{\Omega}^c$ et $x_0 \in \overline{\Omega}^c$] $= \tilde{g}(\tilde{x}(t))$ [par le point 1]. L'égalité démontrée $\tilde{g}(x(t)) = \tilde{g}(\tilde{x}(t))$, la stricte convexité de $\tilde{f}$ et le point 3 impliquent que $x(t) = \tilde{x}(t)$. Par continuité, $x(t) = \tilde{x}(t)$ si $x(t) \notin \Omega$. Par symétrie $x(t) = \tilde{x}(t)$ si $\tilde{x}(t) \notin \Omega$.

Algorithmes de quasi-Newton

11.1. La condition $Y^\mathsf{T}S \in \mathcal{S}^n_{++}$ est nécessaire, car si $Y = \bar{M}S$, on a $Y^\mathsf{T}S = S^\mathsf{T}\bar{M}S$, qui est symétrique définie positive lorsque $\bar{M} \in \mathcal{S}^n_{++}$ et S est injective. Le calcul effectif d'une solution du problème montrera que la condition est aussi suffisante. On suppose donc à présent que $Y^\mathsf{T}S \in \mathcal{S}^n_{++}$.

On sait que le critère est strictement convexe (démonstration de la proposition 11.2) et donc que le problème a au plus une solution. On calcule une solution du problème par ses conditions d'optimalité, ce qui prouvera l'existence d'une solution. On note $\Lambda \in \mathbb{R}^{n\times p}$ le multiplicateur de Lagrange associé à la contrainte $Y = \bar{M}S$. Le lagrangien s'écrit $\ell(\bar{M}, \Lambda) = \psi(W^{1/2}\bar{M}W^{1/2}) + \operatorname{tr}\Lambda^\mathsf{T}(Y - \bar{M}S)$. Après un travail similaire à celui mené dans la démonstration de la proposition 11.2, la condition $\ell'_{\bar{M}}(\bar{M}, \Lambda)\cdot H = 0$ pour tout $H \in \mathcal{S}^n$ conduit à

$$\bar{W} = W - \frac{1}{2}(S\Lambda^\mathsf{T} + \Lambda S^\mathsf{T}). \tag{D.38}$$

Pour calculer Λ, on utilise la contrainte $Y = \bar{M}S$ et la symétrie de $Y^\mathsf{T}S$:

$$S = WY - \frac{1}{2}\Big(S(\Lambda^\mathsf{T}Y) + \Lambda(Y^\mathsf{T}S)\Big). \tag{D.39}$$

En multipliant à gauche par Y^T, on trouve l'équation en $X = \Lambda^\mathsf{T}Y$ suivante : $AX + X^\mathsf{T}A = B$, où $A = Y^\mathsf{T}S \in \mathcal{S}^n_{++}$ et $B = -2Y^\mathsf{T}(S - WY) \in \mathcal{S}^n$. Ses solutions sont de la forme $X = \frac{1}{2}A^{-1}B + K$, où K est une solution arbitraire de l'équation homogène $AK + K^\mathsf{T}A = 0$ ($K = A^{-1}Z$ avec Z antisymétrique). On a donc $\Lambda^\mathsf{T}Y = -(Y^\mathsf{T}S)^{-1}$ $Y^\mathsf{T}(S - WY) + K$, qui réinjecté dans (D.39) donne la valeur de Λ :

$$\Lambda = -\Big(2(S - WY) - S(Y^\mathsf{T}S)^{-1}Y^\mathsf{T}(S - WY) + SK\Big)(Y^\mathsf{T}S)^{-1}.$$

On trouve alors $\bar{W}$ par (D.38) (K disparaît) et $\bar{M}$.

Il reste à montrer que la matrice $\bar{M}$ obtenue est bien dans $\mathcal{S}^n_{++}$. Quel que soit $v \in \mathbb{R}^n$, on a $v^\mathsf{T}Y(Y^\mathsf{T}S)^{-1}Y^\mathsf{T}v \geqslant 0$ et

$$v^\mathsf{T}MS(S^\mathsf{T}MS)^{-1}S^\mathsf{T}Mv = u^\mathsf{T}R^\mathsf{T}(RR^\mathsf{T})^{-1}Ru \leqslant \|u\|_2^2 = v^\mathsf{T}Mv,$$

où on a noté $R = S^\mathsf{T}M^{1/2}$ et $u = M^{1/2}v$; et utilisé le fait que $P := R^\mathsf{T}(RR^\mathsf{T})^{-1}R$ est un projecteur orthogonal (donc de norme ℓ_2 inférieure à 1). Donc $\bar{M}$ est semi-définie positive. Enfin, si $v^\mathsf{T}\bar{M}v = 0$, on a $Y^\mathsf{T}v = 0$ et $v \in \mathcal{R}(S)$ (car $Pu = u$); alors $Y^\mathsf{T}S \in \mathcal{S}^n_{++}$ implique que $v = 0$.

11.2. Si $s_k = 0$ et $y_k \neq 0$, le problème (11.26) n'est pas réalisable, il est donc sans solution. Si $s_k = y_k = 0$, le problème (11.26) est réalisable et, son critère étant quadratique (en prenant le carré) fortement convexe, il a au plus une solution. Le critère est nul en $M = M_k$; cette matrice est donc bien l'unique solution.

Considérons à présent le cas où $s_k \neq 0$. Alors, le problème (11.26) a une solution et une seule, puisque son ensemble admissible est un convexe fermé (c'est un sous-espace affine) non vide (il contient la matrice M_{k+1} donnée par la formule (11.27)) et le critère est quadratique fortement convexe. Pour calculer cette solution explicitement, on utilise les conditions d'optimalité du premier ordre du problème (11.26). Soient $\lambda \in \mathbb{R}^n$ et $\Lambda \in \mathbb{R}^{n\times n}$ les multiplicateurs de Lagrange associés respectivement à l'équation de quasi-Newton et à la condition de symétrie. Le lagrangien de (11.26) s'écrit $\ell(M, \lambda, \Lambda) = \frac{1}{2}\|M - M_k\|_F^2 + \lambda^\mathsf{T}(y_k - Ms_k) + \operatorname{tr}\Lambda^\mathsf{T}(M - M^\mathsf{T})$. En annulant la dérivée de ℓ par rapport à M, on trouve que la solution s'écrit $M_{k+1} = M_k + \lambda s_k^\mathsf{T} + \Lambda - \Lambda^\mathsf{T}$. En exprimant que M_{k+1} est symétrique, on obtient $M_{k+1} = M_k + (\lambda s_k^\mathsf{T} + s_k\lambda^\mathsf{T})/2$. L'équation de quasi-Newton permet alors de déterminer $\lambda = (2/\|s_k\|_2^2)(y_k - M_k s_k - s_k^\mathsf{T}(y_k - M_k s_k)/(2\|s_k\|_2^2)s_k)$. On en déduit (11.27).

11.3. On minimise d'abord en M puis en σ. D'après l'exercice 11.2, la minimisation en $M \in \mathbb{R}^{n\times n}$ donne

$$M = \sigma I + \frac{(y_1 - \sigma s_1)s_1^\mathsf{T} + s_1(y_1 - \sigma s_1)^\mathsf{T}}{\|s_1\|_2^2} - \frac{(y_1 - \sigma s_1)^\mathsf{T} s_1}{\|s_1\|_2^4} s_1 s_1^\mathsf{T}.$$

Il faut alors résoudre

$$\min_\sigma \left\| \|s_1\|_2^2 (y_1 - \sigma s_1)s_1^\mathsf{T} + \|s_1\|_2^2 s_1(y_1 - \sigma s_1)^\mathsf{T} - ((y_1 - \sigma s_1)^\mathsf{T} s_1)s_1 s_1^\mathsf{T} \right\|_F,$$

qui, après un calcul laborieux, donne la formule de σ_1 dans (11.20). La solution du second problème s'obtient à partir de celle du premier, puisqu'il s'obtient en faisant les substitutions $y_1 \leftrightarrow s_1$ et $M \to W$ dans le premier problème.

11.4. Les propriétés (i)-(iv) sont clairement vraies pour $k = 1$. Supposons les vérifiées pour k et montrons les pour $k + 1$.

(i) Soit $v \in \mathcal{G}_{k+1}$. Par récurrence, $s_k := -\alpha_k W_k g_k \in W_1(\mathcal{G}_k) \subseteq W_1(\mathcal{G}_{k+1})$. Ensuite, soient v_k la projection W_1-orthogonale de v sur $\mathcal{G}_k$ et $v_k^\perp := v - v_k$ qui est W_1-orthogonale à $\mathcal{G}_k$. Par récurrence $W_k v_k \in W_1(\mathcal{G}_k) \subseteq W_1(\mathcal{G}_{k+1})$ et $W_k v_k^\perp = W_1 v_k^\perp \in W_1(\mathcal{G}_{k+1})$, si bien que $W_k v \in W_1(\mathcal{G}_{k+1})$. Enfin, comme $y_k \in \mathcal{G}_{k+1}$, on a aussi $W_k y_k \in W_1(\mathcal{G}_{k+1})$. Alors, d'après (11.10), on a

$$W_{k+1}v \in W_k \left(v - \frac{s_k^\mathsf{T} v}{y_k^\mathsf{T} s_k} y_k \right) + \mathrm{vect}\{s_k\} \in W_1(\mathcal{G}_{k+1}).$$

(ii) Soit $v \perp_{W_1} \mathcal{G}_{k+1}$. Alors $s_k^\mathsf{T} v = 0$, car $s_k \in W_1(\mathcal{G}_{k+1})$, par récurrence (voir le point (i)). De même $y_k^\mathsf{T} W_k v = 0$, car $W_k y_k \in W_1(\mathcal{G}_{k+1})$, par récurrence (voir le point (i)). Alors la formule (11.10) donne $W_{k+1}v = W_k v = W_1 v$, par récurrence puisque $v \perp_{W_1} \mathcal{G}_k$.

(iii) Soit $v \in \mathcal{G}_{k+1}$. On note v_k la projection orthogonale de v sur $\mathcal{G}_k$ et $v_k^\perp := v - v_k$ qui est orthogonale à $\mathcal{G}_k$. Par récurrence, $M_k v_k \in \mathcal{G}_k \subseteq \mathcal{G}_{k+1}$ et $M_k v_k^\perp = \sigma v_k^\perp \in \mathcal{G}_{k+1}$. D'autre part, $y_k \in \mathcal{G}_{k+1}$ et $M_k s_k = -\alpha_k g_k \in \mathcal{G}_k \subseteq \mathcal{G}_{k+1}$. Dès lors, la formule (11.9) donne

$$M_{k+1}v = M_k v + \frac{y_k^\mathsf{T} v}{y_k^\mathsf{T} s_k} y_k - \frac{s_k^\mathsf{T} M_k v}{s_k^\mathsf{T} M_k s_k} M_k s_k \in \mathcal{G}_{k+1}.$$

(iv) Soit $v \perp_I \mathcal{G}_{k+1}$. Par (11.9) et récurrence, on a $M_{k+1}v = M_k v = \sigma v$.

Problèmes non lisses

??. Selon la proposition **??**, il suffit de montrer que x n'est pas une solution isolée de l'inclusion fonctionnelle, c'est-à-dire que x n'est pas un point isolé de X. Comme (QC-MF) est identique à (QC-R) dans le cas qui nous occupe, le théorème 4.58 nous apprend qu'il existe une constante positive μ telle que, pour tout $p > 0$ petit :

$$\mathrm{dist}(x, F^{-1}(\mathbb{R}^m_- - p)) \leqslant \mu \ \mathrm{dist}(F(x), \mathbb{R}^m_- - p). \tag{D.40}$$

Comme $F(x) \in \mathbb{R}^m_-$, en prenant $p = t_k e$ avec $t_k \downarrow 0$, on a $\mathrm{dist}(F(x), \mathbb{R}^m_- - t_k e) = \mathrm{dist}(F(x) + t_k e, \mathbb{R}^m_-) \to 0$, si bien que l'on trouve ainsi par (D.40) une suite $\{x_k\} \subseteq F^{-1}(\mathbb{R}^m_- - t_k e)$, c'est-à-dire qui vérifiant $F(x_k) \leqslant -t_k e \leqslant 0$ donc dans X, et qui converge vers x.

Remarque : On ne peut pas étendre cette conclusion au cas où $X := \{x \in \mathbb{E} : F(x) = 0\}$, avec $F : \mathbb{R}^n \to \mathbb{R}^n$, car alors X est réduit à un seul point.

??. Observons d'abord que

$$x \in \mathrm{Sol}\,(\mathbf{??}) \iff -F(x) \in \mathrm{N}_C(x) \iff x = \mathrm{P}_C(x - F(x)). \tag{D.41}$$

[⇒] Supposons que $x \in \mathrm{Sol}\,(\mathbf{??})$ et posons $z := x - F(x)$. Alors $x = \mathrm{P}_C(z)$ par (D.41). Ceci implique que $z = x - F(x) = \mathrm{P}_C(z) - F(\mathrm{P}_C(z))$, et donc que $F_C(z) = 0$.

[⇐] Soit z est un zéro de F_C et $x := \mathrm{P}_C(z)$. Alors $z = \mathrm{P}_C(z) - F(\mathrm{P}_C(z))$ $[F_C(z) = 0]$ $= x - F(x)$ $[x = \mathrm{P}_C(z)]$. En appliquant P_C, on trouve que $x = \mathrm{P}_C(x - F(x))$. Donc $x \in \mathrm{Sol}\,(\mathbf{??})$ par (D.41).

??. Les normes ℓ_1 et ℓ_∞ sont fortement semi-lisses car affines par morceaux (propositions **??**). Lorsque $1 < p < \infty$, la norme ℓ_p vérifie les conditions (SL$_1$) et (SL$_2$) ; il suffit donc de montrer (SL$'_3$) en $x = 0$ (la norme est c^∞ en-dehors de l'origine). Soit $f(\cdot) := \|\cdot\|_p$. On calcule facilement en $h \neq 0$, $(\partial f/\partial x_i)(h) = |h_i|^{p-1}\,\mathrm{sgn}(h_i)/\|h\|_p^{p-1}$, si bien que $f'(h) \cdot h = \|h\|_p$. Puis $f(h) - f(0) - f'(h) \cdot h = 0$, qui est certainement un $O(\|h\|^2)$.

??. La C-fonction min est le minimum de deux fonctions linéaires (fortement semi-lisses par la proposition **??**), donc fortement semi-lisse (proposition **??**). La C-fonction de Fischer-Burmeister est la somme d'une norme ℓ_2 (fortement semi-lisse par l'exercice **??**) et d'une fonction linéaire (fortement semi-lisse par la proposition **??**), donc fortement semi-lisse (proposition **??**)

Méthodes de l'optimisation avec contraintes

12.1. Par définition, $\bar{x}$ est solution du problème $\inf\{\sum_{i=1}^n (y_i - x_i)^2 : l_i \leqslant x_i \leqslant u_i$ pour tout $i \in [1\,{:}\,n]\}$, qui se décompose en n problèmes indépendants $\inf\{(y_i - x_i)^2 : l_i \leqslant x_i \leqslant u_i\}$, pour tout $i \in [1\,{:}\,n]$. Chacun de ces problèmes consiste à projeter x_i sur l'intervalle $[l_i, u_i]$, dont la solution est bien $\bar{x}_i$.

12.2. On note $g := \nabla f(x)$ et $g^{\mathrm{P}} := g^{\mathrm{P}}(x)$.

1) Par la caractérisation (2.28a) de la projection permet de caractériser g^{P} par le fait que $\langle g^{\mathrm{P}} - g, -d - g^{\mathrm{P}}\rangle \geqslant 0$, pour tout $d \in \mathrm{T}_x X$. Ceci s'écrit aussi : $\langle -g^{\mathrm{P}} + g, d + g^{\mathrm{P}}\rangle \geqslant 0$, pour tout $d \in \mathrm{T}_x X$. Cette dernière expression caractérise le fait que $-g^{\mathrm{P}} = P_{\mathrm{T}_x X}(-g)$.

2) On a en effet les équivalences suivantes :

$$\begin{aligned} g \in (\mathrm{T}_x X)^+ &\iff \forall\, d \in \mathrm{T}_x X : \langle g, d\rangle \geqslant 0 \\ &\iff \forall\, d \in -\mathrm{T}_x X : \langle 0 - g, d - 0\rangle \geqslant 0 \\ &\iff 0 = P_{-\mathrm{T}_x X} g = g^{\mathrm{P}}, \end{aligned}$$

car $-\mathrm{T}_x X$ est convexe (proposition 2.55).

3) Par définition de g^{P}, quel que soit $d \in \mathrm{T}_x X$, on a $\langle -d - g^{\mathrm{P}}, g^{\mathrm{P}} - g\rangle \geqslant 0$. En prenant $d = 0 \in \mathrm{T}_x X$, on trouve l'inégalité.

4) [⇒] Clair car la projection appartient à X. [⇐] Comme $g^{\mathrm{P}} = \mathrm{T}_{-\mathrm{T}_x X}\, g$ et que $-\mathrm{T}_x X$ est un cône, on a $\langle g - g^{\mathrm{P}}, d\rangle \geqslant 0$ quel que soit $d \in \mathrm{T}_x X$. Soit $y \in X$. Comme X est convexe, $y - x \in \mathrm{T}_x X$, si bien que $\langle g - g^{\mathrm{P}}, y - x\rangle \geqslant 0$ et donc aussi $\langle (x - \alpha g^{\mathrm{P}}) - (x - \alpha g), y - x\rangle \geqslant 0$. Alors, $x - \alpha g^{\mathrm{P}} \in X$ implique que $P_X(x - \alpha g) = x - \alpha g^{\mathrm{P}}$.

On suppose maintenant que X est polyédrique. Observons que $-g^{\mathrm{P}} \in \mathrm{T}_x X$. Alors, $x - \alpha g^{\mathrm{P}} \in X$ pour tout $\alpha > 0$ petit (exercice 2.43).

5) La formule de $\mathrm{T}_x[l,u]$ peut se déduire de l'exercice 2.43. On peut aussi donner la démonstration directe suivante. On note $T := \{d \in \mathbb{R}^n : d_i \geqslant 0 \text{ si } x_i = l_i,\ d_i \leqslant 0 \text{ si } x_i = u_i\}$. Il s'agit de montrer que $T = \mathrm{T}_x[l,u]$. [$\subseteq$] Si $d \in T$, alors $x + td \in [l,u]$ pour $t > 0$ suffisamment petit, si bien que $d \in \mathrm{T}_x^a[l,u] \subseteq \mathrm{T}_x[l,u]$ (définition 4.4 du cône des directions admissibles et 4.1 du cône tangent). [$\supseteq$] Si $d \in \mathrm{T}_x[l,u]$, on sait, par la définition équivalente (4.5) d'une direction tangente, qu'il existe une suite $\{x_k\} \subseteq [l,u]$ convergeant vers x et une suite $\{t_k\} \downarrow 0$ telles que $(x_k - x)/t_k \to d$. Si $x_i = l_i$, alors $(x_k)_i - x_i \geqslant 0$, ce qui donne $d_i \geqslant 0$ à la limite. De même $x_i = u_i$ implique que $d_i \leqslant 0$. Donc $d \in T$.

6) Se déduit de la projection sur le pavé de $\mathbb{R}^n$ qu'est $\mathrm{T}_x[l,u]$ (exercice 12.1).

7) On a $g^{\mathrm{P}} = 0$ par le point 2. On en déduit les conditions par le point 5.

Pénalisation

13.1. 1) Il suffit de montrer que $\min\{\|c_I(x) + s\|_2^2 : s \in \mathbb{R}_+^{m_I}\} = \|c_I(x)^+\|_2^2$, ce qui se déduit du fait que le problème à gauche de l'égalité consiste à projeter $-c_I(x)$ sur l'orthant positif $\mathbb{R}_+^{m_I}$. Comme le projeté est $(-c_I(x))^+ = c_I(x)^-$ (exercice 12.1), la valeur optimale du problème vaut $c_I(x) + c_I(x)^- = c_I(x)^+$.

2) Seule la dérivée de $\frac{1}{2}\|c_I(x)^+\|_2^2 = \frac{1}{2}\sum_{i\in I}[c_i(x)^+]^2$ est moins immédiate. Il suffit de voir $x \mapsto [c_i(x)^+]^2$ comme la composition de deux fonctions différentiables $x \mapsto c_i(x)$ et $t \mapsto (t^+)^2$, dont les dérivées respectives sont $c_i'(x)$ et $2t^+$, si bien que sa dérivée vaut $2c_i(x)^+c_i'(x)$. On en déduit la formule proposée (théorème C.6).

13.2. *Résultat et démonstration adaptés de* [218 ; proposition 2.3].

On sait que $H_r \succ 0$ pour r assez grand (lemme de Finsler), si bien que, pour de tels r, $\kappa_2(H_r) := \lambda_{\max}(H_r)/\lambda_{\min}(H_r)$. Donnons d'abord un encadrement de $\lambda_{\max}(H_r)$. On note $S := \{u : \|u\|_2 = 1\}$ la sphère unité. On a

$$\lambda_{\max}(H_r) = \sup_{v\in S} v^{\mathsf{T}} H_r v = \sup_{v\in S} \left(v^{\mathsf{T}} H v + r\|Av\|_2^2\right) \leqslant \lambda_{\max}(H) + r\|A\|_2^2.$$

D'autre part, on peut trouver un r_0 tel que $H_{r_0} \succcurlyeq \alpha I$, avec $\alpha > 0$ (lemme de Finsler). On peut aussi trouver un $\bar{v} \in S$ tel que $\|A\bar{v}\|_2 = \|A\|_2$. On a

$$\lambda_{\max}(H_r) \geqslant \bar{v}^{\mathsf{T}} H_r \bar{v} = \bar{v}^{\mathsf{T}} H_{r_0} \bar{v} + (r - r_0)\|A\|_2^2 \geqslant \alpha + (r - r_0)\|A\|_2^2.$$

Les deux estimations précédentes montrent que

$$\lim_{r\to+\infty} \frac{\lambda_{\max}(H_r)}{r} = \|A\|_2^2.$$

Il reste à montrer que $\lim_{r\to+\infty} \lambda_{\min}(H_r) = \sigma$. Soit v_r un vecteur propre unitaire correspondant à la valeur propre minimale de H_r. On a

$$v_r^{\mathsf{T}} H v_r \leqslant \lambda_{\min}(H_r) = v_r^{\mathsf{T}} H v_r + r\|Av_r\|_2^2 \leqslant \sigma.$$

En extrayant une sous-suite au besoin, on peut supposer que $v_r \to \bar{v}$. Comme dans la démonstration du lemme de Finsler, on voit que $\bar{v} \in \mathcal{N}(A) \cap S$, si bien que $\bar{v}^{\mathsf{T}} H \bar{v} \geqslant \sigma$. En passant à la limite ci-dessus, on trouve que $\lambda_{\min}(H_r) \to \sigma$.

13.5. Les courbes de niveaux du tracé B ne sont pas différentiables. Il ne peut donc s'agir que de Θ_r^3, qui est la seule fonction de pénalisation à ne pas être différentiable.

Le minimum de la fonction de pénalisation du tracé A vérifie la contrainte : il ne peut donc s'agir que de Θ_r^2 (exactitude du lagrangien augmenté si r est assez grand, ce qui doit être le cas ici).

Le tracé C a été obtenu avec Θ_r^1. C'est compatible; le minimum de Θ_r^1 ne vérifie pas la contrainte (c'est le cas quel que soit $r > 0$ si $\nabla f(x_*) \neq 0$).

Dualité

14.1. [$\geqslant$] Comme $D \subseteq X$, on a pour tout $\alpha \in \Delta_p$

$$\inf_{x\in X} \sum_{\substack{i\in[1:p]\\ \alpha_i\neq 0}} \alpha_i f_i(x) \leqslant \inf_{x\in D} \sum_{i\in[1:p]} \alpha_i f_i(x),$$

où l'on a pu enlever la condition $\alpha_i \neq 0$ à droite, car tous les $f_i(x) \in \mathbb{R}$ lorsque $x \in D$. En prenant le supremum en $\alpha \in \Delta_p$, on obtient

$$\begin{aligned}
\sup_{\alpha\in\Delta_p} \inf_{x\in X} \sum_{\substack{i\in[1:p]\\ \alpha_i\neq 0}} \alpha_i f_i(x) &\leqslant \sup_{\alpha\in\Delta_p} \inf_{x\in D} \sum_{i\in[1:p]} \alpha_i f_i(x) \\
&\leqslant \inf_{x\in D} \sup_{\alpha\in\Delta_p} \sum_{i\in[1:p]} \alpha_i f_i(x) \qquad \text{[dualité faible (14.6)]} \\
&= \inf_{x\in D} \max_{i\in[1:p]} f_i(x) \qquad \text{[exercice 2.3]} \\
&= \inf_{x\in X} \max_{i\in[1:p]} f_i(x),
\end{aligned}$$

puisque l'infimum en $x \in X$ a certainement lieu pour des $x \in D$, car sur $X \setminus D$ le critère vaut $+\infty$.

[$\leqslant$] On a

$$\begin{aligned}
\sup_{\alpha\in\Delta_p} \inf_{x\in X} \sum_{\substack{i\in[1:p]\\ \alpha_i\neq 0}} \alpha_i f_i(x) &\geqslant \sup_{\alpha\in\Delta_p} \sum_{\substack{i\in[1:p]\\ \alpha_i\neq 0}} \inf_{x\in X} \Big(\alpha_i\, f_i(x)\Big) \qquad \text{[exercice 1.2]} \\
&= \sup_{\alpha\in\Delta_p} \sum_{\substack{i\in[1:p]\\ \alpha_i\neq 0}} \alpha_i \left(\inf_{x\in X} f_i(x)\right) \qquad [\alpha_i > 0] \\
&= \max_{i\in[1:p]} \inf_{x\in X} f_i(x) \qquad \text{[exercice 2.3]}.
\end{aligned}$$

14.2. Par la propriété de dualité faible (14.6) et le fait que φ_i ne dépend que de la composante i de x, l'inégalité "$\geqslant$" a lieu dans (14.71). Montrons l'inégalité inverse. Soit $\varepsilon > 0$. Pour tout $i \in [1:p]$, on peut trouver un $\bar{x}_i \in X_i$ tel que

$$\varphi_i(\bar{x}_i) \leqslant \inf_{x_i\in X_i} \varphi_i(x_i) + \varepsilon.$$

Dès lors,

$$\max_{i\in[1:p]} \varphi_i(\bar{x}_i) \leqslant \max_{i\in[1:p]} \inf_{x_i\in X_i} \varphi_i(x_i) + \varepsilon. \tag{D.42}$$

C'est ici que la situation particulière de l'exercice intervient. Puisque le membre de gauche de (D.42) est la valeur en $\bar{x} := (\bar{x}_1, \ldots, \bar{x}_p)$ de la fonction $x = (x_1, \ldots, x_p) \in X \mapsto \max_{i\in[1:p]} \varphi_i(x_i)$, la première inégalité ci-dessous a certainement lieu (la seconde est (D.42))

$$\inf_{x\in X} \max_{i\in[1:p]} \varphi_i(x_i) \leqslant \max_{i\in[1:p]} \varphi_i(\bar{x}_i) \leqslant \max_{i\in[1:p]} \inf_{x_i\in X_i} \varphi_i(x_i) + \varepsilon.$$

Puisque $\varepsilon > 0$ est arbitraire, l'inégalité "$\leqslant$" dans (14.71) a lieu.

14.3. 1) [⇒] En prenant dans (14.72), pour chaque indice k, le supremum en $y \in Y$ à gauche et l'infimum en $x \in X$ à droite, on obtient

$$\sup_{y\in Y} \varphi(x_k, y) \leqslant \varepsilon_k + \inf_{x\in X} \varphi(x, y_k).$$

On en déduit que

$$\inf_{x\in X} \sup_{y\in Y} \varphi(x, y) \leqslant \varepsilon_k + \sup_{y\in Y} \inf_{x\in X} \varphi(x, y).$$

En prenant d'abord la limite en $k \to \infty$ et en utilisant ensuite l'inégalité de dualité faible (14.6), on en déduit qu'il n'y a pas de saut de dualité.

[⇐] Par définitioin de l'infimum et du supremum, quelle que soit la suite $\{\varepsilon_k\} \downarrow 0$, il existe une suite $\{(x_k, y_k)\}_k \subseteq X \times Y$ telle que pour tout $(x, y) \in X \times Y$ on ait

$$\sup_{y'\in Y} \inf_{x'\in X} \varphi(x', y') \leqslant \inf_{x'\in X} \varphi(x', y_k) + \frac{\varepsilon_k}{2} \leqslant \varphi(x, y_k) + \frac{\varepsilon_k}{2},$$
$$\inf_{x'\in X} \sup_{y'\in Y} \varphi(x', y') \geqslant \sup_{y'\in Y} \varphi(x_k, y') - \frac{\varepsilon_k}{2} \geqslant \varphi(x_k, y) - \frac{\varepsilon_k}{2}.$$

Comme les membres de gauche sont égaux, par l'absence de saut de dualité, on en déduit (14.72).

2) De (14.72), on déduit

$$\inf_{x\in X} \sup_{y\in Y} \varphi(x, y) \leqslant \sup_{y\in Y} \varphi(x_k, y) \leqslant \varepsilon_k + \inf_{x\in X} \varphi(x, y_k) \leqslant \varepsilon_k + \varphi(x_k, y_k),$$
$$\varepsilon_k + \sup_{y\in Y} \inf_{x\in X} \varphi(x, y) \geqslant \varepsilon_k + \inf_{x\in X} \varphi(x, y_k) \geqslant \sup_{y\in Y} \varphi(x_k, y) \geqslant \varphi(x_k, y_k).$$

En combinant, on trouve

$$\inf_{x\in X} \sup_{y\in Y} \varphi(x, y) - \varepsilon_k \leqslant \varphi(x_k, y_k) \leqslant \sup_{y\in Y} \inf_{x\in X} \varphi(x, y) + \varepsilon_k.$$

L'absence de saut de duality permet de conclure.

14.4. Le dual de Fenchel s'obtient par perturbation de Ax dans (P_F). On peut donc retrouver le dual associé en récrivant le problème (P_F) comme suit

$$\inf_{\substack{(x,y)\in\mathbb{E}\times\mathbb{F}\\ Ax=y}} f(x) + g(y)$$

et en dualisant la contrainte d'égalité par le lagrangien. Le problème primal s'écrit

$$\inf_{(x,y)\in\mathbb{E}\times\mathbb{F}} \sup_{\lambda\in\mathbb{F}} f(x) + g(y) - \langle \lambda, Ax - y\rangle.$$

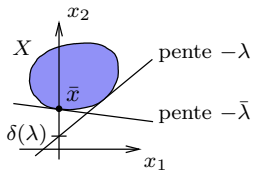
Son dual min-max s'écrit

$$\begin{aligned}
&\sup_{\lambda\in\mathbb{F}} \inf_{(x,y)\in\mathbb{E}\times\mathbb{F}} f(x) + g(y) - \langle \lambda, Ax - y\rangle \\
&\quad = \sup_{\lambda\in\mathbb{F}} \left(\inf_{x\in\mathbb{E}} \left[f(x) - \langle A^*\lambda, x\rangle \right] + \inf_{y\in\mathbb{F}} \left[g(y) - \langle -\lambda, y\rangle \right] \right) \\
&\quad = \sup_{\lambda\in\mathbb{F}} \left(-f^*(A^*\lambda) - g^*(-\lambda) \right),
\end{aligned}$$

qui est le dual de Fenchel (D_F) dans (14.27).

14.5. 1) Le problème primal s'écrit $\inf_{x\in X} \sup_{\lambda\in\mathbb{R}} \ell(x,\lambda)$, où $\ell(x,\lambda) = \lambda x_1 + x_2$. Le dual est donc

$$(D) \quad \sup_{\lambda\in\mathbb{R}} \inf_{x\in X} \lambda x_1 + x_2$$



Le lagrangien vaut la constante a le long de la droite de pente $-\lambda$ passant par $(0,a)$. Dès lors, la valeur de la fonction duale $\delta(\lambda) = \inf_{x\in X} \lambda x_1 + x_2$ est la borne inférieure des a, correspondant aux droites de pente $-\lambda$ intersectant l'ensemble X ou encore la plus grande valeur de a tel que la droite de pente $-\lambda$ passant par $(0,a)$ soit en dessous de X. Le problème dual consiste à maximiser $\delta(\lambda)$ en λ, donc à trouver une droite sous X et en contact avec X qui intersecte l'axe des ordonnées le plus haut possible.

2) Le problème primal a pour solution le point $\bar{x} = (0,\bar{x}_2)$, où $\bar{x}_2$ est la borne inférieure de $\{x_2 \in \mathbb{R}\colon (0,x_2) \in X\}$, qui est un fermé (projection sur l'axe des ordonnées du fermé $\{x \in X : x_1 = 0\}$).

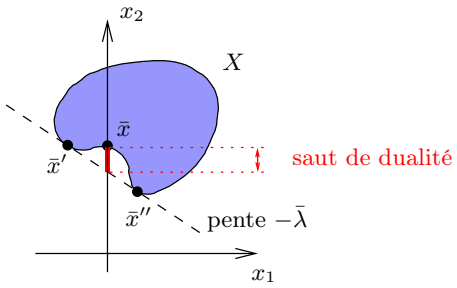
Montrons que le dual a une solution. Par optimalité du primal, $\bar{x} \in \partial X$, si bien que $\mathrm{N}_{\bar{x}}X$ contient une normale non nulle $\nu \in \mathbb{R}^2$ (proposition 2.29) : $\nu^{\mathsf{T}} x \leqslant \nu^{\mathsf{T}}\bar{x} = \nu_2\bar{x}_2$, pour tout $x \in X$. Comme X contient deux points x^- et x^+ avec $x_1^- < 0$ et $x_1^+ > 0$, on voit que $\nu_2 \neq 0$ (sinon ν serait nul). Si $\nu^{\mathsf{T}} x = \nu_2\bar{x}_2$ pour tout $x \in X$, le signe de ν_2 peut être positif ou négatif. Par contre, s'il existe un $\tilde{x} \in X$ tel que $\nu^{\mathsf{T}}\tilde{x} < \nu_2\bar{x}_2$, alors $\nu_2 < 0$ (en effet, on peut trouver un triplet $(\alpha^-,\alpha^+,\tilde{\alpha}) \in \Delta_3$ tel que $\alpha^- x_1^- + \alpha^+ x_1^+ + \tilde{\alpha}\tilde{x}_1 = 0$; alors le point $x = \alpha^- x^- + \alpha^+ x^+ + \tilde{\alpha}\tilde{x} \in X$ vérifie $x_1 = 0$ et donc $\nu_2 x_2 < \nu_2\bar{x}_2$; comme $\bar{x}_2 \leqslant x_2$ par optimalité de $\bar{x}$, on a $\nu_2 < 0$). Dès lors, en posant $\bar{\lambda} := \nu_1/\nu_2$, on a $\bar{x}_2 \leqslant \bar{\lambda}x_1 + x_2$, pour tout $x \in X$. Ceci montre que X est « au-dessus » de la droite $\{x \in \mathbb{R}^2 : \bar{\lambda}x_1 + x_2 = \bar{x}_2\}$ et même $\delta(\bar{\lambda}) = \bar{x}_2$. Enfin pour λ arbitraire dans $\mathbb{R}$, on a

$$\delta(\lambda) = \inf_{x\in X} (\lambda x_1 + x_2) \leqslant \inf_{\substack{x\in X \\ x_1=0}} (\lambda x_1 + x_2) = \bar{x}_2 = \delta(\bar{\lambda}),$$

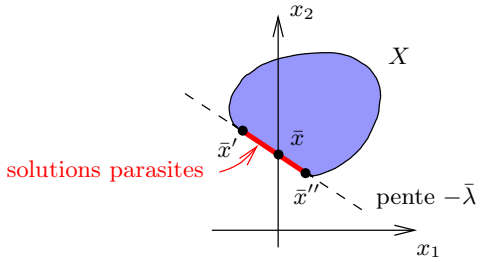
si bien que $\bar{\lambda}$ est solution du problème dual.

Les valeurs optimales primales et duales sont toutes deux égales à $\bar{x}_2$ (il n'y a pas de saut de dualité). Par le théorème 14.3, le lagrangien a un point-selle en $(\bar{x},\bar{\lambda})$.

3.a) Il y a un saut de dualité lorsque l'ensemble X est comme ci-dessous. Les solutions du problème de Lagrange sont les deux points $\bar{x}'$ et $\bar{x}''$ qui ne sont pas solutions de (P) (ils ne vérifient pas la contrainte dualisée).



3.b) Lorsque l'ensemble X est comme ci-dessous, le lagrangien a un point-selle (car les problèmes primal et dual ont une solution, $\bar{x}$ et $\bar{\lambda}$ respectivement, et il n'y a pas de saut de dualité), mais tout le segment $[\bar{x}', \bar{x}'']$ est solution du problème de Lagrange. Celui-ci contient l'unique solution primale $\bar{x}$ (comme le prédit la théorie), mais aussi d'autres points qui sont donc des *solutions importunes*, indésirables.

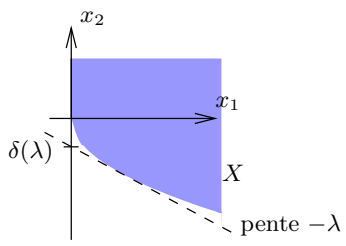


3.c) Soit $X = \{0\}$. Alors, pour tout $\lambda \in \mathbb{R}$, $\delta(\lambda) = 0$. Donc tout $\lambda \in \mathbb{R}$ est solution duale. Plus généralement, il en est de même si $X \subseteq \{x \in \mathbb{R}^2 : x_1 = 0,\ x_2 \geqslant b\}$ pour un réel b donné.

3.d) Considérons le cas où $X = \{x \in \mathbb{R}^2 : x_1 \geqslant 0,\ x_2 > -\sqrt{x_1}\}$ (représenté dans la figure ci-dessous). Clairement, le problème primal n'a pas de solution (en prenant $x_2 \geqslant -\sqrt{x_1}$ dans la définition de X, il en aurait une) et la valeur optimale primale est nulle. La fonction duale s'écrit

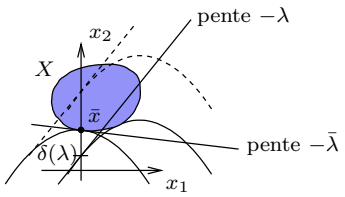
$$\delta(\lambda) = \inf_{\substack{x_1 \geqslant 0 \\ x_2 \geqslant -\sqrt{x_1}}} (\lambda x_1 + x_2) = \inf_{x_1 \geqslant 0} (\lambda x_1 - \sqrt{x_1}) = \begin{cases} -\infty & \text{si } \lambda \leqslant 0 \\ -\frac{1}{4\lambda} & \text{si } \lambda > 0. \end{cases}$$

Dès lors le problème dual n'a pas de solution (sa valeur optimale nulle n'est pas atteinte). Comme les valeurs optimales primale et duale sont toutes les deux nulles, il n'y a pas de saut de dualité.



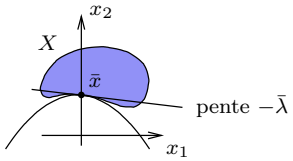
4) Le problème primal s'écrit aussi $\inf_{x \in X} \sup_{\lambda \in \mathbb{R}} \ell_r(x, \lambda)$, où $\ell_r(x, \lambda) = \lambda x_1 + x_2 + \frac{r}{2} x_1^2$. Le dual associé est donc

$$(D) \quad \sup_{\lambda\in\mathbb{R}} \inf_{x\in X} \lambda x_1 + x_2 + \frac{r}{2}x_1^2$$



Le lagrangien augmenté vaut la constante a le long de la parabole concave passant par $(0,a)$, de pente $-\lambda$ en ce point et de courbure $-r$. Dès lors, la valeur de la fonction duale $\delta_r(\lambda) = \inf_{x\in X} \lambda x_1 + x_2 + \frac{r}{2}x_1^2$ est la borne inférieure des a tels que la parabole $x_1 \mapsto -\frac{r}{2}x_1^2 - \lambda x_1 + a$ intersecte X. Le problème dual consiste à maximiser $\delta_r(\lambda)$ en λ, donc à trouver une parabole de courbure $-r$, qui soit sous X, en contact avec X et qui intersecte l'axe des ordonnées le plus haut possible.

Cette dualisation ne donne pas de saut de dualité pour des ensembles qui ne sont pas nécessairement convexes, comme dans l'exemple ci-dessous, alors qu'il n'en serait pas de même pour la dualisation par le lagrangien ordinaire.



14.7. 1) La variable duale $y \in \mathbb{R}^m$ suggère de dualiser la contrainte d'égalité de (14.73), en récrivant ce problème sous la forme équivalente suivante

$$\inf_{x\geqslant 0} \sup_{y\in\mathbb{R}^m} \frac{1}{2}\|x-z\|^2 - y^\mathsf{T}(Ax-b). \tag{D.43}$$

Le dual s'écrit alors

$$\begin{aligned}
&\sup_{y\in\mathbb{R}^m} \inf_{x\geqslant 0} \frac{1}{2}\|x-z\|^2 - y^\mathsf{T}(Ax-b)\\
&= \sup_{y\in\mathbb{R}^m} \inf_{x\geqslant 0} \frac{1}{2}\|x-(A^\mathsf{T}y+z)\|^2 - \frac{1}{2}\|A^\mathsf{T}y+z\|^2 + \frac{1}{2}\|z\|^2 + b^\mathsf{T}y\\
&= \sup_{y\in\mathbb{R}^m} \frac{1}{2}\|(A^\mathsf{T}y+z)^-\|^2 - \frac{1}{2}\|A^\mathsf{T}y+z\|^2 + \frac{1}{2}\|z\|^2 + b^\mathsf{T}y\\
&= \sup_{y\in\mathbb{R}^m} -\frac{1}{2}\|(A^\mathsf{T}y+z)^+\|^2 + \frac{1}{2}\|z\|^2 + b^\mathsf{T}y,
\end{aligned}$$

où l'on a noté $v^- := \max(0,v)$ et utilisé $\|v\|^2 = \|v^+\|^2 + \|v^-\|^2$.

2) Utilisons les conditions d'optimalité du problème primal (14.73) (on pourrait aussi utiliser les conditions d'optimalité du problème dual (14.74)). Comme il s'agit d'un problème de projection, il a une solution unique, disons $\bar{x}$. Ses contraintes de (14.73) étant qualifiées (car affines), il existe $\bar{y} \in \mathbb{R}^m$ et $\bar{s} \in \mathbb{R}^n$ tels que

$$\begin{cases} \bar{x} - z - A^\mathsf{T}\bar{y} - \bar{s} = 0 \\ A\bar{x} = b \\ 0 \leqslant \bar{x} \perp \bar{s} \geqslant 0. \end{cases} \tag{D.44}$$

Par complémentarité, lorsque $\bar{x}_i > 0$, on a $\bar{s}_i = 0$ et donc $\bar{x}_i = (A^\mathsf{T}\bar{y}+z)_i$; et lorsque $\bar{x}_i = 0$, $(A^\mathsf{T}\bar{y}+z)_i = -\bar{s}_i \leqslant 0$; par conséquent,

$$\bar{x} = (A^\mathsf{T}\bar{y} + z)^+. \tag{D.45}$$

Par ailleurs,

$$b^\mathsf{T}\bar{y} = \bar{x}^\mathsf{T}A^\mathsf{T}\bar{y} = \bar{x}^\mathsf{T}(\bar{x} - z - \bar{s}) = \bar{x}^\mathsf{T}(\bar{x} - z). \tag{D.46}$$

On a alors

$$\begin{aligned}
\mathrm{val}(D) &\geqslant -\frac{1}{2}\|(A^\mathsf{T}\bar{y} + z)^+\|^2 + b^\mathsf{T}\bar{y} + \frac{1}{2}\|z\|^2 \\
&= -\frac{1}{2}\|\bar{x}\|^2 + \bar{x}^\mathsf{T}(\bar{x} - z) + \frac{1}{2}\|z\|^2 \qquad [(\text{D.45}) \text{ et } (\text{D.46})] \\
&= \frac{1}{2}\|\bar{x} - z\|^2 \\
&= \mathrm{val}(P).
\end{aligned}$$

Comme $\mathrm{val}(D) \leqslant \mathrm{val}(P)$ par dualité faible, les valeurs optimales primale et duale sont égales.

3) Le problème (14.74) peut aussi s'écrire comme le problème de la minimisation de la fonction convexe $\varphi : \mathbb{R}^m \to \mathbb{R}$ définie en $y \in \mathbb{R}^m$ par

$$\varphi(y) := \frac{1}{2}\|(A^\mathsf{T}y + z)^+\|^2 - b^\mathsf{T}y - \frac{1}{2}\|z\|^2.$$

$[(a) \Rightarrow (b)]$ Si (b) n'a pas lieu, on peut trouver une direction $p \in \mathbb{R}^m$ telle que $p \neq 0$, $A^\mathsf{T}p \leqslant 0$ et $b^\mathsf{T}p \geqslant 0$. Si $\bar{y}$ minimise φ (donc est solution de (14.74)) et $t \geqslant 0$, on a $A^\mathsf{T}(\bar{y} + tp) + z \leqslant A^\mathsf{T}\bar{y} + z$ composante par composante, si bien que $0 \leqslant [A^\mathsf{T}(\bar{y} + tp) + z]^+ \leqslant [A^\mathsf{T}\bar{y} + z]^+$ et donc $\varphi(\bar{y} + tp) \leqslant \varphi(\bar{y})$. Donc $\bar{y} + tp$ est encore solution de (14.74). Comme $p \neq 0$, l'ensemble des solutions de (14.74) n'est pas borné.

$[(b) \Rightarrow (c)]$ Il est clair que A est surjective, sinon on pourrait trouver $p \neq 0$ tel que $A^\mathsf{T}p = 0$; comme ce $p \in \mathcal{R}(A)^\perp$ et $b \in \mathcal{R}(A)$, on aurait $b^\mathsf{T}p = 0$, ce qui contredirait (b). Par ailleurs, (b) s'écrit aussi $\{p : A^\mathsf{T}p \leqslant 0,\ b^\mathsf{T}p \geqslant 0\} = \{0\}$ ou $\{p : (-A^\mathsf{T}p, b^\mathsf{T}p) \in \mathbb{R}^{n+1}_+\}^+ = \mathbb{R}^m$. Par le lemme de Farkas, cela revient à dire que $\{-Ax + \alpha b : (x, \alpha) \in \mathbb{R}^{n+1}_+\} = \mathbb{R}^m$. On peut donc trouver $x_0 \geqslant 0$ et $\alpha_0 \geqslant 0$ tels que $A(x_0 + e) = \alpha b$. Alors $x := (\bar{x} + x_0 + e)/(\alpha + 1)$ vérifie $x > 0$ et $Ax = b$.

$[(c) \Rightarrow (a)]$ La fonction asymptotique de φ prend en $p \in \mathbb{R}^m$ la valeur $-b^\mathsf{T}p$ si $A^\mathsf{T}p \leqslant 0$ et la valeur $+\infty$ autrement. Supposons que $\varphi^\infty(p) \leqslant 0$. Alors $A^\mathsf{T}p \leqslant 0$ et $b^\mathsf{T}p \geqslant 0$. Par (b), on peut trouver un $x > 0$ tel que $Ax = b$. En multipliant par p, on trouve $0 \leqslant b^\mathsf{T}p = p^\mathsf{T}Ax \leqslant 0$, si bien que $A^\mathsf{T}p = 0$; donc $p = 0$ par l'injectivité de A^T. On conclut en utilisant l'équivalence $(iii) \Leftrightarrow (iv)$ de la proposition 3.29.

4.1) Si $b \in \mathcal{R}(A)$, $Ax - b \in \mathcal{R}(A)$ et, au lieu de (D.43), on peut écrire (14.73) comme suit

$$\inf_{x \geqslant 0} \sup_{y \in \mathcal{R}(A)} \frac{1}{2}\|x - z\|^2 - y^\mathsf{T}(Ax - b). \tag{D.47}$$

En effet, si $Ax \neq b$, le suppremum en $y \in \mathcal{R}(A)$ sera $+\infty$ en prenant $y = t(Ax - b) \in \mathcal{R}(A)$ et $t \to \infty$. Comme au point 1, la dualisation de ce problème conduit au dual (14.75) de (14.73).

4.2) Il suffit de montrer que l'on peut trouver une solution $(\bar{x}, \bar{y}, \bar{s})$ de (D.44) avec $\bar{y} \in \mathcal{R}(A)$. Soit $(\bar{x}, \bar{y}_0, \bar{s})$ une solution de (D.44) et $\bar{y}$ la projection de $\bar{y}_0$ sur $\mathcal{R}(A)$. Alors $\bar{y}_0 - \bar{y} \in \mathcal{R}(A)^\perp = \mathcal{N}(A^\mathsf{T})$. Dès lors, $(\bar{x}, \bar{y}, \bar{s})$ est encore solution de (D.44).

4.3) Le problème (14.75) peut aussi s'écrire comme le problème de la minimisation de la fonction convexe $\varphi : \mathbb{R}^m \to \mathbb{R}$ définie en $y \in \mathbb{R}^m$ par

$$\varphi(y) := \frac{1}{2}\|(A^{\mathsf{T}}y + z)^+\|^2 - b^{\mathsf{T}}y - \frac{1}{2}\|z\|^2 + \mathcal{I}_{\mathcal{R}(A)}(y).$$

$[(a) \Rightarrow (b)]$ Si (b) n'a pas lieu, on peut trouver une direction $p \in \mathcal{R}(A)$ telle que $p \neq 0$, $A^{\mathsf{T}}p \leqslant 0$ et $b^{\mathsf{T}}p \geqslant 0$. Si $\bar{y}$ minimise φ (donc est solution de (14.75)) et $t \geqslant 0$, on a comme précédemment $\varphi(\bar{y} + tp) \leqslant \varphi(\bar{y})$. Donc $\bar{y} + tp$ est encore solution de (14.75). Comme $p \neq 0$, l'ensemble des solutions de (14.75) n'est pas borné.

$[(b) \Rightarrow (c)]$ (b) s'écrit aussi $\{p \in \mathcal{R}(A) : A^{\mathsf{T}}p \leqslant 0,\ b^{\mathsf{T}}p \geqslant 0\} = \{0\}$ ou

$$\begin{aligned}
\mathbb{R}^m &= \{p \in \mathcal{R}(A) : (-A^{\mathsf{T}}p, b^{\mathsf{T}}p) \in \mathbb{R}^{n+1}_+\}^+ \\
&= [\{p : (-A^{\mathsf{T}}p, b^{\mathsf{T}}p) \in \mathbb{R}^{n+1}_+\} \cap \mathcal{R}(A)]^+ \\
&= \overline{\{p : (-A^{\mathsf{T}}p, b^{\mathsf{T}}p) \in \mathbb{R}^{n+1}_+\}^+ + \mathcal{R}(A)^+} \qquad [(2.35)] \\
&= \{p : (-A^{\mathsf{T}}p, b^{\mathsf{T}}p) \in \mathbb{R}^{n+1}_+\}^+ + \mathcal{N}(A^{\mathsf{T}}) \qquad [\text{convexe dense}] \\
&= \{-Ax + \alpha b : (x, \alpha) \in \mathbb{R}^{n+1}_+\} + \mathcal{N}(A^{\mathsf{T}}) \qquad [\text{lemme de Farkas}].
\end{aligned}$$

En prenant Ae comme élément de $\mathbb{R}^m$ à gauche de l'identité obtenue, on peut trouver $x_0 \geqslant 0$, $\alpha_0 \geqslant 0$ et $h \in \mathcal{N}(A^{\mathsf{T}})$ tels que $A(x_0 + e) = \alpha b + h$. Comme $b \in \mathcal{R}(A)$, il vient que $h \in \mathcal{N}(A^{\mathsf{T}}) \cap \mathcal{R}(A) = \{0\}$; donc $h = 0$ pour ce choix de $Ae \in \mathbb{R}^m$. Alors $x := (\bar{x} + x_0 + e)/(\alpha + 1)$ vérifie $x > 0$ et $Ax = b$.

$[(c) \Rightarrow (a)]$ La fonction asymptotique de φ prend en $p \in \mathbb{R}^m$ la valeur $-b^{\mathsf{T}}p$ si $p \in \mathcal{R}(A)$ et $A^{\mathsf{T}}p \leqslant 0$ et prend la valeur $+\infty$ autrement. Supposons que $\varphi^\infty(p) \leqslant 0$. Alors $p \in \mathcal{R}(A)$, $A^{\mathsf{T}}p \leqslant 0$ et $b^{\mathsf{T}}p \geqslant 0$. Par (b), on peut trouver un $x > 0$ tel que $Ax = b$. En multipliant par p, on trouve $0 \leqslant b^{\mathsf{T}}p = p^{\mathsf{T}}Ax \leqslant 0$, si bien que $A^{\mathsf{T}}p = 0$. Donc $p \in \mathcal{N}(A^{\mathsf{T}}) \cap \mathcal{R}(A)$, ce qui implique que $p = 0$. On conclut en utilisant l'équivalence $(iii) \Leftrightarrow (iv)$ de la proposition 3.29.

14.8. 1) Le problème (14.76) s'écrit $\inf_{x\in K} \sup_{y\in\mathbb{F}} \ell(x, y)$, si bien que son dual lagrangien s'écrit $\sup_{y\in\mathbb{F}} \inf_{x\in K} \ell(x, y) = \sup_{y\in\mathbb{F}}(\langle b, y\rangle + \inf_{x\in K} \langle c - A^*y, x\rangle)$. Il est clair que

$$\inf_{x\in K} \langle c - A^*y, x\rangle = \begin{cases} 0 & \text{si } c - A^*y \in K^+ \\ -\infty & \text{sinon.} \end{cases}$$

En effet, si $c - A^*y \in K^+$, l'infimum est $\geqslant 0$ et en prenant $x = tx_0$ avec $x_0 \in K \neq \emptyset$ et $t \downarrow 0$, on voit que l'infimum est nul; si $c - A^*y \notin K^+$, il existe un $x_0 \in K$ tel que $\langle c - A^*, x_0\rangle < 0$ et en prenant $x = tx_0$ avec $t \to +\infty$, on voit que l'infimum est $-\infty$. On en déduit que le dual s'écrit $\sup_{y\in\mathbb{F}} \{\langle b, y\rangle : c - A^*y \in K^+\}$, qui se récrit comme en (14.77).

2) Comme $K^{++} = K$, on a

$$\sup_{s\in K^+} -\langle s, x\rangle = \begin{cases} 0 & \text{si } x \in K \\ +\infty & \text{sinon.} \end{cases}$$

Dès lors, le problème (14.76) peut s'écrire

$$\inf_{x\in\mathbb{E}} \sup_{\substack{y\in\mathbb{F} \\ s\in K^+}} \ell(x, y) - \langle s, x\rangle.$$

Celui-ci a pour dual

$$\sup_{\substack{y\in\mathbb{F} \\ s\in K^+}} \inf_{x\in\mathbb{E}} \ell(x, y) - \langle s, x\rangle = \sup_{\substack{y\in\mathbb{F} \\ s\in K^+}} \inf_{x\in\mathbb{E}} \langle b, y\rangle + \langle c - A^*y - s, x\rangle = \sup_{\substack{y\in\mathbb{F} \\ A^*y+s=c \\ s\in K^+}} \langle b, y\rangle.$$

On a donc retrouvé (14.77).

3) Montrons que $\inf((14.77)) = 0$. Pour que x soit admissible, il faut que $1+x_2^2 \leqslant x_3^2$, ce qui s'écrit aussi $1 \leqslant (x_3 - x_2)(x_2 + x_3)$. On voit qu'il suffit de prendre $\varepsilon \downarrow 0$, $x_2 + x_3 = \varepsilon$ et $x_3 - x_2 = 1/\varepsilon$ (ce qui donne $x_2 = (\varepsilon^2 + 1)/(2\varepsilon)$ et $x_3 = (\varepsilon^2 - 1)/(2\varepsilon)$). Cependant cette borne inférieure n'est pas atteinte, puisque l'inégalité $1 \leqslant (x_3 - x_2)(x_2 + x_3)$ ne peut être vérifiée avec $x_2 + x_3 = 0$. D'autre part, le dual n'a qu'un seul point admissible ($y = 0$ et $s = (0, 1, 1)$) qui donne la valeur optimale duale $\sup((14.77)) = 0$.

14.9. Soit $\bar{x}$ une solution du problème (il en exsite une par le théorème de Weierstrass, car le critère est continu et l'ensemble admissible est un compact non vide). On introduit le lagrangien $\ell : (x, \lambda) \in \mathbb{R}^n \times \mathbb{R} \mapsto \ell(x, \lambda) = g^\mathsf{T} x + \frac{1}{2} x^\mathsf{T} M x + \frac{1}{2}\lambda(\|x\|^2 - \Delta^2)$. Par le théorème 9.2, on sait qu'il existe un multiplicateur optimal $\bar{\lambda} \in \mathbb{R}$ associé à la contrainte et que $\bar{x}$ minimise $\ell(\cdot, \bar{\lambda})$. Le fait que $\bar{\lambda}$ maximise $\ell(\bar{x}, \cdot)$ se montre aisément comme au point 2 de la proposition 14.25. Dès lors, le problème dual $\sup_{\lambda \geqslant 0} \inf_{x \in \mathbb{R}^n} \ell(x, \lambda)$ a une solution et il n'y a pas de saut de dualité (théorème 14.3). Il reste à exprimer le problème dual. Le problème interne $\inf_{x \in \mathbb{R}^n} \ell(x, \lambda)$ consiste à minimiser une fonction quadratique sur $\mathbb{R}^n$. Sa valeur optimale est finie si, et seulement si, sa hessienne $H + \lambda I$ est semi-définie positive et son gradient en zéro g est dans l'image de sa hessienne $H + \lambda I$ (point 2 de l'exercice 4.6) ; elle vaut alors $-g^\mathsf{T}(H + \lambda I)^\dagger g$. On a montré (14.78).

Si $\Delta = 0$, le membre de gauche de (14.78) vaut zéro (clair), alors que le membre de droite est strictement négatif lorsque $H \succ 0$ et $g \neq 0$. En effet, le problème est réalisable (n'importe quel $\lambda \geqslant 0$ est admissible car alors $H + \lambda I \succcurlyeq 0$ et $g \in \mathcal{R}(H + \lambda I)$), tandis que le critère du problème est $\leqslant -g^\mathsf{T} H g \leqslant -\lambda_{\min}(H)\|g\|^2 < 0$.

14.10. 1) Si (P) et/ou (D) n'est pas réalisable, l'inégaité est claire du fait de la valeur optimale infinie du problème. Dans le cas contraire, on peut trouver $(x, \lambda) \in \mathbb{R}^n \times \mathbb{R}^m_+$ tel que $\nabla_x \ell(x, \lambda) = 0$ et $\tilde{x} \in \mathbb{R}^n$ tel que $c(\tilde{x}) \leqslant 0$. Comme f et les c_i sont convexes et que $\lambda \geqslant 0$, $x' \mapsto \ell(x', \lambda)$ est convexe. Cette dernière fonction est aussi minimisée en x car $\nabla_x \ell(x, \lambda) = 0$. Ceci conduit à la première inégalité ci-dessous

$$\ell(x, \lambda) \leqslant \ell(\tilde{x}, \lambda) \leqslant f(\tilde{x}),$$

où, pour la seconde inégalité, on a utilisé $c(x) \leqslant 0$ et $\lambda \geqslant 0$. On en déduit que $\mathrm{val}(D) \leqslant \mathrm{val}(P)$.

2) Comme x_* est une solution de (P) et que la contrainte c est qualifiée en x_*, il existe un multiplicateur $\lambda_* \in \mathbb{R}^m$ tel que l'on ait (4.33), à savoir

$$\nabla_x \ell(x_*, \lambda_*) = 0 \qquad \text{et} \qquad 0 \leqslant \lambda_* \perp c(x_*) \leqslant 0.$$

Donc x_* est admissible pour (P) et (x_*, λ_*) est admissible pour (D). Par le point 1, on a

$$\ell(x_*, \lambda_*) \leqslant \mathrm{val}(D) \leqslant \mathrm{val}(P) \leqslant f(x_*).$$

Mais $\ell(x_*, \lambda_*) = f(x_*)$ par la relation de complémentarité $\lambda_* \perp c(x_*)$, si bien que l'on a égalité partout dans le systeme d'inégalités ci-dessus. On en déduit que (x_*, λ_*) est solution de (D) et que $\mathrm{val}(D) = \mathrm{val}(P)$.

3) Supposons que la contrainte s'écrive $c(x) = Ax - b$.

Montrons d'abord qu'il existe un $\mu \in \mathbb{R}^m_+$ tel que $A^\mathsf{T}\mu = 0$ et $b^\mathsf{T}\mu = -1$. Cela revient à montrer que

$$\begin{pmatrix} 0 \\ -1 \end{pmatrix} \in \begin{pmatrix} A^\mathsf{T} \\ b^\mathsf{T} \end{pmatrix} (\mathbb{R}^m_+) = \{(x, \alpha) \in \mathbb{R}^n \times \mathbb{R} : Ax + \alpha b \geqslant 0\}^+,$$

où la dernière égalité provient du lemme de Farkas (proposition 2.45). Si cette appartenance n'était pas vérifiée, il existerait $(x,\alpha)\in\mathbb{R}^n\times\mathbb{R}$ tel que $Ax+\alpha b\geqslant 0$ et $\alpha>0$. Mais alors on aurait $A(-x/\alpha)-b\leqslant 0$ et le problème (P) serait réalisable.

Si la contrainte de (D) est réalisable, il existe $(x,\lambda)\in\mathbb{R}^n\times\mathbb{R}^m_+$ tel que $\nabla_x\ell(x,\lambda)=0$. Avec le μ trouvé ci-dessus et $t\geqslant 0$, $(x,\lambda+t\mu)$ réalise encore les contraintes de (D) et $\ell(x,\lambda+t\mu)=\ell(x,\lambda)+t$ tend vers $+\infty$ lorsque $t\to+\infty$. Donc $\mathrm{val}(D)=+\infty$.

4) Il suffit de montrer que $\mathrm{val}(P)=+\infty$ lorsque $f=0$ dans (P). Le dual de Wolfe de (P) est alors le problème $\sup\{\lambda^\mathsf{T}c(x):\lambda\geqslant 0,\ c'(x)^\mathsf{T}\lambda=0\}=\mathrm{val}(D)$, qui est strictement positif par l'hypothèse donnée. Alors, les contraitnes $c(x)\leqslant 0$ ne peuvent être admissibles sinon l'on aurait $0<\mathrm{val}(D)\leqslant\mathrm{val}(P)\leqslant 0$.

14.11. Le problème (P) peut s'écrire $\inf_{x\in\mathbb{R}^n}\sup_{\lambda\geqslant 0}\ell_h(x,\lambda)-b^\mathsf{T}\lambda$, si bien que le dual lagrangien de (P) s'écrit $\sup_{\lambda\geqslant 0}\inf_{x\in\mathbb{R}^n}\ell_h(x,\lambda)-b^\mathsf{T}\lambda$. Calculons la valeur optimale du problème interne du dual pour $\lambda\geqslant 0$ fixé. Par homogénéité, de $\ell_h(\cdot,\lambda)$, on a

- s'il existe un $x_0\in\mathbb{R}^n$ tel que $\ell_h(x_0,\lambda)<0$, alors le problème interne du dual prend la valeur $-\infty$, car $\inf_{x\in\mathbb{R}^n}\ell_h(x,\lambda)\leqslant\inf_{t>0}\ell_h(tx_0,\bar\lambda)=\inf_{t>0}t^p\ell_h(x_0,\bar\lambda)=-\infty$,
- sinon la valeur optimale du problème interne du dual vaut $-b^\mathsf{T}\lambda$, car $\ell_h(\cdot,\lambda)$ est positif et $\ell_h(0,\lambda)=0$.

Le dual lagrangien de (P) est donc bien (D). Ce dernier est convexe car son critère est linéaire et son ensemble admissible est convexe (intersection de demi-espaces fermés).

14.12. 1) Il est clair que x_λ est admissible pour le problème perturbé. Par ailleurs, si x est admissible pour le problème perturbé, on a

$$\ell(x_\lambda,\lambda)\leqslant\ell(x,\lambda)=f(x)+\sum_{\substack{i\in E\\ i\in I:\lambda_i>0}}\lambda_i c_i(x)\leqslant f(x)+\sum_{\substack{i\in E\\ i\in I:\lambda_i>0}}\lambda_i c_i(x_\lambda).$$

On en déduit que $f(x_\lambda)\leqslant f(x)$, si bien que x_λ est solution du problème perturbé.

2) Par hypothèse, $\bar x$ est admissible pour (P_{EI}). Par ailleurs, d'après le point 1, $\bar x$ est alors solution du problème $\min\{f(x):c_E(x)=0,\ c_i(x)\leqslant 0$ pour tout $i\in I$ tel que $\bar\lambda_i>0\}$ (on a utilisé le fait que $c_E(\bar x)=0$ et $c_i(\bar x)=0$ si $i\in I$ et $\bar\lambda_i>0$). Comme l'ensemble admissible de ce problème contient celui de (P_{EI}), $\bar x$ est solution de (P_{EI}).

14.13. 1) Pour $\bar x^*\in\partial\,\mathrm{d}_C(x)$, qui est non vide (proposition 3.62) et contenu dans $\bar B_{\mathrm{D}}$ (point 4 de l'exercice 3.36), et pour $y\in C$, on a $0=\mathrm{d}_C(y)\geqslant\mathrm{d}_C(x)+\langle\bar x^*,y-x\rangle$; dès lors

$$d_C(x)\leqslant\inf_{y\in C}\langle\bar x^*,x-y\rangle\leqslant\sup_{x^*\in\bar B_{\mathrm{D}}}\inf_{y\in C}\langle x^*,x-y\rangle.$$

Ensuite, par la dualité faible et le fait que la norme biduale est identique à la norme originale (point 2 de l'exercice 3.30), on a

$$\sup_{x^*\in\bar B_{\mathrm{D}}}\inf_{y\in C}\langle x^*,x-y\rangle\leqslant\inf_{y\in C}\sup_{x^*\in\bar B_{\mathrm{D}}}\langle x^*,x-y\rangle=\inf_{y\in C}\|x-y\|=\mathrm{d}_C(x).$$

Les deux inégalités des relations exposées ci-dessus sont donc des égalités.

2) Des calculs précédents, on déduit que l'ensemble des solutions primales $S_P=\arg\min\{\|x-y\|:y\in C\}$ et que l'ensemble des solutions duales $S_D\supseteq\partial\,\mathrm{d}_C(x)$. Montrons l'inclusion réciproque. Soient $\bar x^*\in S_D$ et $\bar x\in C$ tel que $\mathrm{d}_C(x)=\|x-\bar x\|$. Bien sûr, $\bar x^*\in\bar B_{\mathrm{D}}$. Par ailleurs $\mathrm{d}_C(x)=\inf\{\langle\bar x^*,x-y\rangle:y\in C\}$ [$\bar x^*\in S_D$ et

$\mathrm{d}_C(x)$ est la valeur-selle] $\leqslant \langle \bar{x}^*, x-\bar{x}\rangle$ [$\bar{x}\in C$] $\leqslant \|x-\bar{x}\|$ [Cauchy-Schwarz et $\|\bar{x}^*\|_{\mathrm{D}} \leqslant 1$] $= \mathrm{d}_C(x)$; dès lors $\langle \bar{x}^*, x-\bar{x}\rangle = \|x-\bar{x}\|$. Enfin, pour tout $y\in C$: $\langle \bar{x}^*, y-\bar{x}\rangle = \langle \bar{x}^*, y-x\rangle + \langle \bar{x}^*, x-\bar{x}\rangle \leqslant -\mathrm{d}_C(x) + \|x-\bar{x}\| = 0$, ce qui montre que $\bar{x}^*\in \mathrm{N}_C(\bar{x})$. Ces trois résultats montrent que $\bar{x}^*\in \partial\, \mathrm{d}_C(x)$.

14.14. On sait déjà que $\bar{x}$ est unique. L'unicité de $\bar{\lambda}$ vient de ce que $(\bar{x},\bar{\lambda})$ est solution primale-duale de (14.64) et de ce que la qualification (QC-IL) assure l'unicité du multiplicateur (voir la section 4.4.3). Rappelons que la démonstration de la proposition 14.37 a montré que $\{\lambda_k\}$ est bornée. Soient $I_*^0 := \{i : c_i(\bar{x}) = 0\}$ et $I_*^- := \{i : c_i(\bar{x}) < 0\}$. Considérons d'abord les indices $i\in I_*^-$. La règle de mise à jour des variables duales s'écrit : $(\lambda_{k+1})_i = ((\lambda_k)_i + \alpha_k c_i(x_k))^+$. Comme $\alpha_k \geqslant \epsilon > 0$, $c_i(x_k)\to c_i(\bar{x}) < 0$ et $\{\lambda_k\}$ est bornée, on a $(\lambda_k)_i = 0$, pour k assez grand. Il reste à montrer que $(\lambda_k)_{I_*^0} \to (\bar{\lambda})_{I_*^0}$. Pour cela, on passe à la limite dans $\nabla_x \ell(x_k,\lambda_k) = 0$ pour constater que l'unique point d'adhérence de $\{\lambda_k\}$ est $\bar{\lambda}$.

14.15. 1) Pour la première identité, on écrit

$$\|A\|^2 = \sup_{\|x\|_2\leqslant 1} x^{\mathsf{T}} A^{\mathsf{T}} A x = \lambda_{\max}(A^{\mathsf{T}}A) = \|\sigma(A)\|_\infty^2.$$

Pour la seconde, on utilise la décomposition en valeurs singulières $A = V\Sigma U^{\mathsf{T}}$, dans laquelle U et V sont orthogonales et Σ est diagonale de rang $r = \mathrm{rg}(A)$ et contient les valeurs singulières de A :

$$\|A\|_F = (\mathrm{tr}\, A^{\mathsf{T}}A)^{1/2} = \mathrm{tr}\, U\Sigma \underbrace{V^{\mathsf{T}}V}_{I_m} \Sigma U^{\mathsf{T}} = \mathrm{tr}\, \Sigma^2 \underbrace{U^{\mathsf{T}}U}_{I_n} = \|\sigma(A)\|_2^2.$$

La dernière identité est plus difficile à démontrer ; on suit la piste suggérée dans les indications.

(*a*) En prenant $B := VU^{\mathsf{T}}$, on a $\|B\| = \|\sigma(B)\|_\infty = 1$ et donc

$$\|A\|_* \geqslant \langle A, B\rangle = \mathrm{tr}\, A^{\mathsf{T}} V U^{\mathsf{T}} = \mathrm{tr}\, U\Sigma \underbrace{V^{\mathsf{T}}V}_{I_m} U^{\mathsf{T}} = \mathrm{tr}\, \Sigma \underbrace{U^{\mathsf{T}}U}_{I_n} = \|\sigma(A)\|_1.$$

(*b*) On montre que

$$\|B\| \leqslant t \iff t^2 I_m - B^{\mathsf{T}}B \succcurlyeq 0 \iff \begin{pmatrix} tI_n & B \\ B^{\mathsf{T}} & tI_m \end{pmatrix} \succcurlyeq 0$$

La première équivalence s'obtient comme suit

$$\begin{aligned} \|B\| \leqslant t &\iff \|B\|^2 \leqslant t^2 \\ &\iff \sup_{\|x\|_2=1} \|Bx\|_2^2 \leqslant t^2 \\ &\iff \sup_{\|x\|_2=1} x^{\mathsf{T}}(B^{\mathsf{T}}B - t^2 I)x \leqslant 0 \\ &\iff t^2 I_m - B^{\mathsf{T}}B \succcurlyeq 0. \end{aligned}$$

La seconde équivalence est claire si $t = 0$ car chaque côté requiert $B = 0$ (clair pour $-B^{\mathsf{T}}B \succcurlyeq 0$ et, pour le côté droit, il faut $u^{\mathsf{T}}Bv \geqslant 0$ pour tout u et v, c'est-à-dire $B = 0$). Si $t > 0$, on utilise l'équivalence (20.44) utilisant le complément de Schur de tI_n :

$$\begin{pmatrix} tI_n & B \\ B^{\mathsf{T}} & tI_m \end{pmatrix} \succcurlyeq 0 \iff tI_m - \frac{1}{t} B^{\mathsf{T}}B \succcurlyeq 0 \iff t^2 I_m - B^{\mathsf{T}}B \succcurlyeq 0.$$

(*c*) Grâce au point (*b*), on peut écrire

$$\|A\|_* = \sup_{\|B\|\leqslant 1} \langle A, B\rangle = \sup_{\begin{pmatrix} I_n & B \\ B^\mathsf{T} & I_m\end{pmatrix}\succcurlyeq 0} \langle A, B\rangle$$

Grâce à l'autodualité du cône des matrices semi-définie positive (point 3 de l'exrcice 2.35), $\|A\|_*$ peut s'écrire

$$\sup_{B\in\mathbb{R}^{m\times n}} \inf_{\substack{X\in\mathbb{R}^{m\times m},\ Y\in\mathbb{R}^{n\times n},\ Z\in\mathbb{R}^{m\times n} \\ \begin{pmatrix} X & Z \\ Z^\mathsf{T} & Y\end{pmatrix}\succcurlyeq 0}} \left[\langle A, B\rangle + \left\langle \begin{pmatrix} I_n & B \\ B^\mathsf{T} & I_m\end{pmatrix}, \begin{pmatrix} X & Z \\ Z^\mathsf{T} & Y\end{pmatrix}\right\rangle\right].$$

On conclut en observant que le dernier terme s'écrit $\operatorname{tr}(X + BZ^\mathsf{T}) + \operatorname{tr}(B^\mathsf{T}Z + Y) = \langle B, 2Z\rangle + \langle X, I_m\rangle + \langle Y, I_n\rangle$.

(*d*) Par dualité faible, $\|A\|_*$ est majoré par

$$\begin{aligned} &\inf_{\substack{X\in\mathbb{R}^{m\times m},\ Y\in\mathbb{R}^{n\times n},\ Z\in\mathbb{R}^{m\times n} \\ \begin{pmatrix} X & Z \\ Z^\mathsf{T} & Y\end{pmatrix}\succcurlyeq 0}} \sup_{B\in\mathbb{R}^{m\times n}} \langle A + 2Z, B\rangle + \langle X, I_m\rangle + \langle Y, I_n\rangle. \\ = &\inf_{\substack{X\in\mathbb{R}^{m\times m},\ Y\in\mathbb{R}^{n\times n},\ Z\in\mathbb{R}^{m\times n} \\ \begin{pmatrix} X & Z \\ Z^\mathsf{T} & Y\end{pmatrix}\succcurlyeq 0 \\ A+2Z=0}} (\operatorname{tr} X + \operatorname{tr} Y) \\ = &\inf_{\substack{X\in\mathbb{R}^{m\times m},\ Y\in\mathbb{R}^{n\times n} \\ \begin{pmatrix} X & -A \\ -A^\mathsf{T} & Y\end{pmatrix}\succcurlyeq 0}} \frac{1}{2}(\operatorname{tr} X + \operatorname{tr} Y). \end{aligned}$$

(*e*) On majore encore $\|A\|_*$ en prenant ci-dessus $X = V\Sigma V^\mathsf{T}$ et $Y = U\Sigma U^\mathsf{T}$, qui vérifient la contrainte de semi-définie positivité, puisque

$$\begin{pmatrix} V\Sigma V^\mathsf{T} & -V\Sigma U^\mathsf{T} \\ -U\Sigma V^\mathsf{T} & U\Sigma U^\mathsf{T}\end{pmatrix} = \begin{pmatrix} V \\ -U\end{pmatrix} \Sigma \begin{pmatrix} V \\ -U\end{pmatrix}^\mathsf{T} \succcurlyeq 0.$$

On obtient finalement

$$\|A\|_* \leqslant \frac{1}{2}\Big(\operatorname{tr}(V\Sigma V^\mathsf{T}) + \operatorname{tr}(U\Sigma U^\mathsf{T})\Big) = \operatorname{tr}\Sigma = \|\sigma(A)\|_1.$$

2) Cela se déduit des relations entre les normes vectorielles de $\sigma(A)$, à savoir $\|\sigma(A)\|_\infty \leqslant \|\sigma(A)\|_2 \leqslant \|\sigma(A)\|_1 \leqslant r^{1/2}\|\sigma(A)\|_2 \leqslant r\|\sigma(A)\|_\infty$ (car $\sigma(A)$ n'a que $r = \operatorname{rg}(A)$ valeurs singulières non nulles) et des identités (14.82).

3) On a $\operatorname{rg}^* = \mathcal{I}_{\{0\}}$. En effet $\operatorname{rg}^* \geqslant 0$ (en prenant $A = 0$ dans la définition de rg^*) et $\operatorname{rg}^*(0) = \sup_A [-\operatorname{rg}(A)] \leqslant 0$, si bien que $\operatorname{rg}^*(0) = 0$. Par ailleurs, si $A^* \neq 0$, $\operatorname{rg}^*(A^*) \geqslant \sup_A \langle A^*, A\rangle - \min(m, n) = +\infty$ (en prenant $A = tA^*$ avec $t \to \infty$). On déduit de l'expression trouvée de rg^* que

$$\operatorname{rg}^{**}(A) := \sup_{A^*} [\langle A^*, A\rangle - \operatorname{rg}^*(A^*)] = \langle 0, A\rangle = 0.$$

4) Par définition de la conjuguée, $f^*(A^*) = \sup\{\langle A^*, A\rangle - \operatorname{rg}(A) : \|A\| \leqslant 1\}$. On obtient une minoration de $f^*(A^*)$ comme suit. Soit $A^* = V\Sigma U^\mathsf{T}$ la factorisation

SVD de A^*. On introduit $V_s := (v_1 \cdots v_s)$ et $U_s = (u_1 \cdots u_s)$, où s est le plus grand indice tel que $\sigma_i(A) > 1$. On peut prendre $A = V_s U_s^\mathsf{T}$ dans la définition de $f^*(A^*)$, car $\|V_s U_s^\mathsf{T}\| = 1$ par (14.82) (les valeurs singulières non nulles de cette matrice valent 1), ce qui donne

$$f^*(A^*) \geqslant \operatorname{tr} A^* U_s V_s^\mathsf{T} - s = \operatorname{tr} V_s^\mathsf{T} V \Sigma U^\mathsf{T} U_s - s = \sum_{i=1}^{s} (\sigma_i(A^*) - 1) = \|(\sigma(A^*) - e)^+\|_1.$$

Pour la majoration de $f^*(A^*)$, on utilise l'inégalité de trace de von Neumann (B.38) et le fait que, pour tout A tel que $\|A\| \leqslant 1$, on a $\sigma_i(A) \leqslant 1$ et donc

$$\begin{aligned}
\sigma(A^*)^\mathsf{T}\sigma(A) - \operatorname{rg}(A) &= \sum_{\sigma_i(A^*)>1} \Big(\sigma_i(A^*) \underbrace{\sigma_i(A)}_{\leqslant 1} - 1\Big) + \sum_{\substack{\sigma_i(A^*)\leqslant 1 \\ \sigma_i(A)\neq 0}} \Big(\underbrace{\sigma_i(A^*)\sigma_i(A)}_{\leqslant 1} - 1\Big) \\
&\leqslant \sum_{\sigma_i(A^*)>1} \Big(\sigma_i(A^*) - 1\Big) \\
&= \|(\sigma(A^*) - e)^+\|_1.
\end{aligned}$$

5) Par définition, la biconjuguée s'écrit

$$f^{**}(A) = \sup_{A^*} \langle A^*, A\rangle - \|(\sigma(A^*) - e)^+\|_1. \tag{D.48}$$

[$f^{**}(A) \leqslant \|A\|_*$, si $\|A\| \leqslant 1$] Par l'inégalité de trace de von Neumann (B.38), on a

$$\begin{aligned}
f^{**}(A) &\leqslant \sup_{A^*} \Big(\sigma(A^*)^\mathsf{T}\sigma(A) - \|(\sigma(A^*) - e)^+\|_1\Big) \\
&= \|A\|_* + \sup_{A^*} \Big((\sigma(A^*) - e)^\mathsf{T} \underbrace{\sigma(A)}_{\geqslant 0} - \|(\sigma(A^*) - e)^+\|_1\Big) \\
&= \|A\|_* + \sup_{A^*} \Big(\big((\sigma(A^*) - e)^+\big)^\mathsf{T} \sigma(A) - \|(\sigma(A^*) - e)^+\|_1\Big) \\
&= \|A\|_* + \sup_{A^*} \big((\sigma(A^*) - e)^+\big)^\mathsf{T} \underbrace{\big(\sigma(A) - e\big)}_{\leqslant 0} \\
&\leqslant \|A\|_*.
\end{aligned}$$

[$f^{**}(A) \geqslant \|A\|_*$, si $\|A\| \leqslant 1$] Soit $A = V\Sigma U^\mathsf{T}$ la factorisation SVD de A. On prend $A^* = VU^\mathsf{T}$ dans (D.48). Comme les valeurs singulières non nulles de ce A^* valent 1, $\|(\sigma(A^*) - e)^+\|_1 = 0$, si bien que

$$f^{**}(A) \geqslant \operatorname{tr} AUV^\mathsf{T} = \operatorname{tr} V^\mathsf{T} V \Sigma U^\mathsf{T} U = \operatorname{tr} \Sigma = \|\sigma(A)\|_1 = \|A\|_*.$$

[$f^{**}(A) = +\infty$, si $\|A\| > 1$] Dans ce cas, la plus grande valeur singulière $\sigma_1(A) > 1$. Soient $A = V\Sigma U^\mathsf{T}$ la factorisation SVD de A, v_1 la première colonne de V et u_1 la première colonne de U. En prenant pour A^* dans (D.48), la matrice $A_t^* = tv_1u_1^\mathsf{T}$, avec $t > 1$, on trouve $\sigma(A_t^*) = te^1$ et donc

$$f^{**}(A) \geqslant t(\operatorname{tr} Au_1v_1^\mathsf{T}) - (t-1) \geqslant t(\sigma_1(A) - 1) + 1,$$

qui tend vers $+\infty$ lorsque $t \to +\infty$.

14.16. Comme le problème (14.84) s'écrit aussi $\inf_{x\in\mathbb{R}^n} \sup_{y\in\mathbb{R}^m} (\|x\|_1 - y^\mathsf{T}(Ax - b))$, son dual lagrangien s'écrit

$$\begin{aligned}
&\sup_{y\in\mathbb{R}^m}\ \inf_{x\in\mathbb{R}^n}\ \Big(\|x\|_1 - y^{\mathsf{T}}(Ax-b)\Big)\\
&\quad= \sup_{y\in\mathbb{R}^m}\ \Big(b^{\mathsf{T}}y - \sup_{x\in\mathbb{R}^n}\ (A^{\mathsf{T}}y)^{\mathsf{T}}x - \|x\|_1\Big)\\
&\quad= \sup_{y\in\mathbb{R}^m}\ \Big(b^{\mathsf{T}}y - \mathcal{I}_{\bar{B}_\infty}(A^{\mathsf{T}}y)\Big)\\
&\quad= \sup_{\|A^{\mathsf{T}}y\|_\infty\leqslant 1}\ b^{\mathsf{T}}y,
\end{aligned}$$

où $\bar{B}_\infty$ est la boule unité fermée de $\mathbb{R}^m$ pour la norme ℓ_∞ (on a utilisé l'expression de la conjuguée de la norme de l'exercice 3.30).

14.17. Le problème (14.86) s'écrit aussi

$$\inf_{X\in\mathbb{R}^{m\times n}}\ \sup_{y\in\mathbb{R}^p}\ \|X\|_* - \langle y, \mathcal{A}(X)-b\rangle. \tag{D.49}$$

Le dual lagrangien du problème (14.86) s'écrit donc

$$\begin{aligned}
&\sup_{y\in\mathbb{R}^p}\ \inf_{X\in\mathbb{R}^{m\times n}}\ \|X\|_* - \langle y, \mathcal{A}(X)-b\rangle\\
&\quad= \sup_{y\in\mathbb{R}^p}\ \left(\langle b, y\rangle - \sup_{X\in\mathbb{R}^{m\times n}}\ \langle \mathcal{A}^*(y), X\rangle - \|X\|_*\right).
\end{aligned}$$

Par la conjuguée de la norme (exercice 3.30), on a

$$\sup_{X\in\mathbb{R}^{m\times n}}\ \langle \mathcal{A}^*(y), X\rangle - \|X\|_* = \begin{cases} 0 & \text{si } \|\mathcal{A}^*(y)\|\leqslant 1\\ +\infty & \text{sinon,}\end{cases}$$

où l'on a utilisé le fait que la norme duale de $\|\cdot\|$ est la norme $\|\cdot\|_*$ et donc que la norme duale de $\|\cdot\|_*$ est la norme $\|\cdot\|$. On en déduit que le problème dual lagrangien s'écrit comme en (14.87).

Méthodes newtoniennes

15.1. On note $g_x = \nabla f(x)$, $H_x = \nabla^2 f(x)$, $A_x = c'(x)$ et $L_z = \nabla^2\ell(x,\lambda)$. Si A_x^- est un inverse à droite de A_x (qui existe si A_x est surjective) et Z_x^- a ses colonnes qui forment une base de $\mathcal{N}(A_x)$, la solution primale d_x de (15.16) vérifie on a $d_x = -A_x^- c_x + Z_x^- h_x$, avec h_x qui minimise $(g_x - H_x A_x^- c_x)^{\mathsf{T}} Z_x^- h_x + \frac{1}{2} h_x^{\mathsf{T}} \bar{H}_x h_x$, où $\bar{H}_x = Z_x^{-\mathsf{T}} H_x Z_x^-$. En supposant $\bar{H}_x$ inversible, on trouve $d_x = -A_x^- c_x - Z_x^- \bar{H}_x^{-1}(\bar{g}_x - Z_x^{-\mathsf{T}} H_x A_x^- c_x)$, où on a noté $\bar{g}_x = Z_x^{-\mathsf{T}} g_x$ le *gradient réduit*. Dès lors, l'algorithme s'écrit $x_+ = \Phi(x)$ avec $\Phi(x) = x - A_x^- c_x - Z_x^- \bar{H}_x^{-1}(\bar{g}_x - Z_x^{-\mathsf{T}} H_x A_x^- c_x)$. En observant que $\bar{g}'(x_*) = Z_*^{-\mathsf{T}} L_*$, on trouve

$$\Phi'(x_*) = I - A_*^- A_* - Z_*^- \bar{H}_*^{-1} Z_*^{-\mathsf{T}}(L_* - H_* A_*^- A_*) = Z_*^- \bar{H}_*^{-1} Z_*^{-\mathsf{T}}(H_* - L_*).$$

On observe que $\Phi'(x_*)$ serait nul si H_* était égal à L_*, mais n'a pas raison d'avoir ses valeurs propres dans $[-1,1]$ en général. Dans l'exemple proposé, on a $x_* = (0,-1)$, $\lambda_* = 1$ et

$$A_* = \begin{pmatrix}0 & 2\end{pmatrix},\quad Z_*^- = \begin{pmatrix}1\\0\end{pmatrix},\quad H_* = \begin{pmatrix}-2a & 0\\ 0 & 0\end{pmatrix},\quad \text{et}\quad L_* = \begin{pmatrix}2(1-a) & 0\\ 0 & 2\end{pmatrix}.$$

Donc $\bar{H}_* = -2a$ et $\Phi'(x_*) = \mathrm{Diag}(\frac{1}{a}, 0)$ a une valeur propre > 1.

Méthodes de points intérieurs

16.1. L'ellipsoïde $x + \mathscr{E}_A$, centré en x, est dans $\mathbb{R}^n_+$ si $x + v \geqslant 0$ pour tout $v \in \mathscr{E}_A$ ou encore si $x + A^{-1/2}u \geqslant 0$ pour tout $u \in \bar{B}$ (la boule unité fermée pour la norme euclidienne). Ceci s'exprime aussi en disant que, pour tout $i \in [1:n]$ et tout $u \in \bar{B}$: $x_i + (e^i)^\mathsf{T} A^{-1/2}u \geqslant 0$ (e^i désigne le i-ième vecteur de base de $\mathbb{R}^n$). L'infimum en $u \in \bar{B}$ de cette dernière quantité est obtenu pour $u = -A^{-1/2}e^i/\|A^{-1/2}e^i\|_2$ et vaut donc $x_i - \|A^{-1/2}e^i\|_2$. La condition $x + \mathscr{E}_A \subseteq \mathbb{R}^n_+$ s'exprime donc par les n inégalités $\|A^{-1/2}e^i\|_2 \leqslant x_i$, $i = 1, \dots, n$. Par ailleurs, le volume de $\mathscr{E}_A$ (ou de $x + \mathcal{E}_A$) s'écrit

$$V(\mathscr{E}_A) = \int_{\mathscr{E}_A} 1 \, \mathrm{d}v = \int_{\bar{B}} |\det A^{-1/2}| \, \mathrm{d}u = \sigma_n \det A^{-1/2},$$

où σ_n est le volume de $\bar{B}$. En faisant le changement de variable $M = A^{-1/2}$, on voit que l'on cherche à évaluer la fonction

$$f(x) = \inf_{\substack{M \succ 0 \\ \|Me^i\|_2 \leqslant x_i, \forall i}} -\log\det M.$$

Si $x > 0$, ce problème a une solution unique $M \succ 0$ (le critère est strictement convexe, fermé et les contraintes d'inégalité définissent un ensemble convexe compact non vide). Si on note λ_i les multiplicateurs associés aux contraintes (élevées au carré) et $\Lambda = \mathrm{Diag}(\lambda_1, \dots, \lambda_n)$, les conditions d'optimalité s'écrivent :

$$-M^{-1} + \Lambda M + M\Lambda = 0 \quad \text{et} \quad \forall i : \ \|Me^i\|_2 \leqslant x_i, \ \lambda_i \geqslant 0, \ \lambda_i(\|Me^i\|_2 - x_i) = 0.$$

On a clairement $\Lambda \succ 0$, car si $\lambda_i = 0$, l'élément (i,i) de la première équation s'écrirait $(M^{-1})_{ii} = 0$, ce qui n'est pas compatible avec le fait que $M \succ 0$. Cherchons M diagonale (le système ayant au plus une solution, si on en trouve une avec M diagonale, c'est la bonne). La première équation accepte une solution de cette forme : $M = (2\Lambda)^{-1/2}$. Par ailleurs, comme $\lambda > 0$, on a $\|Me^i\|_2 = x_i$ par complémentarité, si bien que $M = \mathrm{Diag}(x_1, \dots, x_n)$ et les $\lambda_i = 1/(2x_i^2) > 0$. Dès lors $f(x) = -\log\det \mathrm{Diag}(x_1, \dots, x_n) = -\sum_i \log x_i = \mathrm{lb}(x)$.

Comme la matrice A optimale est X^{-2}, l'ellipsoïde le plus volumineux contenu dans $\mathbb{R}^n_+$, centré en $x > 0$, est donc $\{v \in \mathbb{R}^n : v^\mathsf{T} X^{-2} v \leqslant 1\}$ qui n'est autre que la boule unité $\bar{B}_x(0,1)$ pour la norme locale en x associée à lb.

16.2. 1) La vérification de l'autoconcordance immédiate, car $\mathrm{dom}\, f = \mathbb{R}^n$ et $\|v\|_x^2 = v^\mathsf{T} A v$ est indépendant de x. Mais f n'est pas une barrière, car $\|g_x(x)\|_x^2 = x^\mathsf{T} A x + 2b^\mathsf{T} x + b^\mathsf{T} A^{-1} b$ n'est pas borné.

Problèmes de moindres-carrés

19.1. 1) On peut voir (P) comme le problème $\inf\{\|y - b\|_\mathbb{F} : y \in \mathcal{R}(A)\}$, dans lequel on projète b sur $\mathcal{R}(A)$. Il ne s'agit pas d'une projection au sens habituel, car la norme $\|\cdot\|_\mathbb{F}$ ne dérive pas nécessairement d'un produit scalaire. Cette projection existe (mais elle n'est pas nécessairement unique) puisque le critère tend vers l'infini à l'infini et que l'ensemble admissible est fermé non vide (corollaire 1.4).

2) Soit f_ε le critère de (P_ε). Il est clair que f_ε est continue et que $f_\varepsilon(x) \to \infty$ si $\|x\|_\mathbb{E} \to \infty$; donc (P_ε) a au moins une solution (corollaire 1.4).

3) Par optimalité de $\bar{x}_\varepsilon$, $f_\varepsilon(\bar{x}_\varepsilon) \leqslant f_\varepsilon(0)$, ce qui conduit à $\|\bar{x}_\varepsilon\|_\mathbb{E}^\beta \leqslant \|b\|_\mathbb{F}^\alpha/\varepsilon$, montrant que $\bar{x}_\varepsilon \to 0$ lorsque $\varepsilon \to \infty$.

4) Par la proposition 13.2, lorsque $\varepsilon \downarrow 0$, $\|\bar{x}_\varepsilon\|_{\mathbb{E}}^{\beta}$ croît, donc aussi $\|\bar{x}_\varepsilon\|_{\mathbb{E}}$.

Soit $\hat{x}$ une solution de (P). D'une part $f_\varepsilon(\bar{x}_\varepsilon) \leqslant f_\varepsilon(\hat{x})$ par optimalité de $\bar{x}_\varepsilon$ et, d'autre part, $\|A\hat{x} - b\|_{\mathbb{F}} \leqslant \|A\bar{x}_\varepsilon - b\|_{\mathbb{F}}$ par optimalité de $\hat{x}$. En additionnant ces inégalités, on obtient $\|\bar{x}_\varepsilon\|_{\mathbb{E}} \leqslant \|\hat{x}\|_{\mathbb{E}}$, ce qui montre que $\{\bar{x}_\varepsilon\}$ est bornée.

Le fait que les points d'adhérence de $\{\bar{x}_\varepsilon\}$ sont solutions de $\min\{\|x\|_{\mathbb{E}} : x \in S\}$, se déduit de la proposition 13.3.

5) La hessienne de f_ε s'écrit $\nabla^2 f_\varepsilon(x) = 2(A^*A + \varepsilon I)$. Comme elle est définie positive, f_ε est strictement convexe et (P_ε) a une solution unique.

La solution $\bar{x}_\varepsilon$ vérifie $A^*(A\bar{x}_\varepsilon - b) + \varepsilon\bar{x}_\varepsilon = 0$. Si $\varepsilon \to \infty$, $\bar{x}_\varepsilon \to 0$ par le point 3 et cette dernière équation montre que $\varepsilon\bar{x}_\varepsilon \to A^*b \neq 0$.

Comme le problème $\min\{\|x\|_{\mathbb{E}} : x \in S\}$ a aussi une unique solution (projection de zéro sur S), toute la suite $\{\bar{x}_\varepsilon\}_{\varepsilon\downarrow 0}$ converge vers cette solution de norme minimale.

19.2. 1) Il s'agit de la minimisation d'une fonction continue sur un compact non vide (théorème 1.2).

2) On peut récrire le problème comme suit : $\min\{\frac{1}{2}\|Ax - b\|^2 : \frac{1}{2}(\|x\|^2 - \Delta^2) \leqslant 0\}$. Sa contrainte est qualifiée (0 est un point de Slater) et les fonctions qui le définissent sont différentiables, si bien que l'on peut écrire ses conditions de KKT : il existe un multiplicateur λ tel que

$$\begin{cases} (A^\mathsf{T}A + \lambda I)x = A^\mathsf{T}b \\ \|x\| \leqslant \Delta \\ \lambda \geqslant 0 \\ \lambda(\|x\| - \Delta) = 0. \end{cases} \tag{D.50}$$

Ces conditions sont nécessaires et suffisantes pour l'optimalité, car le critère et la contrainte sont convexes.

3) On utilise les conditions d'optimalité (D.50). Si $\lambda > 0$, alors $x = A^\mathsf{T}(b - Ax)/\lambda$ montre que $x \in \mathcal{R}(A^\mathsf{T})$. Si $\lambda = 0$, alors $A^\mathsf{T}(Ax - b) = 0$ et la solution de $\min\{\|x\| : A^\mathsf{T}(Ax - b) = 0\}$ sera certainement aussi solution de (D.50) avec un multiplicateur nul (elle sera aussi de norme $\leqslant \Delta$) ; c'est la solution de norme minimale d'un problème de moindres-carrés linéaire, qui est dans $\mathcal{R}(A^\mathsf{T})$, voir (19.5).

4) [$\Rightarrow$] On montre la contraposée en supposant qu'il existe deux points $x_1 \neq x_2$ dans $E := \bar{B}(0, \Delta) \cap \{x : A^\mathsf{T}(Ax - b) = 0\}$. Alors ces deux points sont solutions de (D.50) avec $\lambda = 0$, donc solutions de (19.59).

[$\Leftarrow$] On montre à nouveau la contraposée en supposant qu'il existe deux solutions $x_1 \neq x_2$. Comme le problème est convexe, l'ensemble des solutions primales-duales forment un produit cartésien, si bien qu'il existe un multiplicateur λ commun aux deux solutions. Par la première équation de (D.50) et le fait que $A^\mathsf{T}A + \lambda I \succ 0$ si $\lambda > 0$, on a nécessairement $\lambda = 0$. Alors $x_1 \neq x_2$ sont clairement dans E.

19.3. 1) On voit (19.60) comme le problème $\inf\{\|y - b\|_\infty : y \in \mathcal{R}(A)\}$, dans lequel on « projette » le point b sur $\mathcal{R}(A)$. Cette projection existe (mais elle n'est pas nécessairement unique) puisque le critère est coercif et que l'ensemble admissible est fermé non vide (corollaire 1.4).

2) Soit x_{tch} une solution de (19.60). On a $\|r_{\text{mc}}\|_\infty \leqslant \|r_{\text{mc}}\|_2 \leqslant \|Ax_{\text{tch}} - b\|_2$ [optimalité du problème de MCL] $\leqslant \sqrt{m}\,\|Ax_{\text{tch}} - b\|_\infty = \sqrt{m}\,v_{\text{tch}}$.

3) Il est clair que (19.62) a une solution, car le critère est continu et l'ensemble admissible est compact non vide.

Première technique de dualisation. La norme ℓ_∞ est la norme duale de la norme ℓ_1, si bien que le problème (19.60) s'écrit

$$\inf_{x\in\mathbb{R}^n} \sup_{\substack{y\in\mathbb{R}^m \\ \|y\|_1\leqslant 1}} y^\mathsf{T}(b-Ax).$$

On peut donc prendre comme dual min-max

$$\sup_{\substack{y\in\mathbb{R}^m \\ \|y\|_1\leqslant 1}} \inf_{x\in\mathbb{R}^n} y^\mathsf{T}(b-Ax) = \sup_{\substack{y\in\mathbb{R}^m \\ \|y\|_1\leqslant 1}} \begin{cases} b^\mathsf{T}y & \text{si } A^\mathsf{T}y=0 \\ -\infty & \text{sinon,} \end{cases}$$

qui est bien (19.62).

Deuxième technique de dualisation. On récrit d'abord le problème (19.60) comme le problème avec contrainte

$$\begin{cases} \inf_{(x,z)\in\mathbb{R}^n\times\mathbb{R}^m} \|z\|_\infty \\ b-Ax=z. \end{cases}$$

La dualisation lagrangienne de la contrainte permet alors d'écrire le problème (19.60) comme un infsup :

$$\inf_{(x,z)\in\mathbb{R}^n\times\mathbb{R}^m} \sup_{y\in\mathbb{R}^m} \|z\|_\infty + y^\mathsf{T}(b-Ax-z).$$

Le dual de celui-ci s'écrit alors

$$\sup_y \inf_{(x,z)} \|z\|_\infty + y^\mathsf{T}(b-Ax-z) = \sup_y \Big(b^\mathsf{T}y + \inf_x (-y^\mathsf{T}Ax) + \inf_z (\|z\|_\infty - y^\mathsf{T}z) \Big).$$

D'une part, $\inf\{-y^\mathsf{T}Ax : x\in\mathbb{R}^n\}$ est nul si $A^\mathsf{T}y=0$ et $-\infty$ autrement. D'autre part, $\inf\{\|z\|_\infty - y^\mathsf{T}z : z\in\mathbb{R}^m\} = -\sup\{y^\mathsf{T}z - \|z\|_\infty : z\in\mathbb{R}^m\}$ est nul si $\|y\|_1\leqslant 1$ et $-\infty$ autrement (exercice 3.30). Cela conduit à (19.62).

4) Le point $-r_{\mathrm{mc}}/\|r_{\mathrm{mc}}\|_1$ est admissible pour (19.62), car $A^\mathsf{T}r_{\mathrm{mc}}=0$ (c'est l'équation normale du problème de moindres-carrés linéaire). Alors $\|r_{\mathrm{mc}}\|_2^2/\|r_{\mathrm{mc}}\|_1 = -b^\mathsf{T}r_{\mathrm{mc}}/\|r_{\mathrm{mc}}\|_1$ $[A^\mathsf{T}r_{\mathrm{mc}}=0] \leqslant \mathrm{val}((19.62))$ [optimalité dans (19.62)] $\leqslant v_{\mathrm{tch}}$ [dualité faible]. Pour montrer que $\|r_{\mathrm{mc}}\|_\infty/\sqrt{m} \leqslant \|r_{\mathrm{mc}}\|_2^2/\|r_{\mathrm{mc}}\|_1$, on utilise le fait que, quel que soit $y\in\mathbb{R}^m$, on a $\|y\|_\infty \leqslant \|y\|_2$ (clair) et $\|y\|_1 \leqslant \sqrt{m}\,\|y\|_2$ (Cauchy-Schwarz).

19.4. 1) On a en effet $\partial f(x) = A^\mathsf{T}\partial(\|\cdot\|)(Ax-b)$ (proposition 3.72) et en utilisant (3.82), on obtient (19.63).

2) On peut s'y prendre d'au moins deux manières différentes.

- Dans la première approche, on profite du fait que l'on peut écrire la norme $\|\cdot\|$ comme un supremum. En effet, puisque $\|\cdot\| = \|\cdot\|_{\mathrm{DD}}$ (point 3 de l'exercice 3.30), on a $\|v\| = \sup\{y^\mathsf{T}v : \|y\|_{\mathrm{D}} \leqslant 1\}$ et donc le problème (P) s'écrit

$$\inf_x \sup_{\|y\|_{\mathrm{D}}\leqslant 1} y^\mathsf{T}(b-Ax).$$

Alors, on peut prendre comme dual min-max de (P) :

$$\sup_{\|y\|_{\mathrm{D}}\leqslant 1} \inf_x y^\mathsf{T}(b-Ax) = \sup_{\|y\|_{\mathrm{D}}\leqslant 1} \begin{cases} b^\mathsf{T}y & \text{si } A^\mathsf{T}y=0 \\ -\infty & \text{sinon.} \end{cases}$$

Ce problème est équivalent au problème (D).

- Dans la seconde approche, on récrit (P) comme un problème avec contrainte

$$(\tilde{P}) \qquad \begin{cases} \inf_{(x,v)\in\mathbb{R}^n\times\mathbb{R}^m} \|v\| \\ Ax - b + v = 0 \end{cases}$$

et on dualise sa contrainte par le lagrangien. Cela permet d'écrire (P) comme le problème

$$\inf_{(x,v)\in\mathbb{R}^n\times\mathbb{R}^m} \sup_{y\in\mathbb{R}^m} \|v\| - y^\mathsf{T}(Ax - b + v).$$

Alors, on peut prendre comme dual min-max de (P) :

$$\begin{aligned} &\sup_{y\in\mathbb{R}^m} \inf_{(x,v)\in\mathbb{R}^n\times\mathbb{R}^m} \|v\| - y^\mathsf{T}(Ax - b + v) \qquad (D.51)\\ &\quad = \sup_{y\in\mathbb{R}^m} \left(b^\mathsf{T} y - \sup_x (A^\mathsf{T} y)^\mathsf{T} x - \sup_v \left[y^\mathsf{T} v - \|v\| \right] \right) \\ &\quad = \sup_{y\in\mathbb{R}^m} \begin{cases} b^\mathsf{T} y & \text{si } A^\mathsf{T} y = 0 \text{ et } \|y\|_\mathrm{D} \leqslant 1 \\ -\infty & \text{sinon,} \end{cases} \end{aligned}$$

où l'on a utilisé la conjuguée $\mathcal{I}_{\bar{B}_\mathrm{D}}$ de la norme (point 2 de l'exercice 3.30). Ce dernier problème est équivalent au problème (D).

3) Le problème (P) peut aussi s'écrire $\inf\{\|y - b\| : y \in \mathcal{R}(A)\}$ qui a une solution $y \in \mathcal{R}(A)$ (car le critère est continu coercif et l'ensemble admissible est fermé non vide), donc il existe un $x \in \mathbb{R}^n$ tel que $y = Ax$ et ce x est solution de (P).

Le problème (D) consiste à maximiser une fonction continue sur un compact non vide ; il a donc une solution.

Typiquement, lorsque (P) et (D) ont une solution, l'absence de saut de dualité se montre au moyen des conditions d'optimalité (voir les propositions 14.25 et 14.31).

- La condition d'optimalité de (P) en une solution x s'écrit $0 \in \partial f(x)$. Alors, par (19.63), il existe un $y' \in \mathbb{R}^m$ tel que $A^\mathsf{T} y' = 0$, $\|y'\|_\mathrm{D} \leqslant 1$ et $(y')^\mathsf{T}(Ax - b) = \|Ax - b\|$ ou $-(y')^\mathsf{T} b = \|Ax - b\| = \mathrm{val}(P)$. On voit alors que $y = -y'$ vérifie les contraintes de (D) avec un coût égal à $\mathrm{val}(P)$. Comme $\mathrm{val}(D) \leqslant \mathrm{val}(P)$, on en déduit que y est solution de (D) et qu'il n'y a pas de saut de dualité.
- On peut aussi utiliser les conditions d'optimalité de (D). Comme les contraintes de (D) sont qualifiées ($y = 0$ est un point de Slater), il existe des multiplicateurs $x \in \mathbb{R}^n$ et $\sigma \in \mathbb{R}$ tels que

$$\begin{cases} 0 \in \partial_y(-b^\mathsf{T} y + (A^\mathsf{T} y)^\mathsf{T} x + \sigma(\|y\|_\mathrm{D} - 1) \\ A^\mathsf{T} y = 0 \\ 0 \leqslant \sigma \perp (\|y\|_\mathrm{D} - 1) \leqslant 0. \end{cases}$$

Si $\sigma = 0$, on obtient $Ax = b$, si bien que $\mathrm{val}(P) = 0$. Comme $\mathrm{val}(D) \leqslant \mathrm{val}(P) = 0$ et que $y = 0$ est admissible pour (D), $y = 0$ est solution de (D), ce qui montre que $\mathrm{val}(D) = 0$. Il n'y a donc pas de saut de dualité.
Si $\sigma > 0$, on obtient

$$\begin{cases} \frac{1}{\sigma}(b - Ax) \in \partial(\|\cdot\|_\mathrm{D})(y) \\ A^\mathsf{T} y = 0 \\ \|y\|_\mathrm{D} = 1, \end{cases}$$

ce qui s'écrit aussi

$$\begin{cases} \|Ax - b\| \leqslant \sigma, \quad (b - Ax)^\mathsf{T} y = \sigma\|y\|_\mathrm{D} \\ A^\mathsf{T} y = 0 \\ \|y\|_\mathrm{D} = 1. \end{cases}$$

On en déduit que $b^\mathsf{T} y = \sigma \leqslant \mathrm{val}(D)$ et $\mathrm{val}(P) \leqslant \|Ax - b\| \leqslant \sigma$. Donc $\mathrm{val}(D) = \mathrm{val}(P)$ par dualité faible et il n'y a pas de saut de dualité.

4) Si $x \in \mathrm{Sol}(P)$, on peut écrire $x = x_0 + x_1$ avec $x_0 \in \mathcal{N}(A)$ et $x_1 \in \mathcal{N}(A)^\perp = \mathcal{R}(A^\mathsf{T})$. Comme $Ax - b = Ax_1 - b$, il s'ensuit que $x_1 \in \mathcal{R}(A^\mathsf{T}) \cap \mathrm{Sol}(P) = C$. Donc $x \in C + \mathcal{N}(A)$. Inversement, si $x = x_1 + x_0$ avec $x_1 \in C$ et $x_0 \in \mathcal{N}(A)$, on a $Ax - b = Ax_1 - b$, si bien que $\|Ax - b\| = \mathrm{val}(P)$, ce qui signifie que $x \in \mathrm{Sol}(P)$.

L'ensemble C contient les solutions de (P) qui sont dans l'image de A^T. Lorsque $\|\cdot\| = \|\cdot\|_2$, cet ensemble contient l'unique solution de norme minimale de (P).

5) En effet, (P) se récrit alors comme suit

$$\begin{cases} \min_{(x,t)\in\mathbb{R}^n\times\mathbb{R}^m} \sum_{i=1}^m t_i \\ -t_i \leqslant (Ax-b)_i \leqslant t_i, \quad \forall i \in [1:m], \end{cases}$$

qui est bien un problème d'optimisation linéaire (minimisation d'une fonction linéaire sur un polyèdre convexe).

Optimisation linéaire : théorie et algorithme du simplexe

17.1. Dans le système d'optimalité, on remplace les conditions de complémentarité (non linéaires) par une condition équivalente linéaire, à savoir l'absence de saut de dualité. Ceci conduit au système suivant

$$\begin{cases} Ax = b, \quad x \geqslant 0 & \text{(admissibilité primale)} \\ A^\mathsf{T} y + s = c, \quad s \geqslant 0 & \text{(admissibilité duale)} \\ c^\mathsf{T} x = b^\mathsf{T} y & \text{(pas de saut de dualité).} \end{cases}$$

17.2. (i) En effet, l'application $c \mapsto \min\{c^\mathsf{T} x : Ax = b,\ x \geqslant 0\}$ s'obtient comme l'infimum en $x \in \{Ax = b,\ x \geqslant 0\}$ de fonctions linéaires $c \mapsto c^\mathsf{T} x$ (proposition 3.34).

(ii) On peut supposer que l'ensemble admissible X de (P_L) n'est pas un singleton, sinon le résultat est immédiat. Soient $\mathrm{T}_{\bar{x}} X$ le cône tangent à X en $\bar{x}$, S la sphère unité de $\mathbb{R}^n$ et $\epsilon := \frac{1}{2}\inf\{c^\mathsf{T} d : d \in \mathrm{T}_{\bar{x}} X \cap S\}$. On a $\epsilon > 0$. En effet, comme $\mathrm{T}_{\bar{x}} X \cap S$ est un compact non vide (car X n'est pas un singleton), le problème qui définit ϵ a une solution, disons $\bar{d}$, et si $\epsilon \leqslant 0$, on aurait $c^\mathsf{T}\bar{d} \leqslant 0$, $\bar{d} \neq 0$ et un $t > 0$ tel que $\bar{x}' := \bar{x} + t\bar{d} \in X$ (d'après l'exercice 2.43, $\mathrm{T}_{\bar{x}} X$ est l'ensemble des directions d telles que $\bar{x} + \mathbb{R}_{++} d$ rencontre X), si bien que $\bar{x}'$ serait une autre solution de (P_L). Soit alors $\tilde{c}$ tel que $\|\tilde{c} - c\| \leqslant \epsilon$ et montrons que $\bar{x}$ est encore l'unique solution de (P_L) avec c remplacé par $\tilde{c}$. Soit $x \neq \bar{x}$, avec $x \in X$, et posons $d = x - \bar{x}$. On a $d/\|d\| \in \mathrm{T}_{\bar{x}} X \cap S$ et donc $c^\mathsf{T} d/\|d\| \geqslant 2\epsilon$, si bien que $\tilde{c}^\mathsf{T} d/\|d\| \geqslant \epsilon > 0$ et $\tilde{c}^\mathsf{T} x > \tilde{c}^\mathsf{T}\bar{x}$. Dès lors x n'est pas solution du problème perturbé.

Inversement, si $\bar{x}'$ est une autre solution de (P_L) ($\bar{x}$ n'est pas l'unique solution) et $\tilde{c} = c + \epsilon(\bar{x} - \bar{x}')$, alors $\bar{x}$ n'est plus solution du problème (P_L) avec c remplacé par $\tilde{c}$ et cela quel que soit $\epsilon > 0$. En effet, on a $\tilde{c}^\mathsf{T}(\bar{x} - \bar{x}') = \epsilon\|\bar{x} - \bar{x}'\|^2 > 0$.

(iii) D'après le point (ii), $\varphi(c) = c^\mathsf{T}\bar{x}_0$ pour c voisin de c_0. On en déduit le résultat.

17.3. xxx

17.4. Autre démonstration (plus rapide). On utilise la proposition 17.13 : il existe alors $s > 0$ et y tels que $A^\mathsf{T} y + s = c$. Alors pour tout $x \in \mathcal{F}_\mathrm{P}$, on a $c^\mathsf{T} x = y^\mathsf{T} b + s^\mathsf{T} x$, qui tend vers l'infini lors que $\|x\| \to \infty$ dans $\mathcal{F}_\mathrm{P}$.

17.5. $[(i) \Rightarrow (ii)]$ Soient $i \in [1:m_B]$ et $b_i^\mathsf{T} \in \mathbb{R}^{1\times n}$ la i-ième ligne de $B \in \mathbb{R}^{m_B\times n}$. Considérons le problème d'optimisation linéaire en $x \in \mathbb{R}^n$ suivant

$$\begin{cases} \inf_x \ -b_i^\mathsf{T} x \\ Ax \leqslant a. \end{cases}$$

Par hypothèse, ce problème est réalisable (car $P_A \neq \varnothing$)) si bien que sa valeur optimale n'est pas $+\infty$. Par ailleurs, si (i) a lieu, sa valeur optimale est $\geqslant -\beta_i$, où β_i est la i-ième composante de b, car chaque fois que $Ax \leqslant a$, on a $Bx \leqslant b$ et donc $b_i^\mathsf{T} x \leqslant \beta_i$. Le problème a donc une solution, disons $\bar{x}_i \in \mathbb{R}^n$. Le lagrangien du problème s'écrit

$$(x, y_i) \in \mathbb{R}^n \times \mathbb{R}^{m_A} \mapsto -b_i^\mathsf{T} x + y_i^\mathsf{T}(Ax - a).$$

Par les conditions d'optimalité, on a l'existence d'un $y_i \in \mathbb{R}^{m_A}$, tel que

$$b_i^\mathsf{T} = y_i^\mathsf{T} A, \qquad 0 \leqslant y_i \perp (A\bar{x}_i - a) \leqslant 0.$$

Alors, on a

$$\begin{array}{rcll} y_i^\mathsf{T} a & = & y_i^\mathsf{T} A\bar{x}_i & \text{[complémentarité]} \\ & = & b_i^\mathsf{T} \bar{x}_i & [b_i^\mathsf{T} = y_i^\mathsf{T} A] \\ & \leqslant & \beta_i & \text{[valeur optimale } \geqslant -\beta_i]. \end{array}$$

On obtient la condition (ii) en prenant la i-ième ligne de Y égale à y_i^T.

$[(ii) \Rightarrow (i)]$ Supposons que (ii) ait lieu et que $x \in P_A$, si bien que $Ax \leqslant a$. Alors, en utilisant les propriétés de Y données dans (ii), on a

$$Bx = \underbrace{Y}_{\geqslant 0} \underbrace{Ax}_{\leqslant a} \leqslant Ya \leqslant b.$$

Donc $x \in P_B$ et (i) a lieu.

17.6. 1) $[(i) \Rightarrow (ii)]$ Lorsque $A \subseteq B$, on a pour tout $d \in \mathbb{E}$: $\sup\{\langle d, x\rangle : x \in A\} \leqslant \sup\{\langle d, x\rangle : x \in B\}$, c'est-à-dire $\sigma_A \leqslant \sigma_B$.

$[(ii) \Rightarrow (i)]$ Comme $\sigma_A = \mathcal{I}_A^*$ (conjuguée de l'indicatrice de A), l'inégalité $\sigma_A \leqslant \sigma_B$ s'écrit $\mathcal{I}_A^* \leqslant \mathcal{I}_B^*$. En prenant la conjuguée, on obtient $\mathcal{I}_B^{**} \leqslant \mathcal{I}_A^{**}$ ou $\mathcal{I}_B \leqslant \mathcal{I}_A$ (point 3 de la proposition 3.46 et le fait que les idicatrices sont dans $\overline{\mathrm{Conv}}(\mathbb{E})$ car les ensembles sont convexes fermés non vides) ou encore $A \subseteq B$.

$[(ii) \Rightarrow (iii) \Rightarrow (iv)]$ Clair.

$[(iv) \Rightarrow (ii)]$ Soit $d \in \mathbb{E}$. Observons d'abord que la non-vacuité de B implique que $\sigma_B(d) > -\infty$. On distingue trois cas (on utilise chaque fois l'homogénéité positive de la fonction d'appui).

- Si $\sigma_B(d) > 0$, alors $\sigma_B(d/\sigma_B(d)) = 1$, si bien que $\sigma_A(d/\sigma_B(d)) \leqslant 1$ par l'inclusion de (iv) avec $\alpha = 1$, d'où l'on déduit que $\sigma_A(d) \leqslant \sigma_B(d)$.
- Si $\sigma_B(d) = 0$, alors, pour tout $t > 0$, $\sigma_B(td)) = t\sigma_B(d)) \leqslant 1$. Par l'inclusion de (iv) avec $\alpha = 1$, on en déduit que $\sigma_A(td)) \leqslant 1$ ou $\sigma_A(d)) \leqslant 1/t$, pour tout $t > 0$. On déduit que $\sigma_A(d)) \leqslant 0$ et donc que $\sigma_A(d) \leqslant \sigma_B(d)$.
- Si $\sigma_B(d) < 0$, alors $\sigma_B(d/(-\sigma_B(d))) = -1$, si bien que $\sigma_A(d/(-\sigma_B(d))) \leqslant -1$ par l'inclusion de (iv) avec $\alpha = -1$, d'où l'on déduit que $\sigma_A(d) \leqslant \sigma_B(d)$.

2) $[\subseteq]$ Si d est dans le membre de gauche de (17.22), on a

$$\sup_{Ax \leqslant b} d^\mathsf{T} x \leqslant \alpha.$$

Le problème d'optimisation à gauche de l'inégalité est linéaire borné et réalisable. Il a donc une solution x et ses contraintes sont qualisifées par (QC-A),

si bien qu'il existe un multiplicateur optimal $y \in \mathbb{R}^m$ tel que $d = A^\mathsf{T} y$ et $0 \leqslant y \perp (Ax - b) \leqslant 0$. Alors, la valeur optimale du problème vaut $d^\mathsf{T} x = y^\mathsf{T} Ax = b^\mathsf{T} y$. Comme celle-ci doit être $\leqslant \alpha$, d est dans l'ensemble à droite dans (17.22).

[$\supseteq$] Inversement, si $d = A^\mathsf{T} y$ avec $y \geqslant 0$ et $b^\mathsf{T} y \leqslant \alpha$, alors, pour tout $x \in P$, on a $d^\mathsf{T} x = y^\mathsf{T} Ax \leqslant y^\mathsf{T} b \leqslant \alpha$, si bien que d est dans l'ensemble à gauche dans (17.22).

3) [$\subseteq$] Si d est dans le membre de gauche de (17.23), on a

$$\sup_{x^\mathsf{T} Hx \leqslant 1} d^\mathsf{T} x \leqslant 1.$$

Le problème d'optimisation de gauche a une solution x (minimisation d'une fonction continue sur un compact non vide) et sa contrainte est qualisifée par (QC-S), si bien qu'il existe un multiplicateur optimal $\lambda \in \mathbb{R}$ tel que $d = \lambda Hx$ et $0 \leqslant \lambda \perp (x^\mathsf{T} Hx - 1) \leqslant 0$. Si $d \neq 0$, on en déduit que $d^\mathsf{T} x = \lambda(x^\mathsf{T} Hx) = \lambda$, puis $d^\mathsf{T} H^{-1} d = \lambda d^\mathsf{T} x = \lambda^2$, d'où l'on déduit que $d^\mathsf{T} x = (d^\mathsf{T} H^{-1} d)^{1/2}$, qui doit être $\leqslant 1$. Donc $d \in \mathscr{E}(H^{-1})$.

[$\supseteq$] Inversement, si $d^\mathsf{T} H^{-1} d \leqslant 1$ et $x^\mathsf{T} Hx \leqslant 1$, on a $d^\mathsf{T} x = d^\mathsf{T} H^{-1/2} H^{1/2} x \leqslant \|H^{-1/2} d\| \, \|H^{1/2} x\| = (d^\mathsf{T} H^{-1} d)^{1/2} (x^\mathsf{T} Hx)^{1/2} \leqslant 1$. Donc d est dans le membre de gauche de (17.23).

4) L'équivalence (17.24a) (resp. (17.24b)) se déduit directement de l'équivalence $(i) \Leftrightarrow (iii)$ (resp. $(i) \Leftrightarrow (iv)$) et de l'identité (17.22).

Retrouvons l'équivalence de l'exercice 17.5 à partir de (17.24).

- [(ii) de 17.5 $\Rightarrow$ (17.24a)] Soient $\alpha \in \mathbb{R}$ et $z \in \mathbb{R}_+^{m_B}$ tel que $b^\mathsf{T} z \leqslant \alpha$. Avec $Y \in \mathbb{R}_+^{m_B \times m_A}$ donné par la clause (ii) de l'exercice 17.5, on a $B^\mathsf{T} z = A^\mathsf{T} Y^\mathsf{T} z$. Alors $y = Y^\mathsf{T} z$ est tel que $B^\mathsf{T} z = A^\mathsf{T} y$, $y \geqslant 0$ et $a^\mathsf{T} y = a^\mathsf{T} Y^\mathsf{T} z \leqslant b^\mathsf{T} z \leqslant \alpha$.
- [(17.24b) $\Rightarrow$ (ii) de 17.5] Il suffit de construire $Y \in \mathbb{R}_+^{m_B \times m_A}$ ligne par ligne et considère sa ligne $i \in [1 : m_B]$ ci-dessous.
  - Si $b_i > 0$, $z = e^i / b_i$ est positif et vérifie $b^\mathsf{T} z = 1$. En prenant $\alpha = 1$ dans (17.24b), $B^\mathsf{T} z = B_{i:}^\mathsf{T} / b_i$ peut s'écrire $A^\mathsf{T} \tilde{y}$ avec $\tilde{y} \geqslant 0$ et $a^\mathsf{T} \tilde{y} \leqslant 1$. Alors $y = b_i \tilde{y}$ est positif et satisfait $B_{i:} = y^\mathsf{T} A$ et $y^\mathsf{T} a \leqslant b_i$. Ce y peut donc être pris comme la i-ième ligne de Y.
  - Si $b_i = 0$, $z = te^i$ est positif et vérifie $b^\mathsf{T} z = 0 \leqslant 1$. En prenant $\alpha = 1$ dans (17.24b), $B^\mathsf{T} z = t B_{i:}^\mathsf{T}$ peut s'écrire $A^\mathsf{T} \tilde{y}_t$ avec $\tilde{y}_t \geqslant 0$ et $a^\mathsf{T} \tilde{y}_t \leqslant 1$. En posant $y_t := \tilde{y}/t$, on a $B_{i:}^\mathsf{T} = A^\mathsf{T} y_t$, $y_t \geqslant 0$ et $a^\mathsf{T} y_t \leqslant 1/t$.
    - S'il existe un $t > 0$ tel que $a^\mathsf{T} y_t \leqslant 0$, ce $y := y_t$ permet d'écrire $B_{i:}^\mathsf{T} = A^\mathsf{T} y$, avec $y \geqslant 0$ et $a^\mathsf{T} y \leqslant 0 = b_i$, comme souhaité.
    - Dans le cas contraire, la suite $\{(A^\mathsf{T} y_t, a^\mathsf{T} y_t)\}_{t \geqslant 1}$, qui prend ses valeurs dans $\{0\} \times [0, 1]$, est bornée. On peut alors en extraire une sous-suite convergence, dont la limite est nécessairement dans le (polyèdre convexe) fermé $(A^\mathsf{T}, a^\mathsf{T})(\mathbb{R}_+^m)$. On a donc à nouveau trouvé un $y \in \mathbb{R}_+^m$ tel que $B_{i:}^\mathsf{T} = A^\mathsf{T} y$, avec $y \geqslant 0$ et $a^\mathsf{T} y \leqslant 0 = b_i$, comme souhaité.
  - Si $b_i < 0$, $z = e^i / (-b_i)$ est positif et vérifie $b^\mathsf{T} z = -1$. En prenant $\alpha = -1$ dans (17.24b), $B^\mathsf{T} z = B_{i:}^\mathsf{T} / (-b_i)$ peut s'écrire $A^\mathsf{T} \tilde{y}$ avec $\tilde{y} \geqslant 0$ et $a^\mathsf{T} \tilde{y} \leqslant -1$. Alors $y = -b_i \tilde{y}$ est positif et satisfait $B_{i:} = y^\mathsf{T} A$ et $y^\mathsf{T} a \leqslant b_i$. Ce y peut donc être pris comme la i-ième ligne de Y.

5) L'équivalence (17.25) se déduit directement de l'équivalence $(i) \Leftrightarrow (iv)$:

- pour $\alpha = 1$, $\{d \in \mathbb{R}^n : \sigma_{\mathscr{E}(H)} \leqslant 1\} = \mathscr{E}(H^{-1})$ par (17.23) et $\{d \in \mathbb{R}^n : \sigma_P \leqslant 1\} = A^\mathsf{T} \{y \in \mathbb{R}_+^m : b^\mathsf{T} y \leqslant 1\}$ par (17.22) ;

- pour $\alpha = 0$, $\{d \in \mathbb{R}^n : \sigma_{\mathscr{E}(H)} \leqslant 0\}$ est réduit à $\{0\}$, si bien que l'inclusion dans (iv) est trivialement vérifiée ;
- pour $\alpha = -1$, $\{d \in \mathbb{R}^n : \sigma_{\mathscr{E}(H)} \leqslant -1\}$ est vide, si bien que l'inclusion dans (iv) est trivialement vérifiée.

17.7. 1) D'après (17.6) et (17.8a), $\mathcal{S}_{\text{P}} = \{x \in \mathbb{R}^n : Ax = b,\ x_{\mathfrak{B}} \geqslant 0,\ x_{\mathfrak{N}} = 0\}$. Comme il existe une solution strictement complémentaire (proposition 17.8), l'ensemble $\{x \in \mathbb{R}^n : Ax = b,\ x_{\mathfrak{B}} > 0,\ x_{\mathfrak{N}} = 0\}$ n'est pas vide. Alors, selon le point 5 de l'exercice 2.20, l'intérieur relatif de $\mathcal{S}_{\text{P}}$ est ce dernier ensemble, qui est celui donné dans la deuxième formule de (17.9a). En en déduit facilement la première formule de (17.9a).

Les formules de $\mathcal{S}_{\text{D}}^{\diamond}$ dans (17.9b) se démontreent de la mième manière.

2) Il est clair que $\mathcal{S}_{\text{PD}}^{\diamond} := (\mathcal{S}_{\text{P}} \times \mathcal{S}_{\text{D}})^{\diamond} = \mathcal{S}_{\text{P}}^{\diamond} \times \mathcal{S}_{\text{D}}^{\diamond}$ par le point 1 de la proposition 2.17. Par ailleurs, $\mathcal{S}_{\text{PD}}^{\text{sc}} = \mathcal{S}_{\text{PD}}^{\diamond}$, parce que l'ensemble des solutions primales-duales strictement complémentaires sont les solutions primales-duales telles que $x_{\mathfrak{B}} > 0$ et $s_{\mathfrak{N}} > 0$.

3) [$\Rightarrow$] Soit $x \in \mathcal{S}_{\text{P}}^{\diamond}$. On prend $(y, s) \in \mathcal{S}_{\text{D}}^{\diamond}$ (qui est nécessairement non vide, car $\mathcal{S}_{\text{D}} \neq \varnothing$). Alors $(x, y, s) \in \mathcal{S}_{\text{P}}^{\diamond} \times \mathcal{S}_{\text{D}}^{\diamond} = \mathcal{S}_{\text{PD}}^{\text{sc}}$, par le point 2. Aussi par le point 2, si $(x, y, s) \in \mathcal{S}_{\text{PD}}^{\text{sc}}$, on a nécessairement $(y, s) \in \mathcal{S}_{\text{D}}^{\diamond}$.

[$\Leftarrow$] Si $(y, s) \in \mathcal{S}_{\text{D}}$ est tel que $(x, y, s) \in \mathcal{S}_{\text{PD}}^{\text{sc}}$, alors, par le point 2, on a nécessairement $x \in \mathcal{S}_{\text{P}}^{\diamond}$ et $(y, s) \in \mathcal{S}_{\text{D}}^{\diamond}$.

4) Se démontre comme le point 3.

17.8. 1) On considère le problème (P) et son dual (D) :

$$(P) \quad \begin{cases} \inf\ 0^{\mathsf{T}} x \\ Ax = b \\ x \geqslant 0, \end{cases} \qquad \text{et} \qquad (D) \quad \begin{cases} \sup\ b^{\mathsf{T}} y \\ A^{\mathsf{T}} y \leqslant 0. \end{cases}$$

L'inclusion $\{Ax : x \geqslant 0\} \subseteq \{y : A^{\mathsf{T}} y \geqslant 0\}^{+}$ est évidente (démonstration de la proposition 2.45), mais elle peut aussi se déduire de l'inégalité de dualité faible $\text{val}(D) \leqslant \text{val}(P)$. Soit $b \in \{Ax : x \geqslant 0\}$. Alors $\{Ax : x \geqslant 0\} \neq \varnothing$, si bien que $\text{val}(P) = 0$. Par dualité faible, $\text{val}(D) \leqslant 0$, ce qui se traduit par : $b^{\mathsf{T}} y \geqslant 0$ pour tout y tel que $A^{\mathsf{T}} y \geqslant 0$, c'est-à-dire $b \in \{y : A^{\mathsf{T}} y \geqslant 0\}^{+}$.

Réciproquement, si $b \in \{y : A^{\mathsf{T}} y \geqslant 0\}^{+}$, alors $\text{val}(D) \leqslant 0$. Dès lors (D) est borné et réalisable ($y = 0$ est admissible); il admet donc une solution (théorème 17.4). Par suite, (P) admet également une solution x (théorème 17.11). En particulier, ce x est admissible pour (P) : $x \geqslant 0$ et $Ax = b$. Donc $b \in \{Ax : x \geqslant 0\}$.

2) [$(i) \Rightarrow (ii)$] Par hypothèse, le problème d'optimisation linéaire $\gamma := \inf\{c^{\mathsf{T}} x - \alpha : Ax \geqslant b\}$ est réalisable et, par (i), $\gamma \geqslant 0$. Le problème a donc une solution (corollaire 17.12). Par dualité forte (proposition 17.11), le problème dual $\sup\{b^{\mathsf{T}} y - \alpha : A^{\mathsf{T}} y = c,\ y \geqslant 0\}$ a une solution et a pour valeur optimale γ. On en déduit (ii).

[$(ii) \Rightarrow (i)$] Par hypothèse, il existe un x tel que $Ax \geqslant b$. Par (ii), on a alors : $c^{\mathsf{T}} x = y^{\mathsf{T}} Ax \geqslant b^{\mathsf{T}} y = \alpha + \gamma \geqslant \alpha$, c'est-à-dire (i).

17.9. 1) Le dual de (D.3) s'écrit

$$\begin{cases} \sup_x\ 0 \\ Ax = a \\ Bx \leqslant b \\ Cx \leqslant c. \end{cases} \tag{D.52}$$

[$(i) \Rightarrow (ii)$] Si (i) a lieu, le x donné vérifie les contraintes de (D.52) et ce problème a ce point x comme solution et la valeur optimale val (D.52) $= 0$. Alors le problème

primal (D.3) a aussi une solution et une valeur optimale nulle. Soit $(\alpha,\beta,\gamma)\in\mathbb{R}^{m_A}\times\mathbb{R}_+^{m_B}\times\mathbb{R}_+^{m_C}$ vérifiant $A^\mathsf{T}\alpha+B^\mathsf{T}\beta+C^\mathsf{T}\gamma=0$ et $a^\mathsf{T}\alpha+b^\mathsf{T}\beta+c^\mathsf{T}\gamma\leqslant 0$ comme dans (ii). Alors (α,β,γ) est solution du problème primal (D.3) et on a $a^\mathsf{T}\alpha+b^\mathsf{T}\beta+c^\mathsf{T}\gamma=0$. Enfin, en retranchant $(A^\mathsf{T}\alpha+B^\mathsf{T}\beta+C^\mathsf{T}\gamma)^\mathsf{T}x=0$ et $a^\mathsf{T}\alpha+b^\mathsf{T}\beta+c^\mathsf{T}\gamma=0$, on obtient

$$\underbrace{(Ax-a)}_{=0}^\mathsf{T}\alpha+\underbrace{(Bx-b)}_{\leqslant 0}^\mathsf{T}\underbrace{\beta}_{\geqslant 0}+\underbrace{(Cx-c)}_{<0}^\mathsf{T}\underbrace{\gamma}_{\geqslant 0}=0.$$

On en déduit que $\gamma=0$. Pour ce dernier point, on aurait aussi pu utiliser la relation de complémentarité entre la contrainte $\gamma\geqslant 0$ et son multiplicateur $z:=c-Cx>0$ (qui dérive de la dualisation) :

$$0\leqslant(c-Cx)\perp z\geqslant 0.$$

$[(ii)\Rightarrow(i)]$ Le problème primal (D.3) est réalisable et sa valeur optimale est négative (en prenant $\alpha=0$, $\beta=0$ et $\gamma=0$). Si (ii) a lieu, alors $\mathrm{val}\,(\mathrm{D}.3)=0$ et le problème primal (D.3) a une solution. Dès lors, le problème dual (D.52) a aussi une solution (théorème 17.11). En prenant des solutions primales-duales strictement complémentaires (proposition 17.8), disons $(\alpha_0,\beta_0,\gamma_0)$ et x_0, on doit avoir $c-Cx_0>0$ (c'est la variable duale associée à la contrainte $\gamma\geqslant 0$), car $\gamma_0=0$ par (ii).

17.10. 1) Si (P_L) est réalisable et non borné, $\mathrm{val}(P_L)=-\infty$. Par l'inégalité de dualité faible, $\mathrm{val}(D_L)=-\infty$, ce qui signifie que (D_L) n'est pas réalisable.

2) Si (D_L) était borné, il aurait une solution (car il est aussi réalisable, théorème 17.4). Donc (P_L) aurait aussi une solution (dualité forte), ce qui contredirait qu'il n'est pas réalisable.

17.11. 1) L'inclusion $\supseteq$ est claire. Inversement, si $r\in\mathcal{S}_{\mathrm{D},s}^\infty$, il existe des suites $\{s_k\}\subseteq\mathcal{S}_{\mathrm{D},s}$ et $\{t_k\}\to\infty$, telles que $s_k/t_k\to r$. Clairement $r\geqslant 0$. D'autre part, il existe une suite $\{y_k\}\subseteq\mathbb{R}^m$ telle que $A^\mathsf{T}y_k+s_k=c$, $b^\mathsf{T}y_k\geqslant\mathrm{val}(D)$. On voit que $A^\mathsf{T}(y_k/t_k)=(c-s_k)/t_k\to -r$ et comme $\mathcal{R}(A^\mathsf{T})$ est fermé, il existe $z\in\mathbb{R}^m$ tel que $A^\mathsf{T}(y_k/t_k)\to A^\mathsf{T}z$. À la limite, on a $A^\mathsf{T}z+r=0$. Enfin, en tenant compte du fait que $b=Ax$ pour un certain $x\in\mathbb{R}^n$ $(\mathcal{F}_\mathrm{P}\neq\varnothing)$, on peut passer à la limite dans $b^\mathsf{T}(y_k/t_k)\geqslant\mathrm{val}(D)/t_k$. On obtient alors $b^\mathsf{T}z\geqslant 0$.

2) L'inclusion $\supseteq$ est claire : si $r\in\mathcal{S}_{\mathrm{D},s}^\infty$, $\alpha\in\mathbb{R}_+$, $h\in\mathcal{N}(A)$ et $p\geqslant 0$, on a $r^\mathsf{T}(-\alpha x_0+h+p)\geqslant -z^\mathsf{T}A(-\alpha x_0+h)=\alpha z^\mathsf{T}b\geqslant 0$. Inversement, si l'inclusion $\subseteq$ n'a pas lieu, il existerait un $d\in(\mathcal{S}_{\mathrm{D},s}^\infty)^+$ n'appartenant pas à $\mathbb{R}_-x_0+\mathcal{N}(A)+\mathbb{R}_+^n$. On sépare strictement le convexe compact $\{d\}$ du convexe fermé $\mathbb{R}_-x_0+\mathcal{N}(A)+\mathbb{R}_+^n$ (c'est un cône polyédrique) : il existe r tel que

$$r^\mathsf{T}d<\inf\{r^\mathsf{T}(-\alpha x_0+h+p):\alpha\in\mathbb{R}_+,\ h\in\mathcal{N}(A),\ p\geqslant 0\}.$$

On en déduit d'une part que $r^\mathsf{T}d<0$ (0 est dans le cône polyédrique) et d'autre part que $r\geqslant 0$ ($p\geqslant 0$ arbitraire), $r=-A^\mathsf{T}z$ (h arbitraire dans $\mathcal{N}(A)$) avec $b^\mathsf{T}z\geqslant 0$ $(0\geqslant r^\mathsf{T}x_0=-z^\mathsf{T}Ax_0=-z^\mathsf{T}b)$, c'est-à-dire $r\in\mathcal{S}_{\mathrm{D},s}^\infty$. On a obtenu une contradiction.

3) $[(i)]$ En effet, dans le cas contraire, pour tout i, il existerait $(y^i,s^i)\in\mathcal{F}_\mathrm{D}$ tel que $s_i^i>0$, ce qui impliquerait que le couple $(\frac{1}{n}\sum_i y^i,\frac{1}{n}\sum_i s^i)$ serait dans $\mathcal{F}_\mathrm{D}^s$.

$[(ii)]$ Par dualité forte, l'affirmation faite en (i) revient à montrer qu'il existe un indice $i\in[1:n]$ tel que le problème $\min\{c^\mathsf{T}d:Ad=0,\ d\geqslant e^i\}$ est réalisable et a une valeur optimale nulle. Enfin, ceci équivaut à dire que l'on a une direction d telle que $Ad=0$, $d\geqslant 0$, $d\neq 0$ et $c^\mathsf{T}d=0$, c'est-à-dire que $\mathcal{S}_\mathrm{P}$ n'est pas borné.

4) [(i)] En effet, dans le cas contraire, pour tout i, il existerait $x^i \in \mathcal{F}_{\text{P}}$ tel que $x^i_i > 0$, ce qui impliquerait que $\frac{1}{n}\sum_i x^i$ serait dans $\mathcal{F}^s_{\text{P}}$.

[(ii)] Par dualité, cela revient au même de dire qu'il existe un indice $i \in [1:n]$ tel que problème

$$\begin{cases} \max b^{\mathsf{T}} y \\ A^{\mathsf{T}} y + s = -e^i \\ s \geqslant 0 \end{cases}$$

est réalisable et a une valeur optimale nulle. Enfin, ceci équivaut à dire que l'on a une direction $(\bar{y}, \bar{s})$ telle que $b^{\mathsf{T}}\bar{y} = 0$, $A^{\mathsf{T}}\bar{y} + \bar{s} = 0$, $\bar{s} \geqslant 0$ et $\bar{s} \neq 0$.

Il reste à montrer que cette dernière condition exprime que $\mathcal{S}_{\text{D},s}$ n'est pas borné. C'est clairement une condition suffisante; montrons qu'elle est nécessaire. Si $\mathcal{S}_{\text{D},s}$ n'est pas borné, il existe une suite $\{s_k\}$ et une suite $\{y_k\}$, telles que $\|s_k\| \to \infty$, $b^{\mathsf{T}}y_k$ est constante, $A^{\mathsf{T}}y_k + s_k = c$ et $s_k \geqslant 0$. On peut supposer que $s_k/\|s_k\| \to \bar{s} \neq 0$. Évidemment $\bar{s} \geqslant 0$. Comme $\mathcal{R}(A^{\mathsf{T}})$ est fermé, $\{A^{\mathsf{T}}y_k/\|s_k\|\}$, qui est bornée, a une sous-suite qui converge vers un élément de la forme $A^{\mathsf{T}}\bar{y}$; on en déduit que $A^{\mathsf{T}}\bar{y}+\bar{s} = 0$. En multipliant $A^{\mathsf{T}}y_k/\|s_k\| \to A^{\mathsf{T}}\bar{y}$ par un $x \in \mathcal{F}_{\text{P}}$, on voit que $b^{\mathsf{T}}y_k/\|s_k\| \to b^{\mathsf{T}}\bar{y}$; donc $b^{\mathsf{T}}\bar{y} = 0$.

17.12. 1) Justifions l'utilisation des opérateurs dans le membre de gauche de (17.26) (la justification est similaire pour ceux du membre de droite). Le problème $\varphi(y) := \inf\{y^{\mathsf{T}}Ax : y \in \Delta_n\}$ a une solution car le coût est continu (linéaire) et le simplexe unitaire est non vide et compact (théorème 1.2). Par ailleurs, par la proposition 3.34, la fonction φ est concave comme enveloppe inférieure de fonctions concaves (même affines) ; comme φ prend des valeurs finies, elle est continue sur Δ_m (proposition 3.14), si bien que le supremum en $y \in \Delta_m$ de φ est atteint (théorème 1.2).

2) Soit $\overline{v} := \min\{v_i : 1 \leqslant i \leqslant n\}$. Si $x \in \Delta_n$, on a $v^{\mathsf{T}}x \geqslant (\overline{v}e)^{\mathsf{T}}x = \overline{v}$. Par ailleurs si $v_i = \overline{v}$, on a $e^i \in \Delta_n$ et $v^{\mathsf{T}}e^i = \overline{v}$. On raisonne de la même manière pour la seconde identité.

3) Soit E la matrice $m \times n$ dont toutes les composantes valent 1. Comme $y^{\mathsf{T}}Ex = 1$ pour $x \in \Delta_n$ et $y \in \Delta_m$, l'ajout de tE à A augmente α et β de la même quantité t. On peut donc ne s'intéresser qu'au cas où $\alpha > 0$ et $\beta > 0$.

4) Soit t un réel strictement positif. On a les équivalences suivantes :

$$\begin{aligned} & t \leqslant \alpha \\ \iff\ & \exists\, y \in \Delta_m : \min\{y^{\mathsf{T}}Ax : x \in \Delta_n\} \geqslant t && \text{[définitions de } \alpha \text{ et } t\text{]} \\ \iff\ & \exists\, y \in \mathbb{R}^m : y \geqslant 0,\ e^{\mathsf{T}}y = 1,\ A^{\mathsf{T}}y \geqslant te && \text{[point 2]} \\ \iff\ & \exists\, y \in \mathbb{R}^m : y \geqslant 0,\ e^{\mathsf{T}}y = 1/t,\ A^{\mathsf{T}}y \geqslant e && [y \curvearrowright ty \text{ et } t > 0] \\ \iff\ & \inf\{e^{\mathsf{T}}y : y \geqslant 0,\ A^{\mathsf{T}}y \geqslant e\} \leqslant 1/t. \end{aligned}$$

Dès lors $\alpha^{-1} = \min\{e^{\mathsf{T}}y : y \geqslant 0,\ A^{\mathsf{T}}y \geqslant e\}$, où l'opérateur « min » est justifié par le fait que le problème d'optimisation est linéaire et a une valeur optimale finie, voir le théorème 17.4. De la même manière, on a les équivalences suivantes :

$$\begin{aligned} & \beta \leqslant t \\ \iff\ & \exists\, x \in \Delta_n : \max\{y^{\mathsf{T}}Ax : y \in \Delta_m\} \leqslant t && \text{[définitions de } \beta \text{ et } t\text{]} \\ \iff\ & \exists\, x \in \mathbb{R}^n : x \geqslant 0,\ e^{\mathsf{T}}x = 1,\ Ax \leqslant te && \text{[point 2]} \\ \iff\ & \exists\, x \in \mathbb{R}^n : x \geqslant 0,\ e^{\mathsf{T}}x = 1/t,\ Ax \leqslant e && [x \curvearrowright tx \text{ et } t > 0] \\ \iff\ & \sup\{e^{\mathsf{T}}x : x \geqslant 0,\ Ax \leqslant e\} \geqslant 1/t. \end{aligned}$$

Dès lors $\beta^{-1} = \max\{e^{\mathsf{T}}x : x \geqslant 0,\ Ax \leqslant e\}$.

5) On conclut par dualité forte (il n'y a pas de saut de dualité, car les problèmes linéaires ont une solution) :

$$\begin{aligned}
\frac{1}{\alpha} &= \min_{\substack{y\geqslant 0\\ A^{\mathsf T}y\geqslant e}} e^{\mathsf T}y \\
&= \min_{y\geqslant 0}\sup_{x\geqslant 0} e^{\mathsf T}y + x^{\mathsf T}(e-A^{\mathsf T}y) \\
&= \max_{x\geqslant 0}\inf_{y\geqslant 0} e^{\mathsf T}x + y^{\mathsf T}(e-Ax) \qquad \text{[pas de saut de dualité]} \\
&= \max_{\substack{x\geqslant 0\\ Ax\leqslant e}} e^{\mathsf T}x \\
&= \frac{1}{\beta}.
\end{aligned}$$

17.13. Le sommet courant $\hat{x}$ vérifie la condition d'optimalité $A^{\mathsf T}y + s = c$, pour un certain $y \in \mathbb{R}^m$ et pour $s \in \mathbb{R}^n$ vérifiant $s_B = 0$ et $s_N = r$ (voir la démonstration de la proposition 17.14). Soit x un point admissible différent de $\hat{x}$. Alors x_N est positif (car admissible) et non nul (car $x \neq \hat{x}$). Dès lors, $c^{\mathsf T}x = y^{\mathsf T}Ax + r^{\mathsf T}x_N > y^{\mathsf T}A\hat{x}$ $[Ax = A\hat{x} = b$ et $r > 0] = c^{\mathsf T}\hat{x}$ $[s^{\mathsf T}\hat{x} = 0]$, si bien que x ne peut être solution de (P_L).

17.14. $[(i)]$ On définit pour $\varepsilon > 0$:

$$A_\varepsilon := \begin{pmatrix} \varepsilon & -1 \\ \varepsilon & 1 \end{pmatrix}, \qquad P_\varepsilon := \{z \in \mathbb{R}^2 : A_\varepsilon z \leqslant 0\} \qquad \text{et} \qquad x := (1,0).$$

Clairement, la projection de x sur P_ε est le point $(0,0)$, si bien que la distance de x à P_ε en norme ℓ_2 vaut 1. Par ailleurs, $\|(A_\varepsilon x)^+\|_2 = \varepsilon\|(1,1)\|_2 \downarrow 0$, si bien que $h \to \infty$ quand $\varepsilon \downarrow 0$. Voir la figure D.3.

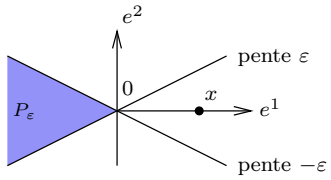


Fig. D.3. Illustration d'une constante de Hoffman tendant vers $+\infty$ (exercice 17.14)

$[(ii)]$ (a) Comme la norme biduale est égale à la norme originale (exercice 3.30), on a $\|x-z\| = \|x-z\|_{\text{DD}} = \sup\{(x-z)^{\mathsf T}u : \|u\|_{\text{D}} \leqslant 1\}$, si bien que, par la dualité faible (proposition 14.2), on peut écrire

$$\begin{aligned}
\mathrm{d}_P(x) &= \inf_{z\in P}\sup_{\|u\|_{\text{D}}\leqslant 1} (x-z)^{\mathsf T}u \\
&\geqslant \sup_{\|u\|_{\text{D}}\leqslant 1}\inf_{z\in P} (x-z)^{\mathsf T}u \qquad \text{[dualité faible]} \\
&= \sup_{\|u\|_{\text{D}}\leqslant 1}\left(x^{\mathsf T}u - \sup_{Az\leqslant b} z^{\mathsf T}u\right). \qquad \text{(D.53)}
\end{aligned}$$

Puis, en prenant $u \in \bar{B}_D$ de la forme $u = A^{\mathsf T}y$ avec $y \geqslant 0$ et $\|A^{\mathsf T}y\|_{\text{D}} \leqslant 1$, l'inégalité (D.53) devient

$$\mathrm{d}_P(x) \geqslant \sup_{\substack{y\geqslant 0\\ \|A^{\mathsf{T}}y\|_{\mathrm{D}}\leqslant 1}} \left(x^{\mathsf{T}}A^{\mathsf{T}}y - \sup_{Az\leqslant b} z^{\mathsf{T}}A^{\mathsf{T}}y\right) \geqslant \sup_{\substack{y\geqslant 0\\ \|A^{\mathsf{T}}y\|_{\mathrm{D}}\leqslant 1}} (Ax-b)^{\mathsf{T}}y, \tag{D.54}$$

car pour $Az \leqslant b$ et $y \geqslant 0$, $z^{\mathsf{T}}A^{\mathsf{T}}y \leqslant b^{\mathsf{T}}y$.

On montre maintenant une inégalité inverse en séparant les ensembles convexes non vides disjoints $C_1 := \{v \in \mathbb{R}^n : \|v\| < \mathrm{d}_P(x)\}$ et $C_2 := \{v \in \mathbb{R}^n : Av \geqslant Ax - b\}$. Ils sont disjoints car si $v \in C_2$, alors $A(x-v) \leqslant b$, donc $x - v \in P$, si bien que $\mathrm{d}_P(x) \leqslant \|(x-v)-x\| = \|v\|$, ce qui montre que $v \notin C_1$. Par la proposition 2.38, on peut séparer (non strictement) C_1 et C_2 :

$$\exists\, u \in \mathbb{R}^n : \quad \|u\|_{\mathrm{D}} = 1 \quad \text{et} \quad \sup_{v\in C_1} u^{\mathsf{T}}v \leqslant \inf_{v\in C_2} u^{\mathsf{T}}v. \tag{D.55}$$

Le membre de gauche de l'inégalité dans (D.55) s'écrit

$$\sup_{v\in C_1} u^{\mathsf{T}}v = \sup_{\|v\|<\mathrm{d}_P(x)} u^{\mathsf{T}}v = \sup_{\|v\|\leqslant \mathrm{d}_P(x)} u^{\mathsf{T}}v = \mathrm{d}_P(x)\,\|u\|_{\mathrm{D}} = \mathrm{d}_P(x). \tag{D.56}$$

Le membre de droite de l'inégalité dans (D.55) s'écrit

$$\begin{aligned}
\inf_{v\in C_2} u^{\mathsf{T}}v &= \inf_{v:\ Av\geqslant(Ax-b)} u^{\mathsf{T}}v && \text{(D.57)}\\
&= \inf_{v}\sup_{y\geqslant 0}\left(u^{\mathsf{T}}v + y^{\mathsf{T}}\,(A(x-v)-b)\right)\\
&= \sup_{y\geqslant 0}\inf_{v}\left(u^{\mathsf{T}}v + y^{\mathsf{T}}\,(A(x-v)-b)\right) && \text{(D.58)}\\
&= \sup_{\substack{y\geqslant 0\\ A^{\mathsf{T}}y=u}} (Ax-b)^{\mathsf{T}}y. && \text{(D.59)}
\end{aligned}$$

L'égalité en (D.58) est justifiée par le fait que le problème d'optimisation linéaire (D.57) a une solution (il est réalisable car $P \neq \varnothing$ et, grâce à (D.55) et (D.56), sa valeur optimale est bornée par $\mathrm{d}_P(x)$) et donc il n'y a pas de saut de dualité (proposition 17.11). Alors (D.55), (D.56) et (D.59) conduisent à (D.60).

$$\exists\, u \in \mathbb{R}^n : \quad \|u\|_{\mathrm{D}} = 1 \quad \text{et} \quad \mathrm{d}_P(x) \leqslant \sup_{\substack{y\geqslant 0\\ A^{\mathsf{T}}y=u}} (Ax-b)^{\mathsf{T}}y. \tag{D.60}$$

En combinant (D.60) et (D.54), on obtient

$$\mathrm{d}_P(x) \leqslant \sup_{\substack{y\geqslant 0\\ A^{\mathsf{T}}y=u}} (Ax-b)^{\mathsf{T}}y \leqslant \sup_{\substack{y\geqslant 0\\ \|A^{\mathsf{T}}y\|_{\mathrm{D}}\leqslant 1}} (Ax-b)^{\mathsf{T}}y \leqslant \mathrm{d}_P(x).$$

On a donc égalité partout et donc (17.28).

(*b*) L'affirmation du point (*a*) implique que le problème d'optimisation linéaire dans (17.28) a une solution (théorème 17.4). A priori, celle-ci dépend de A, b et x. Cependant, on peut montrer que l'on peut *choisir* une solution $\bar{y}$ dont la norme est majorée par une constante ne dépendant que de A. Observons d'abord que l'ensemble $\mathscr{B}$ des sous-matrices B de A obtenues en sélectionnant des lignes linéairement indépendantes de A est fini (il y a au plus 2^m matrices de ce type), si bien que

$$h := \sup_{\substack{B\in\mathscr{B}\\ \|v\|_{\mathrm{D}}=1}} \|(BB^{\mathsf{T}})^{-1}Bv\|_2$$

est un nombre fini. Comme le problème d'optimisation linéaire dans (17.28) a une solution, il en a une $\bar{y}$ sur un sommet du polyèdre convexe $\{y \in \mathbb{R}^m_+ : A^\mathsf{T} y = u\}$ (proposition 17.5). Dès lors, les colonnes i de A^T telles que $\bar{y}_i > 0$ sont linéairement indépendantes (proposition 2.21). Donc, il existe une matrice surjective $B \in \mathscr{B}$ telle que $B^\mathsf{T}\bar{y}_I = u$ et $\bar{y}_{I^c} = 0$. On en déduit que $\bar{y}_I = (BB^\mathsf{T})^{-1}Bu$, $\bar{y}_{I^c} = 0$ et $\|\bar{y}\|_2 = \|\bar{y}_I\|_2 = \|(BB^\mathsf{T})^{-1}Bu\|_2 \leqslant h$, car $\|u\|_\mathrm{D} = 1$.

(c) Alors en utilisant la positivité de $\bar{y}$ et l'inégalité de Cauchy-Schwarz généralisée sur (17.28), on obtient $\mathrm{d}_P(x) \leqslant \bar{y}^\mathsf{T}(Ax-b) \leqslant \bar{y}^\mathsf{T}(Ax-b)^+ \leqslant \|\bar{y}\|_2\,\|(Ax-b)^+\|_2 \leqslant h\,\|(Ax-b)^+\|_2$, ce qui démontre (17.27).

Optimisation linéaire : algorithmes de points intérieurs

18.1. [(i)] On écrit le problème (D_μ) sous la forme $\inf \psi_\mu(y,s)$, où

$$\psi(y,s) := -b^\mathsf{T} y - \mu \sum_{i=1}^n \log s_i + \mathcal{I}_{\{(y,s):A^\mathsf{T}y+s=c\}}(y,s).$$

On vérifie aisément que $\psi_\mu^\infty(dy,ds) = -b^\mathsf{T}dy$ si $ds \geqslant 0$ et $A^\mathsf{T}dy + ds = 0$ et $\psi_\mu^\infty(dy,ds) = +\infty$ sinon. Dès lors $\psi_\mu^\infty(dy,ds) \leqslant 0$ implique que $b^\mathsf{T}dy \geqslant 0$, $ds \geqslant 0$ et $A^\mathsf{T}dy + ds = 0$, ce qui implique que $(dy,ds) = 0$ (on utilise la proposition 17.13 et le fait que $\mathcal{S}_\mathrm{D}$ est borné lorsque $\mathcal{F}_\mathrm{P}^s \neq \varnothing$ et A est surjective). Dès lors, (D_μ) a un ensemble non vide et borné de solutions. La solution est en fait unique : unique en s du fait de la stricte convexité du critère en s (et sa linéarité en y), puis unique en y du fait de la contrainte et de la surjectivité de A.

[(ii)] En écrivant les conditions d'optimalité de (D_μ), sachant qu'il existe un multiplicateur optimal x parce que la contrainte est affine, on voit que $(x,y,s) = (x_\mu, y_\mu, s_\mu)$.

18.2. On considère le problème d'optimisation

$$\begin{cases} \inf\ c^\mathsf{T}x - \sum_{i=1}^n w_i \log x_i \\ Ax = b \quad (x > 0). \end{cases}$$

Par hypothèse son ensemble admissible est non vide. Comme son critère est strictement convexe, il a au plus une solution. De plus, par convexité du problème, celui-ci a une solution x ssi ses conditions d'optimalité sont vérifiées. Si on introduit le lagrangien $\ell(x,y) = c^\mathsf{T}x - \sum_{i=1}^n w_i \log x_i - y^\mathsf{T}(Ax-b)$, les conditions d'optimalité s'écrivent $c - X^{-1}w - A^\mathsf{T}y = 0$ et $Ax = b$. En posant $s = X^{-1}w$, on trouve le système de l'énoncé. On est donc ramené à montrer que le problème ci-dessus a au moins une solution. On introduit

$$\psi(x) = c^\mathsf{T}x - \sum_{i=1}^n w_i \log x_i + \mathcal{I}_{\{x\in\mathbb{R}^n : Ax=b\}}(x).$$

Cette fonction est dans $\mathrm{C}\overline{\mathrm{onv}}(\mathbb{R}^n)$ et $\psi^\infty(d)$ s'écrit comme dans la démonstration de la proposition 18.1. Pour la même raison, $\psi^\infty(d) > 0$ pour tout $d \neq 0$ et donc ψ a un minimum.

18.3. [(i)] Les conditions d'optimalité de (P) donnent $\bar{s} = 0$ (complémentarité), donc $c = A^\mathsf{T}\bar{y}$, c'est-à-dire $c \perp \mathcal{N}(A)$.

[(ii)] Si x est admissible, alors le triplet $(x,\bar{y},0)$ vérifie les conditions d'optimalité, donc x est solution.

[(*iii*)] Le point central primal x_μ vérifie $Ax_\mu = b$, $x_\mu > 0$ et $X_\mu^{-1}e = \frac{1}{\mu}s_\mu = A^\mathsf{T}(y - v) \in \mathcal{R}(A^\mathsf{T})$, où v sont les « composantes » de $c \in \mathcal{R}(A^\mathsf{T})$. Ces conditions caractérisent le centre analytique de $\mathcal{F}_\text{P}$.

[(*iv*)] Les conditions d'optimalité de (P) donnent $\bar{x} = 0$ (complémentarité), donc $b = A\bar{x} = 0$.

[(*v*)] Si $(y, s) \in \mathcal{F}_\text{D}$, alors le triplet $(\bar{x}, y, s)$ vérifie les conditions d'optimalité, donc (y, s) est solution duale.

[(*vi*)] Le point central dual (y_μ, s_μ) vérifie $A^\mathsf{T}y_\mu + s_\mu = c$, $s_\mu > 0$ et $S_\mu^{-1}e = \frac{1}{\mu}x_\mu \in \mathcal{N}(A)$ (car $Ax_\mu = 0$). Ces conditions caractérisent le centre analytique de $\mathcal{F}_\text{D}$.

18.5. On peut récrire $X_\alpha = \{x \in \mathcal{F}_\text{P} : c^\mathsf{T}x + \mu = \alpha,\ \mu \geqslant 0\}$, dont le centre analytique est solution de $\max\{\sum_{i=1}^n \log x_i + \log\mu : x \in \mathcal{F}_\text{P}^s,\ c^\mathsf{T}x + \mu = \alpha,\ \mu > 0\}$. Ce problème a une solution car son ensemble admissible est non vide et borné (on utilise le fait que $X_\alpha^\infty = \mathcal{S}_\text{P}^\infty$ et que $\mathcal{S}_\text{P}^\infty = \{0\}$ par (18.2)). Ses conditions d'optimalité stipulent l'existence de multiplicateurs $y_0 \in \mathbb{R}^m$ et $\beta \in \mathbb{R}$ (les contraintes sont linéaires donc qualifiées), tels que

$$\begin{cases} -X^{-1}e - A^\mathsf{T}y_0 + \beta c = 0 \\ -\frac{1}{\mu} + \beta = 0. \end{cases}$$

En posant $y = \mu y_0$ et $s = \mu X^{-1}e$, on retrouve les équations qui définissent univoquement le triplet μ-centré (x, y, s).

18.6. La dérivée (dx, dy, ds) vérifie

$$\begin{cases} A^\mathsf{T}dy + ds = 0 \\ Adx = 0 \\ Sdx + Xds = e. \end{cases}$$

D'après la deuxième équation, dx est de la forme $dx = Z^- du$, pour un certain $du \in \mathbb{R}^{n-m}$. On observe ensuite que $Z^{-\mathsf{T}}ds = 0$ et que $X^{-1}e = (1/\mu)s$. La dernière équation ci-dessus permet alors de déterminer du puis

$$dx = \frac{1}{\mu} Z^- (Z^{-\mathsf{T}}X^{-1}SZ^-)^{-1}Z^{-\mathsf{T}}s$$

et enfin $c^\mathsf{T}dx$ (on note que $A^\mathsf{T}y + s = c$ implique que $Z^{-\mathsf{T}}s = Z^{-\mathsf{T}}c$). On s'y prend de la même manière pour $b^\mathsf{T}dy$: les deux dernières équations donnent $AS^{-1}Xds = AS^{-1}e$; la première donne alors $AS^{-1}XA^\mathsf{T}dy = -AS^{-1}e = -b/\mu$ qui conduit à $b^\mathsf{T}dy$. Pour la dernière relation, on multiplie scalairement les deux membres de la dernière équation ci-dessus par e.

18.7. 1) Comme le chemin central, qui existe sous l'hypothèse (18.2), est un chemin continu (proposition 18.4), il suffit de montrer qu'un point quelconque $z^1 \in V_p(\theta)$ [resp. $z^1 \in V_{-\infty}(\theta)$] peut être joint à un point central par un chemin continu contenu dans $V_p(\theta)$ [resp. $V_{-\infty}(\theta)$]. On note $\bar{\mu}^1 := \bar{\mu}(z^1)$ et $w^1 := X^1 s^1 \in \mathbb{R}^n_{++}$ (car $\theta < 1$). Pour $t \in [0, 1]$, on considère la solution unique $z^t := (x^t, y^t, s^t)$ de (voir l'exercice 18.2)

$$\begin{cases} A^\mathsf{T}y^t + s^t = c, \quad s^t > 0 \\ Ax^t = b, \quad x^t > 0 \\ (X^t)^\mathsf{T}s^t = (1-t)\bar{\mu}^1 e + tw^1. \end{cases}$$

Clairement z^0 est sur le chemin central et la solution de ce système pour $t = 1$ n'est autre que z^1 (unicité de la solution). D'autre part, $z^t \in V_p(\theta)$ [resp. $z^t \in V_{-\infty}(\theta)$],

car $\bar\mu(z^t) = (1-t)\bar\mu^1 + t\bar\mu^1 = \bar\mu^1$ et $\|(X^t)^\mathsf{T} s^t - \bar\mu(z^t)e\|_p = \|(X^t)^\mathsf{T} s^t - \bar\mu^1 e\|_p = t\|w^1 - \bar\mu^1 e\|_p \leqslant t\theta\bar\mu^1 \leqslant \theta\bar\mu(z^t)$ [resp. $X^t s^t = (1-t)\bar\mu^1 e + tw^1 \geqslant (1-t)\bar\mu^1 e + t(1-\theta)\bar\mu^1 e \geqslant (1-\theta)\bar\mu^1 e$]. Il reste donc à montrer que $t \in [0,1] \mapsto z^t$ est continue. Cela résulte du théorème des fonctions implicites (théorème C.14) appliqué au système ci-dessus (la jacobienne est inversible, car $x^t > 0$ et $s^t > 0$, voir la proposition 18.4).

2) Soit $z^1 \in V_p(\theta)$ [resp. $z^t \in V_{-\infty}(\theta)$]. Comme $\|\cdot\|_\infty \leqslant \|\cdot\|_p$, on a $|x_i^1 s_i^1 - \bar\mu(z^1)| \leqslant \theta\bar\mu(z^1)$ pour tout i et donc $x_i^1 s_i^1 \geqslant (1-\theta)\bar\mu(z^1) > 0$ pour tout i. Il reste à montrer que $x^1 > 0$. Cela résulte de la connexité de $V_p(\theta)$ [resp. $V_{-\infty}(\theta)$] et du fait que l'on peut relier z^1 à un point central $z^0 = (x^0, y^0, s^0)$ par un chemin continu restant dans $V_p(\theta)$ [resp. $V_{-\infty}(\theta)$]. Comme $x^0 > 0$ et que l'on vient de montrer que $Xs > 0$ le long du chemin, les composantes de x ne peuvent s'annuler le long du chemin et donc $x^1 > 0$.

3) Soient $n \geqslant 2$ et $\varepsilon \in]0, n/(n-1)[$. D'après l'exercice 18.2, on peut trouver $z \in \mathcal{F}^s$ tel que $x_1 s_1 = n - (n-1)\varepsilon > 0$ et $x_i s_i = \varepsilon > 0$ pour $i \geqslant 2$. On a $\bar\mu(z) = 1$ et $\|Xs - \bar\mu(z)e\|_p^p = ((n-1)^{p-1} + 1)(n-1)|1-\varepsilon|^p$ qui est > 1 pour $\varepsilon > 0$ assez petit; donc $z \in \mathcal{F}^s \setminus V_p(\theta)$.

Supposons à présent que $n \geqslant 3$. Pour le même point $z \in \mathcal{F}^s$ que ci-dessus, qui vérifie $\bar\mu(z) = 1$, on a pour $\varepsilon > 0$ petit : $\|Xs - \bar\mu(z)e\|_\infty = (n-1)(1-\varepsilon) > 1$. Donc $z \in \mathcal{F}^s \setminus V_\infty(1)$.

Enfin, si $z \in \mathcal{F}^s$, on a $Xs > 0$ et donc on peut trouver un $\theta \in [0,1[$ tel que $Xs \geqslant (1-\theta)\bar\mu(z)e$. Dès lors, $z \in V_{-\infty}(\theta)$.

Optimisation semi-définie

20.1. [$(i) \Rightarrow (ii)$] On peut supposer que $n \geqslant 1$ et que F n'est pas la face triviale $\{0\}$. Soit $M \in F^\oplus$. Alors $M \neq 0$ (sinon $F = \{0\}$) et on peut écrire $M = \sum_{i=1}^r u_i u_i^\mathsf{T}$. On note $E := \operatorname{vect}\{u_1, \ldots, u_r\}$ et $F_E := \operatorname{cone}\{vv^\mathsf{T} : v \in E\}$. Montrons que $F = F_E$. [$\subseteq$] Soit $A \in F$. Il suffit de montrer que $\mathcal{R}(A) \subseteq E$, car alors $A = \sum_i v_i v_i^\mathsf{T}$ avec des $v_i \in E$ et donc $A \in F_E$. Si ce n'est pas le cas, il existe un $v \in E^\perp$ tel que $v^\mathsf{T} A v > 0$ (sinon $E^\perp \subseteq \mathcal{N}(A)$ ou $\mathcal{R}(A) \subseteq E$). Alors, quel que soit $t > 1$, $A_t := (1-t)A + tM \notin \mathcal{S}_+^n$ (car $v^\mathsf{T} A_t v = (1-t)v^\mathsf{T} A v < 0$ puisque $v^\mathsf{T} M v = 0$), ce qui contredit le fait que $A \in F$ (puisque $A \in F$ et $M \in F^\oplus$ impliquent que $A_t \in F$ pour $t > 1$ petit). [$\supseteq$] Soit $A \in F_E$. Alors $\mathcal{R}(A) \subseteq E = \mathcal{R}(M)$, avec M et $A \in \mathcal{S}_+^n$. Dès lors, il existe un $t > 1$ tel que $A_t := (1-t)A + tM \in \mathcal{S}_+^n$ (point 2 de l'exercice B.13). On a donc écrit M comme combinaison convexe stricte de A et $A_t \in \mathcal{S}_+^n$. Comme M est dans la face F, il en est de même de A.

On a $\operatorname{aff} F = \operatorname{aff}\operatorname{cone}\{vv^\mathsf{T} : v \in E\} = \operatorname{vect}\{vv^\mathsf{T} : v \in E\} =: \mathcal{A}_F$ (point 4 de l'exercice 2.2). Donc $\dim F = \dim \mathcal{A}_F$. Soit $\{\bar u_1, \ldots, \bar u_d\}$ une base orthonormale de E et montrons que les $\{\bar u_i \bar u_j^\mathsf{T} + \bar u_j \bar u_i^\mathsf{T} : 1 \leqslant i \leqslant j \leqslant d\}$ forment une base de $\mathcal{A}_F$, ce qui démontrera l'affirmation sur la dimension de F. *La famille est libre* : si $\sum_{i \leqslant j} \alpha_{ij}(\bar u_i \bar u_j^\mathsf{T} + \bar u_j \bar u_i^\mathsf{T}) = 0$, alors en multipliant à gauche par $\bar u_k^\mathsf{T}$ et à droite par $\bar u_l$, $k \leqslant l$, on trouve que $\alpha_{kl} = 0$. *La famille est génératrice* : il suffit de montrer que si $v \in E$, alors vv^T s'écrit comme combinaison linéaire des éléments de la famille ; on a dans ce cas $v = \sum_i \alpha_i \bar u_i$, donc $vv^\mathsf{T} = \sum_{i=j} \alpha_i^2 \bar u_i \bar u_i^\mathsf{T} + \sum_{i<j} \alpha_i \alpha_j (\bar u_i \bar u_j^\mathsf{T} + \bar u_j \bar u_i^\mathsf{T})$.

[$(ii) \Rightarrow (iii)$] Un élément A de $F = \operatorname{cone}\{vv^\mathsf{T} : v \in E\}$ est de la forme $A = \sum_{i=1}^k v_i v_i^\mathsf{T}$ avec $k \in \mathbb{N}^*$ et des $v_i \in E$. Clairement $A \in \mathcal{S}_+^n$ et $\mathcal{R}(A) \subseteq E$. Inversement, $A \in \mathcal{S}_+^n$ et $\mathcal{R}(A) \subseteq E$, la factorisation spectrale de A s'écrit $A = \sum_{i=1}^r \lambda_i \tilde v_i \tilde v_i^\mathsf{T}$, avec $r = \operatorname{rg}(A)$, des $\lambda_i > 0$ et des $\tilde v_i \in E$. En posant $v_i := \sqrt{\lambda_i}\tilde v_i \in E$, on obtient $A = \sum_{i=1}^r v_i v_i^\mathsf{T} \in$

$\mathrm{cone}\{vv^\mathsf{T} : v \in E\}$. On vérifie aisément que $\mathrm{vect}\{vv^\mathsf{T} : v \in E\} = \{A \in \mathcal{S}^n : \mathcal{R}(A) \subseteq E\}$, ce qui conduit à la formule de $\mathrm{aff}\, F$ proposée.

$[(iii) \Rightarrow (iv)]$ Comme $\mathcal{R}(A) \subseteq E$ si, et seulement si, $E^\perp \subseteq \mathcal{N}(A)$, on a $F = \{A \in \mathcal{S}^n_+ : E^\perp \subseteq \mathcal{N}(A)\}$. Par ailleurs, on a bien $\dim E^\perp = n - \dim E$. Trivialement, $\{A \in \mathcal{S}^n : \mathcal{R}(A) \subseteq E\} = \{A \in \mathcal{S}^n : E^\perp \subseteq \mathcal{N}(A)\}$, ce qui, après le changement de notation $E^\perp \to E$, conduit à la formule de $\mathrm{aff}\, F$ proposée.

$[(iv) \Rightarrow (i)]$ L'ensemble F est clairement un cône. Pour montrer que F est convexe, on prend A et $B \in F$ et on observe que $E \subseteq \mathcal{N}(A) \cap \mathcal{N}(B) = \mathcal{N}(A+B)$ (point 1 de l'exercice B.13), si bien que $A + B \in F$. Pour montrer que F est une face de $\mathcal{S}^n_+$, on prend A et $B \in \mathcal{S}^n_+$ tels que $\frac{1}{2}(A+B) \in F$; alors $E \subseteq \mathcal{N}(A+B) = \mathcal{N}(A) \cap \mathcal{N}(B)$; donc A et $B \in F$.

1) On peut écrire $F = \{A \in \mathcal{S}^n_+ : \mathcal{R}(A) \subseteq E\}$, où E est un sous-espace vectoriel de $\mathbb{R}^n$ (point (iii)). Soit $\{w_1, \ldots, w_p\}$ une base de $E^\perp$ et $S = \sum_{i=1}^p w_i w_i^\mathsf{T}$. Montrons que l'on a $F = \arg\min\{\langle S, A\rangle : A \in \mathcal{S}^n_+\}$. $[\subseteq]$ Si $A \in F$, on a $\mathcal{R}(A) \subseteq E$ et donc $\langle S, A\rangle = \sum_i w_i^\mathsf{T} A w_i = 0$, car $w_i \in E^\perp \subseteq \mathcal{N}(A)$. Par ailleurs, pour tout $A \in \mathcal{S}^n_+$, on a $\langle S, A\rangle = \sum_i w_i^\mathsf{T} A w_i \geqslant 0$. $[\supseteq]$ Clairement, $\min\{\langle S, A\rangle : A \in \mathcal{S}^n_+\} = 0$. Donc si $A \in \arg\min\{\langle S, A\rangle : A \in \mathcal{S}^n_+\}$, on a $\langle S, A\rangle = 0$, si bien que $w_i^\mathsf{T} A w_i = 0$ ou $w_i \in \mathcal{N}(A)$ pour tout $i \in [1\,{:}\,p]$. On a donc montré que $E^\perp \subseteq \mathcal{N}(A)$ ou $\mathcal{R}(A) \subseteq E$, ce qui montre que $A \in F$.

2) Les deux expressions de $F(A)$ sont équivalentes, car $\mathcal{R}(B) \subseteq \mathcal{R}(A)$ si, et seulement si, $\mathcal{N}(A) \subseteq \mathcal{N}(B)$. Montrons que $F(A) = F := \{B \in \mathcal{S}^n_+ : \mathcal{R}(B) \subseteq \mathcal{R}(A)\}$. D'après le point (iii), F est une face de $\mathcal{S}^n_+$. Il est aussi clair que $A \in F$. Il reste à montrer que F est la plus petite face de $\mathcal{S}^n_+$ contenant A. Soit donc F' une face de $\mathcal{S}^n_+$ contenant A et montrons que $F \subseteq F'$. Pour cela, il suffit de montrer que $B \in F'$, lorsque $B \in F$. Si $B \in F$, alors $\mathcal{R}(B) \subseteq \mathcal{R}(A)$. Par le point 2 de l'exercice B.13, il existe un $t > 1$ tel que $C_t := (1-t)B + tA \in \mathcal{S}^n_+$. Donc A est combinaison convexe stricte de B et $C_t \in \mathcal{S}^n_+$ et A est dans la face F'. Alors $B \in F'$.

20.3. Si le couple (X, y) est un point-selle du lagrangien ℓ défini en (20.4) sur $\mathcal{S}^n_+ \times \mathbb{R}^m$, on a pour tout $X' \succcurlyeq 0$ et tout $y' \in \mathbb{R}^m$:

$$\ell(X, y') \leqslant \ell(X, y) \leqslant \ell(X', y).$$

L'inégalité de gauche et la linéarité de ℓ en $y' \in \mathbb{R}^m$ montrent que $X \in \mathcal{F}_\mathrm{P}$. L'inégalité de droite donne pour tout $X' \succcurlyeq 0 : \langle S, X\rangle \leqslant \langle S, X'\rangle$, où $S := C - \mathcal{A}^*(y)$. On en déduit que $\langle X, S\rangle = 0$ (en prenant $X' = 0$ et $X' = 2X$) et puis $S \succcurlyeq 0$ (proposition 20.1). La réciproque est évidente.

20.4. $[(i') \Rightarrow (ii')]$ Comme le système dual (20.13) est fortement réalisable, C est intérieur à $\mathcal{A}^*(\mathbb{R}^m) + \mathcal{S}^n_+$. Alors pour $H \succ 0$ et $t > 0$ assez petit, $C - tH \in \mathcal{A}^*(\mathbb{R}^m) + \mathcal{S}^n_+$. Ceci implique l'existence de $(y_0, S_0) \in \mathbb{R}^m \times \mathcal{S}^n_+$ tel que $\mathcal{A}^*(y_0) + S_0 = C - tH$ ou encore $\mathcal{A}^*(y_0) + (S_0 + tH) = C$. Comme $S_0 + tH \succ 0$, $\mathcal{F}^s_\mathrm{D}$ est non vide.

$[(ii') \Rightarrow (iii')]$ Soit $(y_0, S_0) \in \mathcal{F}_\mathrm{D}$. On choisit de montrer que (QC-R) a lieu en (y_0, S_0) dans son expression (4.101c), qui s'écrit ici sous « forme matricielle »

$$\begin{pmatrix} \mathcal{A}^* & I \\ 0 & I \end{pmatrix} \begin{pmatrix} \mathbb{R}^m \\ \mathcal{S}^n \end{pmatrix} - \begin{pmatrix} 0_{\mathcal{S}^n} \\ \mathrm{T}_{S_0}\, \mathcal{S}^n_+ \end{pmatrix} = \begin{pmatrix} \mathcal{S}^n \\ \mathcal{S}^n \end{pmatrix}. \tag{D.61}$$

Soit $(M_1, M_2) \in \mathcal{S}^n \times \mathcal{S}^n$ un élément du membre de droite de (D.61) qu'il s'agit de représenter, grâce à (ii'), par un élément du membre de gauche. Il s'agit donc de trouver $(y_t, S_t) \in \mathbb{R}^m \times \mathcal{S}^n$ tel que

$$\mathcal{A}^*(y_t) + S_t = M_1 \qquad \text{et} \qquad S_t - M_2 \in \mathrm{T}_{S_0}\,\mathcal{S}^n_+. \tag{D.62}$$

Par l'hypothèse (ii'), il existe un couple $(\hat{y}, \hat{S})$ tel que $\mathcal{A}^*(\hat{y}) + \hat{S} = C$ et $\hat{S} \succ 0$. Alors, en prenant $y_t := 0 + t(\hat{y} - y_0)$ et $S_t := M_1 + t(\hat{S} - S_0)$, on vérifie la première équation de (D.62), quel que soit $t \in \mathbb{R}$, ainsi que la seconde équation, pour $t > 0$ assez grand. En effet, grâce à (2.57), il suffit de constater que, pour tout $v \in \mathcal{N}(S_0)$, on a $v^{\mathsf{T}}(S_t - M_2)v = v^{\mathsf{T}}(M_1 - M_2 + t\hat{S})v \geqslant 0$ pour t assez grand.

$[(iii') \Rightarrow (i')]$ Supposons que (QC-R) ait lieu en $(y_0, S_0) \in \mathcal{F}_{\mathrm{D}}$. Par l'expression (4.100) de (QC-R), on a

$$0 \in \operatorname{int}\left\{ \begin{pmatrix} 0 \\ S_0 \end{pmatrix} + \begin{pmatrix} \mathcal{A}^* & I \\ 0 & I \end{pmatrix} \begin{pmatrix} \mathbb{R}^m \\ \mathcal{S}^n \end{pmatrix} - \begin{pmatrix} 0_{\mathcal{S}^n} \\ \mathcal{S}^n_+ \end{pmatrix} \right\}.$$

Ceci implique que pour tout $(Z, 0) \in \mathcal{S}^n \times \mathcal{S}^n$ petit (pris à gauche de l'inclusion), on peut trouver un $(h, H) \in \mathbb{R}^m \times \mathcal{S}^n$ tel que $\mathcal{A}^*(h) + H = Z$ et $S_0 + H \succcurlyeq 0$. Pour tout $Z \in \mathcal{S}^n$ petit, on a donc trouvé $y := y_0 + h$ et $S := S_0 + H \succcurlyeq 0$ tel que $\mathcal{A}^*(y) + S = C + Z$. Autrement dit $C \in \operatorname{int}(\mathcal{A}^*(\mathbb{R}^m) + \mathcal{S}^n_+)$, qui est (i').

$[(ii') \Rightarrow (iv')]$ Supposons que $D \in \mathcal{S}^n$ vérifie $\langle C, D\rangle \leqslant 0$, $\mathcal{A}(D) = 0$ et $D \succcurlyeq 0$. Soit $(y, S) \in \mathcal{F}^s_{\mathrm{D}}$, qui est non vide par hypothèse, donc $\mathcal{A}^*(y) + S = C$ et $S \succ 0$. En multipliant scalairement $\mathcal{A}^*(y) + S = C$ par D, on trouve $0 \leqslant \langle S, D\rangle = \langle C, D\rangle - \langle y, \mathcal{A}(D)\rangle \leqslant 0$, si bien que $\langle S, D\rangle = 0$. Alors $S \succ 0$ et $D \succcurlyeq 0$ impliquent que $D = 0$ (point 2 du lemme 20.1).

$[(iv') \Rightarrow (ii')]$ La condition (iv') peut s'écrire

$$\{D \in \mathcal{S}^n : \mathscr{A}^*(D) \in K^+\} = \{0\}, \tag{D.63}$$

si l'on définit l'application linéaire $\mathscr{A} : \mathbb{R} \times \mathbb{R}^m \times \mathcal{S}^n \to \mathcal{S}^n$ et le cône K par

$$\mathscr{A}(\alpha, y, S) = \alpha C + \mathcal{A}^*(y) + S \qquad \text{et} \qquad K = \mathbb{R}_- \times \mathbb{R}^m \times \mathcal{S}^n_+.$$

En prenant le dual de chaque membre de l'identité (D.63) et en utilisant le lemme de Farkas, on trouve $\overline{\mathscr{A}(K)} = \mathcal{S}^n$. En prenant $C - I$ dans le membre de droite de cette dernière identité, quel que soit $\varepsilon > 0$, on peut trouver $(\alpha_\varepsilon, y_\varepsilon, S_\varepsilon) \in K$ tel que

$$\alpha_\varepsilon C + \mathcal{A}^*(y_\varepsilon) + S_\varepsilon = C - I + E_\varepsilon, \tag{D.64}$$

où l'erreur $E_\varepsilon \in \mathcal{S}^n$ tend vers zéro avec $\varepsilon \downarrow 0$. L'identité (D.64) permet alors d'écrire $\mathcal{A}^*(y'_\varepsilon) + S'_\varepsilon = C$ avec

$$y'_\varepsilon := \frac{y_\varepsilon}{1 - \alpha_\varepsilon} \qquad \text{et} \qquad S'_\varepsilon := \frac{I + S_\varepsilon - E_\varepsilon}{1 - \alpha_\varepsilon}.$$

La matrice S'_ε est définie positive si $\varepsilon > 0$ est pris assez petit, car $1 - \alpha_\varepsilon \geqslant 1$, $S_\varepsilon \succcurlyeq 0$ et $E_\varepsilon \to 0$ quand $\varepsilon \downarrow 0$.

20.5. [509] Il suffit d'utiliser la factorisation gaussienne par bloc de la matrice d'ordre $2n$:

$$M := \begin{pmatrix} I_n & A \\ A^{\mathsf{T}} & X \end{pmatrix} = \begin{pmatrix} I_n & 0 \\ A^{\mathsf{T}} & I_n \end{pmatrix} \begin{pmatrix} I_n & 0 \\ 0 & X - A^{\mathsf{T}}A \end{pmatrix} \begin{pmatrix} I_n & A \\ 0 & I_n \end{pmatrix}$$

Si $X \succcurlyeq 0$, alors il existe une matrice A telle que $X = A^{\mathsf{T}}A$ (factorisation spectrale ou choleskienne), si bien que $\operatorname{rg} M = n$. In versement si $\operatorname{rg}(M) \leqslant n$, alors $X = A^{\mathsf{T}}A \succcurlyeq 0$.

20.6. On note $\mathcal{L} = \mathcal{R}(\mathcal{A}^*)$. Si $b \in \mathcal{R}(\mathcal{A})$, il existe $D \in \mathcal{S}^n$ tel que $\mathcal{A}(D) = b$. Alors X vérifie $\mathcal{A}(X) = b$ ssi $X - D \in \mathcal{N}(\mathcal{A}) = \mathcal{L}^\perp$, ce qui s'écrit aussi $X \in D + \mathcal{L}^\perp$. Cela

conduit à la définition proposée du problème primal. D'autre part, pour le problème dual, on observe que $\langle b, y\rangle = \langle \mathcal{A}(D), y\rangle = \langle D, \mathcal{A}^*(y)\rangle = \langle D, C - S\rangle$. Comme $\langle D, C\rangle$ est une constante, le problème dual revient à minimiser $\langle D, S\rangle$ sous la contrainte que $S = C - \mathcal{A}^*(y)$, c'est-à-dire $S \in C + \mathcal{L}$.

20.7. Soit K la matrice dans (**??**).

[⇒] Clairement $A \succcurlyeq 0$ en multipliant à gauche et à droite la matrice K par des vecteurs de la forme $(x, 0) \in \mathbb{R}^n \times \mathbb{R}^m$. Si $\mathcal{R}(B) \not\subseteq \mathcal{R}(A)$, il existe un vecteur $y \in \mathbb{R}^m$ tel que $By = x_0 + Ax_1$ avec x_0 non nul dans $\mathcal{N}(A)$ et $x_1 \in \mathbb{R}^n$; en multipliant à gauche et à droite la matrice K par des vecteurs de la forme $(tx_0, y) \in \mathbb{R}^n \times \mathbb{R}^m$, on obtient $2t\|x_0\|_2^2 + y^\mathsf{T} Cy$, qui n'est pas positif si t est pris suffisamment négatif.

[⇔] On peut maintenant supposer que $A \succcurlyeq 0$ et que $\mathcal{R}(B) \subseteq \mathcal{R}(A)$. Soit $A = V\Sigma V^\mathsf{T}$ la factorisation spectrale de A, avec V injective et Σ diagonale définie positive. On peut écrire $B = VB_0$, ce qui conduit à la factorisation suivante de K :

$$K = \begin{pmatrix} A & B \\ B^\mathsf{T} & C \end{pmatrix} = \begin{pmatrix} V & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} \Sigma & B_0 \\ B_0^\mathsf{T} & C \end{pmatrix} \begin{pmatrix} V^\mathsf{T} & 0 \\ 0 & I \end{pmatrix}.$$

Comme la matrice à droite est surjective, $K \succcurlyeq 0$ si, et seulement si, la matrice du milieu du membre de droite est $\succcurlyeq 0$, ce qui, par (20.19), est le cas si, et seulement si, $C - B_0^\mathsf{T}\Sigma^{-1}B_0 \succcurlyeq 0$ ou $C - B_0^\mathsf{T}V^\mathsf{T}(V\Sigma^{-1}V^\mathsf{T})VB_0 \succcurlyeq 0$ ou encore $C - B^\mathsf{T}A^\dagger B \succcurlyeq 0$.

20.8. Soit $m := |E| + |I|$ le nombre de contraintes dans (20.26a). On introduit les fonctions affines $A : \mathbb{R}^m \to \mathbb{R}^{n\times n}$, $b : \mathbb{R}^m \to \mathbb{R}^n$ et $c : \mathbb{R}^m \to \mathbb{R}$, définies en $\lambda \in \mathbb{R}^m$ par $A(\lambda) = A_0 + \sum_{i=1}^m \lambda_i A_i$, $b(\lambda) = b_0 + \sum_{i=1}^m \lambda_i b_i$ et $c(\lambda) = c_0 + \sum_{i=1}^m \lambda_i c_i$. Le lagrangien $\ell : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ du problème (20.26a) s'écrit alors en $(x, \lambda) \in \mathbb{R}^n \times \mathbb{R}^m$:

$$\ell(x, \lambda) = x^\mathsf{T} A(\lambda)x + 2b(\lambda)^\mathsf{T} x + c(\lambda).$$

Le problème dual lagrangien s'écrit alors

$$\sup_{\lambda\in\Lambda} \inf_{x\in\mathbb{R}^n} \ell(x, \lambda),$$

où $\Lambda := \{\lambda \in \mathbb{R}^m : \lambda_I \geqslant 0\}$. Si $A(\lambda)$ n'est pas semi-définie positive, l'infimum en $x \in \mathbb{R}^n$ du lagrangien vaut $-\infty$. Il en est de même si $A(\lambda) \succcurlyeq 0$ et $b(\lambda) \notin \mathcal{R}(A(\lambda))$. Si $A(\lambda) \succcurlyeq 0$ et $b(\lambda) \in \mathcal{R}(A(\lambda))$, l'infimum en $x \in \mathbb{R}^n$ du lagrangien est atteint en $x_\lambda := -A(\lambda)^\dagger b(\lambda)$, où $A(\lambda)^\dagger$ désigne le pseudo-inverse de $A(\lambda)$ et prend pour valeur $\ell(x_\lambda, \lambda) = -b(\lambda)^\mathsf{T} A(\lambda)^\dagger b(\lambda) + c(\lambda)$. Le problème dual s'écrit donc $\sup\{\ell(x_\lambda, \lambda) : \lambda \in \Lambda,\ A(\lambda) \succcurlyeq 0,\ b(\lambda) \in \mathcal{R}(A(\lambda))\}$ ou

$$\begin{cases} \sup_{(\lambda,t)\in\Lambda\times\mathbb{R}} -t \\ -b(\lambda)^\mathsf{T} A(\lambda)^\dagger b(\lambda) + c(\lambda) \geqslant -t \\ A(\lambda) \succcurlyeq 0 \\ b(\lambda) \in \mathcal{R}(A(\lambda)). \end{cases}$$

Comme tout dual lagrangien, c'est un problème convexe, dans le sens où le critère est convexe (même linéaire) et le domaine admissible est convexe (même si cela n'apparaît pas clairement à cause de la première contrainte). Avant de le dualiser, on écrit ses contraintes sous la forme d'une condition de semi-définie positivité de matrice qui mettra en évidence le caractère convexe de l'ensemble admissible. Les trois contraintes sont en fait équivalentes à l'unique expression

$$\begin{pmatrix} A(\lambda) & b(\lambda) \\ b(\lambda)^\mathsf{T} & c(\lambda) + t \end{pmatrix} \succcurlyeq 0.$$

En effet, d'après l'exercice 20.7, cette dernière condition est équivalente à $A(\lambda) \succcurlyeq 0$, $b(\lambda) \in \mathcal{R}(A(\lambda))$ et $c(\lambda) + t - b(\lambda)^\mathsf{T} A(\lambda)^\dagger b(\lambda) \geqslant 0$. Le problème dual de (20.26) s'écrit donc comme en (D.4).

Pour écrire le dual lagrangien du problème (D.4), on récrit ce dernier avec les matrices $\tilde{A}_i$ définies en (20.27) comme le problème en $(\lambda, t) \in \mathbb{R}^m \times \mathbb{R}$ suivant

$$\begin{cases} \sup_{(\lambda,t)} -t \\ \tilde{A}_0 + \sum_{i=1}^m \lambda_i \tilde{A}_i + tE^{n+1,n+1} \succcurlyeq 0 \\ \lambda_I \geqslant 0, \end{cases}$$

où $E^{n+1,n+1} \in \mathcal{S}^{n+1}$ a son élément $(n+1, n+1)$ valant 1 et ses autres éléments nuls. Ce problème s'écrit aussi

$$\sup_{(\lambda,t)} \inf_{\substack{X \succcurlyeq 0 \\ s \geqslant 0}} -t + \langle X, \tilde{A}_0 + \textstyle\sum_{i=1}^m \lambda_i \tilde{A}_i + tE^{n+1,n+1} \rangle + s^\mathsf{T} \lambda_I.$$

Ce dernier est le dual de

$$\inf_{\substack{X \succcurlyeq 0 \\ s \geqslant 0}} \sup_{(\lambda,t)} \langle \tilde{A}_0, X \rangle + \sum_{i \in E} \lambda_i \langle \tilde{A}_i, X \rangle + \sum_{i \in I} \lambda_i (\langle \tilde{A}_i, X \rangle + s_i) + t(X_{n+1,n+1} - 1),$$

qui n'est autre que (20.28).

20.9. C'est clairement une application bilinéaire. D'autre part, si on note $E^s_{ij} \in \mathcal{S}^n$ la matrice dont les éléments (i,j) et (j,i) valent 1 et les autres valent zéro, les matrices E^s_{ij}, avec $i \leqslant j$, forment une base de $\mathcal{S}^n$. D'après le théorème de la base, il suffit de montrer que les $E^s_{ij} \otimes E^s_{kl}$, avec $i \leqslant j$ et $k \leqslant l$, forment une base de $L(\mathcal{S}^n, \mathcal{S}^n)$. On a

$$(E^s_{ij} \otimes E^s_{kl})(H) = \frac{1}{2}(H_{ik}E^s_{jl} + H_{il}E^s_{jk} + H_{jk}E^s_{il} + H_{jl}E^s_{ik}).$$

Si $i < j$ et $k < l$, l'opérateur $E^s_{ij} \otimes E^s_{kl}$ envoie

$$E^s_{ik} \mapsto \frac{1}{2}E^s_{jl}, \quad E^s_{il} \mapsto \frac{1}{2}E^s_{jk}, \quad E^s_{jk} \mapsto \frac{1}{2}E^s_{il} \quad \text{et} \quad E^s_{jl} \mapsto \frac{1}{2}E^s_{ik}.$$

Si $i = j$ et $k < l$, on a $(E^s_{ij} \otimes E^s_{kl})(H) = H_{ik}E^s_{il} + H_{il}E^s_{ik}$, si bien que l'opérateur $E^s_{ij} \otimes E^s_{kl}$ envoie

$$E^s_{ik} \mapsto E^s_{il} \quad \text{et} \quad E^s_{il} \mapsto E^s_{ik}.$$

Si $i < j$ et $k = l$, on a $(E^s_{ij} \otimes E^s_{kl})(H) = H_{ik}E^s_{jk} + H_{jk}E^s_{ik}$, si bien que l'opérateur $E^s_{ij} \otimes E^s_{kl}$ envoie

$$E^s_{ik} \mapsto E^s_{jk} \quad \text{et} \quad E^s_{jk} \mapsto E^s_{ik}.$$

Enfin, si $i = j$ et $k = l$, on a $(E^s_{ij} \otimes E^s_{kl})(H) = 2H_{ik}E^s_{ik}$, si bien que l'opérateur $E^s_{ij} \otimes E^s_{kl}$ envoie

$$E^s_{ik} \mapsto E^s_{ik}.$$

Par conséquent, les opérateurs $E^s_{ij} \otimes E^s_{kl}$, avec $i \leqslant j$ et $k \leqslant l$, envoient tous les éléments de la base de $\mathcal{S}^n$ sur les éléments de la base de $\mathcal{S}^n$ et forment ainsi une base de $L(\mathcal{S}^n, \mathcal{S}^n)$. Donc $\otimes$ est bien un produit tensoriel.

20.10. Let $X \in \operatorname{dom} \mathrm{ld} = \mathcal{S}^n_{++}$.

- Soit $D \in \mathcal{S}^n \setminus \mathcal{S}^n_+$. Comme $\mathcal{S}^n_+$ est fermé, $\frac{1}{t}X + D \notin \mathcal{S}^n_+$ pour $t > 0$ assez grand. Donc $D \notin (\operatorname{dom} \mathrm{ld})^\infty$, ce qui implique que $D \notin \operatorname{dom} \mathrm{ld}^\infty$ (point (iii) de la proposition 3.27) ou $\mathrm{ld}(D) = +\infty$.

- Soit $D \in \mathcal{S}_+^n$. Alors pour $t \geqslant 0$, $\det(X+tD) \geqslant \det(X)$ (corollaire B.5), impliquant que $\mathrm{ld}(X+tD) \leqslant \mathrm{ld}(X)$. Donc $\mathrm{ld}^\infty(D) \leqslant 0$.
- Soit à nouveau $D \in \mathcal{S}_+^n$. Fixons $t_0 \geqslant 0$. Par convexité de ld et $\mathrm{ld}'(X) = -X^{-1}$ en $X \succ 0$, on a pour tout $t \in \mathbb{R}$: $\mathrm{ld}(X+tD) = \mathrm{ld}(X+t_0D) - (t-t_0)\langle (X+t_0D)^{-1}, D\rangle$, d'où l'on déduit que

$$\begin{aligned} \mathrm{ld}^\infty(D) &\geqslant -\langle (X+t_0D)^{-1}, D\rangle \\ &= -\frac{1}{t_0}\langle (X+t_0D)^{-1}, (X+t_0D) - X\rangle \\ &= -\frac{n}{t_0} + \frac{1}{t_0}\langle (X+t_0D)^{-1}, X\rangle. \end{aligned}$$

En utilisant

$$|\langle (X+t_0D)^{-1}, X\rangle| \leqslant n\|(X+t_0D)^{-1}\|_2 \, \|X\|_2 = \frac{n\|X\|_2}{\lambda_{\min}(X+t_0D)} \leqslant \frac{n\|X\|_2}{\lambda_{\min}(X)},$$

on trouve alors

$$\mathrm{ld}^\infty(D) \geqslant -\frac{n}{t_0}\left(1 + \frac{\|X\|_2}{\lambda_{\min}(X)}\right).$$

En prenant la limite lorsque $t_0 \to \infty$ conduit à $\mathrm{ld}^\infty(D) \geqslant 0$.

Espaces topologiques

A.1. Comme $\liminf_k b_k > -\infty$, il existe un $b \in \mathbb{R}$ tel que $b_k - b \geqslant 0$ pour tout k. Alors les suites $\{a_k\}$ et $\{b_k - b\}$ sont positives et en appliquent (A.3), on obtient

$$\liminf_{k\to\infty} a_k(b_k - b) \geqslant a\left(\liminf_{k\to\infty}(b_k - b)\right).$$

Comme $a_k b \to ab$, on a $\liminf_k a_k(b_k - b) = (\liminf_k a_k b_k) - ab$ (voir (A.2a) et le commentaire qui suit). Pour la même raison, $\liminf_k (b_k - b) = (\liminf_k b_k) - b$. Alors l'inégalité ci-dessus conduit à (A.12).

A.2. [$(i) \Rightarrow (iii)$] Soit $\{x_k\} \subseteq N_\alpha(\varphi)$ (ensemble de sous-niveau $\alpha \in \mathbb{R}$ de φ) une suite convergeant vers x ; il faut montrer que $x \in N_\alpha(\varphi)$. La suite $\{(x_k, \alpha)\} \subseteq \mathrm{epi}\,\varphi$ converge vers (x, α). Comme $\mathrm{epi}\,\varphi$ est supposé fermé, $(x,\alpha) \in \mathrm{epi}\,\varphi$, c'est-à-dire $x \in N_\alpha(\varphi)$.

[$(iii) \Rightarrow (ii)$] Soient $\{x_k\}$ une suite convergeant vers x et $l := \liminf_k \varphi(x_k)$; il faut montrer que $\varphi(x) \leqslant l$.

- Si $l = +\infty$, le résultat est clair.
- Si $l = -\infty$, $\varphi(x_k) = -\infty$ pour tout k; donc $x_k \in N_\alpha(\varphi)$, quel que soit $\alpha \in \mathbb{R}$. Comme $N_\alpha(\varphi)$ est fermé et $x_k \to x$, $x \in N_\alpha(\varphi)$, quel que soit $\alpha \in \mathbb{R}$; donc $\varphi(x) = -\infty$.
- Si $l \in \mathbb{R}$, par définition de l'infimum, pour tout $k \geqslant 1$, on peut trouver un x_{j_k} avec $j_k \geqslant k$ tel que $\inf_{j\geqslant k} \varphi(x_j) \leqslant \varphi(x_{j_k}) \leqslant \inf_{j\geqslant k} \varphi(x_j) + 1/k$. En passant à la limite dans ces inégalités, on obtient $\lim_{k\to\infty} \varphi(x_{j_k}) = l$ et $x_{j_k} \to x$. Soit $\varepsilon > 0$. Pour tout $k \geqslant 1$ assez grand, $\varphi(x_{j_k}) \leqslant l + \varepsilon$ ou $x_{j_k} \in N_{l+\varepsilon}(\varphi)$ qui est fermé; donc $x \in N_{l+\varepsilon}(\varphi)$ ou $\varphi(x) \leqslant l + \varepsilon$. Comme $\varepsilon > 0$ est arbitraire, $\varphi(x) \leqslant l$.

[$(ii) \Rightarrow (i)$] Soit $\{(x_k, \alpha_k)\} \subseteq \mathrm{epi}\,\varphi$ telle $(x_k, \alpha_k) \to (x, \alpha)$; il faut montrer que $(x,\alpha) \in \mathrm{epi}\,\varphi$. On a $\varphi(x) \leqslant \liminf_k \varphi(x_k)$ [car φ est s.c.i. et $x_k \to x$] $\leqslant \liminf_k \alpha_k$ [car $(x_k, \alpha_k) \in \mathrm{epi}\,\varphi$] $= \alpha$ [car $\alpha_k \to \alpha$].

A.3. En effet, soit $\{x_k\} \to x$. On a $\inf_{k\geqslant l} f(x_k) + \inf_{k\geqslant l} g(x_k) \leqslant \inf_{k\geqslant l}(f(x_k) + g(x_k))$ et donc $f(x) + g(x) \leqslant \liminf_k f(x_k) + \liminf_k g(x_k) \leqslant \liminf_k (f(x_k) + g(x_k))$.

A.4. f est s.c.s. comme enveloppe inférieure de fonctions s.c.s..

Supposons à présent que $\mathbb{E}$ soit un espace métrique, que $\mathbb{F}$ soit compact et que φ soit continue. Il reste à montrer que f est s.c.i. en un point $x_0 \in X$ arbitraire. Si ce n'est pas le cas, il existe $\delta > 0$ et une suite $x_n \to x_0$ dans $\mathbb{E}$ tels que $f(x_n) \leqslant f(x_0) - \delta$. Comme $\mathbb{F}$ est compact, il existe $y_n \in \mathbb{F}$ tel que $f(x_n) = \varphi(x_n, y_n)$ et en extrayant une sous-suite au besoin, on peut supposer que $\{y_n\}$ converge, disons vers y_0. Par continuité de φ, on obtient $\varphi(x_0, y_0) \leqslant f(x_0) - \delta$, ce qui contredit la définition de $f(x_0)$.

A.5. Pour tout x, y, $z \in \mathbb{E}$, on a $d(x,z) \leqslant d(x,y) + d(y,z)$. En prenant l'infimum pour $z \in P$, on trouve que $\mathrm{d}_P(x) \leqslant d(x,y) + \mathrm{d}_P(y)$. Par symétrie $|\,\mathrm{d}_P(x) - \mathrm{d}_P(y)| \leqslant d(x,y)$, donc d_P est lipschitzienne de module 1.

A.6. Supposons que $A + C \subseteq B$ et soit $x \in A$. Il faut montrer que x est de la forme $y - z$ avec $y \in B$ et $z \in C$. On prend z quelconque dans C. L'inclusion $A + C \subseteq B$ implique qu'il existe un $y \in B$ tel que $x + z = y$, donc $x = y - z$ comme attendu.

Par ailleurs, si $C = A = [0,1] \subseteq \mathbb{R}$ et $B = \{0\}$, on a bien $A \subseteq B + C$, alors que $A - C = [-1,1] \not\subseteq B$.

A.7. [$(i) \Rightarrow (ii)$] Alors A s'écrit $x_0 + \mathbb{E}_0$, où x_0 est dans un espace vectoriel $\mathbb{E}$ et $\mathbb{E}_0$ est un sous-espace vectoriel de $\mathbb{E}$. Si x et $y \in A$, on a $x = x_0 + u$, $y = x_0 + v$ avec u, $v \in \mathbb{E}_0$ et donc pour tout $t \in \mathbb{R}$, $tx + (1-t)y = x_0 + tu + (1-t)v \in x_0 + \mathbb{E}_0 = A$.

[$(ii) \Rightarrow (iii)$] On note que $0 \in A - A$, car A est non vide ; ensuite si $t \in \mathbb{R}$ et a, $b \in A$, on a $t(a-b) = [ta + (1-t)b] - b \in A - A$; enfin si a, b, c, $d \in A$, on a $(a-b) + (c-d) = 2(\frac{1}{2}a + \frac{1}{2}c) - 2(\frac{1}{2}b + \frac{1}{2}d) \in A - A$; donc $A - A$ est un sous-espace vectoriel. On a bien sûr $A \subseteq A + (A - A)$ et inversement, si a, b, $c \in A$, on a $a + (b - c) = t(\frac{1}{2}a + \frac{1}{2}b) + (1-t)c \in A$, avec $t = 2$.

[$(iii) \Rightarrow (i)$] Soit $x_0 \in A$ et $\mathbb{E}_0 := A - A$ qui est un sous-espace vectoriel. On a $x_0 + \mathbb{E}_0 \subseteq A$, car $A + (A - A) = A$. Inversement, si $a \in A$, on a $a = x_0 + (a - x_0) \in x_0 + \mathbb{E}_0$.

A.8. Soit $x \in \cap_i A_i$, supposé non vide. Alors $A_i = x + \mathbb{E}(A_i)$, d'où l'on déduit que $\cap_i A_i = x + \cap_i \mathbb{E}(A_i)$ ([$\subseteq$] : si $y \in \cap_i A_i$, alors $y - x \in \cap_i \mathbb{E}(A_i)$, donc $y \in x + \cap_i \mathbb{E}(A_i)$; [$\supseteq$] : clair). Comme l'intersection des sous-espaces vectoriels $\mathbb{E}(A_i)$ est un sous-espace vectoriel, $\cap_i A_i$ est un sous espace affine et $\mathbb{E}(\cap_{i\in I} A_i) = \cap_{i\in I} \mathbb{E}(A_i)$.

A.10. On peut supposer que $\langle x, y\rangle \neq 0$, sinon l'inégalité est trivialement vérifiée. Soit $t \in \mathbb{R}$ arbitraire. Par la semi-définie positivité, on a $0 \leqslant \langle x + ty, x + ty\rangle = \|x\|^2 + 2t\langle x, y\rangle + t^2\|y\|^2$. On en déduit que $y \neq 0$, sinon cette dernière inégalité ne pourrait être vérifiée pour tout t (on a supposé que $\langle x, y\rangle \neq 0$). Le minimum à droite est obtenu pour $t = -\langle x, y\rangle / \|y\|^2$ et prend la valeur $\|x\|^2 - \langle x, y\rangle^2 / \|y\|^2$ qui par sa positivité entraîne l'inégalité de Cauchy-Schwarz.

A.11. (i) Clairement, $\|y\|_{\mathrm{D}} = 0$, si $y = 0$. Inversement, si $y \neq 0$, on peut prendre $x = y/\|y\|_{\mathrm{P}}$ comme argument du supremum dans la définition de $\|y\|_{\mathrm{D}}$, ce qui donnerait $\|y\|_{\mathrm{D}} \geqslant \|y\|^2/\|y\|_{\mathrm{P}} > 0$. (ii) Si $\alpha \in \mathbb{R}$, on a $\|\alpha y\|_{\mathrm{D}} = \sup\{|\langle \alpha y, x\rangle| : \|x\|_{\mathrm{P}} \leqslant 1\} = |\alpha| \sup\{|\langle y, x\rangle| : \|x\|_{\mathrm{P}} \leqslant 1\} = |\alpha|\,\|y\|_{\mathrm{D}}$. (iii) $\|y + z\|_{\mathrm{D}} = \sup\{\langle y + z, x\rangle : \|x\|_{\mathrm{P}} \leqslant 1\} = \sup\{\langle y, x\rangle + \langle z, x\rangle : \|x\|_{\mathrm{P}} \leqslant 1\} \leqslant \sup\{\langle y, x\rangle : \|x\|_{\mathrm{P}} \leqslant 1\} + \sup\{\langle z, x\rangle : \|x\|_{\mathrm{P}} \leqslant 1\}$ [exercice 1.2] $= \|y\|_{\mathrm{D}} + \|z\|_{\mathrm{D}}$.

Algèbre linéaire

B.1. 1) Soit $\mathcal{B} := \{u_1, \dots, u_k\}$ une base de $\mathbb{E}_0 \cap \mathbb{E}_1$. On peut la compléter en une base $\mathcal{B}_0 := \mathcal{B} \cup \{v_1^0, \dots, v_p^0\}$ de $\mathbb{E}_0$ et une base $\mathcal{B}_1 := \mathcal{B} \cup \{v_1^1, \dots, v_q^1\}$ de $\mathbb{E}_1$ (théorème B.1 de la base incomplète). Il suffit de montrer que $\mathcal{B}_+ := \mathcal{B} \cup \{v_1^0, \dots, v_p^0\} \cup \{v_1^1, \dots, v_q^1\}$ est une base de $\mathbb{E}_0 + \mathbb{E}_1$. La famille $\mathcal{B}_+$ est génératrice car si $x_l = \sum_i \alpha_i^l u_i + \sum_j \beta_j^l v_j^l \in \mathbb{E}_l$, $l = 0, 1$, on a $x_0 + x_1 = \sum_i (\alpha_i^0 + \alpha_i^1) u_i + \sum_j \beta_j^0 v_j^0 + \sum_k \beta_k^1 v_k^1 \in \mathbb{E}_0 + \mathbb{E}_1$, qui s'écrit bien dans la base $\mathcal{B}_+$. La famille $\mathcal{B}_+$ est libre, car si $\sum_i \alpha_i u_i + \sum_j \beta_j^0 v_j^0 + \sum_k \beta_k^1 v_k^1 = 0$, on a $\sum_i \alpha_i u_i + \sum_j \beta_j^0 v_j^0 = -\sum_k \beta_k^1 v_k^1 \in \mathbb{E}_0 \cap \mathbb{E}_1$, donc $\sum_k \beta_k^1 v_k^1 = \sum_i \alpha_i' u_i$ pour certains α_i', ce qui implique que les $\beta_k^1 = 0$ (car $\mathcal{B}_1$ est une base). Donc, $\sum_i \alpha_i u_i + \sum_j \beta_j^0 v_j^0 = 0$, ce qui implique que les $\alpha_i = 0$ et les $\beta_j^0 = 0$ (car $\mathcal{B}_0$ est une base).

2) Se déduit de la formule de Grassmann, car si $\dim \mathbb{E}_0 + \dim \mathbb{E}_1 > \dim \mathbb{E}$, celle-ci donne $\dim \mathbb{E} < \dim \mathbb{E}_0 + \dim \mathbb{E}_1 \leqslant \dim(\mathbb{E}_0 + \mathbb{E}_1) + \dim(\mathbb{E}_0 \cap \mathbb{E}_1) = \dim \mathbb{E} + \dim(\mathbb{E}_0 \cap \mathbb{E}_1)$, si bien que $\dim(\mathbb{E}_0 \cap \mathbb{E}_1) > 0$ ou encore $\mathbb{E}_0 \cap \mathbb{E}_1 \neq \varnothing$.

B.2. 1) Clairement $\mathcal{R}\binom{A}{B} \subseteq \mathcal{R}(A) \times \mathcal{R}(B)$ et donc $\dim \mathcal{R}\binom{A}{B} \leqslant \dim(\mathcal{R}(A) \times \mathcal{R}(B)) = \dim \mathcal{R}(A) + \dim \mathcal{R}(B)$.

2) D'après (B.2), $\dim \mathcal{N}(A) + \dim \mathcal{R}(A) = \dim \mathbb{E} = \dim \mathcal{N}\binom{A}{B} + \dim \mathcal{R}\binom{A}{B}$. En utilisant le point 1, on obtient $\dim \mathcal{N}(A) + \dim \mathcal{R}(A) \leqslant \dim \mathcal{N}\binom{A}{B} + \dim \mathcal{R}(A) + \dim \mathcal{R}(B)$, qui conduit au résultat.

B.3.

B.4. Soit $\{A_k\}_{k \geqslant 1} \subseteq \mathbb{R}^{m \times n}$ une suite convergeant vers une matrice $A \in \mathbb{R}^{m \times n}$. Si $\operatorname{rg}(A) = r$, A possède une sous-matrice carrée A_{IJ} inversible d'ordre r ($|I| = |J| = r$), si bien que $\det A_{IJ} \neq 0$. Par la continuité du déterminant (c'est un polynôme de ses coefficients), $\det(A_k)_{IJ} \neq 0$ pour k assez grand. Dès lors, pour k assez grand, $(A_k)_{IJ}$ est inversible et $\operatorname{rg} A_k \geqslant r$.

B.5. 1)

2) Si $x \in \mathbb{E}_0$ et $y \in \mathbb{E}_0^\perp$, on a $\langle x, y \rangle = 0$. Dès lors x est orthogonal à tout élément de $\mathbb{E}_0^\perp$, si bien que $x \in (\mathbb{E}_0^\perp)^\perp$. Donc $\mathbb{E}_0 \subseteq (\mathbb{E}_0^\perp)^\perp$. Inversement, soit $x \in (\mathbb{E}_0^\perp)^\perp$ que l'on décompose en $x = x_0 + x_1$ avec $x_0 \in \mathbb{E}_0$ et $x_1 \in \mathbb{E}_0^\perp$. En prenant le produit scalaire de cette identité avec x_1, on obtient que $x_1 = 0$. Donc $x_0 \in \mathbb{E}_0$.

3) Clair.

4) Soit $\bar{x} = \bar{x}_0 + \bar{x}_1$ avec $\bar{x}_i \in \mathbb{E}_i^\perp$ et $x \in \mathbb{E}_0 \cap \mathbb{E}_1$. Alors $\langle \bar{x}_i, x \rangle = 0$ (car $\bar{x}_i \in \mathbb{E}_i^\perp$ et $x \in \mathbb{E}_i$) et donc $\langle \bar{x}, x \rangle = \langle \bar{x}_0, x \rangle + \langle \bar{x}_1, x \rangle = 0$. On a montré l'inclusion $\mathbb{E}_0^\perp + \mathbb{E}_1^\perp \subseteq (\mathbb{E}_0 \cap \mathbb{E}_1)^\perp$. Inversement, d'après les points 2 et 3, il suffit de montrer que $(\mathbb{E}_0^\perp + \mathbb{E}_1^\perp)^\perp \subseteq \mathbb{E}_0 \cap \mathbb{E}_1$. Soit $x \in (\mathbb{E}_0^\perp + \mathbb{E}_1^\perp)^\perp$. Tout $\bar{x}_i \in \mathbb{E}_i^\perp$ est aussi dans $\mathbb{E}_0^\perp + \mathbb{E}_1^\perp$. Dès lors $\langle x, \bar{x}_i \rangle = 0$. On en déduit que $x \in (\mathbb{E}_i^\perp)^\perp = \mathbb{E}_i$. Donc $x \in \mathbb{E}_0 \cap \mathbb{E}_1$.

B.6. On a $n = \dim \mathcal{N}(A) + \dim \mathcal{R}(A)$ et $n = \dim \mathcal{N}(A) + \dim \mathcal{N}(A)^\perp = \dim \mathcal{N}(A) + \dim \mathcal{R}(A^\mathsf{T})$. On en déduit que $\dim \mathcal{R}(A) = \dim \mathcal{R}(A^\mathsf{T})$, qui est le résultat recherché.

B.7. Notons E^{ij} la matrice carrée dont l'élément (i, j) vaut 1 et les autres éléments sont nuls. Alors $\det'(I) \cdot H = \sum_{ij} H_{ij} \det'(I) \cdot E^{ij} = \sum_{ij} H_{ij} \delta_{ij} = \operatorname{tr} H$. Ensuite, comme l'ensemble des matrices inversibles est dense et que le déterminant est de classe C^1, on peut supposer que A est inversible. On considjere alors l'application $B \mapsto \det(B) = \det(A) \det(A^{-1}B)$, que l'on différentie en A pour trouver que $\det'(A) \cdot H = \det(A) \det'(I) \cdot (A^{-1}H) = \det(A) \operatorname{tr}(A^{-1}H) = \det(A) \langle A^{-\mathsf{T}}, H \rangle$. Dès lors $\nabla \det(A) = (\det(A)) A^{-\mathsf{T}} = \operatorname{cof}(A)$.

B.8. Commençons par l'inégalité de gauche. Soit $W = (w_1 \cdots w_i)$ la matrice orthogonale $r \times i$ formée des vecteurs propres w_k de $U^\mathsf{T}AU$, rangés par ordre croissant des valeurs propres associées. On a par le principe du min-max de Courant-Fisher (l'inégalité vient du fait que UW est $n \times i$ et orthogonale) :

$$\lambda_i(U^\mathsf{T}AU) = \max_{\|x\|_2=1} x^\mathsf{T}W^\mathsf{T}U^\mathsf{T}AUWx \geqslant \min_{\substack{Q:n\times i \\ Q^\mathsf{T}Q=I_i}} \max_{\|x\|_2=1} x^\mathsf{T}Q^\mathsf{T}AQx = \lambda_i(A).$$

Pour l'inégalité de droite, on considère la matrice $W = (w_i \cdots w_r)$, qui est orthogonale et $r \times (r-i+1)$. L'autre aspect du principe du min-max de Courant-Fisher donne (l'inégalité vient du fait que UW est $n \times (r-i+1)$ et orthogonale) :

$$\lambda_i(U^\mathsf{T}AU) = \min_{\|x\|_2=1} x^\mathsf{T}W^\mathsf{T}U^\mathsf{T}AUWx \leqslant \max_{\substack{Q:n\times(r-i+1) \\ Q^\mathsf{T}Q=I_{r-i+1}}} \min_{\|x\|_2=1} x^\mathsf{T}Q^\mathsf{T}AQx = \lambda_{n-r+i}(A),$$

car $r-i+1 = n-(n-r+i)+1$.

B.10. 1) La meilleure constante $\alpha > 0$ est la plus petite valeur propre de A. En effet, soit $A = V\Lambda V^\mathsf{T}$ la factorisation spectrale de A, où $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ est la matrice des valeurs propres de A ($0 < \lambda_1 \leqslant \cdots \leqslant \lambda_n$) et V est une matrice orthogonale. Alors, quel que soit x, $x^\mathsf{T}Ax = \|\Lambda^{1/2}V^\mathsf{T}x\|_2^2 \geqslant \lambda_1\|V^\mathsf{T}x\|_2^2 = \lambda_1\|x\|_2^2$, si bien que l'on peut prendre $\alpha = \lambda_1$. D'autre part, en prenant pour x un vecteur propre associé à λ_1, on voit que la constante α doit être $\leqslant \lambda_1$.

De manière plus rapide, moins précise, mais en faisant apparaître un problème d'optimisation, on note que le quotient de Rayleigh $q(x) := (x^\mathsf{T}Ax)/\|x\|_2^2$ est continu sur la sphère unité S. Comme cette dernière est compacte, q atteint sa valeur minimale sur S, disons $q(x_*) = x_*^\mathsf{T}Ax_* =: \alpha$, avec $x_* \in S$. Comme $x_* \neq 0$ et A est définie positive, on a $\alpha > 0$ et par conséquent $x^\mathsf{T}Ax \geqslant \alpha\|x\|_2^2$, pour tout x.

2) Si A est nulle, on peut prendre $\alpha > 0$ quelconque; sinon, la meilleure constante $\alpha > 0$ est la plus petite valeur singulière de $A \in \mathbb{R}^{m\times n}$. En effet, soit $A = U\Sigma V^\mathsf{T}$ la décomposition de A en valeurs singulières, où $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_p)$ est la matrice des valeurs singulières de A ($0 < \sigma_1 \leqslant \cdots \leqslant \sigma_p$) et $U \in \mathbb{R}^{m\times p}$ et $V \in \mathbb{R}^{n\times p}$ sont telles que $U^\mathsf{T}U = V^\mathsf{T}V = I_p$. Alors, quel que soit x, $\|A^\mathsf{T}Ax\|_2 = \|V\Sigma^2V^\mathsf{T}x\|_2 = \|\Sigma^2V^\mathsf{T}x\|_2 \geqslant \sigma_1\|\Sigma V^\mathsf{T}x\|_2 = \sigma_1\|U\Sigma V^\mathsf{T}x\|_2 = \sigma_1\|Ax\|_2$, si bien que l'on peut prendre $\alpha = \sigma_1$. D'autre part, en prenant pour x un vecteur propre de $A^\mathsf{T}A$ associé à σ_1^2, on a $\|A^\mathsf{T}Ax\|_2 = \sigma_1^2\|x\|_2$ et $\|Ax\|_2 = (xA^\mathsf{T}Ax)^{1/2} = \sigma_1\|x\|_2$, si bien que la constante α doit être $\leqslant \sigma_1$.

On peut aussi se ramener à un problème de minimisation. On écrit $\|A^\mathsf{T}Ax\|_2^2 = (Ax)^\mathsf{T}(AA^\mathsf{T})(Ax)$, ce qui suggère de regarder le problème $\min\{v^\mathsf{T}(AA^\mathsf{T})v : v \in \mathcal{R}(A) \cap S\}$, où S est la sphère unité de $\mathbb{R}^m$. Comme $\mathcal{R}(A) \cap S$ est un compact, ce problème a une solution $v_* = Ax_* \in \mathcal{R}(A) \cap S$. De plus, la valeur minimale α^2 est strictement positive. En effet, dans le cas contraire, on aurait $v_*(AA^\mathsf{T})v_* = 0$, donc $A^\mathsf{T}v_* = A^\mathsf{T}Ax_* = 0$, ou encore $Ax_* = 0$, ce qui contredirait le fait que $v_* = Ax_* \in S$. On en déduit que $\|A^\mathsf{T}Ax\|_2^2 \geqslant \alpha^2\|Ax\|_2^2$ et le résultat.

Encore plus rapide : la démonstration par l'absurde. Si l'affirmation est fausse, pour tout indice k, on peut trouver x_k tel que $\|A^\mathsf{T}Ax_k\|_2 < \frac{1}{k}\|Ax_k\|_2$. On a donc $Ax_k \neq 0$. On extrait une sous-suite telle que $Ax_k/\|Ax_k\|_2$ converge, disons vers Ay (la limite est nécessairement dans $\mathcal{R}(A)$, car $\mathcal{R}(A)$ est fermé). À la limite on trouve $A^\mathsf{T}Ay = 0$, ce qui contredit le fait que $Ay \neq 0$.

Le fait que $\|A^\mathsf{T}\cdot\|$ soit une norme sur $\mathcal{R}(A)$ devient alors clair. Si $\|A^\mathsf{T}x\| = 0$ et $x = Av \in \mathcal{R}(A)$, on a par équivalence des normes : $0 = \|A^\mathsf{T}Av\|_2 \geqslant \sigma\|Av\|_2$; donc $x = Av = 0$. Les autres propriétés faisant de $\|A^\mathsf{T}\cdot\|$ une norme sont faciles à vérifier.

B.11. Soit $A = V\Lambda V^\mathsf{T}$ la décomposition spectrale de A. Alors (λ, v) est un couple propre de AB si, et seulement si, $(\lambda, \Lambda^{-1/2}V^\mathsf{T}v)$ est un couple propre de $\Lambda^{1/2}V^\mathsf{T}BV\Lambda^{1/2} \in \mathcal{S}^n_{++}$. On en déduit le résultat.

B.13. 1) Les relations sont complémentaires. Démontrons la seconde. Quelles que soient les matrices A et $B \in \mathbb{R}^{m\times n}$, on a toujours $\mathcal{N}(A+B) \supseteq \mathcal{N}(A)\cap\mathcal{N}(B)$. Si A et $B \in \mathcal{S}^n_+$ et si $x \in \mathcal{N}(A+B)$, alors $0 = x^\mathsf{T}(A+B)x = x^\mathsf{T}Ax + x^\mathsf{T}Bx$. Comme $x^\mathsf{T}Ax \geqslant 0$ et $x^\mathsf{T}Bx \geqslant 0$, on en déduit que $x^\mathsf{T}Ax = x^\mathsf{T}Bx = 0$ et donc que $x \in \mathcal{N}(A)\cap\mathcal{N}(B)$.

2) Soit $M_t := (1-t)A+tB$. Comme $\mathcal{N}(B) \subseteq \mathcal{N}(A)$, on a $\mathcal{N}(B) \subseteq \mathcal{N}(M_t)$, si bien que les plus petites valeurs propres non nulles de A, B et M_t sont chacune associée à un vecteur propre appartemant à $\mathcal{N}(B)^\perp = \mathcal{R}(B)$. Si λ_1^+ est le plus petite valeur propre strictement positive de B, on a pour tout vecteur unitaire $v \in \mathcal{R}(B)$: $v^\mathsf{T}Bv \geqslant \lambda_1^+$ et $v^\mathsf{T}Av \leqslant \lambda_{\max}(A)$. On en déduit que, pour $t > 1$: $v^\mathsf{T}M_tv \geqslant (1-t)\lambda_{\max}(A) + t\lambda_1^+$ qui est > 0 pour $t > 1$ assez petit.

B.14. 1) On prend une factorisation de Cholesky de $A = LL^\mathsf{T} \succ 0$ et spectrale de $L^{-1}BL^{-\mathsf{T}} = V\Lambda V^\mathsf{T}$. Le résultat s'en ensuit en prenant $R = LV$.

2) Soit $A = V\Lambda V^\mathsf{T}$ la factorisation spectrale de A, avec $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ où les λ_i sont croissants. Si A et B commutent, $S := V^\mathsf{T}BV$ vérifie $\Lambda S = S\Lambda$, si bien que $S_{ij} = 0$ lorsque $\lambda_i \neq \lambda_j$. Dès lors $S = \mathrm{diag}(S_1, \ldots, S_r)$ est diagonale par blocs, l'ordre des blocs symétriques S_i étant égal à la multiplicité de la i-ième valeur propre. Soit $S_i = W_iD_iW_i^\mathsf{T}$ la factorisation spectrale de S_i, $W = W_1 \oplus \cdots \oplus W_r$ et $D = D_1\oplus\cdots\oplus D_r$. Clairement $\mathcal{O} = VW$ est orthogonale, $B = \mathcal{O}D\mathcal{O}^\mathsf{T}$ et $A = \mathcal{O}\Lambda\mathcal{O}^\mathsf{T}$. Cette dernière identité vient du fait que $W\Lambda W^\mathsf{T} = \Lambda$, parce que chaque blocs orthogonal W_i correspond à une même valeur propre.

Inversement, si $A = V\Lambda_AV^\mathsf{T}$ et $B = V\Lambda_BV^\mathsf{T}$ avec Λ_A et Λ_B diagonales et V orthogonale, on a $AB = V\Lambda_A\Lambda_BV^\mathsf{T} = V\Lambda_B\Lambda_AV^\mathsf{T} = BA$.

B.15. On note $M_r := M + rA^\mathsf{T}A$.

$[(i) \Rightarrow (ii)]$ Si (ii) n'a pas lieu, il existe une suite $\{r_k\} \to \infty$ et une suite $\{v_k\}$ de vecteurs propres de M_{r_k} correspondant à une valeur propre strictement négative. On a

$$v_k^\mathsf{T}M_{r_k}v_k = v_k^\mathsf{T}Mv_k + r_k\|Av_k\|^2 < 0.$$

En extrayant une sous-suite au besoin, on peut supposer que $v_k \to v$, avec $\|v\| = 1$. En divisant l'inégalité ci-dessus par $r_k > 0$ et en prenant la limite, on trouve que $Av = 0$. D'autre part $v_k^\mathsf{T}Mv_k < 0$, si bien qu'à la limite on a $v^\mathsf{T}Mv \leqslant 0$, et même $v^\mathsf{T}Mv = 0$ (M est semi-définie positive sur $\mathcal{N}(A)$). Alors (i) implique $Mv = 0$. Observons à présent que v et v_k des vecteurs propres de M_{r_k} correspondant à des valeurs propres différentes (la première est nulle, la seconde est négative). Donc $v_k^\mathsf{T}v = 0$, qui à la limite donne $v = 0$, en contradiction avec $\|v\| = 1$. Cette contradiction montre (ii).

$[(ii) \Rightarrow (i)]$. Soit $r \in \mathbb{R}$ tel que M_r soit semi-définie positive et prenons $v \in \mathcal{N}(A)$ tel que $v^\mathsf{T}Mv = 0$. Alors $v^\mathsf{T}M_rv = 0$ et, puisque M_r est semi-définie positive, $M_rv = 0$. Maintenant $v \in \mathcal{N}(A)$ implique que $Mv = 0$.

$[(ii) \Rightarrow (iii)]$. Par homogénéité positive,

$$\inf_{Av\neq 0} \frac{v^{\mathsf T} M v}{\|Av\|_2^2} = \inf_{\|Av\|_2=1} v^{\mathsf T} M v.$$

En utilisant le r donné par (ii), on a

$$\inf_{\|Av\|_2=1} v^{\mathsf T} M v = \inf_{\|Av\|_2=1} v^{\mathsf T}(M + rA^{\mathsf T}A)v - r \geqslant -r.$$

$[(iii) \Rightarrow (ii)]$. D'après (iii), il existe un $r \in \mathbb{R}$ tel que $\inf\{v^{\mathsf T} M v : \|Av\|_2 = 1\} \geqslant -r$. Donc

$$\inf_{\|Av\|_2=1} v^{\mathsf T}(M + rA^{\mathsf T}A)v \geqslant 0.$$

Montrons que

$$\inf_{v} v^{\mathsf T}(M + rA^{\mathsf T}A)v \geqslant 0.$$

Soit $\{v_k\}$ une suite minimisante. Si $Av_k \neq 0$ infiniment souvent, l'inégalité précédente montre que l'infimum ci-dessus est $\geqslant 0$. Dans le cas contraire, $Av_k = 0$ pour k grand et la positivité de l'infimum vient de ce que M est semi-définie positive sur $\mathcal{N}(A)$.

Comme exemple, on peut prendre

$$M = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad \text{et} \quad A = \begin{pmatrix} 1 & 1 \end{pmatrix}.$$

Le noyau de A est engendré par le vecteur $\begin{pmatrix} 1 & -1 \end{pmatrix}^{\mathsf T}$, si bien que M est semi-définie positive sur $\mathcal{N}(A)$. D'autre part, M est inversible, si bien que (i) n'a pas lieu, ni donc (ii). La dernière affirmation peut d'ailleurs se vérifier directement: les valeurs propres de

$$M_r = M + rA^{\mathsf T}A = \begin{pmatrix} r+1 & r \\ r & r-1 \end{pmatrix}$$

sont $r \pm \sqrt{r^2+1}$, si bien que pour tout $r \in \mathbb{R}$ l'une d'entre elles est strictement négative. Un vecteur propre correspondant à cette dernière est $((1 - \sqrt{r^2+1})/r, 1)$. La démonstration ci-dessus a mis en évidence le fait que, lorsque $r \to \infty$, ce vecteur propre doit tourner de $(0, 1)$ (pour $r = 0$) vers un vecteur de $\mathcal{N}(A)$, ici $(-1, 1)$.

B.16. (i) Soit $\lambda \in \mathbb{C}$ une valeur propre de $I_n + uv^{\mathsf T}$ et $\varphi \in \mathbb{R}^n$ un vecteur propre associé (non nul donc). On a $\varphi + (v^{\mathsf T}\varphi)u = \lambda\varphi$. Si $\lambda \neq 1$, cette relation montre que $v^{\mathsf T}\varphi \neq 0$ et que φ est colinéaire à u. En prenant le produit scalaire avec v, on a $(v^{\mathsf T}\varphi)(1 + v^{\mathsf T}u - \lambda) = 0$ et donc $\lambda = 1 + v^{\mathsf T}u$ (car $v^{\mathsf T}\varphi \neq 0$). Dès lors, les seules valeurs propres de $I_n + uv^{\mathsf T}$ sont 1 et $1 + v^{\mathsf T}u$. Pour en déduire que $\det(I_n + uv^{\mathsf T}) = 1 + v^{\mathsf T}u$, il reste à montrer que la valeur propre $1 + v^{\mathsf T}u$ est simple, ce qui sera le cas si $\dim\mathcal{N}([(1-\lambda)I_n + uv^{\mathsf T}]^2) = 1$. Soient φ_1 et $\varphi_2 \in \mathcal{N}([(1-\lambda)I_n + uv^{\mathsf T}]^2)$. Alors les $\xi_i = (1-\lambda)\varphi_i + (v^{\mathsf T}\varphi_i)u$ vérifient $[(1-\lambda)I_n + uv^{\mathsf T}]\xi_i = 0$. Donc les ξ_i sont parallèles à u $(\lambda \neq 1)$, ce qui implique aussi que les φ_i sont parallèles à u, donc colinéaires. On a donc montré que $\det(I_n + uv^{\mathsf T}) = 1 + v^{\mathsf T}u$. Si ce déterminant est non nul, on vérifie aisément que $(I_n + uv^{\mathsf T})(I_n - (uv^{\mathsf T})/(1+v^{\mathsf T}u)) = I_n$.

(ii) On écrit $A + uv^{\mathsf T} = A(I_n + A^{-1}uv^{\mathsf T})$ et on applique (i).

(iii) Supposons que $1 + v_1^{\mathsf T}u_1 \neq 0$. Alors $A := I_n + u_1v_1^{\mathsf T}$ est inversible et la formule du déterminant s'obtient en appliquant (ii). Si $1 + v_1^{\mathsf T}u_1 = 0$, la formule reste vraie par continuité.

(iv) Il suffit de montrer que les valeurs propres différentes de 1 de $I_n + UV^{\mathsf T}$ et de $I_m + V^{\mathsf T}U$ sont identiques, comptées avec leur multiplicité. Soit $\lambda \neq 1$ une valeur propre de $I_n + UV^{\mathsf T}$ et $\varphi \in \mathbb{R}^n$ un vecteur propre associé (non nul donc). Alors

$\varphi + UV^\mathsf{T}\varphi = \lambda\varphi$. Donc $V^\mathsf{T}\varphi \neq 0$ et en appliquant V^T à gauche, on voit que λ est valeur propre de $I_m + V^\mathsf{T}U$ avec $V^\mathsf{T}\varphi$ comme vecteur propre associé. En échangeant les rôles de U et V^T, le même raisonnement montre que si $\lambda \neq 1$ est une valeur propre de $I_m + V^\mathsf{T}U$ et si $\psi \in \mathbb{R}^m$ est un vecteur propre associé (non nul donc), alors λ est une valeur propre de $I_n + UV^\mathsf{T}$ avec $U\psi$ comme vecteur propre associé. Il reste à examiner la question de la multiplicité des valeurs propres. Soit $\lambda \neq 1$ une valeur propre commune de $I_n + UV^\mathsf{T}$ et $I_m + V^\mathsf{T}U$. On sait que sa multiplicité pour chacune des matrices (comme racine du polynôme caractéristique) est aussi égale à chacun des membres de l'égalité ci-dessous, qu'il s'agit de démontrer:

$$\dim N\Big(\Big[(1-\lambda)I_n + UV^\mathsf{T}\Big]^p\Big) = \dim N\Big(\Big[(1-\lambda)I_m + V^\mathsf{T}U\Big]^p\Big),$$

pourvu que l'entier p soit assez grand (plus grand que la puissance de $(\xi - \lambda)$ dans le polynôme minimal $\check{p}(\xi)$ annihilant de chaque matrice). En fait, on peut montrer que cette relation a lieu quel que soit $p \geqslant 1$. Soient $\varphi_1, \dots \varphi_k$ des vecteurs linéairement indépendants dans $\mathcal{N}([(1-\lambda)I_n + UV^\mathsf{T}]^p)$. En utilisant la relation $V^\mathsf{T}[(1-\lambda)I_n + UV^\mathsf{T}]^p = [(1-\lambda)I_n + V^\mathsf{T}U]^pV^\mathsf{T}$, on voit que les $V^\mathsf{T}\varphi_i \in N[(1-\lambda)I_n + V^\mathsf{T}U]^p$. Montrons qu'ils sont linéairement indépendants, ce qui prouvera l'inégalité '$\leqslant$' dans l'égalité ci-dessus. Soient donc des scalaires α_i tels que $\sum_i \alpha_i V^\mathsf{T}\varphi_i = 0$. Comme on a $[(1-\lambda)I_n + UV^\mathsf{T}]^p\varphi_i = 0$, on a aussi $0 = [(1-\lambda)I_n + UV^\mathsf{T}]^p \sum_i \alpha_i\varphi_i = [(1-\lambda)I_n + UV^\mathsf{T}]^{p-1}(1-\lambda)\sum_i \alpha_i\varphi_i = (1-\lambda)^p \sum_i \alpha_i\varphi_i$. Mais $\lambda \neq 1$, si bien que $\sum_i \alpha_i\varphi_i = 0$. On en déduit que $\alpha_i = 0$ pour tout i. L'inégalité inverse '$\geqslant$' se démontre de la même manière, en échangeant les rôles de U et de V^T.

Grâce à la première étape, on a $\det(A + URV^\mathsf{T}) = \det(A)\det(I_n + A^{-1}URV^\mathsf{T}) = \det(A)\det(I_m + V^\mathsf{T}A^{-1}UR) = \det(A)\det(R^{-1} + V^\mathsf{T}A^{-1}U)\det(R)$.

On vérifie enfin facilement que $(A + URV^\mathsf{T})^{-1}(A^{-1} - A^{-1}UX^{-1}V^\mathsf{T}A^{-1}) = I_n$.

B.17. Soit T d'ordre n, inversible, triangulaire supérieure et $S := T^{-1}$. La colonne j de S, S^j, est solution du système linéaire $TS^j = e^j$, où e^j est le j-ième vecteur de base de $\mathbb{R}^n$. On en déduit que $S_{ij} = 0$ si $i > j$, c'est-à-dire que S est triangulaire supérieure. Le cas où T est triangulaire inférieure se déduit du résultat précédent parce qu'alors $(T^{-1})^\mathsf{T} = (T^\mathsf{T})^{-1}$ est triangulaire supérieure.

Calcul différentiel

C.1. xxx

C.2. 1) C'est un maximum de fonctions linéaires (donc convexes) et donc une fonction convexe (point 1 de la proposition 3.34). La lipschitzianité de μ peut se déduire de sa convexité (proposition 3.14) ou se montrer directement : soient x et $y \in \mathbb{R}^n$; alors, si $i_0 \in I(y)$, on a

$$\mu(y) - \mu(x) \leqslant y_{i_0} - x_{i_0} \leqslant |y_{i_0} - x_{i_0}| \leqslant \|y - x\|_\infty.$$

Par symétrie, $\mu(x) - \mu(y) \leqslant \|x - y\|_\infty$, si bien que $|\mu(y) - \mu(x)| \leqslant \|y - x\|_\infty$.

2) Si $I(x)$ est un singleton, la fonction max est linéaire dans le voisinage de x, donc Fréchet différentiable. Plus précisément, si $I(x) = \{i\}$, alors $\mu(y) = y_i$ pour y voisin de x et on a alors $\mu'(x)\cdot h = h_i$.

3) Si k et l sont deux indices distincts de $I(x)$, alors $\mu'(x; e^k) = \mu'(x; e^l) = \mu'(x; e^k + e^l) = 1$, si bien que $\mu'(x; \cdot)$ n'est pas linéaire, ce qui empêche la Fréchet différentiabilité (proposition C.4).

4) Soit $h \in \mathbb{R}^n$. Alors, pour $t > 0$ suffisamment petit, on a

$$\begin{aligned} \mu(x+th) &= \max_{i\in I(x)} (x+th)_i \qquad [t>0 \text{ petit et nombre fini d'indices}] \\ &= \mu(x) + t \max_{i\in I(x)} h_i \qquad [x_i = \mu(x) \text{ pour tout } i \in I(x), \text{ puis } t>0]. \end{aligned}$$

On en déduit (C.4).

C.3. xxx

C.4. 1) Pour $h \in \mathbb{E}$, on a $l(x+h) = l(x) + l(h)$. Comme le second terme du membre de droite est linéaire continu en h, l est différentiable en x et $l'(x) \cdot h = l(h)$. Comme $x \mapsto l'(x)$ est constant, $l'' = 0$.

2) Par la bilinéarité de b, on a $b(x_1+h_1, x_2+h_2) = b(x_1, x_2+h_2) + b(h_1, x_2+h_2) = b(x_1,x_2) + b(h_1,x_2) + b(x_1,h_2) + b(h_1,h_2)$. Par continuité de b dont on note $\|b\|$ la norme, le dernier terme peut être majoré en norme comme suit : $\|b(h_1,h_2)\| \leqslant \|b\|\,\|h_1\|\,\|h_2\| \leqslant \|b\|\,(\|h_1\|+\|h_2\|)^2 = o(\|h\|)$, où $h = (h_1,h_2)$. D'autre part, par la bilinéarité et la continuité de b, $h \mapsto b(h_1,x_2) + b(x_1,h_2)$ est linéaire continue, de norme $\|b\|\max(\|x_1\|,\|x_2\|)$. Dès lors b est différentiable en $x = (x_1,x_2)$ et $b'(x)\cdot h = b(h_1,x_2) + b(x_1,h_2)$. La dérivée seconde s'obtient en dérivant $x \mapsto b(x).h$.

C.5. Comme pour $h \in \mathbb{R}^n$ on a le développement suivant

$$f(x+h) = f(x) + \left(\frac{1}{2}A^{\mathsf{T}}x + \frac{1}{2}Ax - b\right)^{\mathsf{T}} h + \frac{1}{2}h^{\mathsf{T}}Ah,$$

le théorème C.18 permet d'écrire pour tout h :

$$\nabla f(x)^{\mathsf{T}}h = f'(x)\cdot h = \left(\frac{1}{2}A^{\mathsf{T}}x + \frac{1}{2}Ax - b\right)^{\mathsf{T}} h \qquad [\text{partie linéaire en } h], \tag{D.65}$$

$$h^{\mathsf{T}}\nabla^2 f(x)h = f''(x)\cdot h^2 = h^{\mathsf{T}}Ah \qquad [\text{partie quadratique en } h \text{ fois } 2]. \tag{D.66}$$

De (D.65), on déduit le gradient. Par ailleurs, par symétrie de $\nabla^2 f(x)$, on a quels que soient h et k :

$$\begin{aligned} (h+k)^{\mathsf{T}}\nabla^2 f(x)(h+k) &= h^{\mathsf{T}}\nabla^2 f(x)h + 2h^{\mathsf{T}}\nabla^2 f(x)k + k^{\mathsf{T}}\nabla^2 f(x)k \\ (h+k)^{\mathsf{T}}A(h+k) &= h^{\mathsf{T}}Ah + h^{\mathsf{T}}(A+A^{\mathsf{T}})k + k^{\mathsf{T}}Ak. \end{aligned}$$

Alors $h^{\mathsf{T}}\nabla^2 f(x)k = \frac{1}{2}h^{\mathsf{T}}(A+A^{\mathsf{T}})k$ par (D.66), d'où l'on déduit la hessienne.

C.6. On a $f(0) = f(1) = (1,0)$ et $f'(x) = 2\pi(-\sin 2\pi x, \cos 2\pi x) \neq (0,0) = f(1) - f(0)$ quel que soit $x \in [0,1]$.

C.7. Comme f est C^1, on a $f(y) - f(x) = \int_0^1 \langle \nabla f(x+t(y-x)), y-x\rangle \,\mathrm{d}t$. En retranchant $\langle \nabla f(x), y-x\rangle$ et en prenant la valeur absolue, on a

$$\begin{aligned}
\big|f(y) - f(x) - \langle \nabla f(x), y - x\rangle\big| &= \left|\int_0^1 \langle \nabla f(x + t(y-x)) - \nabla f(x), y - x\rangle \,\mathrm{d}t\right| \\
&\leqslant \int_0^1 \big|\langle \nabla f(x + t(y-x)) - \nabla f(x), y - x\rangle\big| \,\mathrm{d}t \\
&\leqslant \int_0^1 \big\|\nabla f(x + t(y-x)) - \nabla f(x)\big\| \,\big\|y - x\big\| \,\mathrm{d}t \\
&\leqslant \int_0^1 Lt\|y - x\|^2 \,\mathrm{d}t \\
&= \frac{L}{2}\|y - x\|^2,
\end{aligned}$$

où la première inégalité vient de celle de Fenchel, la seconde de celle de Cauchy-Schwarz et la dernière de la lipschitzianité de ∇f.

C.9. 1) Supposons que $x \in P$. Alors $d_P(x) = 0$ et, comme d_P est différentiable en x, on a pour tout $h \in \mathbb{E}$: $d'_P(x) \cdot h = \lim_{t\downarrow 0} d_P(x + th)/t \geqslant 0$. Dès lors $d'_P(x) = 0$, ce qui contredit l'hypothèse.

2) Comme d_P est non expansive (exercice A.5), $d'_P(x)\cdot h = \lim_{t\downarrow 0}[d_P(x+th) - d_P(x)]/t \leqslant \|h\|$, si bien que $\|\nabla d_P(x)\| \leqslant 1$. Par ailleurs, pour $t \in [0, 1]$, $\bar{x}$ est aussi une projection de $x_t := x + t(\bar{x} - x)$ (en effet, s'il existait un point $\bar{x}_t \in P$ tel que $\|x_t - \bar{x}_t\| < \|x_t - \bar{x}\|$, on aurait $\|x - \bar{x}_t\| \leqslant \|x - x_t\| + \|x_t - \bar{x}_t\| < t\|x - \bar{x}\| + (1-t)\|x - \bar{x}\| = \|x - \bar{x}\|$, ce qui contredirait le fait que $\bar{x}$ est une projection de x sur P). Alors, $d'_P(x)\cdot(\bar{x} - x) = \lim_{t\downarrow 0}[d_P(x_t) - d_P(x)]/t = \lim_{t\downarrow 0}[(1-t)\|\bar{x} - x\| - \|\bar{x} - x\|]/t = -\|\bar{x} - x\|$. Par Cauchy-Schwarz et ce qui précède, on trouve que $\|\bar{x} - x\| = \langle -\nabla d_P(x), \bar{x} - x\rangle \leqslant \|\nabla d_P(x)\| \,\|\bar{x} - x\| \leqslant \|\bar{x} - x\|$. On a donc égalité partout, ce qui implique que $\nabla d_P(x)$ est parallèle à $\bar{x} - x$. On en déduit le résultat.

3) La formule $\nabla d_P(x) = (x - \bar{x})/\|x - \bar{x}\|$ implique que $\bar{x}$ est déterminé de manière unique.

Lexiques

Cette annexe est un petit dictionnaire français-anglais et anglais-français donnant la traduction de quelques mots courants de l'optimisation. Nous avons étiqueté certains mots anglais par (US) lorsque l'orthographe américaine différait de l'anglaise. Comme de coutume, les commentaires (entre parenthèses et en italique) du lexique A-B s'adressent au lecteur de langue A.

Lexique français-anglais

Signification des indicateurs utilisés.

- *adj* : adjectif,
- *nf* : nom féminin,
- *nm* : nom masculin,
- *prép* : préposition,
- *vt* : catégorie transitive d'un verbe.

actif *adj* active ♦ **contrainte active** active constraint ♦ **ensemble ~** (*ensemble d'indices*) active set.
activation *nf* → *méthode*.
admissible *adj* feasible ♦ → *ensemble, point*.
algorithme *nm* algorithm ♦ **~ de BFGS** BFGS algorithm ♦ **~ de la sécante** secant algorithm ♦ **~ de points intérieurs** interior point algorithm ♦ **~ de quasi-Newton** quasi-Newton algorithm ♦ **~ de suivi de chemin** path-following algorithm ♦ **~ de Newton** Newton algorithm ♦ **~ de Newton inexact** inexact Newton algorithm ♦ **~ de Newton semi-lisse** semi-smooth Newton algorithm ♦ **~ de Newton tronqué** truncated Newton algorithm ♦ **~ du gradient**, **de la plus profonde descente** gradient, steepest descent algorithm ♦ **~ du gradient projeté** gradient projection algorithm ♦ **~ du gradient conjugué** conjugate gradient algorithm ♦ **~ du simplexe** simplex algorithm ♦ **~ du simplexe révisé** revised simplex algorithm ♦ **~ OQS (optimisation quadratique successive)** SQP (sequential quadratic programming) algorithm ♦ → *méthode*.
arête *nf* (*d'un polyèdre convexe*) edge.
autoconcordance *nf* self-concordancy, → *fonction*.
base *nf* basis ♦ **~ d'indices** (*en optimisation linéaire*) basis.
calcul différentiel *nm adj* calculus.
cas difficile *nm adj* (*dans les méthodes à régions de confiance*) hard case.
centre *nm* center ♦ **~ analytique** analytic center.
combinaison *nf* combination ♦ **~ affine** affine combination ♦ **~ conique** conical combination ♦ **~ convexe** convex combination.
commande optimale *nf adj* optimal control.
commander *vt* (*un système*) to control.
communication *nf* communication ♦ **~ directe** direct communication ♦ **~ inverse** reverse communication.
compact → *ensemble*.
compatible *adj* → *contrainte*.

complémentarité *nf* complementarity ♦ **~ stricte** strict complementarity ♦ **conditions de ~** complementarity conditions/slackness.
complexité *nf* complexity ♦ **classe de ~** complexity class.
cône *nm* cone ♦ **~ asymptotique** asymptotic cone ♦ **~ critique** critical cone ♦ **~ de récession** recession cone ♦ **~ dual** dual cone ♦ **~ épointé** blunt cone ♦ **~ normal** normal cone ♦ **~ pointé** pointed cone ♦ **~ saillant** salient cone ♦ **~ tangent** tangent cone.
conditionnement *nm* (*le phénomène*) conditioning ♦ (*la valeur qui estime le phénomène*) condition number ♦ **mauvais ~** ill-conditioning.
conjugué *adj* conjugate ♦ → *algorithme*.
contraction *nm* → *fonction contractante*.
contrainte *nf* constraint ♦ **~ active** active constraint ♦ **~ compatible** feasible constraint ♦ **~ d'égalité** equality constraint ♦ **~ de borne** bound constraint ♦ **~ d'inégalité** inequality constraint ♦ **~ inactive** inactive constraint ♦ **~ non réalisable/compatible** infeasible constraint ♦ **~ quasi-réalisable** weakly infeasible constraint ♦ **~ réalisable** feasible constraint.
convexe *nm*, *adj* convex ♦ → *combinaison*, *enveloppe*.
déplacement *nm* step, displacement.
direction *nf* direction ♦ **~ à courbure négative** negative curvature direction ♦ **~ conjuguée** conjugate direction ♦ **~ de descente** descent direction ♦ **~ de montée** ascent direction.
domaine *nm* (*d'une fonction*) domain.
dualité *nf* duality ♦ **saut de ~** duality gap.
égalité *nf* equality.
ensemble *nm* set ♦ **~ actif** (*ensemble d'indices*) active set ♦ **~ admissible** feasible set ♦ **~ compact** compact set ♦ **~ fermé** closed set ♦ **~ de sous-niveau** sublevel set.
enveloppe *nf* hull ♦ **~ affine** (*d'un ensemble*) affine hull ♦ **~ convexe** (*d'un ensemble*) convex hull ♦ **~ supérieure** (*de fonctions*) pointwise supremum, max-function.
erreur *nf* error ♦ **~ amont** backward error ♦ **~ aval** forward error.
face *nf* (*d'un ensemble convexe*) face.
facteur *nm* (*d'un produit*) factor ♦ **~ d'inexactitude** (*dans les algorithmes de Newton inexacts*) forcing term.
famille de problèmes *nf nm* (*en théorie de la complexité*) problem ♦ **~ NP-complets** NP-complete problem ♦ **~ NP-ardus** NP-hard problem.
fermé → *ensemble*.
fonction *nf* function ♦ **~ auto-concordante** self-concordant function ♦ **~ contractante** nonexpansive function ♦ **~ indicatrice** indicator function ♦ **~ marginale** marginal function ♦ **~ strictement contractante** contraction mapping.
gradient *nm* gradient.
hessienne *nf* Hessian (*avec majuscule*).
hyperplan *nm* hyperplane ♦ **~ d'appui** supporting hyperplane.
inégalité *nf* inequality ♦ **~ matricielle linéaire (IML)** linear matrix inequality (LMI).
IML → inégalité matricielle linéaire.
inf-convolution *nf* infimal convolution.
intérieur *nm* (*d'un ensemble*) interior ♦ **~ relatif** relative interior.
lagrangien *nm* Lagrangian (*avec majuscule*) ♦ **~ augmenté** augmented Lagrangian.
matrice *nf* matrix ♦ **~ antisymétrique** skew symmetric matrix ♦ **~ creuse** sparse matrix ♦ **~ symétrique** symmetric matrix.
méthode *nf* method ♦ **~ d'activation (de contraintes)** active set method ♦ **~ de BFGS** BFGS method ♦ **~ de pivotage** pivoting method ♦ **~ de faisceaux** bundle method ♦ → *algorithme*.
minimum *nm* (*un point*) minimum ♦ **~ saillant** sharp minimum.
multiplicateur *nm* multiplier.
négatif *adj* (*pour un nombre réel*) nonpositive ♦ **strictement ~** negative.
optimalité *nf* optimality ♦ **condition d'~** optimality condition.
optimisation *nf* optimisation, (*US*) optimization, programming (*devient obsolète*) ♦ **~ convexe** convex programming ♦ **~ linéaire** linear programming ♦ **~ multicritère** multicriteria optimization ♦ **~ non linéaire** nonlinear programming ♦ **~ quadratique** quadratic programming ♦ **~**

semi-définie positive semi-definite programming.
pas *nm* (*de recherche linéaire*) step-size, stepsize, step-length, steplength.
pénalisation *nf* penalisation, (*US*) penalization.
pivotage *nm* → *méthode*.
pli *nm* (*d'une fonction convexe*) kink.
point *nm* point, (*en optimisation linéaire*) solution ♦ **~ admissible** feasible point, (*en optimisation linéaire*) feasible solution.
point-selle *nm* saddle-point.
polyèdre *nm* polyhedron.
polyédrique *adj* polyhedral.
positif *adj* (*pour un nombre réel*) nonnegative ♦ **strictement ~** positive.
produit *nm* (*dans un réseau*) commodity, (*opération mathématique*) product ♦ **~ cartésien** Cartesian (*avec majuscule*) product ♦ **~ scalaire** inner product ♦ **~ scalaire euclidien** Euclidean (*avec majuscule*) inner product.
proximalité *nf* proximation.
quasi-réalisable *adj* → *contrainte*.
racine *nf* (*d'un polynôme*) root ♦ **~ carrée** (*d'un nombre*) square root.
réalisable *adj* → *contrainte*.
rebroussement *nm* (*en recherche linéaire*) backtracking.
recherche linéaire *nf adj* line-search, linesearch ♦ → *pas*, *rebroussement*.
région de confiance *nf nf* trust region.
saillant *adj* → *minimum*.
semi-continue *adj* (*fonction*) semi-continuous ♦ **~ inférieurement** lower semi-continuous ♦ **~ supérieurement** upper semicontinuous.
semi-continuité *nf* (*d'une fonction*) semicontinuity.
semi-lisse *adj* (*fonction*) semi-smooth.
sommet *nm* (*d'un polyèdre convexe*) vertex ♦ (*en optimisation linéaire*) vertex, basic feasible solution.
sous-différentiel *nm* subdifferential.
sous-niveau → *ensemble*.
suite *nf* sequence.
terme *nm* (*d'une somme*) term.
variable *nf* variable ♦ **~ basique** (*en optimisation linéaire*) basic variable ♦ **~ d'écart** slack variable ♦ **~ non basique** (*en optimisation linéaire*) nonbasic variable.

Lexique anglais-français

active → *constraint*, *method*, *set*.
affine affine, → *combination*, *hull*.
algorithm algorithme ♦ **BFGS ~** algorithme de BFGS ♦ **conjugate gradient ~** algorithme du gradient conjugué ♦ **gradient**, **steepest descent ~** algorithme du gradient, de la plus profonde descente ♦ **gradient projection ~** algorithme du gradient projeté ♦ **inexact Newton ~** algorithme de Newton inexact ♦ **interior point ~** algorithme de points intérieurs ♦ **Newton ~** algorithme de Newton ♦ **path-following ~** algorithme de suivi de chemin ♦ **quasi-Newton ~** algorithme de quasi-Newton ♦ **secant ~** algorithme de la sécante ♦ **semi-smooth Newton ~** algorithme de Newton semi-lisse ♦ **simplex ~** algorithme du simplexe ♦ **SQP (sequential quadratic programming) ~** algorithme OQS (optimisation quadratique successive) ♦ **truncated Newton ~** algorithme de Newton tronqué ♦ → *method*.
backtracking (*in line-search*) rebroussement.
basis base, (*in linear optimization*) base d'indices.
calculus calcul différentiel.
center centre ♦ **analytic ~** centre analytique.
closed → *set*.
combination combinaison ♦ **affine ~** combinaison affine ♦ **conical ~** combinaison conique ♦ **convex ~** combinaison convexe.
commodity (*in a network*) produit.
communication communication ♦ **direct ~** communication directe ♦ **reverse ~** communication inverse.
compact → *set*.
complementarity complémentarité ♦ **~ conditions, slackness** conditions de com-

plémentarité ♦ **strict ~** complémentarité stricte.
complexity complexité ♦ **~ class** classe de complexité.
condition number conditionnement.
conditioning conditionnement ♦ **ill-~** mauvais conditionnement.
conjugate conjugué ♦ → *algorithm*.
cone cône ♦ **asymptotic ~** cône asymptotique ♦ **blunt ~** cône épointé ♦ **critical ~** cône critique ♦ **dual ~** cône dual ♦ **normal ~** cône normal ♦ **pointed ~** cône pointé ♦ **recession ~** cône de récession ♦ **salient ~** cône saillant ♦ **tangent ~** cône tangent.
constraint contrainte ♦ **active ~** contrainte active ♦ **bound ~** contrainte de borne ♦ **equality ~** contrainte d'égalité ♦ **feasible ~** contrainte réalisable *ou* compatible ♦ **inactive ~** contrainte inactive ♦ **inequality ~** contrainte d'inégalité ♦ **infeasible ~** contrainte non réalisable ♦ **weakly infeasible ~** contrainte quasi-réalisable.
contraction (mapping) fonction strictement contractante.
control *vt* (*a system*) commander.
convex convexe ♦ → *combination*, *hull*.
convolution convolution ♦ **infimal ~** inf-convolution.
direction direction ♦ **ascent ~** direction de montée ♦ **conjugate ~** direction conjuguée ♦ **descent ~** direction de descente ♦ **negative curvature ~** direction à courbure négative.
domain (*of a function*) domaine.
duality dualité ♦ **~ gap** saut de dualité.
edge (*of a convex polyhedron*) arête.
equality égalité.
error erreur ♦ **backward ~** erreur amont ♦ **forward ~** erreur aval.
face (*of a convex set*) face.
factor (*of a product*) facteur.
feasible → *constraint, point, set*.
function fonction ♦ **indicator ~** fonction indicatrice ♦ **marginal ~** fonction marginale ♦ **nonexpansive ~** fonction contractante ♦ **self-concordant ~** fonction autoconcordante.
gradient gradient.
hard case (*in trust region methods*) cas difficile.
Hessian hessienne (*no capital letter*).
hull enveloppe (*literally* coque) ♦ **convex ~** (*of a set*) enveloppe convexe ♦ **affine ~** (*of a set*) enveloppe affine.
hyperplane hyperplan ♦ **supporting ~** hyperplan d'appui.
inequality inégalité ♦ **linear matrix ~ (LMI)** inégalité matricielle linéaire (IML).
infeasible → *constraint, point*.
interior (*of a set*) intérieur ♦ **relative ~** intérieur relatif.
kink (*of a convex function*) pli.
Lagrangian lagrangien (*no capital letter*) ♦ **augmented ~** lagrangien augmenté.
LMI → inequality > linear matrix.
line-search, **linesearch** recherche linéaire ♦ → *backtracking*, *step-size*.
matrix matrice ♦ **skew symmetric ~** matrice antisymétrique ♦ **sparse ~** matrice creuse ♦ **symmetric ~** matrice symétrique.
method méthode ♦ **active set ~** méthode d'activation (de contraintes) ♦ **BFGS ~** méthode de BFGS ♦ **bundle ~** méthode de faisceaux ♦ **pivoting ~** méthode de pivotage ♦ → *algorithm*.
minimum *nm* (*a point*) minimum ♦ **sharp ~** minimum saillant.
multiplier multiplicateur.
negative (*for a real number*) strictement négatif ♦ **non~** positif.
nonexpansive → *function*.
optimal control commande optimale.
optimality optimalité ♦ **~ condition** condition d'optimalité.
optimisation, (*US*) **optimization** optimisation ♦ **multicriteria ~** optimisation multicritère.
penalisation, (*US*) **penalization** pénalisation.
pivoting → *method*.
point point ♦ **feasible ~** point admissible ♦ **infeasible ~** point non admissible.
pointwise supremum (*of functions*) enveloppe supérieure.
polyhedral polyédrique.
polyhedron polyèdre.
positive (*for a real number*) strictement positif ♦ **non~** négatif.
problem (*in complexity theory*) (famille de) problème(s) ♦ **NP-complete ~** (famille de) problème(s) NP-complet(s) ♦ **NP-hard ~** (famille de) problème(s) NP-ardu(s).

product (*mathematical operation*) produit ♦ **Cartesian ~** produit cartésien (*no capital letter*) ♦ **inner ~** produit scalaire ♦ **Euclidean inner ~** produit scalaire euclidien (*no capital letter*).
program (*in optimisation*) problème d'optimisation.
programming (*in optimisation*) optimisation ♦ **convex ~** optimisation convexe ♦ **linear ~** optimisation linéaire ♦ **nonlinear ~** optimisation non linéaire ♦ **quadratic ~** optimisation quadratique ♦ **semi-definite ~** optimisation semi-définie positive.
proximation proximalité.
root (*of a polynomial*) racine ♦ **square ~** (*of a number*) racine carrée.
saddle-point point-selle.
self-concordancy autoconcordance, → *function*.
semi-continuity (*of a function*) semi-continuité.
semi-continuous (*function*) semi-continue ♦ **lower ~** semi-continue inférieurement ♦ **upper ~** semi-continue supérieurement.
semi-smooth semi-lisse.
set ensemble ♦ **active ~** (*an index set*) ensemble actif ♦ **closed ~** ensemble fermé ♦ **compact ~** ensemble compact ♦ **feasible ~** ensemble admissible ♦ **sublevel ~** ensemble de sous-niveau.
sequence suite.
sharp → *minimum*.
solution solution, (*in linear optimization*) point admissible ♦ **basic feasible ~** (*in linear optimization*) sommet.
step déplacement ♦ **step-size**, **stepsize**, **step-length**, **steplength** (*in linesearch*) pas.
subdifferential sous-différentiel.
sublevel → *set*.
term (*of a sum*) terme ♦ **forcing ~** (*in inexact Newton algorithms*) facteur d'inexactitude.
trust region région de confiance.
variable variable ♦ **basic ~** (*in linear optimization*) variable basique ♦ **nonbasic ~** (*in linear optimization*) variable non basique ♦ **slack ~** variable d'écart.
vertex (*of a convex polyhedron*) sommet.
weakly infeasible → *constraint*.

Notations

Notations générales

$\forall$, $\exists$, $\exists!$	quantificateurs signifiant respectivement « pour tout », « il existe », « il existe un(e) unique »
α^+, α^-	(avec $\alpha \in \mathbb{R}$) $\alpha^+ = \max(\alpha, 0)$, $\alpha^- = \max(-\alpha, 0)$; donc $\alpha = \alpha^+ - \alpha^-$
$x \geqslant 0$	pour un vecteur $x \in \mathbb{R}^n$, signifie que *toutes* les composantes de x sont positives ($x_1 \geqslant 0, \ldots, x_n \geqslant 0$); de même, $x \leqslant 0$ signifie que $(-x) \geqslant 0$
$x > 0$	pour un vecteur $x \in \mathbb{R}^n$, signifie que *toutes* les composantes de x sont strictement positives ($x_1 > 0, \ldots, x_n > 0$); de même, $x < 0$ signifie que $(-x) > 0$
$\{x_k\}$	(avec x_k dans un ensemble $\mathbb{E}$) suite de $\mathbb{E}$, c'est-à-dire application de $\mathbb{N}$ dans $\mathbb{E}$ (ou son image)
$x_k \to x$	la suite $\{x_k\}$ converge vers x
$\alpha_k \downarrow \alpha$	la suite de réels $\{\alpha_k\}$ converge vers α, par des valeurs strictement supérieures (pour tout k : $\alpha_k > \alpha$)
$\alpha_k \uparrow \alpha$	la suite de réels $\{\alpha_k\}$ converge vers α, par des valeurs strictement inférieures (pour tout k : $\alpha_k < \alpha$)
$n!$	factorielle de l'entier n, $n! = n(n-1)(n-2)\cdots 2$
$\binom{n}{k}$	combinaison sans répétition de n éléments pris k par k (n et k sont entiers et $k \leqslant n$), $\binom{n}{k} = n!/(k!\,(n-k)!)$

Ensembles particuliers

$\mathrm{AC}(\mathbb{E})$	ensemble des fonctions autoconcordantes sur un espace vectoriel $\mathbb{E}$
B_r	boule ouverte de centre 0 et de rayon r
$\bar{B}_r$	boule fermée de centre 0 et de rayon r
$B(x, r)$	$= x + B_r$ boule ouverte de centre x et de rayon r
$\bar{B}(x, r)$	$= x + \bar{B}_r$ boule fermée de centre x et de rayon r
$\mathrm{BAC}(\mathbb{E})$	ensemble des (fonctions) barrières autoconcordantes sur un espace vectoriel $\mathbb{E}$
$\mathrm{Conv}(\mathbb{E})$	ensemble des fonctions d'un espace vectoriel $\mathbb{E}$ dans $\mathbb{R} \cup \{+\infty\}$, convexes, non identiquement égales à $+\infty$
$\mathrm{C}\overline{\mathrm{onv}}(\mathbb{E})$	ensemble des fonctions d'un espace vectoriel $\mathbb{E}$ dans $\mathbb{R} \cup \{+\infty\}$, convexes, fermées, non identiquement égales à $+\infty$
Δ_n	simplexe unitaire de $\mathbb{R}^n$
$[\alpha, \beta]$, $[\alpha, \beta[$, $]\alpha, \beta]$, $]\alpha, \beta[$, avec $\alpha \leqslant \beta$ dans $\mathbb{R}$	intervalles de $\mathbb{R}$ (fermé, semi-ouvert à droite, semi-ouvert à gauche, ouvert)

$[n_1 : n_2]$, avec $n_1 \leqslant n_2$ dans $\mathbb{N}$	$= [n_1, n_2] \cap \mathbb{N}$, intervalle de $\mathbb{N}$ formé des entiers n_1, $\ldots$, n_2
$\mathcal{L}(\mathbb{E}, \mathbb{F})$	ensemble des applications linéaires continues d'un espace vectoriel topologique $\mathbb{E}$ dans un autre espace vectoriel topologique $\mathbb{F}$
$\mathbb{N}$	ensemble des nombres entiers (naturels), $\mathbb{N} := \{0, 1, 2, \ldots\}$
$\mathbb{N}^*$	ensemble des nombres entiers non nuls, $\mathbb{N}^* := \mathbb{N} \setminus \{0\} = \{1, 2, 3, \ldots\}$
$\mathcal{P}(\mathbb{E})$	ensemble des parties d'un ensemble $\mathbb{E}$
$\mathbb{R}$	ensemble des nombres réels
$\mathbb{R}_+$	ensemble des nombres réels positifs $\{t \in \mathbb{R} : t \geqslant 0\}$
$\mathbb{R}_-$	ensemble des nombres réels négatifs $\{t \in \mathbb{R} : t \leqslant 0\} = -\mathbb{R}_+$
$\mathbb{R}_+^*$, $\mathbb{R}_{++}$	ensemble des nombres réels strictement positifs $\{t \in \mathbb{R} : t > 0\}$
$\mathbb{R}_-^*$, $\mathbb{R}_{--}$	ensemble des nombres réels strictement négatifs $\{t \in \mathbb{R} : t < 0\} = -\mathbb{R}_+^*$
$\overline{\mathbb{R}}$	la « droite achevée », $\mathbb{R} \cup \{-\infty, +\infty\}$
$\mathbb{R}^n$	ensemble des n-uplets $(x_1, \ldots, x_n)$ formés des nombres réels x_1, $\ldots$, x_n
$\mathbb{R}_+^n$	orthant positif de $\mathbb{R}^n$, $\{x \in \mathbb{R}^n : x \geqslant 0\}$
$\mathbb{R}_{++}^n$	intérieur de l'orthant positif de $\mathbb{R}^n$, $\{x \in \mathbb{R}^n : x > 0\}$
$\mathbb{R}_{\leqslant}^n$	simplexe ordonné de $\mathbb{R}^n$, $\{x \in \mathbb{R}^n : x_1 \leqslant \cdots \leqslant x_n\}$
$\mathbb{R}_{\triangledown}^n$	cône du second ordre de $\mathbb{R}^n$, $\{x \in \mathbb{R}^n : x_1^2 + \cdots + x_{n-1}^2 \leqslant x_n^2,\ x_n \geqslant 0\}$
$\mathcal{S}^n$	ensemble des matrices d'ordre n symétriques
$\mathcal{S}_+^n$	cône de $\mathcal{S}^n$ formé des matrices d'ordre n symétriques semi-définies positives
$\mathcal{S}_{++}^n$	cône de $\mathcal{S}^n$ formé des matrices d'ordre n symétriques définies positives
$\mathcal{S}_-^n$	cône de $\mathcal{S}^n$ formé des matrices d'ordre n symétriques semi-définies négatives, $-\mathcal{S}_+^n$

Opérations sur les ensembles

Ci-dessous, P, P_1, P_2 désignent des ensembles quelconques et C désigne une partie convexe d'un espace vectoriel $\mathbb{E}$.

∂P	frontière d'un ensemble P
P°, $\operatorname{int} P$	intérieur d'un ensemble P
P°_r}, $\operatorname{intr} P$	intérieur relatif d'un ensemble P
$\overline{P}$, $\operatorname{adh} P$	adhérence d'un ensemble P
P^+	cône dual d'un ensemble P
P^-	cône dual négatif, $P^- := -P^+$
C^∞	cône asymptotique d'un ensemble convexe fermé C
$C^\infty(x)$	cône asymptotique d'un ensemble convexe C en $x \in C$
$\mathrm{N}_x P$, $\mathrm{N}_P(x)$	cône normal à l'ensemble P en x
$\mathrm{T}_x P$, $\mathrm{T}_P(x)$	cône tangent à l'ensemble P en x
$\mathrm{T}_x^a P$, $\mathrm{T}_P^a(x)$	cône des directions admissibles à l'ensemble P en x
$\operatorname{aff} P$	enveloppe affine d'un ensemble P
$\operatorname{cone} P$	enveloppe conique d'un ensemble P
$\operatorname{co} P$	enveloppe convexe d'un ensemble P
$\overline{\operatorname{co}} P$	enveloppe convexe fermée d'un ensemble P
$\operatorname{ext}(C)$	ensemble des points extrêmes d'un ensemble convexe C
$P_1 + P_2$	somme (de Minkowski) des ensembles P_1 et P_2 d'un espace vectoriel, c'est l'ensemble $\{x + y : x \in P_1,\ y \in P_2\}$

$P_1 - P_2$	somme (de Minkowski) des ensembles P_1 et $-P_2$ d'un espace vectoriel, c'est l'ensemble $\{x - y : x \in P_1,\ y \in P_2\}$
αP	produit du scalaire $\alpha \in \mathbb{R}$ et de l'ensemble P d'un espace vectoriel, c'est l'ensemble $\{\alpha x : x \in P\}$

Espaces vectoriels et calcul matriciel

Ci-dessous, A désigne une matrice.

$\|\cdot\|$	norme
$\|\cdot\|_D$	norme duale de la norme $\|\cdot\|$ pour un produit scalaire donné
$\|\cdot\|_0$	compteur de composante non nulle d'un vecteur
$\|\cdot\|_p$	norme ℓ_p, pour $p \in [1, \infty]$; voir (A.5a)-(A.5b)
$\langle \cdot, \cdot \rangle$	produit scalaire
A^*	adjointe de A, définie par la relation $\langle Au, v\rangle = \langle u, A^* v\rangle$ pour tout u, v
A^{-*}	adjointe de l'opérateur A^{-1} (supposé exister) ou inverse de A^* (ce sont les mêmes opérateurs)
A^T	transposée de A, qui est l'adjointe pour le produit scalaire euclidien : $(A^\mathsf{T})_{ij} = A_{ji}$
$A^{-\mathsf{T}}$	transposée de la matrice A^{-1} (supposée exister) ou inverse de A^T (ce sont les mêmes matrices)
$A \succcurlyeq 0$	la matrice A est symétrique semi-définie positive, c.-à-d., $A \in \mathcal{S}^n_+$
$A \succ 0$	la matrice A est symétrique définie positive, c.-à-d., $A \in \mathcal{S}^n_{++}$
E^{ij}	matrice de base de l'espace vectoriel des matrices d'un type donné : son élément (i, j) vaut 1 et les autres valent zéro
$\mathcal{N}(A)$	noyau de A (ensemble des vecteurs x tels que $Ax = 0$)
$\mathcal{R}(A)$	image de A (ensemble des vecteurs de la forme Ax)
$\lambda(A)$	ensemble ouvecteur (suivant le contexte) des *valeurs propres* de A
$\mathcal{S}^n$	ensemble des matrices d'ordre n symétriques
$\mathcal{S}^n_+$	cône de $\mathcal{S}^n$ formé des matrices d'ordre n symétriques semi-définies positives
$\mathcal{S}^n_{++}$	cône de $\mathcal{S}^n$ formé des matrices d'ordre n symétriques définies positives
$\sigma(A)$	vecteurs des valeurs singulières non nulles de A
$\det A$	déterminant d'une matrice carrée A
$\dim \mathbb{E}$	dimension d'un espace vectoriel $\mathbb{E}$ (nombre maximal de vecteurs linéairement indépendants dans $\mathbb{E}$)
e^i	vecteur de base de $\mathbb{R}^n$: sa composante i vaut 1 et les autres valent zéro
$\operatorname{tr} A$	trace d'une matrice carrée A, $\operatorname{tr} A = \sum_i A_{ii}$
$\operatorname{vect}\{x_1, \ldots, x_p\}$	sous-espace vectoriel engendré par les vecteurs $x_1, \ldots, x_p$

Fonctions particulières et opérations sur les fonctions

$f : \mathbb{E} \to \mathbb{F}$	fonction entre deux ensembles $\mathbb{E}$ et $\mathbb{F}$
$f : \mathbb{E} \multimap \mathbb{F}$	multifonction entre deux ensembles $\mathbb{E}$ et $\mathbb{F}$ (si $x \in \mathbb{E}$, $f(x)$ est un sous-ensemble de $\mathbb{F}$)
$\mathcal{I}_P$	indicatrice d'un ensemble P
$f'(x; d)$	dérivée directionnelle d'une fonction f en x dans la direction d

$f'_+(x)$	dérivée à droite en $x \in \mathbb{R}$ d'une fonction f d'une variable réelle, $f'(x;1)$
$f'_-(x)$	dérivée à gauche en $x \in \mathbb{R}$ d'une fonction f d'une variable réelle, $f'(x;-1)$
$\partial f(x)$	sous-différentiel en x d'une fonction $f \in \mathrm{Conv}(\mathbb{E})$
$f'(x)$	dérivée (de Fréchet ou de Gâteaux) d'une fonction f en x
$f'(x) \cdot d$	valeur dans la direction d de la dérivée (de Fréchet ou de Gâteaux) d'une fonction f en x
$\nabla f(x)$	gradient en x d'une fonction f
$\nabla^2 f(x)$	hessienne en x d'une fonction f
f^∞	fonction asymptotique d'une fonction $f \in \mathrm{C}\overline{\mathrm{onv}}(\mathbb{E})$
$\mathrm{d}(x,y)$	distance entre les points x et y dans un espace métrique
d_P, $\mathrm{dist}(\cdot,P)$	distance à l'ensemble P, $\mathrm{d}_P(x) := \inf\{\mathrm{d}(x,y) : y \in P\}$
$\mathrm{dom}\, f$	domaine d'une fonction $f : \mathbb{E} \to \overline{\mathbb{R}}$
$\mathrm{dom}\, T$	domaine d'une multifonction $T : \mathbb{E} \multimap \mathbb{F}$
$\mathrm{epi}\, f$	épigraphe d'une fonction $f : \mathbb{E} \to \overline{\mathbb{R}}$
$\mathrm{epi}_s\, f$	épigraphe stricte d'une fonction $f : \mathbb{E} \to \overline{\mathbb{R}}$
$\mathcal{G}(T)$	graphe d'une multifonction $T : \mathbb{E} \multimap \mathbb{F}$
P_f	opérateur proximal associé à une fonction $f \in \mathrm{C}\overline{\mathrm{onv}}(\mathbb{E})$
$g \circ f$	fonction de $\mathbb{E}$ dans $\mathbb{G}$, obtenue par composition de la fonction $f : \mathbb{E} \to \mathbb{F}$ et de la fonction $g : \mathbb{F} \to \mathbb{G}$ ($\mathbb{E}$, $\mathbb{F}$ et $\mathbb{G}$ sont des ensembles quelconques); $(g \circ f)(x) := g(f(x))$
$f \uplus g$	inf-convolution de deux fonctions f et $g : \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ définies sur le même espace vectoriel $\mathbb{E}$
$f \veebar A$	inf-image de $f : \mathbb{E} \to \overline{\mathbb{R}}$ sous l'application linéaire $A : \mathbb{E} \to \mathbb{F}$ ($\mathbb{E}$ et $\mathbb{F}$ sont des espaces vectoriels)

Problèmes d'optimisation

$I^0(x)$, I^0_*	ensemble des indices des contraintes d'inégalité actives en x et en x_*
I^{00}_*	ensemble des indices des contraintes d'inégalité faiblement actives en x_*
I^{0+}_*	ensemble des indices des contraintes d'inégalité fortement actives en x_*
$I^+(x)$	ensemble des indices des contraintes d'inégalité strictement positives en x
$\Lambda(x_*) = \Lambda_*$	ensemble des multiplicateurs optimaux associés à un point stationnaire x_*
$S(P)$	ensemble des solutions du problème (P)
$\mathrm{val}(P)$	valeur optimale du problème (P)

Bibliographie

[1] J. Abadie (1967). On the Kuhn-Tucker theorem. In J. Abadie, éditeur, *Nonlinear Programming*, pages 19–36. North-Holland Publishing Company, Amsterdam. 241

[2] C.M. Ablow, G. Brigham (1955). An analog solution of programming problems. *Journal of the Operations Research Society of America*, 3, 388–394. 479, 501

[3] H. Akaike (1959). On a successive transformation of probability distribution and its application to the analysis of the optimum gradient method. *Ann. Inst. Statist. Math. Tokyo*, 11, 1–16. 322

[4] G.E. Alefeld, A. Frommer, G. Heindl, J. Mayer (2004). On the existence theorems of Kantorovich, Miranda and Borsuk. *Electronic Transactions on Numerical Analysis*, 17, 102–111. 429

[5] G.E. Alefeld, F.A. Potra, Z. Shen (2001). On the existence theorems of Kantorovich, Moore and Miranda. *Computing*, 15, 21–28. 429

[6] V. Alexeev, V. Tikhomirov, S. Fomine (1982). *Commande Optimale*. Mir, Moscou. 241

[7] F. Alizadeh (1995). Interior point methods in semidefinite programming with applications to combinatorial optimization. *SIAM Journal on Optimization*, 5, 13–51. 743

[8] F. Alizadeh (2012). An introduction to formally real Jordan algebras and their applications in optimization. In M.F. Anjos, J.B. Lasserre, éditeurs, *Handbook on Semidefinite, Conic and Polynomial Optimization*, International Series in Operations Research and Management Science 166, pages 297–337. Springer. 742

[9] F. Alizadeh, J.-P.A. Haeberly, M.L. Overton (1998). Primal-dual interior-point methods for semidefinite programming: convergence rates, stability and numerical results. *SIAM Journal on Optimization*, 8, 746–768. 742

[10] E.L. Allgower, K. Georg (1990). *Numerical Continuation Methods – An Introduction*. Springer Series in Computational Mathematics 13. Springer-Verlag. 428

[11] E.L. Allgower, K. Georg (1993). Continuation and path following. In *Acta Numerica 1993*, Tome 2, pages 1–64. Cambridge University Press. 428

[12] E.D. Andersen, K.D. Andersen (1999). The MOSEK interior-point optimizer for linear programming: an implementation of the homogeneous algorithm. In S. Zhang, H. Frenk, C. Roos, T. Terlaky, éditeurs, *High Performance Optimization Techniques*, pages 197–232. Kluwer Academic Publishers, Dordrecht, The Netherlands. 678

[13] R. Andreani, E.G. Birgin, J.M. Martínez, M.L. Schuverdt (2007). On augmented Lagrangian methods with general lower-level constraints. *SIAM Journal on Optimization*, 18, 1286–1302. [doi]. 498

[14] N. Andrei (2020). *Nonlinear Conjugate Gradient Methods for Unconstrained Optimization*, Springer Optimization and Its Applications, Tome 158. Springer. 366

[15] M.F. Anjos, J.B. Lasserre (2012). *Handbook on Semidefinite, Conic and Polynomial Optimization*, International Series in Operations Research & Management Science, Tome 166. Springer. [doi]. 743

[16] K.M. Anstreicher, M.H. Wright (2000). A note on the augmented Hessian when the reduced Hessian is semidefinite. *SIAM Journal on Optimization*, 11, 243–253. 786

[17] A.S. Antipin (1992). Controlled proximal differential systems for solving saddle problems. *Differential Equations*, 28, 1498–1510. 551

[18] I.K. Argyros (2008). *Convergence and Applications of Newton-type Iterations*. Springer. 431

[19] I.K. Argyros, S. Hilout (2010). A Kantorovich-type convergence analysis of the Newton-Josephy method for solving variational inequalities. *Numerical Algorithms*, 55, 447–466. 579

[20] P. Armand, J.Ch. Gilbert, S. Jan-Jégou (2002). A BFGS-IP algorithm for solving strongly convex optimization problems with feasibility enforced by an exact penalty approach. *Mathematical Programming*, 92, 393–424. [doi]. 609

[21] L. Armijo (1966). Minimization of functions having Lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics*, 16, 1–3. [journal]. 307, 322

[22] K.J. Arnold, J.V. Beck (1977). *Parameter Estimation in Engineering and Science*. John Wiley & Sons, New York. 714

[23] K.J. Arrow, F.J. Gould, S.M. Howe (1973). A general saddle point result for constrained optimization. *Mathematical Programming*, 5, 225–234. [doi]. 504, 551

[24] K.J. Arrow, R.M. Solow (1958). Gradient methods for constrained maxima with weakened assumptions. In K.J. Arrow, L. Hurwicz, H. Uzawa, éditeurs, *Studies in Linear and Nonlinear Programming*, pages 166–176. Stanford University Press, Standford, Calif. 504

[25] J.-P. Aubin, I. Ekeland (1984). *Applied Nonlinear Analysis*. John Wiley & Sons, New York. 145, 242

[26] A. Auger, O. Teytaud (2007). Continuous lunches are free plus the design of optimal optimization algorithms. *Algorithmica*. 6

[27] A. Auslender, R. Cominetti, M. Haddou (1997). Asymptotic analysis for penalty and barrier methods in convex and linear programming. *Mathematics of Operations Research*, 22, 43–62. 100

[28] A. Auslender, M. Teboulle (2000). Lagrangian duality and related multiplier methods for variational inequality problems. *SIAM Journal on Optimization*, 10, 1097–1115. 552

[29] A. Auslender, M. Teboulle (2003). *Asymptotic Cones and Functions in Optimization and Variational Inequalitites*. Springer Monographs in Mathematics. Springer, New York. 9, 15, 16, 28, 65, 145

[30] A. Auslender, M. Teboulle, S. Ben-Tiba (1999). A logarithmic-quadratic proximal method for variational inequalitites. *Computational Optimization and Applications*, 12, 31–40. 504, 552

[31] A. Auslender, M. Teboulle, S. Ben-Tiba (1999). Interior proximal and multiplier methods based on second order homogeneous functionals. *Mathematics of Operations Research*, 24, 645–668. 504

[32] David Avis, Komei Fukuda (1992). A pivoting algorithm for convex hulls and vertex enumeration of arrangements and polyhedra. *Discrete & Computational Geometry*, 8(3), 295–313. [doi]. 45

[33] J.B. Baillon (2002). Communication personnelle. 609

[34] M.L. Balinski, A. Russakoff (1972). Some properties of the assignment polytope. *Mathematical Programming*, 3, 257–258. 624

[35] R.G. Baraniuk (2007). Compressive sensing [lecture notes]. *IEEE Signal Processing Magazine*, 24(4), 118–121. 499

[36] V. Barbu, T. Precupanu (1975). *Convexity and Optimization in Banach Spaces*. Editura Academiei, Bucarest. 145, 242

[37] Y. Bard (1974). *Nonlinear Parameter Estimation*. Academic. 714

[38] J. Barzilai, J.M. Borwein (1988). Two-point step size gradient methods. *IMA Journal of Numerical Analysis*, 8, 141–148. [doi]. 322

[39] J.-L. Beauvois, R.-V. Joule (1987). *Petit traité de manipulation à l'usage des honnêtes gens*. Presses Universitaires de Grenoble. 473
[40] M. Bellare, P. Rogaway (1995). The complexity of approximating a nonlinear program. *Mathematical Programming*, 69, 4129–441. [doi]. 733
[41] R. Bellman, K. Fan (1963). On systems of linear inequalities in Hermitian matrix variables. In *Proceedings of Symposia in Pure Mathematics, Vol. 7*. AMS. 743
[42] R.E. Bellman, R.R. Kalaba, J. Lockett (1966). *Numerical Inversion of the Laplace Transform*. Elsevier. 323
[43] I. Ben Gharbia, J.Ch. Gilbert (2012). Nonconvergence of the plain Newton-min algorithm for linear complementarity problems with a P-matrix. *Mathematical Programming*, 134, 349–364. [doi]. 412
[44] A. Ben-Tal, A. Nemirovski (2001). *Lectures on Modern Convex Optimization – Analysis, Algorithms, and Engineering Applications*. MPS-SIAM Series on Optimization 2. SIAM. 616, 679, 725, 785
[45] J.F. Benders (1962). Partitioning procedures for solving mixed-variables programming problems. *Numerische Mathematik*, 4, 238–252. [doi]. 552
[46] J.M. Bennet (1965). Triangular factors of modified matrices. *Numerische Mathematik*, 7, 217–221. 785
[47] Commandant Benoît (1924). Note sur une méthode de résolution des équations normales provenant de l'application de la méthode des moindres carrés à un système d'équations linéaires en nombre inférieur à celui des inconnues. Aplication de la méthode à la résolution d'un système défini d'équations linéaires (Procédé du Commandant Cholesky). *Bulletin Géodésique* (Toulouse), 2, 67–77. 785
[48] C. Berg (1987). The multidimensional moment problem and semigroups. In H.J. Landau, éditeur, *Moment in Mathematics*, pages 110–124. AMS, Providence, RI. 729
[49] A. Berman, N. Shaked-Monderer (2003). *Completely Positive Matrices*. World Scientific, River Edge, NJ, USA. 203
[50] D.P. Bertsekas (1976). On the Goldstein-Levitin-Polyak gradient projection method. *IEEE Transactions on Automatic Control*, 21, 174–184. 464, 470
[51] D.P. Bertsekas (1976). Multiplier methods: a survey. *Automatica*, 12, 133–145. 504
[52] D.P. Bertsekas (1982). Projected Newton methods for optimization problems with simple constraints. *SIAM Journal on Control and Optimization*, 20, 221–246. [doi]. 465
[53] D.P. Bertsekas (1982). *Constrained Optimization and Lagrange Multiplier Methods*. Academic Press. 498
[54] D.P. Bertsekas (1995). *Nonlinear Programming*. Athena Scientific. 16, 141, 553
[55] D.P. Bertsekas (1999). *Nonlinear Programming* (seconde édition). Athena Scientific. 498
[56] D.P. Bertsekas (2015). *Convex Optimization Algorithms*. Athena Scientific. 16
[57] D. Bertsimas, J.N. Tsitsiklis (1997). *Introduction to linear optimization*. Athena Scientific, Belmont, Massachusetts. 622, 643
[58] L.T. Biegler, O. Ghattas, M. Heinkenschloss, B. van Bloemen Waanders, éditeurs (2003). *Large-Scale PDE-Constrained Optimization*. Springer, Berlin. 16
[59] E.G. Birgin, R.A. Castillo, J.M. Martínez (2005). Numerical comparison of augmented Lagrangian algorithms for nonconvex problems. *Computational Optimization and Applications*, 31, 31–55. 504
[60] G. Birkhoff (1946). Tres observaciones sobre el algebra lineal. *Universidad Nacionale Tucamán Revista*, 5, 147–151. 67
[61] Å. Björck (1996). *Numerical Methods for Least Squares Problems*. SIAM Publication, Philadelphia. 713
[62] Å. Björck (2004). The calculation of linear least squares problems. In *Acta Numerica 2004*, Tome 13, pages 1–53. Cambridge University Press. 713

[63] Å. Björck, T. Elfving, Z. Strakos (1998). Stability of conjugate gradient and Lanczos methods for linear least squares problems. *SIAM Journal on Matrix Analysis and Applications*, 19, 720–736. 686

[64] R.G. Bland (1977). New finite pivoting rules for the simplex method. *Mathematics of Operations Research*, 2, 103–107. [doi]. 639

[65] G. Blekherman, P.A. Parrilo, R.R. Thomas (2013). *Semidefinite Optimization and Convex Algebraic Geometry*. MOS-SIAM Series on Optimization. SIAM and MPS, Philadelphia. [doi]. 20

[66] L. Blum, F. Cucker, M. Shub, S. Smale (1988). *Complexity and Real Computation*. Springer Verlag. 291

[67] L. Blum, M. Shub, S. Smale (1988). On a theory of computations over the real numbers: NP-completeness, recursive functions and universal machines. In *Proceedings of the 29th Symp. Foundations of Computer Science*, pages 387–397. 260

[68] P.T. Boggs, R.H. Byrd (2019). Adaptive, limited-memory BFGS algorithms for unconstrained optimization. *SIAM Journal on Optimization*, 29(2), 1282–1299. [doi]. 454

[69] I.M. Bomze (2012). Copositive optimization – Recent developments and applications. *European Journal of Operations Research*, 216, 509–520. [doi]. 20, 582

[70] J.F. Bonnans (1989). Asymptotic admissibility of the unit stepsize in exact penalty methods. *SIAM Journal on Control and Optimization*, 27, 631–641. 506

[71] J.F. Bonnans (1989). Local study of Newton type algorithms for constrained problems. In S. Dolecki, éditeur, *Optimization*, Lecture Notes in Mathematics 1405, pages 13–24. Springer. 569

[72] J.F. Bonnans (1994). Local analysis of Newton-type methods for variational inequalities and nonlinear programming. *Applied Mathematics and Optimization*, 29, 161–186. [doi]. 569, 579

[73] J.F. Bonnans, J.Ch. Gilbert, C. Lemaréchal, C. Sagastizábal (2006). *Numerical Optimization – Theoretical and Practical Aspects* (seconde édition). Universitext. Springer Verlag, Berlin. [authors] [editor] [doi]. 16, 322, 569, 579, 678

[74] J.F. Bonnans, A. Shapiro (2000). *Perturbation Analysis of Optimization Problems*. Springer Verlag, New York. 145, 159, 242, 791, 802

[75] K.-H. Borgwardt (1982). The average number of pivot steps required by the simplex-method is polynomial. *Z. Oper. Res.*, (26), A157–A177. 642

[76] K.-H. Borgwardt (1987). *The Simplex Method: A Probabilistic Analysis*. Springer, Berlin. 642

[77] J.M. Borwein (2014). A very complicated proof of the minimax theorem. *Minimax Theory and its Applications*, 1(1). 809

[78] J.M. Borwein, A.S. Lewis (2000). *Convex Analysis and Nonlinear Optimization – Theory and Examples*. CMS Books in Mathematics 3. Springer, New York. 105, 145, 839

[79] J.M. Borwein, J.D. Vanderwerff (2010). *Convex Functions – Constructions, Characterizations and Counterexamples*. Encyclopedia of Mathematics and Its Applications 109. Cambridge University Press. 145

[80] J.M. Borwein, Q.J. Zhu (2010). *Techniques of Variational Analysis*. Computational Management Science. Springer Science+Business Media, Inc., Berlin. 322

[81] A. Bouaricha, R.B. Schnabel (1998). Tensor methods for large sparse systems of nonlinear equations. *Mathematical Programming*, 83, 377–400. 430

[82] S. Boulmier (2019). *Optimisation globale avec LocalSolver*. Thèse de doctorat, Laboratoire Jean Kuntzmann, Université de Grenoble, Grenoble, France. 555, 556

[83] N. Bourbaki (1971). *Éléments de Mathématiques – Topologie Générale*. Hermann, Paris. Réédité par Springer-Verlag en 2007. 103

[84] S. Boyd, L. El Ghaoui, E. Feron, V. Balikrishnan (1994). *Linear Matrix Inequalities in System and Control Theory*. SIAM, Philadelphia, PA, USA. 725

[85] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein (2010). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3, 1–122. [doi]. 499

[86] S. Boyd, L. Vandenberghe (2004). *Convex Optimization*. Cambridge University Press. 16, 715

[87] C.B. Boyer (1968). *A History of Mathematics*. Princeton University Press, Princeton, New Jersey. 241

[88] D. Bremner, K. Fukuda, A. Marzetta (1998). Primal-dual methods for vertex and facet enumeration. *Discrete & Computational Geometry*, 20(3), 333–357. [doi]. 45

[89] C. Brezinski (2006). The life and work of André Cholesky. *Numerical Algorithms*, 43, 279–288. 785

[90] H. Brézis (1973). *Opérateurs Maximaux Monotones et semi-groupes de contractions dans les espaces de Hilbert*. Mathematical Studies 5. North-Holland, Amsterdam. 145, 551

[91] H. Brézis (1983). *Analyse Fonctionnelle Appliquée*. Masson, Paris. 8, 16, 53, 145, 216, 756, 785

[92] K.W. Brodlie, A.R. Gourlay, J. Greenstadt (1973). Rank-one and rank-two corrections to positive definite matrices expressed in product form. *Journal of the Institute of Mathematics and its Applications*, 11, 73–82. 440

[93] A. Brondsted (1964). Conjugate convex functions in topological vector spaces. *Matematiskfysiske Meddelelser udgivet af det Kongelige Danske Videnskabernes Selskab*, 34, 1–26. 145

[94] P.N. Brown (1991). A theoretical comparison of the Arnoldi and GMRES algorithms. *SIAM Journal on Scientific and Statistical Computing*, 12, 58–78. 361

[95] P.N. Brown, Y. Saad (1990). Hybrid Krylov methods for nonlinear systems of equations. *SIAM Journal on Scientific and Statistical Computing*, 11, 450–481. 426

[96] C.G. Broyden (1970). The convergence of a class of double rank minimization algorithms, part I. *Journal of the Institute of Mathematics and its Applications*, 6, 76–90. 440

[97] R.E. Bruck, S. Reich (1977). Nonexpansive projections and resolvents of accretive operators in Banach spaces. *Houston Journal of Mathematics*, 3, 459–470. 499

[98] J.V. Burke (1989). A sequential quadratic programming method for potentially infeasible mathematical programs. *Journal of Mathematical Analysis and Applications*, 139, 319–351. [doi]. 557

[99] J.V. Burke (1991). An exact penalization viewpoint of constrained optimization. *SIAM Journal on Control and Optimization*, 29, 968–998. 473, 505

[100] J.D. Buys (1972). *Dual Algorithms for Constrained Optimization*. Thèse de doctorat, Rijksuniversiteit te Leiden, Leiden, The Netherlands. 504

[101] R.H. Byrd, J.Ch. Gilbert, J. Nocedal (2000). A trust region method based on interior point techniques for nonlinear programming. *Mathematical Programming*, 89, 149–185. [doi]. 679

[102] R.H. Byrd, P. Lu, J. Nocedal, C.Y. Zhu (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(6), 1190–1208. 454

[103] R.H. Byrd, J. Nocedal, R.B. Schnabel (1994). Representations of quasi-Newton matrices and their use in limited memory methods. *Mathematical Programming*, 63, 129–156. 454

[104] P.H. Calamai, J.J. Moré (1987). Projected gradient methods for linearly constrained problems. *Mathematical Programming*, 39, 93–116. [doi]. 470

[105] E.J. Candès, T. Tao (2005). Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(11), 4203–4215. [doi]. 706

[106] C. Carathéodory (1907). Über den Variabilitätsbereich der Koeffizienten von Potenzreihen, die gegebene Werte nicht annehmen. *Mathematische Annalen*, 64, 95–115. 26

[107] C.W. Carroll (1961). The created response surface technique for optimizing nonlinear restrained systems. *Operations Research*, 9, 169–184. 487, 488

[108] J.L. Casti (1996). *Great Theories of 20th-Century Mathematics–and Why They Matter*. John Wiley & Sons, New York. 510

[109] A.L. Cauchy (1847). Méthode générale pour la résolution des systèmes d'équations simultanées. *Comptes Rendus de l'Académie des Sciences de Paris*, t25, 536–538. 297

[110] E. Cavazzuti, J. Morgan (1983). Well-posed saddle point problems. In *Optimization: theory and algorithms (Confolant, 1981)*, Lecture Notes in Pure and Mathematics 86, pages 61–76. Dekker, New York. 552

[111] Y. Censor, S. Zenios (1992). Proximal minimization algorithm with D-functions. *Journal of Optimization Theory and Applications*, 73, 451–464. 504

[112] J.-L. Chabert, Évelyne Barbin, Michel Guillemot, Anne Michel-Pajus, Jacques Borowczyk, Ahmed Djebbar, J.-C. Martzloff (1994). *Histoire d'Algorithmes – Du Caillou à la Puce*. Regards sur la Science. Belin, Paris. 429, 681

[113] F.S. Chaharsooghi, M.J. Emadi, M. Zamanighomi, M.R. Aref (2011). A new method for variable elimination in systems of inequations. In *Proceedings IEEE International Symposium on Information Theory*, pages 1215–1219. 65

[114] V. Chandrasekaran, B. Recht, P.A. Parrilo, A.S. Willsky (2012). The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6), 805–849. [doi]. 558

[115] Y.-Y. Chang (1979). Least index resolution of degeneracy in linear complementarity problems. Technical Report 79-14, Department of Operations Research, Stanford University, Stanford, CA, USA. 641

[116] B.W. Char, K.O. Geddes, G.H. Gonnet, M.B. Monagan, S.M. Watt (1988). Maple reference manuel. Symbolic Computation Group, Department of Computer Science, University of Waterloo, Ontario, Canada. 277

[117] S.S. Chen, D.L. Donoho, M.A. Saunders (1998). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20, 33–61. [doi]. 558, 707

[118] X. Chen, M. Fukushima (1999). Proximal quasi-Newton methods for nondifferentiable convex optimization. *Mathematical Programming*, 85, 313–334. 323

[119] N.V. Chernikova (1964). Algorithm for finding a general formula for the non-negative solutions of system of linear equations. *U.S.S.R. Computational Mathematics and Mathematical Physics*, 4(4), 151–158. 45

[120] N.V. Chernikova (1965). Algorithm for finding a general formula for the non-negative solutions of system of linear inequalities. *U.S.S.R. Computational Mathematics and Mathematical Physics*, 5(2), 228–233. 45

[121] N.V. Chernikova (1968). Algorithm for discovering the set of all solutions of a linear programming problem. *U.S.S.R. Computational Mathematics and Mathematical Physics*, 8(6), 282–293. 45

[122] A. Chiche, J.Ch. Gilbert (2016). How the augmented Lagrangian algorithm can deal with an infeasible convex quadratic optimization problem. *Journal of Convex Analysis*, 23(2), 425–459. [pdf] [editor]. 496, 497, 498, 499, 505, 577

[123] A.L. Cholesky (1910). Sur la résolution numérique des systèmes d'équations linéaires. Manuscrit retrouvé par C. Brezinski dans les archives d'André Louis Cholesky qu'une personne de sa famille, M. Gross-Cholesky, a léguées à l'École Polytechnique. [bibnum]. 776, 785

[124] G. Choquet (1969). *Cours d'Analyse – Tome II : Topologie*. Masson, Paris. 758

[125] V. Chvátal (1983). *Linear Programming*. W.H. Freeman, New York. 71, 643
[126] P.G. Ciarlet (1982). *Introduction à l'Analyse Numérique Matricielle et à l'Optimisation*. Masson, Paris. 16, 342
[127] P.G. Ciarlet (1988). *Introduction à l'Analyse Numérique Matricielle et à l'Optimisation* (seconde édition). Masson, Paris. 643
[128] F.H. Clarke (1983). *Optimization and Nonsmooth Analysis*. John Wiley & Sons, New York. Reprinted in 1990 by SIAM, Classics in Applied Mathematics 5 [doi]. 246, 808
[129] K.A. Cliffe, A. Spence, S.J. Tavener (2000). The numerical analysis of bifurcation problems with application to fluid mechanics. In *Acta Numerica 2000*, pages 39–131. Cambridge University Press. 428
[130] A. Cobham (1965). The intrinsic computational difficulty of functions. In Y. Bar-Hillel, éditeur, *Proc. 1964 International Congress for Lagic, Methodology and Philosophy of Science*, pages 20–30. North-Holland, Amsterdam. 291
[131] G. Cohen (2000). *Convexité et Optimisation*. École Nationale Supérieure des Ponts et Chaussées et INRIA. 323
[132] P.L. Combettes, J.-B. Hiriat-Urruty, M. Thera (2014). Preface. *Mathematical Programming*, 148, 1–4. 19
[133] R. Cominetti (1990). Metric regularity, tangent sets, and second-order optimality conditions. *Applied Mathematics and Optimization*, 21(1), 265–287. [doi]. 219, 242
[134] R. Cominetti (1999). Nonlinear averages and convergence of penalty trajectories in convex programming. In Michel Théra, Rainer Tichatschke, éditeurs, *Ill-posed Variational Problems and Regularization Techniques*, Lecture Notes in Economics and Mathematical System 477, pages 65–78. Springer Verlag, Berlin. 609
[135] A.R. Conn, N.I.M. Gould, A. Sartenaer, Ph.L. Toint (1996). Convergence properties of an augmented Lagrangian algorithm for optimization with a combination of general equality and linear constraints. *SIAM Journal on Optimization*, 6, 674–703. 498
[136] A.R. Conn, N.I.M. Gould, Ph.L. Toint (1991). A globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds. *SIAM Journal on Numerical Analysis*, 28, 545–572. 498
[137] A.R. Conn, N.I.M. Gould, Ph.L. Toint (1992). *LANCELOT: A Fortran Package for Large-Scale Nonlinear Optimization (Release A)*. Computational Mathematics 17. Springer Verlag, Berlin. [doi]. 497
[138] A.R. Conn, N.I.M. Gould, Ph.L. Toint (2000). *Trust-Region Methods*. MPS-SIAM Series on Optimization 1. SIAM and MPS, Philadelphia. [doi]. 16
[139] D. Coppersmith, S. Winograd (1982). On the asymptotic complexity of matrix multiplications. *SIAM Journal on Computing*, 11, 472–492. 365
[140] R.W. Cottle (1974). Manifestations of the Schur complement. *Linear Algebra and its Applications*, 8, 189–211. 776
[141] R.W. Cottle, J.-S. Pang, R.E. Stone (1992). *The Linear Complementarity Problem*. Academic Press, Boston. 182, 460
[142] R.W. Cottle, J.-S. Pang, R.E. Stone (2009). *The Linear Complementarity Problem*. Classics in Applied Mathematics 60. SIAM, Philadelphia, PA, USA. [doi]. 182
[143] R. Courant (1943). Variational methods for the solution of problems of equilibrium and vibrations. *Bulletin of the American Mathematical Society*, 49, 1–23. 479
[144] R. Courant (1956-1957). Calculus of Variational and Supplementary Notes and Exercises (Mimeographed Notes), Supplementary Notes by M. Kruskal and H. Rubin, revised and augmented by J. Moser. New York University. 479
[145] P. Courtier, O. Talagrand (1987). Variational assimilation of meteorological observations with the adjoint vorticity equation. I: Theory. *Quaterly Journal of the Royal Meteorological Society*, 113, 1311–1328. 453
[146] P. Courtier, O. Talagrand (1987). Variational assimilation of meteorological observations with the adjoint vorticity equation. II: Numerical results. *Quaterly Journal of the Royal Meteorological Society*, 113, 1329–1347. 453

[147] D.A. Cox, J.B. Little, D. O'Shea (2007). *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra* (troisième édition). Undergraduate Texts in Mathematics. Springer, New York. 411
[148] J.Ch. Culioli (1994). *Introduction à l'Optimisation*. Ellipses, Paris. 16
[149] H.B. Curry (1944). The method of steepest descent for non-linear minimization problems. *Quaterly of Applied Mathematics*, 2, 258–261. 322
[150] J. Czyzyk, S. Mehrotra, M. Wagner, S.J. Wright (1999). PCx: an interior-point code for linear programming. *Optimization Methods and Software*, 11-12, 397–430. 678
[151] Y.H. Dai (2002). Convergence properties of the BFGS algorithm. *SIAM Journal on Optimization*, 13, 693–701. 454
[152] J.M. Danskin (1967). *The Theory of Max-Min and Its Applications to Weapons Allocation Problems*. Springer-Verlag, New York. 791
[153] G.B. Dantzig (1951). Maximization of a linear function of variables subject to linear inequalities. In Tj.C. Koopmans, éditeur, *Activity Analysis of Production and Allocation*, pages 339–347. Wiley, New York. 620, 634
[154] G.B. Dantzig (1990). Origins of the simplex method. In G. Nash, éditeur, *A History of Scientific Computing*, ACM Press Hist. Ser., pages 141–151. ACM Press, Reading, MA, USA. 619, 633
[155] W.C. Davidon (1959). Variable metric methods for minimization. AEC Research and Development Report ANL-5990, Argonne National Laboratory, Argonne, Illinois. 250, 433
[156] W.C. Davidon (1991). Variable metric methods for minimization. *SIAM Journal on Optimization*, 1(1), 1–17. [doi]. 250, 433
[157] R. Dawkins (1996). *Le Gène Égoïste*. Odile Jacob, Paris. Traduction de *The Selfish Gene*, Oxford University Press, 1976. 249
[158] E. de Klerk (2002). *Aspects of Semidefinite Programming - Interior Point Algorithms and Selected Applications*. Kluwer Academic Publishers, Dordrecht. [doi]. 725, 743
[159] E.J. Dean, R. Glowinski (2006). An augmented Lagrangian approach to the numerical solution of the Dirichlet problem for the elliptic Monge-Ampère equation in two dimensions. *Electronic Transactions on Numerical Analysis*, 22, 71–96. [pdf]. 499
[160] J.-P. Dedieu (2006). *Points Fixes, Zéros et la Méthodes de Newton*. Mathématiques et Applications 54. Springer Verlag, Berlin. 431
[161] F. Delbos, J.Ch. Gilbert (2005). Global linear convergence of an augmented Lagrangian algorithm for solving convex quadratic optimization problems. *Journal of Convex Analysis*, 12(1), 45–69. [preprint] [editor]. 496, 497, 498, 505, 552
[162] R.S. Dembo, S.C. Eisenstat, T. Steihaug (1982). Inexact Newton methods. *SIAM Journal on Numerical Analysis*, 19, 400–408. [doi]. 430
[163] R.S. Dembo, T. Steihaug (1983). Truncated-Newton algorithms for large-scale unconstrained optimization. *Mathematical Programming*, 26, 190–212. 430
[164] J.W. Demmel (1997). *Applied Numerical Linear Algebra*. SIAM. 785
[165] D. den Hertog (1992). *Interior Point Approach to Linear, Quadratic and Convex Programming*. Mathematics and its Applications 277. Kluwer Academic Publishers, Dordrecht. 678
[166] J.E. Dennis, J.J. Moré (1974). A characterization of superlinear convergence and its application to quasi-Newton methods. *Mathematics of Computation*, 28, 549–560. [doi]. 321
[167] J.E. Dennis, H.F. Walker (1984). Inaccuracy in quasi-Newton methods: local improvement theorems. *Mathematical Programming Study*, 22, 70–85. 430
[168] P. Deuflhard (2004). *Newton Methods for Nonlinear Problems – Affine Invariance and Adaptative Algorithms*. Computational Mathematics 35. Springer, Berlin. 430
[169] P. Deuflhard, G. Heindl (1980). Affine invariant convergence theorems for Newton's method and extensions to related methods. *SIAM Journal on Numerical Analysis*, 16, 1–10. 429

[170] P.J.C. Dickinson, L. Gijben (2011). On the computational complexity of membership problems for the completely positive cone and its dual. Rapport de recherche. [Optimization Online]. 203
[171] E.D. Dolan, J.J. Moré (2002). Benchmarking optimization software with performance profiles. *Mathematical Programming*, 91, 201–213. [doi]. 291
[172] S. Dolecki, G.H. Greco (2007). Towards historical roots of necessary conditions of optimality – Regula of Peano. *Control and Cybernetics*, 36, 491–518. 241
[173] A.L. Dontchev, R.T. Rockafellar (2009). *Implicit Functions and Solution Mappings – A View from Variational Analysis*. Springer Monographs in Mathematics. Springer. 578, 802
[174] W.S. Dorn (1961). Duality in quadratic programming. *Quaterly of Applied Mathematics*, 18, 155–162. 509
[175] Z. Dostál (2005). Inexact semimonotonic augmented Lagrangians with optimal feasibility convergence for convex bound and equality constrained quadratic programming. *SIAM Journal on Numerical Analysis*, 43, 96–115. [doi]. 498
[176] Z. Dostál, A. Friedlander, S.A. Santos (1999). Augmented Lagrangians with adaptive precision control for quadratic programming with equality constraints. *Computational Optimization and Applications*, 14, 37–53. [doi]. 498
[177] Z. Dostál, A. Friedlander, S.A. Santos (2003). Augmented Lagrangians with adaptive precision control for quadratic programming with simple bounds and equality constraints. *SIAM Journal on Optimization*, 13, 1120–1140. [doi]. 498
[178] Z. Dostál, A. Friedlander, S.A. Santos, K. Alesawi (2000). Augmented Lagrangians with adaptive precision control for quadratic programming with equality constraints: corrigendum and addendum. *Computational Optimization and Applications*, 23, 127–133. [doi]. 498
[179] Z. Dostál, F.A.M. Gomes, S.A. Santos (2000). Duality based domain decomposition with natural coarse space for variational inequalities. *Journal of Computational and Applied Mathematics*, 126, 397–415. 498
[180] J. Drkošová, A. Greenbaum, M. Rozložník, Z. Strakoš (1995). Numerical stability of GMRES. *BIT*, 35, 309–330. 366
[181] I.S. Duff, J.K. Reid (1982). MA27 – A set of Fortran subroutines for solving sparse symmetric sets of linear equations. Rapport de Recherche AERE R10533, HMSO, London. 685
[182] I.S. Duff, J.K. Reid (1983). The multifrontal solution of indefinite sparse symmetric linear equations. *ACM Transactions on Mathematical Software*, 9, 302–325. 685
[183] I.S. Duff, J.K. Reid (1996). Exploiting zeros on the diagonal in the direct solution of indefinite sparse symmetric linear systems. *ACM Transactions on Mathematical Software*, 22, 227–257. 685
[184] J.-P. Dussault, M. Frappier, J.Ch. Gilbert (2019). Polyhedral Newton-min algorithms for complementarity problems. Rapport de recherche. [hal-02306526]. 430
[185] J.-P. Dussault, J.Ch. Gilbert (2021). A complexity result for the linear complementarity problem with a P-matrix (à paraître). Rapport de recherche. 18
[186] J. Eckstein (1993). Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming. *Mathematics of Operations Research*, 18, 202–226. 145, 504
[187] J. Eckstein, P.J.S. Silva (2010). Proximal methods for nonlinear programming: double regularization and inexact subproblems. *Computational Optimization and Applications*, 46, 279–304. [doi]. 498, 504
[188] J. Eckstein, P.J.S. Silva (2013). A practical relative error criterion for augmented Lagrangians. *Mathematical Programming*, 141, 319–348. [doi]. 498
[189] J. Edmonds (1965). Paths, trees, and flowers. *Canad. J. Math.*, 17, 449–467. 291

[190] I. Ekeland (1974). On the variational principle. *Journal of Mathematical Analysis and Applications*, 47(2), 324–353. [doi]. 322

[191] I. Ekeland, R. Temam (1974). *Analyse Convexe et Problèmes Variationnels*. Dunod, Paris. 145, 242, 551

[192] I.I. Eremin (1965). The relaxation method of solving systems of inequalities with convex functions on the left-hand sides. *Doklady Akad. Nauk SSSR*, 160, 994–996. 292

[193] I.I. Eremin (1969). Fejer mappings and problems of convex optimization. *Sibirsk. Mat. Zh.*, 10, 1034–1047. 292

[194] I.I. Eremin, V.D. Mazurov (1979). *Nestacionarnye Processy Programmirovanija*. Moskva. En russe. 292

[195] A. Ern, V. Giovangigli, D.E. Keyes, M.D. Smooke (1994). Towards polyalgorithmic linear system solvers for nonlinear elliptic systems. *SIAM Journal on Scientific Computing*, 15, 681–703. 428

[196] V. Faber, T. Manteuffel (1984). Necessary and sufficient conditions for the existence of a conjugate gradient method. *SIAM Journal on Numerical Analysis*, 21, 352–361. 360

[197] F. Facchinei, A. Fischer, M. Herrich (2014). An LP-Newton method: nonsmooth equations, KKT systems, and nonidolated solutions. *Mathematical Programming*, 146, 1–36. 430

[198] F. Facchinei, J.-S. Pang (2003). *Finite-Dimensional Variational Inequalities and Complementarity Problems* (deux volumes). Springer Series in Operations Research. Springer. 430

[199] L. Fahrmeir, T. Kneib, S. Lang, B. Marx (2013). *Regression – Models, Methods and Applications*. Springer. [doi]. 681

[200] K. Fan (1949). On a theorem of Weyl concerning the eigenvalues of linear transformations. *Proceedings of the National Academy of the Sciences of U.S.A.*, 35, 652–655. 785

[201] J. Faraut, A. Korányi (1994). *Analysis on Symmetric Cones*. Clarendon Press, Oxford. 742

[202] J. Farkas (1902). Theorie der einfachen ungleichungen. Journal für die reine und angewandte Mathematik, 124, 1–27. 65, 70

[203] W. Fenchel (1949). On conjugate functions. *Canadian Journal of Mathematics*, 1, 73–77. 145, 551

[204] W. Fenchel (1951). *Convex Cones, Sets, and Functions*. Mimeographed Notes. Princeton University. 551

[205] D. Fernández, M.V. Solodov (2012). Local convergence of exact and inexact augmented Lagrangian methods under the second-order sufficient optimality condition. *SIAM Journal on Optimization*, 22(2), 384–407. 498, 505

[206] A.V. Fiacco, G.P. McCormick (1964). The sequential unconstrained minimization technique for nonlinear programing, a primal-dual method. *Management Science*, 10(2). [doi]. 488

[207] A.V. Fiacco, G.P. McCormick (1968). *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*. John Wiley, New York. Reprinted in 1990 by SIAM in the collection "Classics in Applied Mathematics", number 4, [doi]. 16, 241, 473, 679

[208] M.L. Flegel, C. Kanzow (2005). Abadie-type constraint qualification for mathematical programs with equilibrium constraints. *Journal of Optimization Theory and Applications*, 124(3), 595–614. [doi]. 242

[209] R. Fletcher (1970). A new approach to variable metric algorithms. *The Computer Journal*, 13, 317–322. 440

[210] R. Fletcher (1980). *Practical Methods of Optimization. Volume 1: Unconstrained Optimization*. John Wiley & Sons, Chichester. 322

[211] R. Fletcher (1982). A model algorithm for composite nondifferentiable optimization problems. *Mathematical Programming Study*, 17, 67–76. 576
[212] R. Fletcher (1987). *Practical Methods of Optimization* (seconde édition). John Wiley & Sons, Chichester. 16, 497, 643
[213] R. Fletcher (1995). An optimal positive definite update for sparse Hessian matrices. *SIAM Journal on Optimization*, 5, 192–218. 454
[214] R. Fletcher (2010). The sequential quadratic programming method. In G. Di Pillo, F. Schoen, éditeurs, *Numerical Optimization*, Lecture Notes in Mathematics 1989, pages 165–214. Springer. 579
[215] R. Fletcher, M.J.D. Powell (1974). On the modification of LDL^T factorizations. *Mathematics of Computation*, 28, 1067–1087. 785
[216] R. Fletcher, C.M. Reeves (1964). Function minimization by conjugate gradients. *The Computer Journal*, 7, 149–154. 365
[217] A. Forsgren, P.E. Gill, M.H. Wright (2002). Interior methods for nonlinear optimization. *SIAM Review*, 44, 525–597. 679
[218] M. Fortin, R. Glowinski (1982). *Méthodes de Lagrangien Augmenté – Applications à la Résolution Numérique de Problèmes aux Limites*. Méthodes Mathématiques de l'Informatique 9. Dunod, Paris. 499, 504, 880
[219] J.B.J. Fourier (1827). Analyse des Travaux de l'Académie Royale des Sciences pendant l'année 1824, Partie Mathématique. In *Histoire de l'Académie Royale des Sciences de l'Institut de France*, Tome 7, pages xlvii–lv. 38, 65
[220] J.B.J. Fourier (1831). *Analyse des Équations Déterminées*. 429
[221] M. Frank, P. Wolfe (1956). An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3, 95–110. [doi]. 243
[222] M. Fréchet (1911). Sur la notion de différentielle. *C. R. Acad. Sci. Paris*, 152, 845–847. [Gallica]. 793
[223] R. Frisch (1955, mai). The logarithmic potential method for convex programming with particular application to the dynamics of planning for national development. Memorandum, Institut d'Économie, Université d'Oslo, Oslo, Norvège. 487, 488
[224] M. Fu, Z.-Q. Luo, Y. Ye (1998). Approximation algorithms for quadratic programming. *Journal of Combinatorial Optimization*, 2(1), 29–50. 402
[225] A. Fuduli, J.Ch. Gilbert (2003). OPINL: a truncated Newton interior-point algorithm for nonlinear optimization. Note de travail. 679
[226] D. Gabay, B. Mercier (1976). A dual algorithm for the solution of nonlinear variational problems via finite-element approximations. *Computers and Mathematics with Applications*, 2, 17–40. 499
[227] E.M. Gafni, D.P. Bertsekas (1984). Two-metric projection methods for constrained optimization. *SIAM Journal on Control and Optimization*, 22, 936–964. [doi]. 465
[228] F.R. Gantmacher (1959). *The Theory of Matrices*, Tome 1. Chelsea, New York. 774
[229] M.R. Garey, D.S. Johnson (1979). *Computers and Intractability: a Guide to the Theory of NP-Completeness*. W.H. Freeman, San Francisco. 259
[230] W. Gautschi (1997). *Numerical Analysis – An Introduction*. Birkhäuser, Boston. 274
[231] J. Gauvin (1977). A necessary and sufficient regularity condition to have bounded multipliers in nonconvex programming. *Mathematical Programming*, 12, 136–138. 196
[232] J. Gauvin (1992). *Théorie de la programmation mathématique non convexe*. Les Publications CRM, Montréal. 242
[233] J. Gauvin (1995). *Leçons de Programmation Mathématique*. Éditions de l'École Polytechnique de Montréal, Montréal. 16
[234] D.M. Gay (1981). Computing optimal locally constrained steps. *SIAM Journal on Scientific and Statistical Computing*, 2, 186–197. 402
[235] J. Genet (1976). *Mesure et Intégration – Théorie Élémentaire*. Vuibert. 837

[236] A.M. Geoffrion (1971). Duality in nonlinear programming: a simplified applications-oriented development. *SIAM Review*, 13(1), 1–37. [doi]. 151, 509

[237] A.M. Geoffrion (1972). Generalized Benders decomposition. *Journal of Optimization Theory and Applications*, 10(4), 237–260. [doi]. 552

[238] J.Ch. Gilbert (1992). Automatic differentiation and iterative processes. *Optimization Methods and Software*, 1, 13–21. [doi]. 291

[239] J.Ch. Gilbert (2008). Optimisation différentiable. In Editions T.I., éditeur, *Techniques de l'Ingénieur*. Document AF-1-252. 16

[240] J.Ch. Gilbert (2016). *Step-by-step design of an interior point solver in self-dual conic optimization – Application to the Shor relaxation of some small OPF problems*. Lecture notes of the Master-2 "Optimization" at Paris-Saclay University, Paris, France. [hal]. 726

[241] J.Ch. Gilbert, C. Gonzaga, E. Karas (2005). Examples of ill-behaved central paths in convex optimization. *Mathematical Programming*, 103, 63–94. [doi]. 609, 679

[242] J.Ch. Gilbert, X. Jonsson (2008). LIBOPT – An environment for testing solvers on heterogeneous collections of problems. Soumis à *ACM Transactions on Mathematical Software*. 291

[243] J.Ch. Gilbert, G. Le Vey, J. Masse (1991). La différentiation automatique de fonctions représentées par des programmes. Rapport de Recherche 1557, INRIA, BP 105, F-78153 Le Chesnay, France. [internet]. 280, 281, 285, 291

[244] J.Ch. Gilbert, C. Lemaréchal (1989). Some numerical experiments with variable-storage quasi-Newton algorithms. *Mathematical Programming*, 45, 407–435. [doi]. 435, 451

[245] J.Ch. Gilbert, J. Nocedal (1992). Global convergence properties of conjugate gradient methods for optimization. *SIAM Journal on Optimization*, 2, 21–42. [doi]. 357, 359

[246] J.Ch. Gilbert, J. Nocedal (1993). Automatic differentiation and the step computation in the limited memory BFGS method. *Applied Mathematics Letters*, 6(3), 47–50. [doi]. 454

[247] P.E. Gill, G.H. Golub, W. Murray, M.A. Saunders (1974). Methods for modifying matrix factorizations. *Mathematics of Computation*, 28, 505–535. 785

[248] P.E. Gill, W. Murray (1972). Quasi-Newton methods for unconstrained optimization. *Journal of the Institute of Mathematics and its Applications*, 9, 91–108. 449

[249] P.E. Gill, W. Murray (1977). Modification of matrix factorizations after a rank-one change. In D.A.H. Jacobs, éditeur, *The State of the Art in Numerical Analysis*. Academic Press, London. 785

[250] P.E. Gill, W. Murray, M.A. Saunders (2002). SNOPT: an SQP algorithm for large-scale constrained optimization. *SIAM Journal on Optimization*, 12, 979–1006. [doi]. 576

[251] P.E. Gill, W. Murray, M.H. Wright (1981). *Practical Optimization*. Academic Press, New York. 16, 785

[252] R. Glowinski, J.L. Lions, R. Tremolières (1976). *Analyse Numérique des Inéquations Variationnelles - Tome 1 : Théorie Générale, Premières Applications*. Dunod-Bordas, Paris. 94, 336, 339, 340, 504

[253] R. Glowinski, J.L. Lions, R. Tremolières (1981). *Numerical Analysis of Variational Inequalities*. North-Holland, Amsterdam, New York. 504

[254] R. Glowinski, A. Marocco (1975). Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation dualité, d'une classe de problèmes de Dirichlet non linéaires. *Revue Française d'Automatique, Informatique et Recherche Opérationnelle*, R-9, 41–76. 499

[255] M.X. Goemans, D.P. Williamson (1995). Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of Assoc. Comput. Mach.*, 42, 1115–1145. 733

[256] C.J. Goh, X.Q. Yang (2002). *Duality in Optimization and Variational Inequality*. Optimization Theory and Applications 2. Taylor and Francis. 551

[257] D. Goldfarb (1970). A family of variable-metric method derived by variational means. *Mathematics of Computation*, 24, 23–26. 440

[258] D. Goldfarb (1976). Factorized variable metric methods for unconstrained optimization. *Mathematics of Computation*, 30, 796–811. 440, 449

[259] D. Goldfarb, A. Idnani (1983). A numerically stable dual method for solving strictly convex quadratic programs. *Mathematical Programming*, 27, 1–33. 449

[260] D. Goldfarb, M.J. Todd (1989). Linear programming. In G.L. Nemhauser, A.H.G. Rinnooy Kan, M.J. Todd, éditeurs, *Handbooks in Operations Research and Management Science*, Tome 1: Optimization, chapitre 2, pages 73–170. Elsevier Science Publishers B.V., North-Holland. 643, 645

[261] A.J. Goldman, A.W. Tucker (1956). Polyhedral convex cones. In H.W. Kuhn, A.W. Tucker, éditeurs, *Linear Inequalities and Related Systems*, Annals of Mathematics Studies 38, pages 19–40. Princeton University Press, Princeton, NJ.

[262] A.A. Goldstein (1964). Convex programming in Hilbert space. *Bulletin of the American Mathematical Society*, 70, 709–710. 462, 470

[263] A.A. Goldstein (1965). On steepest descent. *SIAM Journal on Control*, 3, 147–151. 307

[264] T. Goldstein, B. O'Donoghue, S. Setzer, R. Baraniuk (2014). Fast alternating direction optimization methods. *SIAM Journal on Imaging Sciences*, 7(3), 1588–1623. [doi]. 500

[265] H.H. Goldstine (1977). *A History of Numerical Analysis From the 16th Through the 19th Century*. Studies in the History of Mathematics and Physical Sciences 2. Springer-Verlag, New York. 713

[266] E.G. Golshtein, N.V. Tretyakov (1996). *Modified Lagrangians and Monotone Maps in Optimization*. Discrete Mathematics and Optimization. John Wiley & Sons, New York. 552

[267] G.H. Golub, C.F. Van Loan (1996). *Matrix Computations* (troisième édition). The Johns Hopkins University Press, Baltimore, Maryland. 785

[268] J. Gondzio (1995). HOPDM (version 2.12) – a fast LP solver based on a primal-dual interior point method. *European Journal of Operations Research*, 85, 221–225. [doi]. 678

[269] C. Gonzaga (2000). Two facts on the convergence of the Cauchy algorithm. *Journal of Optimization Theory and Applications*, 107, 591–600. [doi]. 322

[270] P. Gordan (1873). Über die Auflösung linearer Gleichungen mit reelen Coefficienten. Mathematische Annalen, 6, 23–28. 70

[271] N.I.M. Gould, D. Orban, A. Sartenaer, Ph.L. Toint (2000). Superlinear convergence of primal-dual interior point algorithms for nonlinear programming. *Mathematical Programming*, 87, 215–249. 616

[272] N.I.M. Gould, D. Orban, Ph.L. Toint (2005). Numerical methods for large-scale nonlinear optimization. In *Acta Numerica 2005*, pages 299–361. Cambridge University Press. 16, 579

[273] J.V. Grabiner (1983). The changing concept of change: the derivative from Fermat to Weierstrass. *Mathematics Magazine*, 56, 195–206. 789

[274] B. Gracián (1647). *Oracle manuel et art de prudence*. Seuil. Recueil d'écrits de Baltasar Gracián y Morales (1601-1658), traduits de l'espagnol, introduits et annotés par Benito Pelegrín. 507

[275] H.G. Grassmann (1847). *Geometrische Analyse*. Leipzig. 754

[276] S. Gratton, D. Titley-Peloquin, Ph.L. Toint, J. Tshimanga Ilunga (2014). Differentiating the method of conjugate gradients. *SIAM Journal on Matrix Analysis and Applications*, 35(1), 110–126. [doi]. 365

[277] J. Gray (2002). Adrien-Marie Legendre (1752-1833). *European Mathematical Society Newsletter*, 45, 13. 713
[278] J.F. Grcar (2011). How ordinary elimination became Gaussian elimination. *Historia Mathematica*, 38, 163–218. 773
[279] A. Greenbaum (1989). Behavior of slightly perturbed Lanczos and conjugate-gradient recurrences. *Linear Algebra and its Applications*, 113, 7–63. 365
[280] A. Greenbaum (1994). The Lanczos and conjugate gradient algorithms in finite precision arithmetic. In J.D. Brown, M.T. Chu, D.C. Ellison, R.J. Plemmons, éditeurs, *Proceedings of the Cornelius Lanczos International Centenary Conference*, pages 49–60. SIAM, Philadelphia, PA, USA. 365
[281] A. Greenbaum (1997). *Iterative Methods for Solving Linear Systems*. SIAM, Philadelphia. 366
[282] A. Greenbaum (1997). Estimating the attainable accuracy of recursively computed residual methods. *SIAM Journal on Matrix Analysis and Applications*, 18, 535–551. 365
[283] A. Greenbaum, Z. Strakoš (1992). Predicting the behavior of finite precision Lanczos and conjugate gradient computations. *SIAM Journal on Matrix Analysis and Applications*, 13, 121–137. 365
[284] M. Grevisse (1986). *Le Bon Usage – Grammaire Française*. Duculot, Paris, Louvain-la-Neuve. Douzième édition refondue par A. Goosse. 9
[285] A. Griewank (1985). On solving nonlinear equations with simple singularities or nearly singular solutions. *SIAM Review*, 27, 537–563. 430
[286] A. Griewank (1989). On automatic differentiation. In M. Iri, K. Tanabe, éditeurs, *Mathematical Programming: Recent Developments and Applications*, pages 83–108. Kluwer Academic Publishers, Dordrecht. 285
[287] A. Griewank (1992). Achieving logarithmic growth of temporal and spatial complexity in reverse automatic differentiation. *Optimization Methods and Software*, 1, 35–54. 285
[288] A. Griewank (2000). *Evaluating Derivatives – Principles and Techniques of Algorithmic Differentiation*. SIAM Publication. 276, 291
[289] A. Griewank (2003). A mathematical view of automatic differentiation. In *Acta Numerica 2003*, pages 321–398. Cambridge University Press. 291
[290] A. Griewank, G. Corliss, éditeurs (1991). *Automatic Differentiation of Algorithms: Theory, Implementation, and Application*, number 53 in Proceedings in Applied Mathematics. SIAM, Philadelphia. 291
[291] A. Griewank, A. Walther (2008). *Evaluating Derivatives – Principles and Techniques of Algorithmic Differentiation* (seconde édition). SIAM Publication. 276, 291
[292] L. Grippo, F. Lampariello, S. Lucidi (1986). A nonmonotone line search technique for Newton's method. *SIAM Journal on Numerical Analysis*, 23, 707–716. 322
[293] L. Grippo, M. Sciandrone (2000). On the convergence of the block nonlinear Gauss-Seidel method under convex constraints. *Operations Research Letters*, 26(3), 127–136. [doi]. 340
[294] M. Guignard (1969). Generalized Kuhn-Tucker conditions for mathematical programming problems in a Banach space. *SIAM Journal on Control*, 7(2), 232–241. 242
[295] O. Güler (2010). *Foundations of Optimization*. Graduate Texts in Mathematics 258. Springer. [doi]. 70, 244, 645, 646
[296] J. Guo, A.S. Lewis (2018). Nonsmooth variants of Powell's BFGS convergence theorem. Rapport de Recherche 2. 454
[297] M.H. Gutknecht (1997). Lanczos-type solvers for nonsymmetric linear systems of equations. In *Acta Numerica 1997*, pages 271–397. Cambridge University Press. 366
[298] W. Hackbusch (1994). *Iterative Solution of Large Sparse Systems of Equations*. Applied Mathematical Sciences 95. Springer-Verlag, New York. 342

[299] W.W. Hager (1993). Analysis and implementation of a dual algorithm for constrained optimization. *Journal of Optimization Theory and Applications*, 79, 427–462. 498

[300] A. Halevy, P. Norvig, F. Pereira (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2), 8–12. [doi]. 499

[301] S.-H. Han, J.-S. Pang, N. Rangaraj (1992). Globally convergent Newton methods for nonsmooth equations. *Mathematics of Operations Research*, 17, 586–607. [doi]. 430

[302] S.-P. Han (1976). Superlinearly convergent variable metric algorithms for general nonlinear programming problems. *Mathematical Programming*, 11, 263–282. 579

[303] S.-P. Han (1977). A globally convergent method for nonlinear programming. *Journal of Optimization Theory and Applications*, 22, 297–309. 579

[304] R.J. Hanson, C.L. Lawson (1969). Extensions and applications of the Householder algorithm for solving linear least squares problems. *Mathematics of Computation*, 23, 787–812. 713

[305] G.H. Hardy, J.E. Littlewood, G. Pólya (1952). *Inequalities*. Cambridge University Press, Cambridge, U.K. 769

[306] R. Hauser, J. Nedič (2005). The continuous Newton-Raphson method can look ahead. *SIAM Journal on Optimization*, 15, 915–925. 431

[307] B. He, X. Yuan (2012). On the $O(1/n)$ convergence rate of the Douglas-Rachford alternating direction method. *SIAM Journal on Numerical Analysis*, 50(2), 700–709. [doi]. 500

[308] R. Helgason, J. Kennington (1995). Handbooks in Operations Research and Management Science 7: Netwok Models. North-Holland. 643

[309] C. Helmberg, F. Rendl, R.J. Vanderbei, H. Wolkowicz (1996). An interior-point method for semidefinite programming. *SIAM Journal on Optimization*, 6(2), 342–361. [doi]. 742, 743

[310] J.M. Hendrickx, A. Olshevsky (2010). Matrix p-norms are NP-hard to approximate if $p \neq 1, 2, \infty$. *SIAM Journal on Matrix Analysis and Applications*, 31(5), 2802–2812. [doi]. 764

[311] M.R. Hestenes (1969). Multiplier and gradient methods. *Journal of Optimization Theory and Applications*, 4, 303–320. 504

[312] M.R. Hestenes, E. Stiefel (1952). Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, 49, 409–436. 365, 686

[313] N.J. Higham (2002). *Accuracy and Stability of Numerical Algorithms* (seconde édition). SIAM Publication, Philadelphia. 300, 341, 365, 430, 777, 785

[314] N.J. Higham (2008). Cholesky factorization. MIMS EPrint 2008.116, University of Manchester. 780

[315] J.-B. Hiriart-Urruty (1996). *L'Optimisation*. Que sais-je 3184. Presses Universitaires de France. 16, 195, 242

[316] J.-B. Hiriart-Urruty (2013). *Bases, outils et principes pour l'analyse variationnelle*. Mathématiques et Applications 70. Springer Verlag. 126, 145

[317] J.-B. Hiriart-Urruty, C. Lemaréchal (1993). *Convex Analysis and Minimization Algorithms*. Grundlehren der mathematischen Wissenschaften 305-306. Springer. 62, 101, 105, 106, 133, 134, 139, 145, 323, 370, 545, 551, 552

[318] J.-B. Hiriart-Urruty, A. Seeger (2010). A variational approach to copositive matrices. *SIAM Review*, 52, 593–629. 203

[319] A.J. Hoffman (1952). On approximate solutions of systems of linear inequalities. *Journal of Research of the National Bureau of Standard*, 49, 263–265. 645

[320] R.A. Horn, C. Jonhson (1985). *Matrix Analysis*. Cambridge University Press, Cambridge, U.K. 785

[321] A.S. Houselholder (1964). *The Theory of Matrices in Numerical Analysis*. Blaisdell, New York. 774

[322] S. Van Huffel, J. Vandewalle (1991). *The Total Least Squares Problem: Computational Aspects and Analysis*. Frontiers in Applied Mathematics 9. SIAM, Philadelphia, PA, USA. 714

[323] C. Humes, P.J.S. Silva, B.F. Svaiter (2004). Some inexact hybrid proximal augmented Lagrangian algorithms. *Numerical Algorithms*, 35, 175–184. [doi]. 498

[324] T. Huynh, C. Lassez, J.-L. Lassez (1992). Practical issues on the projection of polyhedral sets. *Annals of Mathematics and Artificial Intelligence*, 6, 295–316. 65

[325] Kh.D. Ikramov, N.V. Savel'eva (2000). Conditionally definite matrices. *Journal of Mathematical Sciences*, 98, 1–50. 203

[326] M. Iri (1984). Simultaneous computation of functions, partial derivatives and estimates of rounding errors, complexity and practicality. *Japan Journal of Applied Mathematics*, 1, 223–252. 281

[327] M. Iri, K. Kubota (1987). Methods of fast automatic differentiation and applications. Research Memorandum RMI 87-02, Department of Mathematical Engineering and Instrumentation Physics, Faculty of Engineering, University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo, Japan. 281

[328] K. Ito, K. Kunisch (2008). *Lagrange Multiplier Approach to Variational Problems and Applications*. Advances in Design and Control. SIAM Publication, Philadelphia. [doi]. 16, 242

[329] A.F. Izmailov, A.S. Kurennoy, M.V. Solodov (2013). The Josephy-Newton method for semismooth generalized equations and semismooth SQP for optimization. *Set-Valued and Variational Analysis*, 21(1), 17–45. [doi]. 579

[330] A.F. Izmailov, M.V. Solodov (2014). *Newton-Type Methods for Optimization and Variational Problems*. Springer Series in Operations Research and Financial Engineering. Springer. [doi]. 431, 579

[331] D.D. Jackson (1972). Interpretation of inaccurate, insufficient and inconsistent data. *gjras*, 28, 97–109. 688

[332] J. Jahn (1985). *Scalarization in multi objective optimization*. Springer, Vienna. 147

[333] G. James, D. Witten, T. Hastie, R. Tibshirani (2013). *An Introduction to Statistical Learning*. Springer Texts in Statistics 103. Springer. [doi]. 681

[334] B. Jansen (1997). *Interior Point Techniques in Optimization – Complementarity, Sensitivity and Algorithms*. Applied Optimization 6. Kluwer Academic Publishers, Dordrecht. 678

[335] J.L.W.V. Jensen (1905). Om Konvekse Funktioner og Uligheder mellem Middelværdier. *Nyt Tidsskr. Math.*, B 16, 49–68. 146

[336] J.L.W.V. Jensen (1906). Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica*, 30, 175–193. 146

[337] H. Jiang, P.A. Forsyth (1995). Robust linear and nonlinear strategies for solution of the transonic Euler equations. *Comput. Fluids*, 24, 753–770. 428

[338] F. John (1948). Extremum problems with inequalities as subsidiary conditions. In K.O. Friedrichs, O.E. Neugebauer, J.J. Stokes, éditeurs, *Studies and Essays, Courant Anniversary Volume*, pages 186–204. Wiley Interscience, New York. 242

[339] C.N. Jones, E.C. Kerrigan, J.M. Maciejowski (2004). Equality set projection: A new algorithm for the projection of polytopes in halfspace representation. CUED Technical Report CUED/FINFENG/TR.463. 65

[340] W. Kahan (1966). Numerical linear algebra. *Canadian Mathematical Bulletin*, 9, 757–801. [doi]. 265

[341] L. Kantorovich (1939). The method of successive approximations for functional equations. *Acta Math.*, 71, 63–97. 429

[342] L. Kantorovich (1948). On Newton's method for functional equations. *Dokl. Akad. Nauk SSSR*, 59, 1237–1240. En russe. 429

[343] L. Kantorovich (1949). On Newton's method. *Trudy Mat. Inst. Steklov*, 28, 104–144. En russe. 429
[344] L.V. Kantorovich (1940). On an efficient method for solving some classes of extremum problems. *Doklady Akad. Nauk SSSR*, 28, 212–215. 241
[345] L.V. Kantorovich, G.P. Akilov (1982). *Functional Analysis* (seconde édition). Pergamon Press, London. 326, 429
[346] N. Karmarkar (1984). A new polynomial-time algorithm for linear programming. *Combinatorica*, 4, 373–395. 620
[347] W.E. Karush (1939). *Minima of Functions of Several Variables with Inequalities as Side Conditions*. Master's thesis, Department of Mathematics, University of Chicago, Chicago. 242
[348] C.T. Kelley (1995). *Iterative Methods for Linear and Nonlinear Equations*. SIAM Publication, Philadelphia. 16, 430
[349] C.T. Kelley (1999). *Iterative Methods for Optimization*. Frontiers in Applied Mathematics 18. SIAM Publication, Philadelphia. 16, 430
[350] C.T. Kelley (2003). *Solving Nonlinear Equations with Newton's Method*. SIAM Publication, Philadelphia. 430
[351] C.T. Kelley, D.E. Keyes (1998). Convergence analysis of pseudo-transient continuation. *SIAM Journal on Numerical Analysis*, 35, 508–523. 428
[352] D.E. Keyes (1995). Aerodynamic applications of Newton-Krylow-Schwarz solvers. In M. Deshpande, S. Desai, R. Narasimha, éditeurs, *Proc. 14th Internat. Conference on Numerical Methods in Fluid Dynamics*, pages 1–20. Springer. 427
[353] K.V. Kim, Yu.E. Nesterov, B.V. Cherkasskiĭ (1984). An estimate of the effort in computing the gradient. *Soviet Math. Dokl.*, 29, 384–387. 281
[354] V. Klee, G.L. Minty (1972). How good is the simplex algorithm ? In O. Shisha, éditeur, *Inequalities III*, pages 159–175. Academic Press, New York. 642
[355] I. Klep, M. Schweighofer (2013). An exact duality theory for semidefinite programming based on sums of squares. *Mathematics of Operations Research*, 38(3), 569–590. [doi]. 724
[356] D.A. Knoll, D.E. Keyes (2004). Jacobian-free Newton-Krylov methods: a survey of approaches and applications. *Journal of Computational Physics*, 193, 357–397. 430
[357] M. Koecher (1999). *The Minnesota Notes on Jordan Algebras and Their Applications*. Lecture Notes in Mathematics 1710. Springer-Verlag, Berlin. 742
[358] M. Kojima, S. Mizuno, A. Yoshise (1989). A primal-dual interior-point method for linear programming. In N. Megiddo, éditeur, *Progress in Mathematical Programming, Interior-point and Related Methods*, pages 29–47. Springer, New York. 678
[359] M. Kojima, S. Mizuno, A. Yoshise (1989). A polynomial-time algorithm for a class of linear complementarity problems. *Mathematical Programming*, 44, 1–26. 678
[360] M. Kojima, S. Shindo (1986). Extension of Newton and quasi-Newton methods to systems of PC^1 equations. *Journal of Operations Research Society of Japan*, 29, 352–375. [doi]. 430
[361] M. Kojima, S. Shindoh, S. Hara (1997). Interior-point methods for the monotone semidefinite linear complementarity problem in symmetric matrices. *SIAM Journal on Optimization*, 7, 86–125. [doi]. 742, 743
[362] N. Kollerstrom (1992). Thomas Simpson and "Newton's method of approximation": an enduring myth. *Brit. J. Hist. Sci.*, 25, 347–354. 429
[363] H. Komiya (1988). Elementary proof for Sion's minimax theorem. *Kodai Mathematical Journal*, 11, 5–7. 551
[364] D. König (1936). *Theorie der Endlichen and Unendlichen Graphen*. Akademische Verlagsgegesellschaft, Leipzig. 67
[365] M.A. Krasnosel'skii, S.G. Krein (1952). An iteration process with minimal residues. *Mat. Sb.*, 31, 315–334. En russe. 322

[366] J. Kruskal (1969). Two convex counterexamples: a discontinuous envelope function and a nondifferentiable nearest-point mapping. *Proceedings of the American Mathematical Society*, 23, 697–703. 47

[367] M. Kubicek, M. Marek (1983). *Computational Methods in Bifurcation Theory and Dissipative Structures*. Springer Series in Comput. Phys. Springer-Verlag, New York. 428

[368] H.W. Kuhn (1976). Nonlinear programming: a historical view. In R.W. Cottle, C.E.Lemke, éditeurs, *Nonlinear Programming*, SIAM-AMS Proceedings IX, pages 1–26. American Mathematical Society, Providence, RI. 242

[369] H.W. Kuhn, A.W. Tucker (1951). Nonlinear programming. In J. Neyman, éditeur, *Proceedings of the second Berkeley Symposium on Mathematical Studies and Probability*, pages 481–492. University of California Press, Berkeley, California. 242

[370] J. Kyparisis (1985). On uniqueness of Kuhn-Tucker multipliers in nonlinear programming. *Mathematical Programming*, 32, 242–246. [doi]. 195

[371] J.-C. Lafaille, B. Heimermann (2003). *Prisonnier de l'Annapurna*. Guérin, Chamonix. xvii

[372] J.-L. Lagrange (1788). *Méchanique Analytique*, Tome 1. [internet]. 157, 241

[373] P. Lascaux, R. Théodor (1986). *Analyse Numérique Matricielle Appliquée à l'Art de l'Ingénieur*. Masson, Paris. 785

[374] J.B. Lasserre (1997). A Farkas lemma without a standard closure condition. *SIAM Journal on Control and Optimization*, 35, 265–272. [doi]. 59

[375] J.B. Lasserre (2001). Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11, 796–817. [doi]. 17, 411, 731

[376] J.B. Lasserre (2010). *Moments Positive Polynomials and Their Applications*. Imperial College Press Optimization Series 1. Imperial College Press. 20, 731

[377] J.B. Lasserre (2015). *An Introduction to Polynomial and Semi-Algebraic Optimization*. Cambridge Texts in Applied Mathematics. Cambridge University Press. 20

[378] J.B. Lasserre, J.-B. Hiriart-Urruty (2002). Mathematical properties of optimization problems defined by positively homogeneous functions. *Journal of Optimization Theory and Applications*, 112, 31–52. [doi]. 556

[379] J.B. Lasserre, T. Netzer (2007). SOS approximations of nonnegative polynomials via simple high degree perturbations. *Mathematische Zeitschrift*, 256, 99–112. 729

[380] C. Lassez, J.-L. Lassez (1992). *Symbolic and Numerical Computation for Artificial Intelligence*, chapitre 4, pages 103–119. Academic Press. 65

[381] J.-L. Lassez (1990). Query constraints. In *Proceedings ACM Conference Principles of Database Systems*, 4, pages 288–298. 65

[382] P.-J. Laurent (1972). *Approximation et Optimisation*. Hermann, Paris. 551

[383] F.-X. Le Dimet, I.M. Navon, D.N. Daescu (2001). Second order information in data assimilation. *Monthly Weather Review*. 291

[384] E.B. Lee, L. Markus (1967). *Foundations of Optimal Control Theory* (première édition). Wiley. 265

[385] A.M. Legendre (1787). Mémoire sur l'intégration de quelques équations aux différences partielles. *Mém. Acad. Sciences*, pages 309–351. 145

[386] Bruno Lemaitre (2015). *An Essay on Science and Narcissism*. À compte d'auteur [http://brunolemaitre.ch]. 403

[387] C. Lemaréchal. Lagrangian relaxation. In M. Jünger, D. Naddef, éditeurs, *Computational Combinatorial Optimization*, pages 115–160. Springer, Heidelberg. 552

[388] C. Lemaréchal (1981). A view of line-searches. In A. Auslender, W. Oettli, J. Stoer, éditeurs, *Optimization and Optimal Control*, Lecture Notes in Control and Information Science 30, pages 59–78. Springer, Heidelberg. 322

[389] K. Levenberg (1944). A method for the solution of certain nonlinear problems in least squares. *Quaterly of Applied Mathematics*, 2, 164–168. [doi]. 401, 714

[390] E.S. Levitin (1994). *Perturbation Theory in Mathematical Programming and its Applications*. Wiley. 242

[391] E.S. Levitin, B.T. Polyak (1966). Constrained minimization problems. *USSR. Comput. Math. and Math. Phys.*, 6, 1–50. 462, 470

[392] A.S. Lewis, M.L. Overton (2013). Nonsmooth optimization via quasi-Newton methods. *Mathematical Programming*, 141, 135–163. 454

[393] D.H. Li, M. Fukushima (2001). A modified bfgs method and its global convergence in nonconvex minimization. *Journal of Computational and Applied Mathematics*, 129, 15–35. 454

[394] X. Li, D. Sun, K.-C. Toh (2020). On the efficient computation of a generalized Jacobian of the projector over the Birkhoff polytope. *Mathematical Programming*, 179, 419–446. [doi]. 554

[395] C.-J. Lin, R. Saigal (1995). A predictor-corrector method for semi-definite linear programming. Technical Report 95-20, Department of Industrial & Operations Engineering, The University of Michigan, Ann Arbor, Michigan 48109-2117, USA. 741

[396] H. Lin, J. Mairal, Z. Harchaoui (2017). A generic quasi-Newton algorithm for faster gradient-based optimization. Rapport de recherche. [arXiv:1610.00960]. 454

[397] Y.Y. Lin, J.-S. Pang (1987). Iterative methods for large scale convex quadratic programs: a survey. *SIAM Journal on Control and Optimization*, 25, 383–411. 552

[398] J.L. Lions (1968). *Contrôle Optimal de Systèmes Gouvernés par des Equations aux Dérivées Partielles*. Etudes Mathématiques. Dunod – Gauthier-Villars, Paris. 265, 271

[399] D.C. Liu, J. Nocedal (1989). On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45, 503–520. 451

[400] S. Lucidi, M. Roma (1978). Nonmonotone conjugate gradient methods for optimization. In J. Henry, J.-P. Yvon, éditeurs, *System Modelling and Optimization*, Lecture Notes in Control and Information Sciences 170, pages 206–214. Springer, Berlin. 322

[401] Z.-Q. Luo, J.-S. Pang, D. Ralph (1996). *Mathematical Programs with Equilibrium Constraints*. Cambridge University Press. 246

[402] Z.-Q. Luo, P. Tseng (1992). On the convergence of the coordinate descent method for convex differentiable minimization. *Journal of Optimization Theory and Applications*, 72, 7–35. 339

[403] S. Mandelbrojt (1939). Sur les fonctions convexes. *C. R. Acad. Sci. Paris*, 209, 977–978. 145

[404] J.-M. Mandosio (2010). Présentation. In *Grammaire Générale et Raisonnée* d'Antoine Arnauld et Claude Lancelot, 1660. Éditions Allia. 3

[405] O.L. Mangasarian, S. Fromovitz (1967). The Fritz John necessary optimality conditions in the presence of equality and inequality constraints. *Journal of Mathematical Analysis and Applications*, 17, 37–47. [doi]. 242

[406] O.L. Mangasarian, J. Ponstein (1965). Minmax and duality in nonlinear programming. *Journal of Mathematical Analysis and Applications*, 11, 504–518. 509

[407] D.W. Marquardt (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11, 431–441. 384, 401, 714

[408] R.K. Martin (1999). *Large Scale Linear and Integer Optimization: a Unified Approach*. Kluwer Academic Publishers, Boston. 643

[409] B. Martinet (1970). Régularisation d'inéquations variationnelles par approximations successives. *Revue Française d'Informatique et Recherche Opérationnelle*, R-3, 154–179. 323

[410] J.M. Martínez (1994). Local minimizers of quadratic functions on Euclidean balls and spheres. *SIAM Journal on Optimization*, 4, 159–176. 402

[411] W.F. Mascarenhas (2004). The BFGS method with exact line searches fails for non-convex objective functions. *Mathematical Programming*, 99, 49–61. 450, 454
[412] W.F. Mascarenhas (2008). Newton's iterates can converge to non-stationary points. *Mathematical Programming*, 112, 327–334. 429
[413] J. Mawhin (1992). *Analyse – Fondements, Techniques, Évolution*. De Boeck. 241, 758
[414] J. Mawhin, N. Rouche (1973). *Équations Différentielles Ordinaires, Tome 1: Théorie Générale*. Masson et C^{ie}, Paris. 344
[415] G.P. McCormick (1983). *Nonlinear Programming. Theory, Algorithms and Applications*. J. Wiley & Sons, New York. 16
[416] L. McLinden (1980). An analogue of Moreau's proximation theorem, with application to the nonlinear complementarity problem. *Pacific Journal of Mathematics*, 88, 101–161. 679
[417] L. McLinden (1982). An application of Ekeland's theorem to minimax problems. *Nonlinear Analysis, Theory, Methods & Applications*, 6(2), 189–196. [doi]. 552
[418] N. Megiddo (1987). On the complexity of linear programming. In T. Bewley, éditeur, *Advances in Economic Theory*, pages 225–268. Cambridge Univ. Press, Cambridge. 642
[419] S. Mehrotra (1992). On the implementation of a primal-dual interior point method. *SIAM Journal on Optimization*, 2(4), 575–601. [doi]. 678
[420] F. Meng, G. Zhao, M. Goh, R. De Souza (2008). Lagrangian-dual functions and Moreau-Yosida regularization. *SIAM Journal on Optimization*, 19, 39–61. 551
[421] H. Minkowski (1896). *Geometrie der Zahlen*. Teubner, Leipzig. 65
[422] M. Minoux (1983). *Programmation Mathématique. Théorie et Algorithmes*. Dunod, Paris. 643
[423] G.J. Minty (1964). On the monotonicity of the gradient of a convex function. *Pacific Journal of Mathematics*, 14, 243–247. 145
[424] S. Mizuno (1992). A new polynomial time method for a linear complementarity problem. *Mathematical Programming*, 56, 31–43. 662
[425] S. Mizuno, M.J. Todd, Y. Ye (1993). On adaptive-step primal-dual interior-point algorithms for linear programming. *Mathematics of Operations Research*, 18(4), 964–981. [doi]. 678
[426] Montaigne (1572-1592). *Les Essais*. Gallimard. Adaptation en français moderne par André Lanly, 2009. 3
[427] R. Monteiro, I. Adler (1989). Interior path following primal-dual algorithms. part I: linear programming. *Mathematical Programming*, 44, 27–41. 678
[428] R.D.C. Monteiro (1997). Primal-dual path-following algorithms for semidefinite programming. *SIAM Journal on Optimization*, 7(3), 663–678. [doi]. 742, 743
[429] R.D.C. Monteiro (2003). First- and second-order methods for semidefinite programming. *Mathematical Programming*, 97, 209–244. [doi]. 743
[430] R.D.C. Monteiro, B.F. Svaiter (2013). Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers. *SIAM Journal on Optimization*, 23(1), 475–507. [doi]. 500
[431] R.D.C. Monteiro, M. Todd (2000). Path-following methods. In H. Wolkowicz, R. Saigal, L. Vandenberghe, éditeurs, *Handbook of Semidefinite Programming – Theory, Algorithms, and Applications*, International Series in Operations Research & Management Science, Tome 27, chapitre 10, pages 267–306. Kluwer Academic Publishers. 743
[432] R.D.C. Monteiro, P.R. Zanjacomo (1997). A note of the existence of the alizadeh-haeberly-overton direction for semidefinite programming. *Mathematical Programming*, 78, 393–396. 742
[433] R.D.C. Monteiro, F. Zhou (1998). On the existence and convergence of the central path for convex programming and some duality results. *Computational Optimization and Applications*, 10, 51–77. 609

[434] E.H. Moore (1919). On the reciprocal of the general algebraic matrix. *Bulletin of the American Mathematical Society*, 26, 394–395. 766
[435] G.H. Moore (2008). The emergence of open sets, closed sets, and limit points in analysis and topology. *Historia Mathematica*, 2008, 220–241. 747
[436] B.S. Mordukhovich (2006). *Variational Analysis and Generalized Differentiation*. Grundlehren der mathematischen Wissenschaften 330-331. Springer. 242
[437] J.J. Moré (1978). The Levenberg-Marquardt algorithm: implementation and theory. In G.A. Watson, éditeur, *Numerical Analysis*, Lecture Notes in Mathematics 630, pages 105–116. Springer, Berlin. 714
[438] J.J. Moré (1983). Recent developments in algorithms and software for trust region methods. In A. Bachem, M. Grötschel, B. Korte, éditeurs, *Mathematical Programming, the State of the Art*, pages 258–287. Springer, Berlin. 265, 402
[439] J.J. Moré (2015). Un hommage à Michael J. D. Powell sur le site http://michaeljd-powell.blogspot.ca. 443
[440] J.J. Moré, D.C. Sorensen (1983). Computing a trust region step. *SIAM Journal on Scientific and Statistical Computing*, 4, 553–572. 391
[441] J.-J. Moreau (1965). Proximité et dualité dans un espace hilbertien. *Bulletin de la Société Mathématique de France*, 93, 273–299. [url]. 68, 128, 145, 853
[442] J.-J. Moreau (1967). Fonctionnelles convexes. *Séminaire Équations aux dérivees partielles, Collège de France*. 145
[443] D.D. Morrison (1960). Methods for nonlinear least squares problems and convergence proofs. *Proceedings of the Seminar on Tracking Programs and Orbit Determination, Jet Propulsion Laboratory, Pasadena, CA, USA*, pages 1–9. 714
[444] T.S. Motzkin (1936). Beiträge zur Theorie der linearen Ungleichungen. University Basel Dissertation, Jerusalem, Israel. 65
[445] T.S. Motzkin, H. Raiffa, G.L. Thompson, R.M. Thrall (1953). The double description method. In *Contributions to the Theory of Games, Vol. 2*, Annals of Mathematics Studies 28, pages 51–73. Princeton University Press, Princeton, NJ. 45
[446] W. Mulder, B.V. Leer (1985). Experiments with implicit upwind methods for the Euler equations. *Journal of Computational Physics*, 59, 232–246. 427
[447] K.G. Murty, S.N. Kabadi (1987). Some NP-complete problems in quadratic and nonlinear programming. *Mathematical Programming*, 39, 117–129. 20, 203
[448] B.K. Natarajan (1995). Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2), 227–234. [doi]. 706, 707
[449] U. Naumann (2008). Optimal jacobian accumulation is np-complete. *Mathematical Programming*, 112, 427–441. 291
[450] L. Nazareth (1994). *The Newton-Cauchy Framework – A unified approach to unconstrained nonlinear minimization*. Lecture Notes in Computer Science 769. Springer. 16
[451] E.D. Nering, A.W. Tucker, éditeurs (1993). Linear Programs and Related Problems. Academic Press. 643
[452] Y. Nesterov (2004). *Introductory Lectures on Convex Optimization – A Basic Course*. Kluwer Academic Publishers. 6, 16, 322, 616
[453] Y.E. Nesterov (1983). A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Doklady Akad. Nauk SSSR*, 269(3), 543–547. 500
[454] Y.E. Nesterov (1997). Quality of semidefinite relaxation for nonconvex quadratic optimization. Discussion Paper 9719, CORE, Université de Louvain, B-1348 Louvain-la-Neuve, Belgique. 733
[455] Y.E. Nesterov, A.S. Nemirovskii (1993). An interior point method for generalized linear-fractional programming. Rapport de recherche. 581, 601, 616
[456] Y.E. Nesterov, A.S. Nemirovskii (1994). *Interior-Point Polynomial Algorithms in Convex Programming*. SIAM Studies in Applied Mathematics 13. SIAM, Philadelphia, PA, USA. 581, 679, 725, 743

[457] Y.E. Nesterov, M.J. Todd (1997). Self-scaled barriers and interior-point methods for convex programming. *Mathematics of Operations Research*, 22(1), 1–42. [doi]. 743
[458] J. Nocedal (1980). Updating quasi-Newton matrices with limited storage. *Mathematics of Computation*, 35, 773–782. [doi]. 451, 454
[459] J. Nocedal, R.A. Waltz (2001). KNITRO 1.00 – User's manual. Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, Il 60208, USA. 679
[460] J. Nocedal, S.J. Wright (2006). *Numerical Optimization* (seconde édition). Springer Series in Operations Research. Springer, New York. 16, 244
[461] G. Nolet, R. Montelli, J. Virieux (1999). Explicit, approximate expressions for the resolution and a posteriori covariance of massive tomographic systems. *Geophysical Journal International*, 138, 36–44. 688
[462] Y. Notay (1993). On the convergence rate of the conjugate gradients in presence of rounding errors. *Numerische Mathematik*, 65, 301–317. 365
[463] J.M. Ortega, W.C. Rheinboldt (1970). *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, New York. Reprinted in 2000 by SIAM, Classics in Applied Mathematics 30, [doi]. 16
[464] J.M. Ortega, M.L. Rockoff (1966). Nonlinear difference equations and Gauss-Seidel type iterative methods. *SIAM Journal on Numerical Analysis*, 3, 497–513. [doi]. 340
[465] M.R. Osborne (1976). Nonlinear least squares – the Levenberg algorithm revisited. *Journal of the American Mathematical Society*, 19 (Series B)(3), 343–357. [doi]. 699, 700, 714
[466] M.R. Osborne, B. Presnell, B.A. Turlach (2000). On the LASSO and its dual. *Journal of Computational and Graphical Statistics*, 9(2), 319–337. [doi]. 711
[467] M.R. Osborne, B. Presnell, B.A. Turlach (2000). A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*, 20(3), 389–403. [doi]. 711
[468] M. Padberg (1999). *Linear Optimization and Extensions* (seconde édition). Springer, Berlin. 643
[469] C.C. Paige, M.A. Saunders (1982). LSQR: an algorithm for sparse linear equations and sparse least squares. *ACM Transactions on Mathematical Software*, 8, 43–71. 686
[470] D.P. Palomar, Y.C. Eldar, éditeurs (2010). *Convex Optimization in Signal Processing and Communications*. Cambridge University Press, Cambridge. 326
[471] J.-S. Pang (1990). Newton's method for B-differentiable equations. *Mathematics of Operations Research*, 15, 311–341. [doi]. 430
[472] J.-S. Pang (1991). A B-differentiable equation-based, globally and locally quadratically convergent algorithm for nonlinear programs, complementarity and variational inequality problems. *Mathematical Programming*, 51(1-3), 101–131. [doi]. 430
[473] C.H. Papadimitriou, K. Steiglitz (1982). *Combinatorial optimization: Algorithms and Complexity*. Prentice-Hall, New Jersey. 259
[474] P.A. Parrilo (2000). *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization*. Thèse de doctorat, California Institute of Technology. 731
[475] V.Th. Paschos (2004). *Complexité et approximation polynomiale*. Lavoisier, Paris. 291
[476] V.Th. Paschos, éditeur (2005). *Optimisation combinatoire 1 – Concepts fondamentaux*. Lavoisier, Paris. 6, 291
[477] G. Peano (1887). *Applicazioni Geometriche del Calcolo Infinitesimale*. Fratelli Bocca Editori, Torino. [Cornell]. 241
[478] G. Peano (1908). *Formulario Mathematico,* Editio V. Fratelli Bocca Editori, Torino. 241
[479] R. Penrose (1955). A generalized inverse for matrices. *Proceedings Cambridge Philos. Soc.*, 51, 406–413. [doi]. 766

[480] R.R. Phelps (1993). *Convex Functions, Monotone Operators and Differentiability*. Lecture Notes in Mathematics 1364. Springer, Berlin. 87, 145
[481] E. Polak (1971). *Computational Methods in Optimization - A Unified Approach*, Mathematics in Science and Engineering, Tome 77. Academic Press, New York. 339
[482] E. Polak (1997). *Optimization – Algorithms and Consistent Approximations*. Applied Mathematical Sciences 124. Springer. 16
[483] M. Polanyi (1962). *Personal Knowledge: Towards a Post-Critical Philosophy*. The University of Chicago Press. 809
[484] B.T. Polyak (1987). *Introduction to Optimization*. Optimization Software, New York. 150
[485] B.T. Polyak (2001). History of mathematical programming in the USSR: analyzing the phenomenon. *Mathematical Programming*, 91, 401–416. 241
[486] M.J.D. Powell (1969). A method for nonlinear constraints in minimization problems. In R. Fletcher, éditeur, *Optimization*, pages 283–298. Academic Press, London. 504
[487] M.J.D. Powell (1970). A hybrid method for nonlinear equations. In P. Rabinowitz, éditeur, *Numerical Methods for Nonlinear Algebraic Equations*, pages 87–114. Gordon and Breach, New York. 401, 429
[488] M.J.D. Powell (1973). On search directions for minimization algorithms. *Mathematical Programming*, 4, 193–201. [doi]. 339
[489] M.J.D. Powell (1975). Convergence properties of a class of minimization algorithms. In O.L. Mangasarian, R.R. Meyer, S.M. Robinson, éditeurs, *Nonlinear Programming 2*, pages 1–27. Academic Press, New York. 402
[490] M.J.D. Powell (1976). Some global convergence properties of a variable metric algorithm for minimization without exact line searches. In R.W. Cottle, C.E. Lemke, éditeurs, *Nonlinear Programming*, SIAM-AMS Proceedings 9. American Mathematical Society, Providence, RI. 454
[491] M.J.D. Powell (1978). Algorithms for nonlinear constraints that use Lagrangian functions. *Mathematical Programming*, 14, 224–248. 579
[492] M.J.D. Powell (1984). Nonconvex minimization calculations and the conjugate gradient method. In *Lecture Notes in Mathematics 1066*, pages 122–141. Springer, Berlin. 454
[493] M.J.D. Powell (1984). On the global convergence of trust region algorithms for unconstrained minimization. *Mathematical Programming*, 29, 297–303.
[494] M.J.D. Powell (1987). Updating conjugate directions by the BFGS formula. *Mathematical Programming*, 38, 29–46. 449
[495] M.J.D. Powell (1991). A view of nonlinear optimization. In J.K. Lenstra, A.H.G. Rinnooy Kan, A. Schrijver, éditeurs, *History of Mathematical Programming, A Collection of Personal Reminiscences*, pages 119–125. CWI North-Holland, Amsterdam. 433, 447, 634
[496] M.J.D. Powell (2000). On the convergence of the DFP algorithm for unconstrained optimization when there are only two variables. *Mathematical Programming*, 87, 281–301. 454
[497] M.J.D. Powell (2003). An interview with M. J. D. Powell. *Bulletin of the International Center for Mathematics*, 14. Interview by Luís Nunes Vicente, University of Coimbra. 488, 647
[498] L. Pronzato, H.P. Wynn, A.A. Zhigljavsky (2001). Renormalised steepest descent in Hilbert space converges to a two-point attractor. *Acta Applicandae Mathematicae*, 67, 1–18. 322
[499] L. Pronzato, H.P. Wynn, A.A. Zhigljavsky (2004). Asymptotic behaviour of a family of gradient algorithms in $\mathbb{R}^d$ and Hilbert spaces. Manuscript. 322
[500] B.N. Pshenichnyj (1994). *The Linearization Method for Constrained Optimization*. Computational Mathematics 22. Springer. 579

[501] M. Putinar (1993). Positive polynomials on compact semi-algebraic sets. *Indiana University Mathematics Journal*, 42, 969–984. 731

[502] L. Qi, X. Chen (1997). A preconditioning proximal Newton method for nondifferentiable convex optimization. *Mathematical Programming*, 76, 411–429. 323

[503] L. Qi, J. Sun (1993). A nonsmooth version of Newton's method. *Mathematical Programming*, 58, 353–367. [doi]. 430

[504] M. Ramana (1997). An exact duality theory for semidefinite programming and its complexity implications. *Mathematical Programming*, 77, 129–162. 743

[505] M. Ramana, L. Tunçel, H. Wolkowicz (1997). Strong duality for semidefinite programming. *SIAM Journal on Optimization*, 7, 641–662. 743

[506] J. Raphson (1690). *Analysis aequationum universalis seu ad aequationes algebraicas resolvendas methodus generalis, et expedita, ex nova infinitarum serierum doctrina deducta ac demonstrata*. London. 429

[507] M. Raussen, C. Skau (2015). Interview with Abel Laureate Louis Nirenberg. *Newsletter of the European Mathematical Society*, 98, 87–121. 809

[508] J. Rawls (1987). *Théorie de la Justice*. Collection Points 354. Seuil. Traduction française de Catherine Audard da la version remaniée en 1975 de *A Theory of Justice*, The Belknap Press of Harvard University Press, 1971. 510

[509] B. Recht, M. Fazel, P. Parrilo (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 53, 471–501. [doi]. 557, 910

[510] S. Reich (1977). On infinite products of resolvents. *Rend. Classe Sci. Fis. Mat. e Nat. Accad. Naz. Lincei Ser. VIII*, LXIII, Fasc. 5. 499

[511] F. Rendl, H. Wolkowicz (1997). A semidefinite framework for trust region subproblems with applications to large scale minimization. *Mathematical Programming*, 77(1), 273–299. [doi]. 402

[512] J. Renegar (2001). *A Mathematical View of Interior-Point Methods in Convex Optimization*. MPS-SIAM Series on Optimization 3. SIAM. 555, 616, 679

[513] S.M. Robinson (1976). Stability theory for systems of inequalities, part II: differentiable nonlinear systems. *SIAM Journal on Numerical Analysis*, 13, 497–513. [doi]. 181, 212, 242

[514] S.M. Robinson (1982). Generalized equations and their solutions, Part II: Applications to nonlinear programming. *Mathematical Programming Study*, 19, 200–221. 247, 565

[515] R.T. Rockafellar (1969). Duality in nonlinear programming. Mathematics of the Decision Sciences, Part I, pages 401–422. American Mathematical Society, Providence. 509

[516] R.T. Rockafellar (1969). Convex functions and duality in optimization problems and dynamics. Lecture Notes in Operations Research and Mathematical Economics 11, pages 117–141. Springer. 551

[517] R.T. Rockafellar (1970). *Convex Analysis*. Princeton Mathematics Ser. 28. Princeton University Press, Princeton, New Jersey. 28, 73, 105, 106, 145, 551, 837, 842

[518] R.T. Rockafellar (1971). New applications of duality in convex programming. In *Proceedings of the 4th Conference of Probability, Brasov, Romania*, pages 73–81. (version écrite d'un exposé donné à différentes conférences, en particulier au "7th International Symposium on Mathematical Programming", La Haye, 1970). 504, 551

[519] R.T. Rockafellar (1973). A dual approach to solving nonlinear programming problems by unconstrained optimization. *Mathematical Programming*, 5, 354–373. [doi]. 542, 549, 551

[520] R.T. Rockafellar (1973). The multiplier method of Hestenes and Powell applied to convex programming. *Journal of Optimization Theory and Applications*, 12, 555–562. [doi]. 492, 504, 552

[521] R.T. Rockafellar (1974). Augmented Lagrange multiplier functions and duality in nonconvex programming. *SIAM Journal on Control*, 12, 268–285. 504
[522] R.T. Rockafellar (1974). *Conjugate Duality and Optimization*. Regional Conference Series in Applied Mathematics 16. SIAM, Philadelphia, PA, USA. 551
[523] R.T. Rockafellar (1976). Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14, 877–898. [doi]. 149, 155, 323
[524] R.T. Rockafellar (1976). Augmented Lagrangians and applications of the proximal point algorithm in convex programming. *Mathematics of Operations Research*, 1, 97–116. 155, 323, 504
[525] R.T. Rockafellar (1981). Proximal subgradients, marginal values, and augmented Lagrangians in nonconvex optimization. *Mathematics of Operations Research*, 6, 424–436. 242
[526] R.T. Rockafellar (1993). Lagrange multipliers and optimality. *SIAM Review*, 35, 183–238. 19, 161, 242
[527] R.T. Rockafellar, R. Wets (1991). Scenarios and policy aggregation in optimization under uncertainty. *Mathematics of Operations Research*, 16, 119–147. 552
[528] R.T. Rockafellar, R. Wets (1998). *Variational Analysis*. Grundlehren der mathematischen Wissenschaften 317. Springer. 507
[529] C. Roos, T. Terlaky, J.-Ph. Vial (1997). *Theory and Algorithms for Linear Optimization – An Interior Point Approach*. John Wiley & Sons, Chichester. 678
[530] A. Ruszczyński (2006). *Nonlinear Optimization*. Princeton University Press, Princeton. 248
[531] Y. Saad (2003). *Iterative Methods for Sparse Linear Systems* (seconde édition). SIAM Publication, Philadelphia. 366
[532] Y. Saad, M.H. Schultz (1986). GMRES: a generalized minimal residual algorithm for solving non symmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 7, 856–869. 360, 366
[533] Y. Saad, H.A. van der Vorst (2000). Iterative solution of linear systems in the 20th century. *Journal of Computational and Applied Mathematics*, 123, 1–33. 366
[534] R. Saigal (1995). *Linear Programming – A Modern Integrated Analysis*. Kluwer Academic Publisher, Boston. 643, 678
[535] J.W. Sawyer (1984). First partial differentiation by computer with an application to categorial data analysis. *The American Statistician*, 38, 300–308. 281
[536] Cplex. [internet]. 678
[537] Mosek. [internet]. 678
[538] SeDuMi. [internet]. 678
[539] D. Schott (1995). Basic properties of Fejer monotone sequences. *Rostocker Mathematisches Kolloquium*, 49, 57–74. 292
[540] A. Schrijver (1986). *Theory of Linear and Integrated Programming*. John Wiley & Sons. 643
[541] L. Schwartz (1991). *Analyse I – Théorie des Ensembles et Topologie*. Hermann, Paris. 758
[542] L. Schwartz (1992). *Analyse II – Calcul Différentiel et Équations Différentielles*. Hermann, Paris. 789, 807
[543] M. Schweighofer (2005). Optimization of polynomials on compact semialgebraic sets. *SIAM Journal on Optimization*, 15, 805–825. 731
[544] R. Shamir (1987). The efficiency of the simplex method: a survey. *Management Science*, 33, 301–334. 642
[545] D.F. Shanno (1970). Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation*, 24, 647–656. 440
[546] A. Shapiro (1990). On concepts of directional differentiability. *Journal of Optimization Theory and Applications*, 66, 477–487. [doi]. 790

[547] A. Shapiro (1994). Directionally nondifferentiable metric projection. *Journal of Optimization Theory and Applications*, 1, 203–204. 47

[548] N.Z. Shor (1987). Quadratic optimization problems. *Soviet Journal of Computer and System Sciences*, 25, 1–11. Traduit de *Izvestiya Akademii Nauk SSSR Tekhnicheskaya Kibernetika*, 1987, no. 25:1, 128-139. 732

[549] G.A. Shultz, R.B. Schnabel, R.H. Byrd (1985). A family of trust-region-based algorithms for unconstrained minimization with strong global convergence properties. *SIAM Journal on Numerical Analysis*, 22, 47–67. 402

[550] D. Siegel (1993). Updating conjugate direction matrices using members of Broyden's family. *Mathematical Programming*, 60, 167–185. 449

[551] P.J.S. Silva, J. Eckstein (2006). Double-regularization proximal methods, with complementarity applications. *Computational Optimization and Applications*, 33, 115–156. 504

[552] T. Simpson (1737). *A New Treatise of Fluxions*. 429

[553] T. Simpson (1740). Essays on several curious and useful subjects in speculative and mix'd mathematiks, illustrated by a variety of examples. *London*. 429

[554] M. Sion (1958). On general minimax theorems. *Pacific Journal of Mathematics*, 8, 171–176. 551

[555] M. Slater (1950). Lagrange multipliers revisited: a contribution to non-linear programming. Cowles Commission Discussion Paper, Math. 403. 242

[556] S. Smale (1983). On the average number of steps of the simplex method of linear programming. *Mathematical Programming*, 27, 241–262. 642

[557] M.V. Solodov, B.F. Svaiter (1999). A hybrid approximate extragradient-proximal point algorithm using the enlargement of a maximal monotone operator. *Set-Valued Analysis*, 7, 323–345. [doi]. 498

[558] M.V. Solodov, B.F. Svaiter (1999). A hybrid projection-proximal point algorithm. *Journal of Convex Analysis*, 6, 59–70. [journal]. 498

[559] M.V. Solodov, B.F. Svaiter (2000). An inexact hybrid generalized proximal point algorithm and some new results on the theory of Bregman functions. *Mathematics of Operations Research*, 25, 214–230. [doi]. 498

[560] G. Sonnevend, J. Stoer, G. Zhao (1989). On the complexity of following the central path of linear programs by linear extrapolation. *Mathematics of Operations Research*, 63, 19–31. 678

[561] G. Sonnevend, J. Stoer, G. Zhao (1991). On the complexity of following the central path of linear programs by linear extrapolation II. *Mathematical Programming*, 52, 527–553. 678

[562] L. Sorber, M. Van Barel, L. De Lathauwer (2012). Unconstrained optimization of real functions in complex variables. *SIAM Journal on Optimization*, 22(3), 879–898. 454

[563] D.C. Sorensen (1982). Newton's method with a model trust region modification. *SIAM Journal on Numerical Analysis*, 19, 409–426. 402

[564] D.C. Sorensen (1982). Collinear scaling and sequential estimation in sparse optimization algorithms. *Mathematical Programming*, 18, 135–159. 454

[565] B. Speelpenning (1980). *Compiling fast partial derivatives of functions given by algorithms*. PhD thesis, Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801. 281

[566] D.A. Spielman, S.-T. Teng (2004). Smoothed analysis of algorithms: why the simplex algorithm usually takes polynomial time. *Journal of Assoc. Comput. Mach.*, 51, 385–463. [doi]. 642

[567] J.E. Spingarn (1987). A projection method for least-squares solutions to overdetermined systems of linear inequalities. *Linear Algebra and its Applications*, 86, 211–236. [doi]. 499

[568] S. Sra, S. Nowozin, S.J. Wright, éditeurs (2011). *Optimization for Machine Learning*. MIT Press, Cambridge. 326

[569] T. Steihaug (1983). The conjugate gradient method and trust regions in large scale optimization. *SIAM Journal on Numerical Analysis*, 20, 626–637.

[570] R.J. Stern, H. Wolkowicz (1995). Indefinite trust region subproblems and nonsymmetric eigenvalue perturbations. *SIAM Journal on Optimization*, 5(2), 286–313. [doi]. 402

[571] E. Stiefel (1952). Ausleichung ohne aufstellung der gausschen normalgleichungen. *Wiss. Z. Technische Hochschule Dresden*, 2, 441–442. 686

[572] J. Stoer (1963). Duality in nonllinear programming and the minimax theorem. *Numerische Mathematik*, 5, 371–379. 509

[573] V. Strassen (1969). Gaussian elimination is not optimal. *Numerische Mathematik*, 13, 354–356. 365

[574] B. Stroustrup (1994). *The Design and Evolution of C++*. Addison-Wesley. iv

[575] J.F. Sturm (1999). Using SeDuMi 1.02, a Matlab toolbox for optimization over symmetric cones. *Optimization Methods and Software*, 11, 625–653. 678, 726

[576] M. Teboulle (1992). Entropic proximal mappings with applications to nonlinear programming. *Mathematics of Operations Research*, 17, 670–681. 145, 504

[577] C.M. Teobald (1975). An inequality for the trace of the product of two symmetric matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, 77(2), 265–266. 785

[578] T. Terlaky (1985). A convergent criss-cross method. *Math. Operationsforsch. und Statist., ser. Optimization*, 16, 683–690. 641

[579] T. Terlaky (1987). A finite criss-cross method for oriented matroids. *Journal of Combinatorial Theory*, 42, 319–327. 641

[580] T. Terlaky, éditeur (1996). *Interior Point Methods of Mathematical Programming*. Kluwer Academic Press, Dordrecht. [doi]. 678

[581] T. Terlaky, S. Zhang (1993). Pivot rules for linear programming – a survey. *Annals of Operations Research*, 46, 203–233. [doi]. 642

[582] S.W. Thomas (1975). *Sequential estimation techniques for quasi-Newton algorithms*. Thèse de doctorat, Cornell University, Ithaca, NY. 402

[583] R.J. Tibshirani (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 58(1), 267–288. 711

[584] R.J. Tibshirani (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 73(3), 273–282. [doi]. 711

[585] R.J. Tibshirani (2013). The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7, 1456–1490. [doi]. 711

[586] A.N. Tikhonov (1963). Solution of ill-posed problems and the regularization method. *Soviet Mathematics Doklady*, 4, 1035–1038. 714

[587] F. Tisseur (2001). Newton's method in floating point arithmetic and iterative refinement of generalized eigenvalue problems. *SIAM Journal on Matrix Analysis and Applications*, 22, 1038–1057. 430

[588] Hans Raj Tiwary (2008). On computing the shadows and slices of polytopes. Rapport de recherche. [arXiv:0804.4150v2] (2012 version). 65

[589] M.J. Todd (2001). Semidefinite optimization. In *Acta Numerica 2001*, pages 515–560. Cambridge University Press. [doi]. 743

[590] M.J. Todd, K.-C. Toh, R.H. Tütüncü (1998). On the Nesterov-Todd direction in semidefinite programming. *SIAM Journal on Optimization*, 8, 769–796. [doi]. 742, 743

[591] M.J. Todd, K.-C. Toh, R.H. Tütüncü (1999). SDPT3 - A MATLAB software package for semidefinite programming, version 1.3. *Optimization Methods and Software*, 11/12(1-4), 545–581. [doi]. 726

[592] Ph.L. Toint (1977). On sparse and symmetric matrix updating subject to a linear equation. *Mathematics of Computation*, 31, 954–961. 454
[593] J.F. Toland (1978). Duality in nonconvex optimization. *Journal of Mathematical Analysis and Applications*, 66, 399–415. [doi]. 529, 551
[594] L.N. Trefethen, D. Bau (1997). *Numerical Linear Algebra*. SIAM Publication, Philadelphia. 785
[595] M. Trnovská (2005). Strong duality conditions in semidefinite programming. *Journal of Electrical Engineering*, 56(12/s), 1–5. 743
[596] K. Ueda, N. Yamashita (2010). On a global complexity bound of the Levenberg-Marquardt method. *Journal of Optimization Theory and Applications*, 147(3), 443–453. [doi]. 704, 714
[597] K. Ueda, N. Yamashita (2012). Global complexity bound analysis of the Levenberg-Marquardt method for nonsmooth equations and its application to the nonlinear complementarity problem. *Journal of Optimization Theory and Applications*, 152, 450–467. [doi]. 701, 714
[598] M. Ulbrich (2011). *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*. MPS-SIAM Series on Optimization 11. SIAM Publications, Philadelphia, PA, USA. [doi]. 431
[599] A. van der Sluis (1969). Condition numbers and equilibration of matrices. *Numerische Mathematik*, 15, 14–23. 355
[600] H.A. van der Vorst (1990). The convergence behaviour of preconditioned CG and CG-S in the presence of rounding errors. In O. Axelsson, L.Y. Kolotilina, éditeurs, *Preconditioned Conjugate Gradient Methods*, Lecture Notes in Mathematics 1457, pages 126–136. Springer, Berlin. 365
[601] H.A. van der Vorst (2000). Krylov subspace iteration. *Computing in Science & Engineering*, 2, 32–37. [doi]. 342, 366
[602] H.A. van der Vorst (2003). *Iterative Krylov Methods for Large Linear Systems*. Cambridge monographs on applied and computational mathematics 13. Cambridge University Press, Oxford. 366
[603] L. Vandenberghe, S. Boyd (1996). Semidefinite programming. *SIAM Review*, 38, 49–95. [doi]. 743
[604] R.J. Vanderbei (1997). *Linear Programming: Foundations and Extensions*. Kluwer Academic Publishers, Boston. 643, 678
[605] R.S. Varga (1962). *Matrix Iterative Analysis*. Prentice-Hall, Upper Saddle River, NJ, USA. 340, 342
[606] S.A. Vavasis (1991). *Nonlinear Optimization – Complexity Issues*. Oxford University Press, New York. 259, 260
[607] Vax Unix Macsyma: Reference manual, version 11 (1985). Symbolics. 276
[608] V. Venkatakrishnan (1989). Newton solution of inviscid and viscous problems. *AIAA J.*, 27, 885–891. 427
[609] J. Ville (1938). Sur la théorie générale des jeux où intervient l'habileté des joueurs. In *Traité du Calcul des Probabilités et de ses Applications,* E. Borel, Tome IV, Fascicule II*, Applications aux jeux de hasard,* J. Ville (ed.), pages 105–113. Gauthier-Villars, Paris. 70
[610] P.M.B. Vitanyi (2009). Turing machine. Scholarpedia, 4(3):6240. 260
[611] V.V. Voevodin (1983). The problem of a non-selfadjoint generalization of the conjugate gradient method has been closed. *USSR Computational Math. and Math. Phys.*, 23, 143–144. 360
[612] J. von Neumann (1928). Zur Theorie der Gesellshaftsspiele. *Mathematische Annalen*, 100, 295–320. 551, 645
[613] J. von Neumann (1937). Some matrix inequalities and metrization of matrix-space. *Tomsk University Review*, 1, 286–300. *Collected Works*, Pergamon, Oxford, 1962, Volume IV, 205-218. 785

[614] M. Walk (1989). *Theory of Duality in Mathematical Programming*. Springer, Wien. 551
[615] Z. Wang (1987). A conformal elimination-free algorithm for oriented matroid programming. *Chin. Ann. Math.*, 8, B1. 641
[616] L.T. Watson (1990). Globally convergent homotopy algorithms for nonlinear systems of equations. *Nonlinear Dynamics*, 1, 143–191. 428
[617] D.T. Whiteside, éditeur (1967-1976). *The Mathematical Papers of Isaac Newton, Volumes I-VII*. Cambridge University Press, Cambridge. 453
[618] R.A. Wiggins (1972). General linear inverse problem – implication of surface waves and free ascillatiions for Earth structure. *Rev. Geophys. Space Phys.*, 10, 251–285. 688
[619] R. Wilkinson, K. Pickett (2013). *Pourquoi l'Égalité est Meilleure Pour Tous*. Les Petits Malins. 682
[620] A.C. Williams (1970). Boundedness relations for linear constraint sets. *Linear Algebra and its Applications*, 3, 129–141.
[621] R.B. Wilson (1963). *A simplicial algorithm for concave programming*. Thèse de doctorat, Graduate School of Business Administration, Harvard University, Cambridge, MA, USA. 579
[622] P. Wolfe (1961). A duality theorem for non-linear programming. *Quaterly of Applied Mathematics*, 19, 239–244. [doi]. 509, 555
[623] P. Wolfe (1969). Convergence conditions for ascent methods. *SIAM Review*, 11, 226–235. [doi]. 322
[624] P. Wolfe (1971). Convergence conditions for ascent methods II: some corrections. *SIAM Review*, 13, 185–188. [doi]. 322
[625] P. Wolfe (1972). On the convergence of gradient methods under constraint. *IBM Journal of Research and Development*, 16, 407–411. 459
[626] H. Wolkowicz, R. Saigal, L. Vandenberghe, éditeurs. *Handbook of Semidefinite Programming – Theory, Algorithms, and Applications*, International Series in Operations Research & Management Science, Tome 27. Kluwer Academic Publishers. 679, 725, 731, 743
[627] D.H. Wolpert, W.G. Macready (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1, 67–82. 6
[628] S. Wright, D. Orban (2002). Properties of the log-barrier function on degenerate nonlinear programs. *Mathematics of Operations Research*, 27, 585–613. 616
[629] S.J. Wright (1997). *Primal-Dual Interior-Point Methods*. SIAM Publication, Philadelphia. 672, 678
[630] S.J. Wright (1999). Modified Cholesky factorizations in interior-point algorithms for linear programming. *SIAM Journal on Optimization*, 9, 1159–1191. 678
[631] T. Yamamoto (2000). Historical developments in convergence analysis for Newton's and Newton-like methods. *Journal of Computational and Applied Mathematics*, 124, 1–23. 429
[632] Y. Ye (1992). A new complexity result on minimization of a quadratic function with a sphere constraint. In *Recent advances in global optimization (Princeton, NJ, 1991)*, pages 19–31. Princeton Ser. Comput. Sci., Princeton Univ. Press, Princeton, NJ. 402
[633] Y. Ye (1997). *Interior Point Algorithms – Theory and Analysis*. Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley. 652, 678
[634] Y. Ye (1997). Approximating quadratic programming with bound constraints. Working paper, Department of Management Science, University of Iowa, Iowa City, USA. 733
[635] Y. Ye (1999). Approximating quadratic programming with bound and quadratic constraints. *Mathematical Programming*, 84, 219–226. [doi]. 733
[636] M. Yourcenar (1968). *L'œuvre au noir*. Gallimard. 507
[637] R. Yousefpour (2015). Combination of steepest descent and BFGS methods for nonconvex nonsmooth optimization. *Numerical Algorithms*. [doi]. 454

[638] T.J. Ypma (1983). Finding a multiple zero by transformations and Newton-like methods. *SIAM Review*, 25, 365–378. 430

[639] T.J. Ypma (1995). Historical development of the Newton-Raphson method. *SIAM Review*, 37, 531–551. 429, 453

[640] A.J. Zaslavski (2010). *Optimization on Metric and Normed Spaces*. Optimization and its Applications 44. Springer, New York. 16, 505

[641] X. Zhan (2002). *Matrix Inequalities*. Springer. 785

[642] Y. Zhang (1998). Solving large-scale linear programs with interior-point methods under the Matlab environment. *Optimization Methods and Software*, 10, 1–31. 678

[643] Ruizue Zhao, Jinyan Fan (2016). Global complexity bound of the levenberg-marquardt method. *Optimization Methods and Software*, 31(4), 805–814. [doi]. 714

[644] G. Zoutendijk (1970). Nonlinear programming, computational methods. In J. Abadie, éditeur, *Integer and Nonlinear Programming*, pages 37–86. North-Holland, Amsterdam. 322

[645] J. Zowe, S. Kurcyusz (1979). Regularity and stability for the mathematical programming problem in Banach spaces. *Applied Mathematics and Optimization*, 5(1), 49–62. 242, 247

Index

adhérence, 748
– enveloppe affine d'une –, 34, 66
– minimisation d'une fonction continue sur une –, 17
adhérence d'une fonction, 108
admissibilité asymptotique du pas unité
– algorithme de Newton pour un système d'équations, 418
affine, *voir* enveloppe, fonction, hyperplan, minorante, sous-espace
affinement indépendants, *voir* vecteurs
algorithme, *voir aussi* méthode
– l-BFGS, 453
– à directions de descente, 299
– à mémoire limitée, 434
– à régions de confiance, 370
– ADMM, *voir* algorithme du lagrangien augmenté à directions alternées
– affine, 656
– avec activation de contraintes, 465
– CGLS (gradient conjugué pour problèmes de moindres-carrés linéaires), 686
– d'Arrow-Hurwicz, 550
– de BFGS, 443
– de Fletcher-Lemaréchal, **310**, 310, 313, 814, 873
– de Gauss-Newton, 304, **695**
– de Gram-Schmidt, 762
– de la plus profonde descente, 301, 326
– de la sécante, 436
– de Levenberg-Morrison-Marquardt, 697, 700
– de minimisation de la fonction duale, 547
– de Moré-Sorensen, 391
– de Newton en optimisation, 302, **409**
– – tronqué, 420, **422**, 430
– de Newton pour système non linéaire, 405
– – globalisation de la convergence par recherche linéaire, 414
– – inexact, 411, 414, 430
– de pénalisation extérieure, 482
– de pénalisation intérieure, 488
– de points intérieurs, 620
– de points-frontière, 620
– de quasi-Newton, 303
– – réduit, 578
– de suivi de chemin, 657
– des directions conjuguées, 347–349
– dogleg de Powell, 387
– du gradient, 301, 326
– du gradient conjugué, 302, 341, **349–350**, 685
– – préconditionné, 353–356, 366, 367
– – projeté, 356, 366
– – réduit, 366
– du gradient projecté, 462
– du lagrangien augmenté, 549
– du lagrangien augmenté à directions alternées, 499
– du premier ordre, 325
– du résidu minimal (GMRES), 342, **359–365**
– du résidu orthogonal, 361
– du simplexe, 620
– du simplexe révisé, 634, **640**
– en croix, 641
– GMRES, 363
– LSQR (Paige et Saunders), 686
– OQS, **568**, 568
– – réduit, 577
– proximal, 328, 340
angle de descente, **298**, 303, 313, 416, 450
anti-cyclage, *voir* règle d'anti-cyclage de l'algorithme du simplexe
antigradient, 326
application linéaire
– adjointe, 753

– auto-adjointe, 756
– continue, 753
– continue inversible, 753
application-coordonnée, 761
arbre de scénarios, 552
arête d'un ensemble convexe, 30
argmin, 5
Armijo, 307
Arnoldi, 361
Arrow, 550
augmentabilité, 786

Bachelard, iii
Banach
– lemme de perturbation, 753
base
– d'indices, 623
– d'indices associée à un sommet, 623
– d'un espace vectoriel, 761
– de Gröbner, 411
Baudelaire, 979
Benders, 643
Borel, 748
borne d'erreur, 212, 645
Borsuk, 429
Borwein, 809
boule, *voir aussi* minimisation
– centre d'une –, 750
– fermée, 750
– ouverte, 750
– rayon d'une –, 750
Bouveresse, iii
Broyden, 440, 455

calcul des solutions, *voir aussi* vérification de solution
– de (P_E), 176–177
– de (P_{EI}), 205–206
Carathéodory, 26
cas difficile, 384
Cauchy, 297, 385, *voir aussi* entrelacement, inégalité, pas de recherche linéaire, point, recherche linéaire, suite
Céline, iii
centrage, 659
centre analytique, 649, 652
– de la face optimale, 653
centre de gravité, 652
certificat d'inconsistance, 555
changement de variables, 11
chemin central
– en optimisation linéaire, 647, 649
chemin continu, 680
Cholesky, 785, *voir aussi* factorisation
cible de Fejér, 292
$\mathcal{C}^n$, 70
– cône dual, 70
$\mathcal{C}^{n+}$, 70
– cône dual, 70
code
– linéaire cotangent, 283
– linéaire tangent, 279
codimension, 762
coercivité, *voir* fonction coercive
comatrice, 786
combinaison
– affine, 23
– conique, 27
– convexe, 25
combinaison d'éléments, 929
combinatoire, 608
combinatoire des problèmes d'optimisation, 179, **206**, 650
commande
– optimale, 265
communication
– directe, 286
– inverse, 287
commutativité
– d'une inf-convolution, 105
– de matrices, 786
compatibilité d'égalités et d'inégalités linéaires, 70
compatible, *voir aussi* réalisable
complément de Schur, 725, **776**
– singulier, 744
complémentarité, 184, 627
– stricte, 184, 736
complémentarité, 498
complexité, 259–264, 365
– en moyenne, 642
– en perturbations moyennées, 642
– itérative, 657, 659
– pire-cas, 642
complexité itérative
– de l'algorithme de Gauss-Newton, 696
– de l'algorithme de Levenberg-Morrison-Marquardt (LMM), 704
composante
– basique, 623
– non basique, 623
compteur de composante non nulle, 153, 706

concordance, 372
condition, *voir aussi* inégalité
– d'Armijo, 306
– de croissance positive à l'infini, 155
– de décroissance linéaire, 306
– de décroissance forte, 389
– de décroissance suffisante, 386
– de Powell, 392
– de Wolfe, 309
– de Zoutendijk, 313
conditionnement, 264, 303, 450
– en pénalisation, 486
conditions d'optimalité, 158
– de Karush-Kuhn-Tucker, 183
– du premier ordre, 158
– du second ordre, 158
– nécessaires
– – du second ordre faibles, 203
– – du second ordre fortes, 203
– – du second ordre semi-fortes, 203
– nécessaires, 158
– – du premier ordre (CN1), **158**, 162–164, 169, 182
– – du second ordre (CN2), **158**, 164, 174, 202
– suffisantes, 158
– – du premier ordre (CS1), **158**, 163, 164, 173, 186, 211
– – du second ordre (CS2), **158**, 165, 175, 204, 243, 244
– – du second ordre diffuse en optimisation sans contrainte, 243
– – du second ordre faibles, 200
– – du second ordre fortes, 199, 205
– – du second ordre semi-fortes, 199, 205
cône, 27
– adhérence, 67
– asymptotique, *voir* cône asymptotique
– autodual, **56**, 69, 718
– bidual, *voir* cône bidual
– contingent, *voir* cône tangent
– critique, **198**
– de Lorentz, *voir* cornet
– de récession, 28, *voir* cône asymptotique
– des directions admissibles, **62**, **159**, 163
– du second ordre, *voir* cornet
– dual, *voir* cône dual
– enveloppe affine, 67
– épointé, 27
– intérieur relatif, 67
– linéarisant, 179, 208
– normal, *voir* cône normal
– pointé, 27
– polyédrique convexe, 42
– sailant, 27
– tangent, *voir* cône tangent
cône asymptotique, 28–30, 66, 97, 99, 100, 146, 632, 633, 654, 655
– d'un cône convexe fermé, 29
– d'un polyèdre convexe, 43, 67, 68
– image par une application linéaire, 29
– image réciproque par une application linéaire, 29
– inclusion, 29
– intersection, 29
– produit cartésien, 29
cône bidual, 56
– expression, 56
cône dual, 55
– d'un produit, 56
– d'une eveloppe convexe, 56
– d'une somme, 56
– d'une union, 56
– de $\mathcal{C}^n$, 70
– de $\mathcal{C}^{n+}$, 70
– de $\mathbb{R}^n_+$, 69
– de $\mathcal{S}^n_+$, 69
– du cornet $\mathbb{R}^{n+1}_\nabla$, 69
– du simplexe ordonné, 69
– et inclusion, 56
– intérieur, 70
– intérieur relatif, 70
– négatif, **56**, 68
cône normal, 161
– à $\mathcal{S}^n_+$, 71
– à un convexe, **48**
– à un produit cartésien, 49
– à une intersection, 49
– sous-différentiel de l'indicatrice, 151
cône tangent, 159
– à $\mathcal{S}^n_+$, 71
– à un convexe, 63
– – transport affine, 243
– à un pavé, 471
– à un polyèdre convexe, 71
– à un produit de convexes, 64
– à une intersection d'ensembles, 242
– à une réunion d'ensembles, 242
conjecture $\mathsf{P} = \mathsf{NP}$, 179
conjuguée
– d'une enveloppe inférieure, 148

– d'une enveloppe supérieure, 149
– d'une fonction quadratique strictement convexe, 148
– d'une indicatrice, 151
– d'une inf-convolution, 119
– d'une inf-image sous une application linéaire, 115
– d'une fonction d'appui, 151
– d'une norme, 151
– de la valeur propre maximale, 153, 154
contractant, *voir* fonction
– strictement, *voir* fonction
contrainte
– active, 178
– affine, 460
– de borne, 460
– faiblement active, 195
– fortement active, 195
– incompatible, 569
– non dégénérée, 469
convergence
– locale, 569, 571
– superlinéaire
– – en 2 pas, 578
convexité
– d'une inf-image sous une application linéaire, 107
– d'une fonction composée, 102
– d'une fonction maximale, 153
– d'une indicatrice, 77
– d'une norme, 151, 152
– de la valeur propre maximale, 153, 154
cornet, 69
– cône dual, 69
correction de Powell, 575
Courant, 767
coût, 619
– réduit, 635
critère, 4, 619
croissance quadratique, 203
Curry, 305
cyclage, 469

Dantzig, 619, 633, 643
Dawkins, 249
décomposition, *voir* factorisation
– de Benders, 643
– de Dantzig-Wolfe, 643
– de Moreau, 68
– lagrangienne, 643
demi-droite, 29
demi-espace fermé, 22
dense, 748
dérivabilité, *voir* différentiabilité
deerivée
– partielle, 794
dérivée, 793
– à droite, 790
– à gauche, 790
dérivée directionnelle
– d'une indicatrice, 151
– d'une composition, 791
– d'une norme, 152
– de la valeur propre maximale, 154
déterminant, 766
– dérivée, 786
diagramme de Voronoï, 65, **820**
différentiabilité
– au second ordre, 802
– au sens de Fréchet, 87, **793**
– au sens de Gâteaux, 87, **791**
– directionnelle, 789
– partielle, 87
différentiation automatique, 276
– en mode direct, 279
– en mode inverse, 283
dimension
– d'image, 763, 785
– d'un ensemble convexe, 24
– d'un espace vectoriel, 761
– d'un sous-espace affine, 752
– de noyau, 763, 785
– finie, 761
direction
– à courbure négative, 388
– admissible, 465
– asymptotique, 28
– conjuguée, 347
– critique, 198
– de descente, 298
– – admissible, 465
– de descente admissible, 459, 465
– de Gauss-Newton, 304
– de la plus profonde descente, 301
– de Newton, 303
– de Newton inexacte, 411
– de non bornitude, 356
– de quasi-Newton, 303
– de récession, 28
– du gradient, 301
– du gradient conjugué, 302
– à courbure quasi-négative, 421
direction admissible, 62

direction de Newton en OSDP
– direction d'AHO, 742
– directions de HKM duale, 743
– directions de HKM primale, 742
– directions de MZ, 742
– directions de NT, 743
distance, **750**
– à un ensemble, 45, 557, 750
– – contractilité, 758
– – différentiabilité, 808
– – sous-différentiabilité, 154
– entre deux ensembles, 69
distance au carré
– convexité, 155
– différentiabilité, 155
divergence
– de Bregman, 504
– φ-divergence, 504
domaine
– d'une fonction, 5, 74
– d'une multifonction, 757
– du sous-différentiel, 131
droite, 601
– achevée, 4
droite achevée, 749
dual, *voir* cône dual, norme duale, problème dual
dualisation de contraintes fonctionnelles
– lagrangienne de (P_{EI}), 531
– wolfienne d'un problème convexe, 555
dualité
– de Fenchel, 524–526
– faible, 508, **512**, 520, 524, 720
– non convexe de Toland, 526–531
– saut de –, 513, **521**, 658, 720

écart en carré, 181
échec d'une itération, 372
élimination
– de Fourier, 38, 65
– gaussienne, 38, 341
élimination gaussienne, 773, 775
– par blocs, 775
ellipsoïde, 583
– de Dikin, 647
– inscrit dans l'orthant positif, 616
ellipsoïde circonscrit, 644
ensemble
– admissible, 4, 157, 620
– – défini par des contraintes abstraites, 207
– – défini par des contraintes d'égalité et d'inégalité X_{EI}, 178
– – défini par des contraintes d'égalité X_E, 165
– affine, *voir* sous-espace
– borné, 751
– connexe par arcs, 680
– convexe, 21
– de sous-niveau, **99**, 146
– des applications linéaires continues inversibles, 753
– des matrices définies positives, *voir* $\mathcal{S}^n_{++}$
– des matrices semi-définies positives, *voir* $\mathcal{S}^n_+$
entrelacement
– de Cauchy, 786
– des valeurs propres, 768
enveloppe
– affine, *voir* enveloppe affine
– conique, *voir* enveloppe conique
– convexe d'un ensemble, *voir* enveloppe convexe d'un ensemble
– convexe fermée d'un ensemble, *voir* enveloppe convexe fermée d'un ensemble
– inférieure, *voir* enveloppe inférieure
– supérieure, *voir* enveloppe supérieure
enveloppe affine, 23, 34
– conservation de l'inclusion, 66
– d'un cône, 67
– d'un convexe, 66
– d'un intérieur relatif, 32, 67
– d'un polyèdre convexe, 67
– d'un produit cartésien, 66
– d'une adhérence, 34, 66
– d'une enveloppe conique, 66
– d'une intersection, 66
– d'une somme, 66
enveloppe conique, 27
– d'un polyèdre convexe, 68
– d'une somme, 66
enveloppe convexe, 24
– d'un ensemble relativement ouvert, 69
– d'un produit cartésien, 27
– d'une somme, 27
– d'une transformation affine, 27
enveloppe convexe fermée, 54
– minimisation d'une fonction linéaire sur une –, 69
enveloppe convexe fermée d'une fonction, 114
enveloppe inférieure, 148

– conjuguée, 149
enveloppe supérieure, 103
– conjuguée, 149
– épigraphe, 103
– sous-differentiel, 138–139
épigraphe, 15, **74**, **749**
– d'une enveloppe supérieure de fonctions, 103
– d'une inf-convolution, 105, 106
– stricte, 74
équation
– adjointe, **268**, 273, 275
– d'état, 577
– de Fermat, 241
– de l'état adjoint, *voir* équation adjointe
– de Newton, 405, 420, 677
– de quasi-Newton, 435
– d'état, 266, 271
– normale, 341, 677, 683
équivalence entre problèmes d'optimisation, 10–15
– changement de variables, 11
– minimisation emboîtée, 12, 14
– minimisation-maximisation, 11
– passage d'un terme du critère en contrainte, 15
erreur, 250
erreur amont, 300
espace
– actif, 466
– compact, 748
– complet, 752
– d'Asplund, 145
– de Banach, 752
– de Hilbert (ou hilbertien), 755
– dual (topologique), 753
– euclidien, 755
– métrique, 750
– – topologie d'un –, 750
– normé, 751
– pré-hilbertien, 755
– topologique, 747
estimation de paramètres, 681
état
– adjoint, 269, 578
– stationnaire, 427
Euler, 241, *voir aussi* équation, schéma
Everett, 556
existence de solution
– d'un problème d'optimisation
– – linéaire, 40
– – par l'approche topologique (Weierstrass), 7
– – par le comportement à l'infini, 9, 100
– d'un système d'équations non linéaires, 407
– revue des méthodes, 9

face
– d'un polyèdre convexe, 67
– de $\mathcal{S}_+^n$, 743
face d'un convexe, 30
– engendrée par une partie, 30
– exposée, 66
– optimale, 652
– propre, 30
face exposée
– d'un polyèdre convexe, 67
facteur
– d'inexactitude, 411
– d'Oren-Luenberger, 449
factorisation
– de Bunch et Kaufman, 677
– de Bunch et Parlett, 677
– de Cholesky, 391, 420, 449, 685, **776**–**782**
– – creuse, 677
– – mise à jour des facteurs, 779, 780
– en valeurs singulières (SVD), 687, **782**–**784**
– gaussienne (LU), 341, 725, **773**–**775**
– QR, 686, **772**
factorisation
– spectrale, 767
famille de problèmes, 259
– CLIQUE, 263
– de classe NP, 261
– de classe NP-ardu, 263
– de classe NPC, 262
– de classe P, 261
– de décision, 260
– évaluation en temps polynomial, 261
– instance, 259
– OL, **259**, 259, 261
– OQ, **259**, 263, 264
– OQC, 261
– polynomialement réductible, 262
– SAT, **263**
– SOUS-SOMME, 263
famille de vecteurs
– libre, 761
Farkas, 59, 70, 645
Fejér, 292, *voir aussi* cible, suite

Fenchel, 111
Fermat, 157, 241
fermé, 747
Finsler, 765
Fisher, 767
Flaubert, iii
Fletcher, 310, 356, 357, 440, 814, 873, *voir aussi* pas de recherche linéaire
fonction
– affine, 77, 752
– als (asymptotically level stable), 65
– asymptotique, 98
– – d'une fonction composée, 100, 101
– autoconcrdante, 584
– B-différentiable, 430
– barrière, 487
– barrière autoconcrdante, 595
– biconjuguée, 113
– – d'une fonction composée, 118
– bilinéaire, 754
– – continue, 754
– – norme, 754
– C-fonction
– – *voir* C-fonction 933
– coercive, **8**, 17
– – forme bilinéaire, 17
– composée, **101**
– – biconjuguée, 118
– – conjuguée, 118
– – convexité, 102
– – fonction asymptotique, 100, 101
– – sous-différentiel, 135
– concave, **74**, 147
– conjuguée, 111
– – d'une fonction composée, 118
– continue, **749**, 750
– contractante, 141, 751, 758
– convexe, 74
– (convexe) polyédrique, 79
– croissante, 102
– d'appui, *voir* fonction d'appui
– de classe $\mathcal{C}^{1,1}$, 808
– de couplage, 511
– de mérite, 413
– de Minkowski, 597
– de moindres-carrés, 413, *voir aussi* problème
– de pénalisation, 474
– de récession, *voir* fonction asymptotique
– dérivée directionnelle, 86
– différentiable, *voir* différentiabilité
– duale, 511, 519, 542
– fermée, 76, 104, **750**
– fortement convexe, **75**, 96
– implicite, 266
– impropre, 75
– indicatrice, *voir* fonction indicatrice
– injective, 749
– intermédiaire, 278
– linéaire, 752, *voir* minimisation
– lipschitzienne, 751
– localement lipschitzienne, 751
– log-barrière (lb), **584**, 595, 616, 650
– log-déterminant (ld), 438, **585**, 595, 737
– marginale, *voir* fonction marginale
– max, **147**
– monotone, 141
– multivoque, *voir* multifonction
– non différentiable, 430
– pénalisante, 474
– polyédrique, *voir* fonction (convexe) polyédrique
– propre, **75**, 146
– quadratique, *voir aussi* fonction quadratique convexe
– quadratique convexe, **146**
– – conjuguée, 148
– – ensembles de sous-niveau, 146
– – régularisée de Moreau-Yosida, 156
– r-pentue, 220
– semi-continue inférieurement (s.c.i.), *voir* semi-continuité inférieure
– semi-continue supérieurement (s.c.s.), *voir* semi-continuité supérieure
– semi-lisse, 430
– séparable, 329
– sous-continue, 220
– sous-différentiable, 124, *voir aussi* sous-différentiabilité
– sous-linéaire, 80
– strictement contractante, 751
– strictement convexe, 75
– surjective, 749
– symétrique, 754
– valeur, *voir* fonction valeur
fonction coût, 4
fonction d'appui, **81**, 151
– conjuguée, 151
– du sous-différentiel, 130, 149
– sous-différentiel, 151
fonction indicatrice, **77**

– conjuguée, 151
– dérivée directionnelle, 151
– sous-différentiel, 151
fonction marginale, 104–106
– continuité, 758
– convexité, 104
– semi-continuité supérieure, 758
– sous-differentiel, 137
fonction max, 790
fonction maximale, 153
– convexité d'une –, 153
fonction objectif, 4
fonction valeur, 106, **228**, 518, 523, 524, 532, 540
– continuité directionnelle, 237–238
– convexité, 232
forme, 753
formule
– de Grassmann, 785
– de Sherman-Morrison-Woodbury, 787
– de Woodbury, 787
– du max, 130
formule de mise à jour, 434
– à mémoire limitée, 451
– de BFGS (Broyden-Fletcher-Goldfarb-Shanno), 440
– PSB (Powell-Symétrique-Broyden), 455
– SR1 (Symétrique de Rang 1), 437
Fourier, 38, 65
Fréchet, 756, 793
Frobenius, 764
Fromovitz, *voir* qualification des contraintes d'inégalité (QC-MF)
frontière, 748
– relative, **32**, 34

Gâteaux, 791
Gauss, 695, *voir aussi* algorithme, élimination, factorisation, méthode de Gauss-Seidel
globalisation de la convergence, 301, 412
– par méthode de continuation, 428
– par recherche linéaire, 413–425
– par régime pseudo-transitoire, 427–428
– par région de confiance, 426
Goldfarb, 440
Goldstein, 308
Gordan, 70
gradient, 88, **792**
– projeté, 336, **470**, 470
– – test du gradient projeté, 470
– réduit, 578
– réduit, 893
Gram, 762
graphe, 621
– connexe, 622
– d'une multifonction, 757
Grassmann, *voir* formule
Gröbner, *voir* base

Heine, 748
hessienne, 807
– réduit du lagrangien, 176
Hilbert, 755
Hoffman, *voir* lemme
Hölder
– inégalité de –, 245, 755
Houellebecq, iv
Hurwicz, 550
hyperplan, 21
– d'appui, 49

identité du parallélogramme, 759
image, 748, 764
– d'une multifonction, 757
– par une application linéaire, *voir* image par une application linéaire
– réciproque, 748
– – par une application linéaire, *voir* image réciproque par une application linéaire
image par une application linéaire, 753
– d'un cône asymptotique, 29
– d'un cône convexe fermé, 68
– d'un convexe, 22
– d'un intérieur relatif de convexe, 34
– d'un polyèdre convexe, 39, 68
– d'une adhérence de convexe, 34
– des points extrêmes, 31
image réciproque par une application linéaire
– d'un cône asymptotique, 29
– d'un convexe, 22
– d'un intérieur relatif de convexe, 35
– d'un polyèdre convexe, 68
– d'une adhérence de convexe, 35
IML, *voir* inégalité matricielle linéaire
inclusions polaires, 644
incompatibilité d'inégalités affines, 71
indicatrice, *voir* fonction indicatrice
indice de saturation d'un sous-espace de Krylov, 343
indépendant, *voir* vecteurs
inégalité, *voir aussi* condition
– de Cauchy-Schwarz, 755, 759

– – généralisée, 755
– de convexité, 75
– de dualité faible, *voir* dualité faible
– de Hölder, 245, 755
– de Jensen, 146
– – version intégrale, 146
– de Kantorovitch, 326
– de trace de von Neumann, 784
– du sous-gradient, 124
– géométrico-arithmétique, 146
– incompatible, 71
– matricielle linéaire, **23**, 720
– triangulaire, 750, 751
inéquation variationnelle, 132
inf-convolution, **105**, 106, 139, 146
– conjuguée d'une –, 119
– de deux formes quadratiques, 147
– épigraphe d'une –, 106
– épigraphe stricte d'une –, 105
inf-image d'une fonction sous une application linéaire, 106
– conjuguée d'une –, 115
– convexité d'une –, 107
– polyédricité d'une –, 107
– propreté et semi-continuité inférieure d'une –, 116
intérieur, 748
– relatif, 31
– – d'un cône, 67
– – enveloppe affine d'un –, 32, 67
intérieur relatif
– d'un polyèdre convexe, 67, 68
itération
– externe, 420
– interne, 421
itéré, 250

jacobienne, 166, 170
Jensen, *voir* inégalité

Kantorovitch, 241, *voir* inégalité
– inégalité, 326
– théorème, 407, 429
Kantorovitch, Leonid, 429
Karush, 182, 183
Kepler, 241
Krylov, 342
Kuhn, 182, 183
Ky Fan, 768

Lagrange, 157, 169, 183, 241, *voir aussi* multiplicateur
lagrangien, **170**, **183**, 537
– associé à des perturbations, 518, 520
– augmenté, 491, 492
– augmenté non différentiable, 506
– modifié, 491
– ordinaire, 269, 502, 532, 625
Lebesgue, 748
Legendre, 111
Leibniz, 241
Lemaitre, 403
Lemaréchal, 310, 814, 873
lemme, *voir aussi* théorème
– de Farkas, 59, 645
– de Finsler, 765
– de Hoffman, 645
– de Mizuno, 662
– de perturbation de Banach, 753
Levenberg, 700, *voir* algorithme
ligne de flux de Newton, 431
limite
– inférieure, 749
– supérieure, 749
linéairement indépendants, *voir* vecteurs
loi de conservation des ennuis, **10**, 18
Löwner, 717
Luenberger, 449
Lyapounov, *voir* opérateur, système linéaire
Lyusternik, 220

machine de Turing, 260
MacLaurin, iii
Mangasarian, *voir* qualification des contraintes d'inégalité (QC-MF)
Marquardt, 700, *voir aussi* algorithme
matrice, 763
– bijective, 764
– carrée, 763
– complètement positive, 70
– copositive, 203
– creuse, 293
– d'incidence d'un réseau, 622
– de covariance des erreurs a posteriori sur la solution de norme minimale, 688
– de covariance des erreurs a priori sur les données, 688
– de résolution d'un problème de moindres-carrés linéaire, 688
– de type $m \times n$, 763
– définie négative, 765
– définie positive ($\mathcal{S}^n_{++}$), 786
– définie positive ($\mathcal{S}^n_{++}$), **765**
– des cofacteurs, 786

– doublement stochastique, 67
– hermitienne, 771
– identité, 763
– injective, 764
– inverse, 18, 764
– inversible, 764
– orthogonale, 772
– positive, 70
– pseudo-inverse, **684**, 687, 766
– semi-définie négative, 765
– semi-définie positive ($\mathcal{S}_+^n$), 69
– semi-définie positive ($\mathcal{S}_+^n$), **765**
– surjective, 764
– symétrique, 765
– symétrique copositive, 70
– transposée-conjuguée, 770
– triangulaire, 772
– triangulaire unitaire, 772
– uniformément injective, 695
– unitaire, 770
maximisation, 10
methode
– méthode
– – de descente par coordonnée, 337
méthode, *voir aussi* algorithme
– d'activation, 460
– de faisceaux, 545, 552
– de Gauss-Seidel
– – en algèbre linéaire, 333–335
– – en optimisation, 336–339
– – pour les systèmes non linéaires, 335
– de l'état adjoint, 269
– de Newton, 403
– de pivotage, 460
– des multiplicateurs, 497, 550
– newtonienne, 566
– primale-duale, 567
– tensorielle, 430
mineur
– principal, 769
– principal de tête, 769
Minimax de von Neumann, 645
minimisation
– d'une fonction linéaire sur une boule, 244
– emboîtée, 12, 14
minimiseur, *voir* minimum
minimum, 4
– faible, 203
– fort, 164, 175
– global, 5
– global strict, 5
– local, 5, 6
– local strict, 5
– saillant, 150
Minkowski, 65, 751
– fonction de –, 597
– norme de –, 752
minorant de la valeur optimale
– par pénalisation extérieure, 482
– par relaxation wolfienne, 556
minorante affine, **77**, 109, 113
– exacte, 78
– pente d'une –, 77
Miranda, 429
Mizuno, 662
module d'une application lipschitzienne, 751
module de complexité, 595
moindres-carrés, 401
monôme, 728
Montaigne, 3
Moore, 429
Moré, 391
Moreau, *voir* décomposition, régularisée
Morrison, 700, 787
Motzkin, 70, 645
multifonction, 757
– convexe, 207
– domaine d'une –, 757
– fortement monotone, 758
– graphe d'une –, 757
– image d'une –, 757
– localement radialement lipschitzienne, 155
– métriquement régulière, 215
– monotone, 758
– monotone maximale, 758
– ouverte, 214
– réciproque, 757
multiplicateur de Lagrange, 169, 183
multiplicateur optimal
– pour (P_E), unicité, 173

Newton, 385, 405, 414, 422, 428, 677, 695, *voir aussi* algorithme
Nirenberg, 809
nombre complexe
– conjugué, 770
– module, 770
nombre imaginaire pur, 770
nombres conjugués, 146, 152, 245, **755**
normale, 48

norme, 151, 152, **751**, 786
– associée à un produit scalaire, 755
– biduale, 151
– conjuguée, 151
– convexité, 151
– convexité d'une –, 152
– de Minkowski ou ℓ_p, 752
– dérivée directionnelle, 152
– duale, 151, 152, 502, 573, **755**, 759
– euclidienne ou ℓ_2, 752
– ℓ_1, 558
– matricielle, *voir* norme matricielle
– sous-différentiel, 152
norme IEEE 754, 257
norme matricielle
– de Frobenius, 557, 717, 764
– de l'inverse, 18
– nucléaire, **557**, 558
– sous-multiplcative, 763
– subordonnée, 763
noyau, 753, 764

OL, *voir* optimisation linéaire
opérateur, *voir aussi* application linéaire
– de Lyapounov, 739
– proximal, 140
– réduction, 262
optimisation, 4
– combinatoire, 6
– conique, 616
– copositive, 582
– en nombres entiers, 6
– globale, 17, 411
– linéaire, 619–680
– multicritère, 6, 147
– quadratique, *voir* optimisation quadratique
– semi-définie positive (SDP), 717–744
– stochastique, 552
optimisation quadratique, 460
– existence de solution, 243
optimiseur, 286
OQ, *voir* optimisation quadratique
ordre, 102
Oren, 449
orthant positif, **22**
OSDP, *voir* optimisation semi-définie positive
ouvert, 747
– relatif, 32

parallélogramme, *voir* identité
Pareto, 147
pas (de recherche linéaire), 299
– admissibilité asymptotique du – unité, 320
– d'activation, 469
– d'Armijo, 307
– de Cauchy, 305
– de Curry, 305
– de Fletcher, 312
– de Goldstein, 308, 324, 873
– de Wolfe, 309
– optimal, 305
passage d'un terme du critère en contrainte, **15**, 581
passage d'un terme du critère en contrainte, 725
pavé
– cône tangent, 471
– projection sur un –, 470
pénalisation
– ℓ_1, 505
– exacte, 475
– extérieure, 478, 486
– facteur de –, 474
– inexacte, 475
– intérieure inverse, 487
– logarithmique, 487
– quadratique, 479
performance, 289
– relative, 289
permutation, 767
petit o, 793
pivot, 775, 777
pivotage, 639
pli d'une fonction convexe, 125
point
– absorbant, 33
– admissible, 4, 620
– critical or stationary
– – regular, 571
– d'accumulation, 748
– d'activation, 459
– de Cauchy, 385
– de Newton, 385
– extrême, 30
– – d'un polyèdre convexe, 68
– – d'une face, 67
– – des boules unités ℓ_1 et ℓ_∞, 67
– moyennement optimal, 147
– Pareto optimal, 147
– proximal, 140

– stationnaire, 170, 183
– – régulier, 569
– strictement admissible, 632
point extrême
– d'un produit cartésien, 31
– image par une application linéaire, 31
point-selle, **144**, 231
Polak, 356, 357
polyèdre convexe, 22, **37**
– borné, 61
– cône asymptotique, 67, 68
– enveloppe affine, 67
– enveloppe conique, 68
– face, 67
– face exposée, 67
– forme standard d'un –, 37
– image par une application linéaire, 68
– image réciproque par une application linéaire, 68
– imbriqué, 644
– intérieur relatif, 67, 68
– point extrême, 68
– produit cartésien, 68
– réduit à un point, 61
– représentation duale d'un –, 37
– représentation primale d'un –, 37
– somme, 67, 68
– sommet, 67
– sous représentation duale, 67
– sous représentation primale, 68
polyédricité
– d'une inf-image sous une application linéaire, 107
polynôme, 411, 728
– annihilant une matrice, 344
– de Motzkin, 729
– degré, 728
– minimal, 344
– quartique, 243
– unitaire, 344
polytope, 37
Pompidou, iii
poursuite de base, 558
Powell, 356, 387, 392, 402, 455
préconditionnement, 265
– de l'algorithme du gradient conjugué, 353–356
préconditionneur, 323, 327
– diagonal, 355
probleme
– LASSO, 711
problème, *voir aussi* famille de problèmes, problème d'optimisation, problème dual
– NP, 261
– NP-ardu, 263
– NP-complet, 262
– à données massives, 499
– barrière primal, 650
– d'identification, 681
– d'inclusion linéaire dans un intervalle, 688
– d'inéquation variationnelle, 430
– d'interpolation linéaire, 681
– de calibration, 681
– de complémentarité, 430
– de Lagrange, 511, 545, 556
– de moindres-carrés, *voir aussi* fonction
– – non linéaire, 303, 312
– de régression, 681
– régression linéaire, 682
– du plus court chemin dans un graphe, 622
– du transport, 622
– dual, *voir* problème dual
– interne, 13, 14
– interne dual, 511
– interne primal, 511
– polynomial, 261
– primal, 508, 511, 518, 524, 526
– quadratique osculateur, 409
problème d'optimisation, 4
– (P_E), 166, 565, 569
– – convexe, 166
– (P_{EI}), **177**, 477, 501, 559
– – convexe, 178
– (P_G), 206
– – convexe, 208
– (P_L), 619
– (P'_L), 620
– (P_X), 157
– (P_X)
– – convexe, 163
– borné, 4, 620
– linéaire, 40, 619–680
– – forme canonique d'un –, 620
– – forme standard d'un –, 619
– non borné, 4
– (P_{SI}), 577
– quadratique, *voir* optimisation quadratique
– réalisable, 4, 620
problème de moindres-carrés, 681–716

– linéaire, 681–688
– – avec région de confiance, 715
– – régularisé, 714
– – total, 714
– non linéaire, 693–706
– polyédrique, 689
problème dual, 508, 511, 519
– de Toland, 527
– de Wolfe, 555
– non borné, 536
problème quadratique osculateur, 568
problème tout quadratique, 732
– relaxation de rang, 744
– relaxation de rang, de Shor ou lagrangienne, 732
produit cartésien, 514, 626
– de deux polyèdres convexes, 68
– enveloppe convexe d'un –, 27
produit de Hadamard, 181
produit scalaire, 754
– euclidien, 755
– produit, 755
produit tensoriel, 434, 737, 744, 784
profil de performance, 289
projection, **750**, 808
– caractérisation, 46
– dérivabilité directionnelle, 47
– en deux temps, 69
– propriétés variationnelles, 68
– sur des contraintes de borne, 470
– sur $\mathcal{S}_+^n$, 69
– sur un cône convexe fermé, 68
– sur un pavé, 470
– sur une somme de sous-espaces affines, 69
projection orthogonale, 45
– est un gradient, 155
projeté, 45
– d'un polyèdre convexe, 68
– existence et unicité, 46
propriété
– de croissance quadratique, 203
– de Heine-Borel-Lebesgue, 748
proximal, *voir* opérateur, point
Putinar, 731

qualification de contraintes, 161, 185, 643
qualification des contraintes
– abstraites, 209
– – (QC-R), de Robinson, 212, 224
– convexes, 34
– d'égalité, 168
– d'inégalité, **181**, 186–195
– – (QC-A), **187**
– – (QC-MF) de Mangasarian-Fromovitz, 212
– – (QC-MF), de Mangasarian-Fromovitz, 193
– – (QC-MFS), de Mangasarian-Fromovitz stricte, 196
– – (QC-S), de Slater, **187**, 651
– – indépendance linéaire (QC-IL), **188**
– – Mangasarian-Fromovitz (QC-MF), **190**
quotient de Rayleigh, 245

rang, 764, 786
– biconjuguée, 557
– resteint
– – biconjuguée, 557
– – conjuguée, 557
– semi-continuité inférieure, 785
Raphson, Joseph, 429
Rawls, 510
Rayleigh, *voir* quotient
rayon de confiance, 370
rayon spectral, 771
réalisabilité, *voir* compatibilité
réalisable, *voir* problème d'optimisation, système
rebroussement, 307
recherche linéaire, 299
– d'Armijo, 372, 464
– de Cauchy, 305
– de Curry, 305
– de Goldstein, 308
– de Wolfe, 309
– de Wolfe forte, **311**, 357
– exacte, 305
– inexacte, 305
– non monotone, 322
recouvrement
– ℓ_1, 558
– nucléaire, 558
redémarrage de Powell, 356
réduction, 262
Reeves, 356, 357
région de confiance, 370
règle d'anti-cyclage de l'algorithme du simplexe, 637, 639
– de Bland, 639
– des petites perturbations, 639
– lexicographique, 639
règle de bascule, 126
règle de pivotage, 639

règle du coût réduit minimal, 637
régularisée de Moreau-Yosida, 142, 328, 539
relatif, *voir* frontière, intérieur
relation binaire, 757
relation de polarisation, 759
relaxation, 732
– lagrangienne, 545
– SDP (semi-définie positive), 725
relaxation de rang
– problème tout quadratique, 732
relaxation de Shor, 732
relaxation lagrangienne
– du problème tout quadratique, 732
réseau, 621
résidu, 303, 359, 677, 693
Ribière, 356, 357
Riesz, 756
Rimbaud, 101, 809
$\mathbb{R}^n_+$
– cône dual, 69
Robinson, *voir* qualification des contraintes abstraites (QC-R)
Rockafellar, 3, 19, 73, 518
Rolle, 799
rotation de Givens, 772

saut de dualité, 508
saut de dualité, *voir* dualité
schéma d'Euler implicite, 274, 426, 427
Schmidt, 762
Schur, *voir* complément
Schwarz, 755
segment, 21, 796
Seidel, *voir* méthode de Gauss-Seidel
semi-continuité inférieure, 749
– du rang, 785
semi-continuité supérieure, 750
séparation de deux convexes, 51
série
– absolument convergente, 752
– convergente, 752
Shadoks, 809
Shanno, 440
Sherman, 787
Shor, *voir* relaxation de rang du problème tout quadratique
simplexe
– unité de $\mathbb{R}^n$, 22
simplexe ordonné, 69
– cône dual, 69
Simpson, Thomas, 429
simulateur, 286, 288
Slater, *voir* qualification des contraintes d'inégalité (QC-S)
$\mathcal{S}^n_+$, 22
– cône dual, 69
– cône normal, 71
– cône tangent, 71
– face, 743
– face engendrée, 744
– face exposée, 744
– projection, 69
$\mathcal{S}^n_{++}$, 22
solution, *voir aussi* existence de solution, unicité de solution, minimum
– d'un problème quadratique osculateur
– – de norme minimale, 570
– – importune, 568, **570**
– de norme minimale, 352, 684
– duale, 170, 511, 626
– hémi-stable de (P_{EI}), 561
– primale, 170, 511, 626
– primale-duale, 170, 183
– semi-stable de (P_{EI}), 561
– strictement complémentaire, 627
somme
– d'un cône convexe fermé et d'un sous-espace vectoriel, 68
– de deux cônes duaux, 60
– de deux convexes, 22
– – fermés, 37
– de deux ensembles, 751
– de deux polyèdres convexes, 67, 68
– de deux sous-espaces affines
– – projection sur une –, 69
– enveloppe affine d'une –, 66
– enveloppe convexe d'une –, 27, 66
sommet d'un polyèdre convexe, **41**, 67
– dégénéré, 623, 643
Sorensen, 391
sous-différentiabilité, 124, 128
sous-différentiel, 124
– constance du –, 155
– d'une fonction composée, 135
– d'une fonction marginale, 137
– d'une enveloppe supérieure, 138–139
– d'une fonction d'appui, 151
– d'une indicatrice, 151
– d'une norme, 152
– d'une somme, 134
– de f et f^{**}, 127
– de la distance à un convexe, 154

– de la valeur propre maximale, 153, 154
– et optimalité, 127, 129, 130
– fonction d'appui du –, 149
– fonction fortement convexe, 149
– propriétés géométrique et topologique du –, 129
– représentation du –, 149
sous-espace affine, 752, 758
– intersection, 23, 759
sous-espace vectoriel
– de Krylov, 342–346
– – saturation d'un –, 343
– orthogonal, 762
– parallèle à un sous-espace affine, 752
sous-espaces vectoriels
– somme de –, 761
– somme directe de –, 761
– supplémentaires, 167, **762**, 786
sous-gradient, 124
sous-problème quadratique
– (*RC*), 370, 375, 402
– osculateur, *voir* problème quadratique osculateur
spectre, 771
stabilité
– d'un ensemble, 191
– d'un ensemble pour de petites perturbations, 213
stabilité d'un ensemble par rapport à des perturbations, 168, 182
Stroustrup, iv
succès d'une itération, 372
suite, 747
– admissible, 459
– convergente, 747
– de Cauchy, 752
– limite, 747
– minimaximisante, 552
– minimisante, **7**
– stationnaire, 4
suite de Fejér, 292
sur-relaxation, 430
système
– fortement réalisable, 722, 723
– quasi-réalisable, 722
système linéaire
– augmenté, 677, 685
– de Lyapounov, 740
– de Sylvester, 740

taux de convergence, 258
Tchebychev, 715
terminaison finie, 659
– de l'algorithme proximal, 340
terminaison quadratique, 447
test du gradient projeté, 470
théorème, *voir aussi* lemme
– d'Everett, 535, **556**
– de Carathéodory, 26
– de l'alternative, 59
– – de Farkas, 70
– – de Gordan, 70
– – de Motzkin homogène, 70
– – de Motzkin non-homogène, 70, 645
– – de Ville, 70
– de l'application ouverte, 785
– de l'image fermée, 754
– de représentation de Riesz-Fréchet, 756
– de Rolle, 799
– de séparation, 52, 53
– des accroissements finis, 796, 799
– des fonctions implicites
– – différentibilité, 801
– – existence, 800
– du rang, 763
topologie, 747
– d'un espace métrique, 750
– induite, 748
– séparée, 747
trace d'une matrice carrée, 763
transformation affine
– enveloppe convexe, 27
transposée, 765
Tucker, 182, 183
Turing, *voir* machine

unicité de solution, 76

valeur, *voir aussi* fonction valeur
– critique, 170, 183
– optimale, 4
valeur propre, 771, 931
valeur propre maximale, 153
– d'une matrice hermitienne, 154
valeur singulière d'une matrice, 245, 686, **784**, *voir aussi* factorisation en valeurs singulières (SVD)
valeur-selle, **144**, 557
variable
– active, 277
– changement de –, 11
– d'écart, 37, 492, 621
– – duale, 626
– d'entrée, 277

– d'état, 577
– de commande, 266, 577
– de sortie, 277
– d'état, 266
– duale, 282
– indépendante, 277
variété, 166
vecteur, *voir aussi* vecteurs
– compteur de composante non nulle, *voir* compteur de composante non nulle
– normal, 161
– tangent, **159**
vecteur propre, 771
– unitaire, 771
vecteurs, *voir aussi* vecteur
– affinement indépendants, 24
– linéairement indépendants, 761
– orthogonaux, 762
vérification de solution de (P_{EI}), 206
Ville, 70
vitesse de convergence
– linéaire
– – taux, 252
– q-cubique, 258
– q-linéaire, 252
– q-ordre, 257
– q-quadratique, 256
– q-quartique, 258
– q-quintique, 258
– q-superlinéaire, 253
– r-linéaire, 258
– sous-quadratique, 613
voisinage, 747
von Neuman, *voir* inégalité, minimax
von Neumann, 645
Voronoï, *voir* diagramme de Voronoï

Weierstrass, 7
Wilkinson, 300
Wittgenstein, iii
Wolfe, 309, 311, 555, 643
– contre-exemple de –, 459
Woodbury, 787

Yosida, *voir* régularisée

zéro, 404
Zoutendijk, 313

Sois sage, ô ma Douleur, et tiens-toi plus tranquille.
Tu réclamais le Soir ; il descend ; le voici :
Une atmosphère obscure enveloppe la ville,
Aux uns portant la paix, aux autres le souci.

Ch. Baudelaire (1868). Recueillement. *Les Fleurs du Mal* [Poème apporté par la troisième édition].