Outils mathématiques et algorithmiques pour le calcul scientifique

Patrick Ciarlet, Erell Jamelot

Outils mathématiques et algorithmiques pour le calcul scientifique

Patrick Ciarlet et Erell Jamelot

Ce polycopié correspond aux notes du cours "Calcul Scientifique Parallèle", tel qu'enseigné de 2014 à 2019 par les auteurs. Ce cours fait partie du cursus Modélisation et Simulation du M2 Analyse, Modélisation, Simulation de l'Université Paris-Saclay et du cursus de 3ème année ModSim de l'ENSTA Paris.

L'objectif principal est de proposer aux étudiants des outils de calcul scientifique permettant d'appréhender les algorithmes adaptés au calcul parallèle, c'est-à-dire pouvant utiliser plusieurs nœuds de calcul simultanément. On abordera essentiellement le calcul parallèle d'un point de vue méthodologique et/ou algorithmique. A partir d'un problème modèle, on présente un certain nombre d'outils et de méthodes permettant de le résoudre numériquement, et on explique comment on peut adapter au calcul parallèle, c'est-à-dire *paralléliser*, les algorithmes associés. On évoque, sans les occulter, tous les aspects de la résolution, qu'ils soient abstraits (point de vue mathématique) ; discrétisation (point de vue numérique) ; algorithmique (mise en œuvre). A noter : le point de vue informatique du calcul parallèle est développé dans la documentation en ligne disponible à l'adresse `https://ams301.pages.math.cnrs.fr/`.

Le polycopié est composé de trois parties et d'annexes.
Dans la première partie, on rappelle quelques problèmes typiques à traiter. Parmi ces problèmes, on se concentrera sur la résolution de l'équation de diffusion des neutrons : résolution mathématique d'une part, et résolution numérique d'autre part. Pour ce second aspect, on introduit deux méthodes de discrétisation : les différences finies et les éléments finis. Les différences finies donnent lieu à des algorithmes de résolution numérique possédant une structure, on parle d'*algorithmes structurés*, alors que les éléments finis conduisent en général à des *algorithmes non-structurés*.
Après discrétisation, l'opération fondamentale à réaliser est la résolution d'un système linéaire. La seconde partie se concentre donc sur l'algèbre linéaire numérique : éléments d'algorithmique numérique, les méthodes de résolution directes et itératives, les méthodes de Krylov et la méthode de la puissance itérée. La prise en compte de la structure, ou de l'absence de structure, joue un rôle déterminant dans la résolution parallèle.
Enfin la troisième partie est une introduction aux méthodes de décomposition de domaine. Le calcul parallèle est naturellement associé à ces méthodes, car on choisit de découper le problème initial en plusieurs sous-problèmes interagissant entre eux, et on discrétise la seconde instance. On reprend comme exemple l'équation de diffusion des neutrons, discrétisée par la méthode des éléments finis. Après une introduction mathématique, on étudiera pour chaque problème deux méthodes de décomposition de domaine : la méthode de Schwarz et la méthode avec contrainte. On s'aidera de l'analyse numérique pour valider nos modèles décomposés.
Les annexes comprennent des rappels en algèbre linéaire, les outils de base pour l'étude et l'approximation de formulations variationnelles en dimension infinie, et enfin quelques outils élémentaires sur les distributions et les espaces fonctionnels de type Sobolev.

Table des matières

Première partie

Modélisation et discrétisation

Le but de cette partie est d'étudier des processus et cheminements qui permettent de résoudre des problèmes issus de la physique, de la mécanique, de la finance, etc. Que comprend un tel processus ? De façon générale, il est possible de le découper en quatre étapes :

1. Identifier *la ou les quantité(s) à calculer.*
   Modéliser *le phénomène de nature physique (ou autre) associé : c'est-à-dire, construire le* **modèle***, composé d'équation(s), inéquation(s), contrainte(s), etc.*
2. *Choisir une* **méthode d'approximation** *(ou* **méthode de discrétisation***) lorsqu'on ne peut pas résoudre le problème exactement... En pratique, ça se passe toujours comme ça ! De plus, on est en général déjà amené à réaliser des* simplifications *pour construire le modèle à l'étape précédente.*
3. *Construire le* problème discrétisé *(ou* problème approché*.)*
   Évaluer la qualité *de la solution approchée par rapport à une solution exacte : on parle aussi de* précision *de la méthode de discrétisation.*
4. Résoudre *le problème discrétisé.*
   Vérifier/visualiser/interpréter *les résultats.*

Nous allons explorer successivement tous ces aspects, de façon plus ou moins approfondie. Plus précisément, nous allons, au chapitre 1, proposer quelques problèmes modèles élémentaires, issus pour la plupart de la physique. Il s'agit d'une présentation très intuitive, qui ne prétend pas respecter les canons mathématiques *! Pour de plus amples détails, ou pour des justifications précises, nous renvoyons le lecteur intéressé à [11, 2], ainsi qu'aux autres cours de mathématiques appliquées, notamment en troisième année. Ensuite, nous allons les discrétiser, à l'aide de la* **méthode des différences finies***. Une fois le problème discrétisé construit, nous répondrons également, dans une certaine mesure, à la question de la qualité de l'approximation de la solution approchée, par comparaison à la quantité de départ à calculer. Ceci sera l'objet du chapitre 2. Nous procédons ensuite à la discrétisation par la* **méthode des éléments finis** *au chapitre 3, où nous proposons de nouvelles réponses aux questions posées ci-dessus. Pour de plus amples précisions, nous renvoyons également à [11, 2].*

Chapitre 1

Considérations générales et modèles

1.1 Problèmes statiques élémentaires

Par problèmes statiques, nous entendons problèmes à l'équilibre, par opposition aux problèmes qui dépendent du temps, évoqués à la prochaine section.

Commençons par un modèle élémentaire, celui du *fil pesant*. Soit donc un fil de longueur unité, fixé en ses deux extrémités. Le but est ici de calculer des déplacements verticaux, dont on suppose qu'ils sont "petits", lorsqu'il est soumis à la gravité. On note $\rho : x \mapsto \rho(x)$ la densité linéique de masse, et $u : x \mapsto u(x)$ le déplacement transversal. D'après les équations de l'élasticité linéaire, on sait que u vérifie

$$-u''(x) = f(x) \text{ sur }]0,1[, \quad \text{avec } f(x) = \rho(x)g. \tag{1.1}$$

Comme les extrémités du fil sont fixées, il est clair que le déplacement vertical est nul en celles-ci, ce que l'on exprime sous la forme

$$u(0) = u(1) = 0. \tag{1.2}$$

Enfin, le fait que l'énergie élastique de déformation soit bornée peut-être exprimé sous la forme

$$\int_0^1 (u'(x))^2 \, dx < \infty. \tag{1.3}$$

On parle ici de **modèle monodimensionnel** ou de **modèle 1D**, puisque les équations (1.1)-(1.2) et l'inéquation (1.3) ne dépendent que de la variable x. Le **domaine de calcul** est dans ce cas l'intervalle *ouvert* $]0,1[$, et on rappelle que l'ensemble $\{0,1\}$ constitue sa **frontière**.

On peut introduire un second modèle 1D : celui de la *poutre*, appuyée en ses extrémités. Le but est encore une fois de déterminer des "petits" déplacements verticaux, lorsqu'elle est soumise à une charge transversale, égale à $f(x)\delta x$, où f est la densité de force par unité de longueur. D'après les équations de l'élasticité linéaire, on retrouve que le modèle est constitué par le même ensemble d'équations et d'inéquations (1.1)-(1.3) que précédemment.

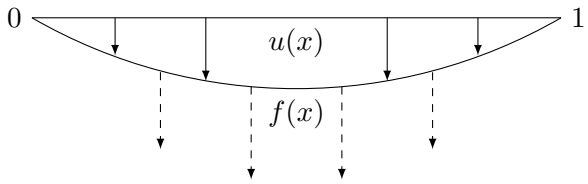


FIGURE 1.1 – Modèles 1D : fil ou poutre

Remarque 1.1 *L'équation (1.1) peut être écrite de façon équivalente*

$$-\frac{d^2u}{dx^2}(x) = f(x) \text{ pour } x \in]0,1[.$$

On peut également concevoir des modèles, provenant de problèmes statiques posés en dimension supérieure.

Commençons par celui de la *membrane élastique* à l'équilibre, dont la frontière est fixée. Encore une fois, on suppose que l'on veut déterminer des "petits" déplacements verticaux, lorsque cette membrane est soumise à une force verticale. Soit D le domaine *ouvert* occupé par la membrane au repos, que l'on suppose être inclus dans le plan Oxy (la membrane est donc horizontale au repos.) Cette fois, le déplacement vertical, toujours noté u, dépend de deux variables, à savoir x et y, pour (x, y) parcourant le domaine de calcul D. Les forces agissant sur la membrane sont de la forme $\tau f(x, y)\delta S$, avec f la densité de force par unité de surface, et τ la tension de la membrane. D'après les équations de l'élasticité linéaire, u vérifie

$$-c(\tau)^2 \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right)(x, y) = f(x, y) \text{ pour } (x, y) \in D. \tag{1.4}$$

Ci-dessus, $c(\tau)$ est un nombre strictement positif, dépendant de τ.
Comme la frontière du domaine, ∂D, est fixée, on écrit

$$u(x, y) = 0, \text{ pour } (x, y) \in \partial D. \tag{1.5}$$

Enfin, le fait que l'énergie élastique de déformation soit bornée (ainsi que l'inégalité dite de Korn) permettent d'écrire

$$\int_D c(\tau)^2 \left(\left[\frac{\partial u}{\partial x}(x, y)\right]^2 + \left[\frac{\partial u}{\partial y}(x, y)\right]^2\right) dxdy < \infty. \tag{1.6}$$

On parle ici de **modèle bidimensionnel** ou de **modèle 2D**, puisque les équations (1.4)-(1.5) et l'inéquation (1.6) dépendent du couple de variables (x, y).

Pour finir, rappelons le modèle associé à la *cavité électrostatique*, incluse dans $\mathbb{R}^3$. Le but est de déterminer le potentiel électrostatique autour d'un système de conducteurs parfaits disjoints, $C_1, \cdots, C_I$, eux-mêmes étant entourés d'un conducteur parfait C_0. Soit C le domaine de calcul *ouvert* inclus dans $\mathbb{R}^3$, constitué de la partie intérieure à ce dernier conducteur parfait, et privé de $\bar{C}_1, \cdots, \bar{C}_I$. Ici, C est la cavité électrostatique. On note

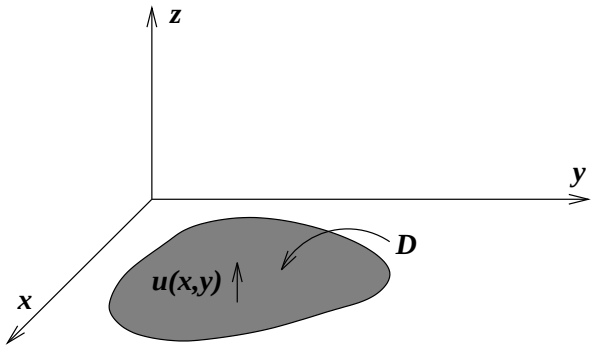


FIGURE 1.2 – Modèle 2D : membrane élastique

$\partial C_1, \cdots, \partial C_I$ les frontières respectives de $C_1, \cdots, C_I$, ainsi que $\partial_{int} C_0$ la frontière interne de C_0 : par construction (cf. Figure 1.3), la frontière ∂C de notre cavité électrostatique est formée par l'union

$$\partial C = \partial C_1 \cup \cdots \cup \partial C_I \cup \partial_{int} C_0.$$

Si on considère que sur la frontière $\partial_{int} C_0$, on se trouve à l'équipotentielle nulle, le potentiel électrostatique est déterminé par la donnée de :

— ρ la densité de charge électrique par unité de volume ;

— V_k la valeur de l'équipotentielle sur la frontière ∂C_k, pour k variant de 1 à I.

On appelle u le potentiel électrostatique à calculer, c'est-à-dire $u(x, y, z)$, pour (x, y, z) parcourant C. D'après les équations de Maxwell, et plus précisément la relation de Coulomb, on sait que

$$-\varepsilon_0 \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \right)(x, y, z) = \rho(x, y, z) \text{ pour } (x, y, z) \in C. \tag{1.7}$$

Ci-dessus, ε_0 est la permittivité électrique du vide.
La frontière de chaque conducteur parfait étant équipotentielle, on écrit

$$u(x, y, z) = V_k, \ (x, y, z) \in \partial C_k, \ 1 \leq k \leq I, \text{ et } V(x, y, z) = 0, \ (x, y, z) \in \partial_{int} C_0. \tag{1.8}$$

Enfin, le fait que l'énergie électrostatique soit bornée signifie que

$$\int_C \varepsilon_0 \left(\left[\frac{\partial u}{\partial x}(x, y, z) \right]^2 + \left[\frac{\partial u}{\partial y}(x, y, z) \right]^2 + \left[\frac{\partial u}{\partial z}(x, y, z) \right]^2 \right) dxdydz < \infty. \tag{1.9}$$

On parle ici de **modèle tridimensionnel** ou de **modèle 3D**, puisque les équations (1.7)-(1.8) et l'inéquation (1.9) dépendent du triplet de variables (x, y, z).

Pour conclure cette section, nous proposons le récapitulatif ci-dessous, qui met en évidence un certain nombre de similitudes entre ces différents modèles statiques.

Remarque 1.2 *Bien évidemment, il existe beaucoup de modèles statiques, qui peuvent être complètement différents, par leur nature, de ceux présentés ici ! De fait, l'"unité" entre tous ces modèles permettra une modélisation* homogène*, comme on le verra au chapitre 2...*

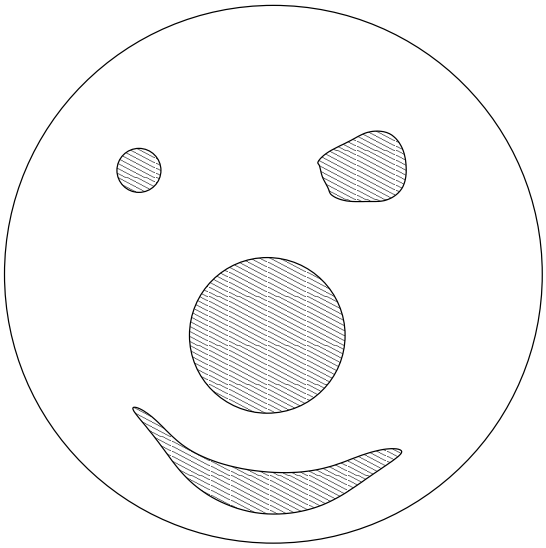


FIGURE 1.3 – Modèle 3D : cavité avec quatre conducteurs internes ($I = 4$)

Commençons par quelques définitions, utilisées fréquemment par la suite.

Définition 1.3 *On utilise le terme de* **condition aux limites** *pour qualifier les équations vérifiées par les inconnues du problème sur la frontière du domaine de calcul.*

Définition 1.4 *On appelle* **Laplacien dans** $\mathbb{R}^d$ $(d \geq 1)$ *l'opérateur scalaire défini par*

$$\Delta_d u = \sum_{k=1}^{d} \frac{\partial^2 u}{\partial {x_k}^2}.$$

Définition 1.5 *On appelle* **gradient dans** $\mathbb{R}^d$ $(d \geq 1)$ *l'opérateur vectoriel défini par*

$$\mathbf{grad}_d\, u = \sum_{k=1}^{d} \frac{\partial u}{\partial x_k} \boldsymbol{e}_k.$$

Pour récapituler :

— Problèmes posés dans des domaines (ouverts bornés) Ω_d de $\mathbb{R}^d$, $d = 1, 2, 3$, par rapport à la variable $\boldsymbol{x} = (x_1, \cdots, x_d)$:

1. $d = 1$: fil, poutre ;
2. $d = 2$: membrane ;
3. $d = 3$: cavité.

— L'opérateur est identique, pour tous les problèmes. Il est composé de :
(I) à l'intérieur, le Laplacien dans $\mathbb{R}^d$;
(F) sur la frontière, une condition aux limites du type $u = 0$ (ou $u = \textit{constante}$).
De plus, même propriété pour l'intégrale du gradient, bornée pour tous les problèmes.

$$\int_{\Omega_d} |\mathbf{grad}_d\, u(\boldsymbol{x})|^2 \, d\boldsymbol{x} < \infty.$$

— Problèmes **linéaires** : si les problèmes avec les données f_1 et f_2 admettent pour solution u_1 et u_2, alors le problème avec pour donnée $\alpha_1 f_1 + \alpha_2 f_2$ admet pour solution $\alpha_1 u_1 + \alpha_2 u_2$.

1.2 Problèmes instationnaires élémentaires

Contrairement aux problèmes de la section précédente, nous nous intéressons ici à des modèles dépendant non seulement de la variable spatiale $\boldsymbol{x}$, mais aussi du temps t. Néanmoins, et pour éviter toute confusion, nous conservons la terminologie modèles 1D, 2D, 3D, par référence à la variable d'espace *uniquement*.

Commençons par un modèle 1D instationnaire, celui de la *corde vibrante*. Soit donc une corde de longueur unité, fixée en ses deux extrémités. Pour simplifier, nous négligeons la gravité, et nous supposons que la densité linéique de masse ρ est constante. Le but est encore une fois de calculer des "petits" déplacements verticaux, à partir d'une **configuration initiale**, connue à l'instant $t = 0$. On note $u : (x, t) \mapsto u(x, t)$ le déplacement transversal. D'après les équations de l'élasticité linéaire, on sait que u vérifie

$$\frac{\partial^2 u}{\partial t^2} - c(\tau)^2 \frac{\partial^2 u}{\partial x^2} = 0 \text{ pour } x \in]0, 1[\text{ et } t > 0. \tag{1.10}$$

(Ci-dessus, si τ est la tension de la corde, on a $c(\tau) = (\tau/\rho)^{1/2}$.)
Comme les extrémités du fil sont fixées, il est clair que le déplacement vertical est toujours nul en celles-ci, ce que l'on exprime sous la forme de conditions aux limites

$$u(0, t) = u(1, t) = 0 \text{ pour } t > 0. \tag{1.11}$$

Le but étant de déterminer les déplacements verticaux de la corde à partir d'une configuration initiale, celle-ci est donc connue : c'est une *donnée*. Pour cela, comme c'est une dérivée seconde en temps qui intervient dans le modèle, nous avons besoin de connaître *à la fois* sa position, ainsi que la dérivée partielle par rapport au temps [1]

$$u(x, 0) = u^0(x) \text{ et } \frac{\partial u}{\partial t}(x, 0) = u^1(x) \text{ pour } x \in]0, 1[. \tag{1.12}$$

La donnée dans l'équation (1.12) est le couple (u^0, u^1) : on parle de **conditions initiales**. Enfin, le fait que l'énergie soit bornée peut-être exprimé sous la forme

$$\int_0^1 \left(\left[\frac{\partial u}{\partial t}(x, t) \right]^2 + c(\tau)^2 \left[\frac{\partial u}{\partial x}(x, t) \right]^2 \right) dx < \infty \text{ pour } t > 0. \tag{1.13}$$

Le **domaine de calcul** est ici égal à $]0, 1[\times]0, +\infty[$: c'est un *ouvert* de $\mathbb{R}^2$, par rapport au couple de variables (x, t). On peut aussi choisir de calculer la solution entre un instant initial $t = 0$ et un instant final $T > 0$, auquel cas il convient de remplacer la condition "pour $t > 0$" par "pour $t \in]0, T[$" dans (1.10)-(1.13). Dans ce cas, le domaine de calcul est égal à $]0, 1[\times]0, T[$.

1. Pour une justification intuitive, nous renvoyons le lecteur à la dernière section.

Remarque 1.6 *Il convient de bien différencier "condition à un instant donné", de "condition aux limites en temps". Pour s'en convaincre, supposons que l'instant final T soit égal à un dans le modèle de la corde vibrante, de solution u_c. On l'a mis en équations sous la forme (1.10)-(1.13) pour (x,t) variant dans $]0,1[\times]0,1[$. Quelles valeurs de u_c connaît-on* a priori *?*

$$u_c(0,t),\ u_c(1,t) \text{ pour } t \in]0,1[\text{ (cf. (1.11))}, \text{ et } u_c(x,0) \text{ pour } x \in]0,1[\text{ (cf. (1.12))}.$$

Par contre, $x \mapsto u_c(x,1)$ est à déterminer !
Si on reprend le modèle 2D statique de la membrane élastique (1.4)-(1.6), de solution u_m, il est également posé dans le domaine $]0,1[\times]0,1[$, avec cette fois la variable $\boldsymbol{x} = (x,y)$. Quelles valeurs de u_m connaît-on a priori *? Celles imposées sur la frontière du domaine par (1.5), c'est-à-dire*

$$u_m(0,y),\ u_m(1,y) \text{ pour } y \in]0,1[, \text{ et } u_m(x,0),\ u_m(x,1) \text{ pour } x \in]0,1[.$$

Ainsi, $x \mapsto u_m(x,1)$ est connue !
Il y a donc là une différence fondamentale*, due à la nature des opérateurs associés à chaque modèle (adimensionnalisé) :*

$$-\frac{\partial^2 u_m}{\partial x^2} \underset{\downarrow}{-} \frac{\partial^2 u_m}{\partial y^2} \quad vs. \quad -\frac{\partial^2 u_c}{\partial x^2} \underset{\downarrow}{+} \frac{\partial^2 u_c}{\partial t^2}.$$

Pour le modèle 2D instationnaire, reprenons celui de la *membrane*. Il s'agit de déterminer des "petits" déplacements verticaux, lorsque cette membrane est soumise à une force verticale qui dépend du temps ; soit $f(\boldsymbol{x},t)$ la densité de force par unité de surface. Le déplacement vertical est toujours noté u, et est lui aussi fonction de $(\boldsymbol{x},t)$. La variable $\boldsymbol{x}$ parcourt D et t est soit positif ($t>0$), soit compris entre zéro et T ($t \in]0,T[$), où l'instant final T est fixé. D'après les équations de l'élasticité linéaire, u vérifie

$$\left(\frac{\partial^2 u}{\partial t^2} - c(\tau)^2 \Delta_2 u\right)(\boldsymbol{x},t) = f(\boldsymbol{x},t) \text{ pour } \boldsymbol{x} \in D \text{ et } t>0. \tag{1.14}$$

Comme la frontière ∂D est fixée, on écrit

$$u(\boldsymbol{x},t) = 0, \text{ pour } \boldsymbol{x} \in \partial D \text{ et } t>0. \tag{1.15}$$

Les conditions initiales s'écrivent cette fois

$$u(\boldsymbol{x},0) = u^0(\boldsymbol{x}) \text{ et } \frac{\partial u}{\partial t}(\boldsymbol{x},0) = u^1(\boldsymbol{x}) \text{ pour } \boldsymbol{x} \in D. \tag{1.16}$$

Enfin, on exprime le fait que l'énergie soit bornée par la relation

$$\int_D \left(\left[\frac{\partial u}{\partial t}(\boldsymbol{x},t)\right]^2 + c(\tau)^2 \left|\mathbf{grad}_2\, u(\boldsymbol{x},t)\right|^2 \right) d\boldsymbol{x} < \infty \text{ pour } t>0. \tag{1.17}$$

Pour un modèle 3D instationnaire, nous nous intéressons à l'*acoustique* d'une salle S. Il s'agit de calculer la propagation du son engendrée par des hauts-parleurs, c'est-à-dire les *variations de pression* de l'air ambiant par rapport à une pression de référence P_{ref}.

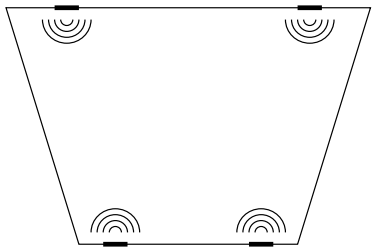


FIGURE 1.4 – Modèle 3D instationnaire : acoustique avec quatre hauts-parleurs

Précisément, c'est le déplacement de la membrane des hauts-parleurs, qui va générer ces variations.

On note ainsi p la variation de pression par rapport à la pression de référence, qui est "petite" par nature (nos tympans ne le supporteraient pas, dans le cas contraire !). C'est une fonction des variables $(\boldsymbol{x}, t)$; la variable $\boldsymbol{x} = (x, y, z)$ parcourt S et t est soit positif, soit dans $]0, T[$, où T est fixé. Le déplacement f des membranes dépend lui aussi des variables $(\boldsymbol{x}, t)$, $\boldsymbol{x}$ parcourant dans ce cas la surface des membranes, appelée HP dans la suite.

Les équations vérifiées par p sont successivement

$$\left(\frac{\partial^2 p}{\partial t^2} - c^2 \Delta_3 p\right)(\boldsymbol{x}, t) = 0 \text{ pour } \boldsymbol{x} \in S \text{ et } t > 0. \tag{1.18}$$

(Pas de forces volumiques dans la salle ; c est la *vitesse du son* dans l'air.)
Pour ce qui concerne les conditions aux limites, elles sont de deux types. Sur les membranes, le flux de pression est supposé proportionnel au déplacement f. Ce flux est égal à $\partial p/\partial n = \mathbf{grad}_3\, p \cdot \boldsymbol{n}$, où $\boldsymbol{n}$ est la normale unitaire sortante à S sur les membranes. Et, aux parois de la salle, $\partial S \setminus HP$, la pression est égale à P_{ref}, soit une variation nulle.

$$\begin{cases} p(\boldsymbol{x}, t) = 0, \text{ pour } \boldsymbol{x} \in \partial S \setminus HP \\ \dfrac{\partial p}{\partial n}(\boldsymbol{x}, t) = \alpha f, \text{ pour } \boldsymbol{x} \in HP \end{cases} \quad \text{et } t > 0. \tag{1.19}$$

L'air étant au repos à l'instant initial, les conditions initiales s'écrivent cette fois

$$p(\boldsymbol{x}, 0) = 0 \text{ et } \frac{\partial p}{\partial t}(\boldsymbol{x}, 0) = 0 \text{ pour } \boldsymbol{x} \in D. \tag{1.20}$$

Enfin, l'énergie acoustique est bornée, ce qui s'écrit

$$\int_S \left(\left[\frac{\partial p}{\partial t}(\boldsymbol{x}, t)\right]^2 + c^2 \left|\mathbf{grad}_3\, p(\boldsymbol{x}, t)\right|^2 \right) d\boldsymbol{x} < \infty \text{ pour } t > 0. \tag{1.21}$$

Quels sont les points communs entre les trois modèles instationnaires (1.10)-(1.13), (1.14)-(1.17) et (1.18)-(1.21) ?

— Problèmes posés dans des domaines $\Omega_d \times]0, T[$ de $\mathbb{R}^{d+1}$, $d = 1, 2, 3$, par rapport aux variables $(\boldsymbol{x}, t)$:

1. $d = 1$: corde ;
2. $d = 2$: membrane ;
3. $d = 3$: acoustique.

— L'opérateur intérieur est identique, pour tous les problèmes :

$$\frac{\partial^2 u}{\partial t^2} - c^2 \Delta_d u.$$

— Les conditions aux limites (à la frontière) présentées jusque là sont de deux types :

Dirichlet : condition sur la valeur de l'inconnue, *id est* $u = \cdots$;

Neumann : condition sur le flux de l'inconnue, i.e. $\dfrac{\partial u}{\partial n} = \cdots$.

— Les conditions initiales sont toujours au nombre deux, pour l'opérateur intérieur présenté ci-dessus : l'une sur u, et l'autre sur $\partial_t u$, à un instant donné.

— Enfin, même propriété pour l'énergie, bornée pour tous les problèmes (voir la section 1.3 pour un résultat plus précis), pour tout t :

$$\int_{\Omega_d} \left(\left[\frac{\partial u}{\partial t}(\boldsymbol{x}, t) \right]^2 + c^2 \left| \mathbf{grad}_d \, u(\boldsymbol{x}, t) \right|^2 \right) d\boldsymbol{x} < \infty.$$

— Problèmes **linéaires** par rapport aux données f, u^0 et u^1.

Pour élargir la perspective, présentons brièvement un autre type de modèle 3D instationnaire, lui aussi basé sur le Laplacien : l'*équation de la chaleur*. Cette fois, on souhaite chauffer la salle S ! Il s'agit de calculer les *variations de température* de l'air ambiant par rapport à une température de référence T_{ref}. Plus précisément, supposons que ces variations soient le fait d'une source volumique de chaleur w. On note θ la variation de température, "petite" (variation de quelques dizaines de degrés au plus, essentiellement au voisinage de la source).

Les équations vérifiées par θ sont successivement

$$\left(\frac{\partial \theta}{\partial t} - \Delta_3 \theta \right) (\boldsymbol{x}, t) = w(\boldsymbol{x}, t) \text{ pour } \boldsymbol{x} \in S \text{ et } t > 0. \tag{1.22}$$

Pour ce qui concerne les conditions aux limites, elles sont d'un seul type, si on suppose que les murs, sol et plafond restent à la température de référence [2]

$$\theta(\boldsymbol{x}, t) = 0, \text{ pour } \boldsymbol{x} \in \partial S \text{ et } t > 0. \tag{1.23}$$

Si la température est égale à T_{ref} à l'instant initial, la condition initiale s'écrit [3]

$$\theta(\boldsymbol{x}, 0) = 0 \text{ pour } \boldsymbol{x} \in S. \tag{1.24}$$

2. Si de l'énergie thermique s'échappait par une ouverture, on aurait sur celle-ci une condition aux limites du type $\lambda\theta + \dfrac{\partial \theta}{\partial n} = 0$, avec $\lambda > 0$, appelée condition aux limites de **Fourier**.

3. Pour ce type d'équations, la donnée de la répartition initiale de température suffit, puisque seule une dérivée partielle première par rapport à t intervient dans le modèle.

Enfin, l'énergie thermique est bornée (voir la section 1.3 pour un résultat plus précis), ce qui s'écrit

$$\int_S (\theta(\boldsymbol{x},t))^2\,d\boldsymbol{x} + \int_0^t \int_S |\mathbf{grad}_3\,\theta(\boldsymbol{x},s)|^2\,d\boldsymbol{x}ds < \infty \ \text{ pour } t>0. \tag{1.25}$$

Remarque 1.7 *On imagine aisément le même type de problème dans une section 2D, pour arriver cette fois à une équation de la chaleur 2D.*

1.3 Classification et propriétés

Du point de vue de la terminologie, on peut classer les équations (celles qui sont définies dans l'intégralité du domaine de calcul) présentées précédemment en trois catégories. Notons tout d'abord qu'il s'agit d'**équations aux dérivées partielles ou EDP** au sens où, si u est fonction de $\boldsymbol{x} = (x_1, \cdots, x_d)$ (et de t), les dérivées partielles de la solution u par rapport à x_1, $\cdots$, x_d (et t) apparaissent. Intéressons-nous maintenant à des modèles dépendant de trois variables, c'est-à-dire des modèles 3D statiques, ou 2D instationnaires. Si nous remplaçons *symboliquement* les dérivations partielles ∂_x, ∂_y, ∂_z, ∂_t respectivement par un facteur X, Y, Z, T, nous arrivons aux symboles (au signe près) :

- $S_1 = X^2 + Y^2 + Z^2$ pour la cavité 3D statique ;
- $S_2 = X^2 + Y^2 - T^2$ pour la membrane 2D instationnaire ;
- $S_3 = X^2 + Y^2 - T$ pour l'équation de la chaleur 2D.

Si nous supposons que S_1, S_2 et S_3 sont des constantes (positive pour S_1), on a successivement l'équation d'une ellipse (ici une sphère), d'un hyperboloïde de révolution, et enfin d'un paraboloïde de révolution. C'est pourquoi, on parle :

- d'**EDP elliptique** pour qualifier les problèmes statiques de la section 1.1 ;
- d'**EDP hyperbolique** pour qualifier les problèmes instationnaires de la section 1.2, faisant intervenir une dérivée partielle seconde par rapport au temps (corde, membrane, acoustique) ;
- d'**EDP parabolique** pour qualifier les problèmes instationnaires de la section 1.2, faisant intervenir une dérivée partielle première par rapport au temps (équations de la chaleur).

Cette terminologie, outre son aspect pratique (classification des EDP en trois catégories), revêt une importance fondamentale, lorsqu'il s'agit d'étudier les propriétés de leurs solutions respectives. En effet, on peut vérifier, que deux solutions d'EDP d'une même classe possèdent un certain nombre de *propriétés identiques*. Donnons-en quelques exemples, d parcourant $\{1,2,3\}$. L'énoncé des résultats est parfois vague (mais c'est volontaire !), et le lecteur est invité à consulter [11, 2] pour les résultats précis et les preuves qui les accompagnent.

Commençons par les problèmes elliptiques

$$-\Delta_d u = f \text{ sur } \Omega_d, \quad u = g \text{ sur } \partial\Omega_d. \tag{1.26}$$

(Ici, nous considérons que la condition aux limites sur la frontière $\partial\Omega_d$ peut être non-constante.)

Théorème 1.8 (Principe de positivité) *Supposons que f et g soient positives, respectivement sur Ω_d et $\partial\Omega_d$. Alors la solution u du problème (1.26) est positive sur Ω_d.*

Théorème 1.9 (Existence et unicité de la solution) *Le problème (1.26) admet une solution et une seule $\boldsymbol{x} \mapsto u(\boldsymbol{x})$ sur le domaine Ω_d, qui dépend continûment des données f et g.*

Remarque 1.10 *On note que si f et g sont nulles, l'unicité de la solution impose que $u = 0$.*

Poursuivons par les problèmes hyperboliques

$$\left\{ \begin{array}{l} \dfrac{\partial^2 u}{\partial t^2} - c^2 \Delta_d u = f \text{ sur } \Omega_d \times]0,T[, \quad u = 0 \text{ sur } \partial\Omega_d \times]0,T[, \\ u|_{t=0} = u^0,\ \dfrac{\partial u}{\partial t}|_{t=0} = u^1 \text{ sur } \Omega_d. \end{array} \right. \tag{1.27}$$

Dans ce cas, on peut établir une égalité d'énergie, en introduisant l'énergie associée à la solution u, somme d'une énergie cinétique et d'une énergie potentielle :

$$E(t) = \frac{1}{2} \int_{\Omega_d} \left(\left[\frac{\partial u}{\partial t}(\boldsymbol{x},t) \right]^2 + c^2 \left| \mathbf{grad}_d\, u(\boldsymbol{x},t) \right|^2 \right) d\boldsymbol{x}. \tag{1.28}$$

Théorème 1.11 (Egalité d'énergie) *A tout instant t $(0 < t < T)$, l'énergie vérifie la relation*

$$E(t) = \int_0^t \int_{\Omega_d} (f \frac{\partial u}{\partial t})(\boldsymbol{x},s)\, d\boldsymbol{x} ds + E(0). \tag{1.29}$$

Ainsi, lorsque la source volumique est nulle, alors l'énergie est conservée au cours du temps.

Corollaire 1.12 (Caractère conservatif) *Supposons que $f = 0$ dans (1.27).*
Pour t $(0 < t < T)$, on a l'égalité

$$E(t) = E(0). \tag{1.30}$$

Théorème 1.13 (Existence et unicité de la solution) *Le problème (1.27) admet une solution et une seule $(\boldsymbol{x},t) \mapsto u(\boldsymbol{x},t)$ sur le domaine $\Omega_d \times]0,T[$, qui dépend continûment des données f, u^0 et u^1.*

Finissons par les problèmes paraboliques

$$\frac{\partial u}{\partial t} - \Delta_d u = f \text{ sur } \Omega_d \times]0,T[, \quad u = 0 \text{ sur } \partial\Omega_d \times]0,T[, \quad u|_{t=0} = u^0 \text{ sur } \Omega_d. \tag{1.31}$$

(Ici, nous considérons que la condition initiale sur Ω_d peut être quelconque.)
Dans ce cas, on peut également établir une égalité d'énergie.

Théorème 1.14 (Egalité d'énergie) *A tout instant t $(0 < t < T)$, la solution u du problème (1.31) vérifie la relation*

$$\begin{aligned} \frac{1}{2} \int_{\Omega_d} (u(\boldsymbol{x},t))^2\, d\boldsymbol{x} + \int_0^t \int_{\Omega_d} |\, \mathbf{grad}_d\, u(\boldsymbol{x},s)|^2\, d\boldsymbol{x} ds \;\; &= \;\; \int_0^t \int_{\Omega_d} (fu)(\boldsymbol{x},s)\, d\boldsymbol{x} ds \\ &\quad + \frac{1}{2} \int_{\Omega_d} (u^0(\boldsymbol{x}))^2\, d\boldsymbol{x}. \end{aligned} \tag{1.32}$$

Si de plus la source volumique est nulle, alors la norme de la solution décroît au cours du temps.

Corollaire 1.15 (Caractère dissipatif) *Supposons que $f = 0$ dans (1.31). Pour t_1 et t_2 tels que $0 \leq t_1 < t_2 < T$ et $\int_{\Omega_d} (u(\boldsymbol{x}, t_1))^2 \, d\boldsymbol{x} \neq 0$, on a l'inégalité stricte*

$$\int_{\Omega_d} (u(\boldsymbol{x}, t_2))^2 \, d\boldsymbol{x} < \int_{\Omega_d} (u(\boldsymbol{x}, t_1))^2 \, d\boldsymbol{x}. \tag{1.33}$$

Théorème 1.16 (Décroissance exponentielle) *Supposons que $f = 0$ dans (1.31). Il existe une constante $\lambda > 0$* indépendante de T *telle que, pour tout t de $[0, T[$,*

$$\int_{\Omega_d} (u(\boldsymbol{x}, t))^2 \, d\boldsymbol{x} \leq \mathrm{e}^{-\lambda t} \int_{\Omega_d} (u^0(\boldsymbol{x}))^2 \, d\boldsymbol{x}. \tag{1.34}$$

Théorème 1.17 (Existence et unicité de la solution) *Le problème (1.31) admet une solution et une seule $(\boldsymbol{x}, t) \mapsto u(\boldsymbol{x}, t)$ sur le domaine $\Omega_d \times]0, T[$, qui dépend continûment des données f et u^0.*

Définition 1.18 *Lorsqu'un problème admet une solution et une seule, qui de plus dépend continûment des données, on dit qu'il est* **bien posé**.

Pour conclure cette courte section sur la classification, notons qu'il existe bien d'autres modèles (avec ou sans Laplacien !), qui font partie d'une de ces trois classes. Nous examinons ci-dessous les équations de Maxwell en milieu hétérogène, ainsi que les équations de la diffusion neutronique.

Commençons par les équations de Maxwell dans la cavité 3D introduite à la section précédente, extérieure à un ensemble de conducteurs parfaits. Elles s'écrivent sous la forme bien connue suivante, composée des relations d'Ampère, de Faraday, de Gauss et de l'absence de monopoles magnétiques libres :

$$\begin{aligned}
&\varepsilon \frac{\partial \boldsymbol{E}}{\partial t} - \mathbf{rot}(\mu^{-1} \boldsymbol{B}) = -\boldsymbol{J} \ \text{ sur } C \times]0, T[, \\
&\frac{\partial \boldsymbol{B}}{\partial t} + \mathbf{rot}\, \boldsymbol{E} = 0 \ \text{ sur } C \times]0, T[, \\
&\mathrm{div}(\varepsilon \boldsymbol{E}) = \rho \ \text{ sur } C \times]0, T[, \\
&\mathrm{div}\, \boldsymbol{B} = 0 \ \text{ sur } C \times]0, T[.
\end{aligned}$$

Ci-dessus, ε et μ sont respectivement la permittivité électrique et la perméabilité magnétique du milieu, et elles peuvent dépendre de la position $\boldsymbol{x}$. Elles sont à valeurs strictement positives, et uniformément bornées inférieurement et supérieurement par des nombres strictement positifs. Les opérateurs différentiels rotationnel **rot** et divergence div sont définis par

$$\mathbf{rot}\, \boldsymbol{v} = (\frac{\partial v_3}{\partial x_2} - \frac{\partial v_2}{\partial x_3}) \boldsymbol{e}_1 + (\frac{\partial v_1}{\partial x_3} - \frac{\partial v_3}{\partial x_1}) \boldsymbol{e}_2 + (\frac{\partial v_2}{\partial x_1} - \frac{\partial v_1}{\partial x_2}) \boldsymbol{e}_3, \quad \mathrm{div}\, \boldsymbol{v} = \sum_{k=1}^{3} \frac{\partial v_k}{\partial x_k}.$$

Les données sont $\boldsymbol{J}$ et ρ, respectivement les densités de courant et de charge. Les inconnues sont $\boldsymbol{E}$ le champ électrique, et $\boldsymbol{B}$, l'induction magnétique : $(\boldsymbol{E}, \boldsymbol{B})$ est le **champ électromagnétique** à déterminer. Qui plus est, on suppose connue la valeur initiale du champ, puisque des dérivées partielles premières par rapport au temps interviennent dans notre modèle. On écrit :

$$\boldsymbol{E}|_{t=0} = \boldsymbol{E}^0, \ \boldsymbol{B}|_{t=0} = \boldsymbol{B}^0 \ \text{ sur } C.$$

Sur la frontière de la cavité, on a la relation

$$\boldsymbol{E} \times \boldsymbol{n} = 0, \quad \text{sur } \partial C \times]0, T[,$$

qui stipule que les composantes tangentielles du champ électrique $\boldsymbol{E}$ s'annulent sur la frontière à tout instant (ci-dessus, $\boldsymbol{n}$ est un vecteur normal unitaire sortant à la frontière de la cavité.)
Et l'énergie électromagnétique est bornée, ce qui s'écrit

$$\int_C \left(\varepsilon(\boldsymbol{x}) \left| \boldsymbol{E}(\boldsymbol{x}, t) \right|^2 + \frac{1}{\mu(\boldsymbol{x})} \left| \boldsymbol{B}(\boldsymbol{x}, t) \right|^2 \right) d\boldsymbol{x} < \infty \text{ pour } t > 0. \tag{1.35}$$

Sous réserve que les densités de charge et de courant ρ et $\boldsymbol{J}$ vérifient la relation de conservation de la charge

$$\frac{\partial \rho}{\partial t} + \operatorname{div} \boldsymbol{J} = 0,$$

ce modèle est *bien posé* (les données sont $\boldsymbol{E}^0$, $\boldsymbol{B}^0$, ρ et $\boldsymbol{J}$). Par ailleurs, on peut prouver qu'il est de type *hyperbolique*.

Dans un milieu homogène, on peut simplifier ces équations : prenons l'exemple du vide. On rappelle que si μ_0 est la perméabilité magnétique du vide, alors c, la vitesse de la lumière, satisfait à $c^2 \varepsilon_0 \mu_0 = 1$.

$$\begin{aligned}
&\frac{\partial \boldsymbol{E}}{\partial t} - c^2 \operatorname{\mathbf{rot}} \boldsymbol{B} = -\frac{1}{\varepsilon_0} \boldsymbol{J} \ \text{ sur } C \times]0, T[, \\
&\frac{\partial \boldsymbol{B}}{\partial t} + \operatorname{\mathbf{rot}} \boldsymbol{E} = 0 \ \text{ sur } C \times]0, T[, \\
&\operatorname{div} \boldsymbol{E} = \frac{1}{\varepsilon_0} \rho \ \text{ sur } C \times]0, T[, \\
&\operatorname{div} \boldsymbol{B} = 0 \ \text{ sur } C \times]0, T[.
\end{aligned}$$

La relation de Gauss est souvent appelée relation de Coulomb dans le cas homogène.

Poursuivons par les équations de la diffusion neutronique. Dans un réacteur nucléaire, que l'on décrira par un ouvert $R \in \mathbb{R}^3$, la densité de neutrons dépend de 7 variables : le temps $t \in \mathbb{R}^+$, l'énergie des neutrons $E \in \mathbb{R}^+$, leur direction des vitesses $\boldsymbol{\Omega} \in \mathbb{S}^2$ ($\mathbb{S}^2$ représente la sphère unité) et leur position $\boldsymbol{x} \in \mathbb{R}^3$. Elle est régie par l'équation de transport des neutrons (qui est une équation intégro-différentielle de type hyperbolique), similaire à l'équation de Boltzmann [15], et dont on ne connaît en général pas de solution explicite. En raison du grand nombre de variable, la version discrète de cette équation est très coûteuse à résoudre. C'est pourquoi on s'intéresse à l'approximation la plus simple de l'équation de transport de neutrons, obtenue en intégrant l'équation de transport sur la variable $\boldsymbol{\Omega}$.

Dans le cas où l'on fait l'intégration également sur la variable énergétique E, on obtient l'équation suivante (qui est une EDP parabolique), appelée *équation cinétique de diffusion neutronique*[4] :

$$\frac{1}{|v|}\frac{\partial \phi}{\partial t} - \operatorname{div} k\,\mathbf{grad}_d\,\phi + \sigma_a \phi = \nu\sigma_f\phi + S \text{ sur } R\times]0,T[. \tag{1.36}$$

La grandeur d'intérêt $\phi(\boldsymbol{x},t)$ est appelée par les neutroniciens le *flux scalaire de neutrons*, et est une quantité positive. $k(\boldsymbol{x},t)$ est appelé coefficient de diffusion, $\sigma_a(\boldsymbol{x},t)$ est la section efficace macroscopique d'absorption des neutrons, $\sigma_f(\boldsymbol{x},t)$ la section efficace macroscopique de fission des neutrons, et ν, le nombre moyen de neutrons émis par fission. La donnée $S(\boldsymbol{x},t)$ représente une possible source externe de neutrons. On suppose de plus que les coefficients et la source sont tels que :

$$\int_R (\phi(\boldsymbol{x},t))^2\,d\boldsymbol{x} < \infty \text{ et } \int_R |\,\mathbf{grad}_d\,\phi(\boldsymbol{x},t)|^2\,d\boldsymbol{x} < \infty \text{ pour tout } t>0.$$

A cette équation, on doit bien sûr adjoindre une condition initiale : $\phi(\boldsymbol{x},0) = \phi_0(\boldsymbol{x})$ sur R, ainsi que des conditions aux limites sur ∂R. La condition aux limites de **Robin** permet de modéliser le fait que la frontière du domaine d'étude n'est pas totalement absorbante[5]. Elle s'écrit :

$$\phi + \alpha\, k\frac{\partial \phi}{\partial n} = 0 \text{ sur } \partial R,$$

où $k\,\partial\phi/\partial n = k\,\mathbf{grad}_d\,\phi\cdot\boldsymbol{n}$, avec $\alpha\in\mathbb{R}^+$.
L'opérateur $\operatorname{div} k\,\mathbf{grad}_d$ est appelé *Laplacien généralisé*, car lorsque k est constant dans R, on a : $\operatorname{div} k\,\mathbf{grad}_d\cdot = k\Delta\cdot$. Concernant la neutronique, en général k (ainsi que les autres coefficients) est constant ou polynômial par morceaux.

1.4 Problèmes aux valeurs propres et problèmes stationnaires

Pour aborder cette dernière section, reprenons le modèle de la membrane 2D instationnaire (1.14)-(1.17), et posons-nous cette question :
Est-ce que la membrane est "stable", pour f dépendant du temps donnée ?
Ou bien, de façon équivalente :
Peut-on s'assurer que les déplacements verticaux restent "petits" ?

A cette question, on peut apporter deux types de réponses :
— L'une, *a posteriori*, consiste en la simulation numérique du comportement de la membrane, issue de la discrétisation de (1.14)-(1.17). Voir le chapitre 2 pour cette approche.
— L'autre, qui consiste en la décomposition du modèle (1.14)-(1.17) en une série de *problèmes aux valeurs propres*, puis en une étude *a priori*.

4. l'équation peut aussi être homogénéisée en espace, c'est pourquoi on considère désormais $R\in\mathbb{R}^d$
5. dans le cas de l'équation de la chaleur, la même condition aux limites est plutôt appelée condition aux limites de Fourier

C'est cette seconde approche que nous allons utiliser ci-après. En effet, il est possible (consulter par exemple [2]) de décomposer la solution $u(\boldsymbol{x})$ de (1.14)-(1.17) selon

$$u(\boldsymbol{x},t) = \sum_{k\in\mathbb{N}} \alpha_k(t)\, u_k(\boldsymbol{x}). \tag{1.37}$$

La somme est infinie[6], puisqu'elle porte sur tout nombre naturel k positif. Pour chaque $k \in \mathbb{N}$, u_k est solution d'un **problème statique 2D aux valeurs propres** : on doit déterminer $\boldsymbol{x} \mapsto u_k(\boldsymbol{x})$ et $\lambda_k > 0$ tels que

$$-c^2\Delta_2 u_k = \lambda_k u_k \text{ sur } D, \quad u_k = 0 \text{ sur } \partial D. \tag{1.38}$$

Les éléments du couple (u_k, λ_k) sont respectivement appelés **mode propre** de la membrane, et **valeur propre** associée λ_k. Pour des raisons pratiques qui vont apparaître immédiatement, nous introduisons la **pulsation propre** ω_k, égale à $\sqrt{\lambda_k}$. On décompose également f, u^0 et u^1, sous la forme

$$f(\boldsymbol{x},t) = \sum_{k\in\mathbb{N}} f_k(t)\, u_k(\boldsymbol{x}), \quad u^m(\boldsymbol{x}) = \sum_{k\in\mathbb{N}} \alpha_k^m\, u_k(\boldsymbol{x}),\ m = 0, 1. \tag{1.39}$$

On trouve donc

$$\sum_{k\in\mathbb{N}} \big(\alpha_k''(t) + \lambda_k\alpha_k(t) - f_k(t)\big)\, u_k(\boldsymbol{x}) = 0 \text{ pour } (\boldsymbol{x},t) \in D\times]0,T[, \tag{1.40}$$

$$\sum_{k\in\mathbb{N}} \big(\alpha_k(0) - \alpha_k^0\big)\, u_k(\boldsymbol{x}) = 0,\ \text{pour } \boldsymbol{x} \in D, \tag{1.41}$$

$$\sum_{k\in\mathbb{N}} \big(\alpha_k'(0) - \alpha_k^1\big)\, u_k(\boldsymbol{x}) = 0 \text{ pour } \boldsymbol{x} \in D. \tag{1.42}$$

Ainsi, une fois $(u_k)_{k\in\mathbb{N}}$ connus, il est équivalent de résoudre (1.14)-(1.17), ou la série ($k \in \mathbb{N}$) d'équations différentielles ordinaires (EDO) :

$$\alpha_k''(t) + \omega_k^2\alpha_k(t) = f_k(t),\ \text{pour } t \in]0,T[,\ \alpha_k(0) = \alpha_k^0,\ \alpha_k'(0) = \alpha_k^1. \tag{1.43}$$

Fixons k.
La *solution générale* ($f_k = 0$) de l'EDO correspondante est de la forme

$$\alpha_k^G(t) = A_k \sin(\omega_k t) + B_k \cos(\omega_k t).$$

A l'aide des *deux conditions initiales*, on peut déterminer les valeurs de A_k et B_k, pour arriver à

$$\alpha_k^G(t) = \frac{\alpha_k^1}{\omega_k}\sin(\omega_k t) + \alpha_k^0 \cos(\omega_k t),\ \text{pour } t \in]0,T[.$$

Remarque 1.19 *On comprend, à la suite de ce calcul, pourquoi* deux *conditions initiales sont nécessaires pour un problème faisant intervenir une dérivée partielle seconde par rapport au temps. De façon similaire, on trouverait qu'*une *condition initiale suffit, pour un problème comprenant uniquement une dérivée partielle première par rapport au temps.*

6. Par définition $u_k \neq 0$, $\forall k \in \mathbb{N}$.

On peut facilement vérifier qu'une *solution particulière* ($\alpha_k^0 = \alpha_k^1 = 0$) de l'EDO est de la forme

$$\alpha_k^P(t) = \frac{1}{\omega_k} \int_0^t \sin(\omega_k(t-s)) f_k(s)\, ds.$$

Ainsi, la solution complète de l'EDO s'écrit

$$\alpha_k(t) = \frac{\alpha_k^1}{\omega_k} \sin(\omega_k t) + \alpha_k^0 \cos(\omega_k t) + \frac{1}{\omega_k} \int_0^t \sin(\omega_k(t-s)) f_k(s)\, ds, \text{ pour } t \in]0, T[. \quad (1.44)$$

La solution de l'EDP instationnaire modélisant la membrane est finalement égale à

$$\begin{aligned} u(\boldsymbol{x}, t) &= \sum_{k\in\mathbb{N}} \left\{ \frac{\alpha_k^1}{\omega_k} \sin(\omega_k t) + \alpha_k^0 \cos(\omega_k t) \right. \\ &\qquad \left. + \frac{1}{\omega_k} \int_0^t \sin(\omega_k(t-s)) f_k(s)\, ds \right\} u_k(\boldsymbol{x}), \text{ pour } (\boldsymbol{x}, t) \in D \times]0, T[. \quad (1.45) \end{aligned}$$

Conclusion : les déplacements verticaux de la membrane sont donc "petits" si, et seulement si, tous les coefficients entre accolades (les $(\alpha_k^G + \alpha_k^P)(t)$) sont eux-mêmes "petits", pour t variant de 0 à T.
Or, l'amplitude des deux premiers termes – dont la somme est égale à $\alpha_k^G(t)$ – ne dépend que de α_k^0 et α_k^1/ω_k, et est par voie de conséquence indépendante de t : ils sont donc "petits" si, et seulement si, les données initiales le sont ! Ceci est intuitif...
Qu'en est-il du troisième et dernier terme, $\alpha_k^P(t)$? La réponse est moins immédiate... Considérons une "petite" sollicitation, prenant la forme simple

$$f(\boldsymbol{x}, t) = f_l(t) u_l(\boldsymbol{x}), \text{ avec } f_l(t) = \mu_l \cos(\omega t),$$

pour une pulsation $\omega \geq 0$ et l donné. Le fait que la sollicitation soit "petite" revient à supposer que $\mu_l \neq 0$ est "petit". Que valent les coefficients $(\alpha_k^P)_{k\in\mathbb{N}}$? Si $k \neq l$, $\alpha_k^P = 0$. Par contre, si $k = l$, deux situations peuvent se produire. En effet, en calculant l'intégrale qui détermine ce coefficient, on arrive à deux résultats différents :

$$\text{si } \omega \neq \omega_l \ : \qquad \alpha_l^P(t) = \frac{\mu_l}{2\omega_l} \left\{ \frac{1}{\omega - \omega_l} + \frac{1}{\omega + \omega_l} \right\} (\cos(\omega_l t) - \cos(\omega t)) ; \quad (1.46)$$

$$\text{si } \omega = \omega_l \ : \qquad \alpha_l^P(t) = \frac{\mu_l}{2\omega_l} t \sin(\omega_l t). \quad (1.47)$$

Ainsi, si $\omega = \omega_l$, l'amplitude des oscillations croît linéairement avec t – on parle de **résonance** – et il n'y a aucune chance que les déplacements verticaux restent "petits" dans ce cas. Si par contre ω est différent (mais proche) de ω_l, le terme de plus grande amplitude est

$$\frac{1}{\omega - \omega_l} \left[\frac{\mu_l}{2\omega_l} (\cos(\omega_l t) - \cos(\omega t)) \right].$$

L'amplitude peut donc être "grande" si $|\omega - \omega_l|$ est "petit".

Bref, ce qui compte avant tout, pour une réponse *a priori*, c'est de déterminer les valeurs propres $(\lambda_k)_{k\in\mathbb{N}}$ ainsi que les modes propres associés, solution des *problèmes aux valeurs propres* (1.38), avec la condition

$$\int_C |\mathbf{grad}_2\, u_k(\boldsymbol{x})|^2\, d\boldsymbol{x} < \infty.$$

A partir de là, on choisira la pulsation ω de la sollicitation f aussi "éloignée" que possible des $(\sqrt{\lambda_k})_{k\in\mathbb{N}}$. En particulier, il sera très intéressant de déterminer avec précision la plus petite valeur $\omega_{min} = \min_k \sqrt{\lambda_k}$, puisqu'un choix de ω dans $]0, \omega_{min}[$ permettra à coup sûr d'éviter les résonances.

Abordons brièvement, pour clore cette section, la notion de **problème stationnaire**. Pour l'obtenir, il suffit de considèrer un problème dépendant du temps t, pour lequel la dépendance par rapport à t est connue explicitement : oscillations forcées en $\sin(\nu t)$ ou $\cos(\nu t)$ avec $\nu \neq 0$, par exemple. On peut alors remplacer la dérivation par rapport au temps par une multiplication par $-\nu^2$, pour aboutir à un problème du type

$$-\nu^2 u - c^2 \Delta_2 u = g \text{ sur } D, \quad u = 0 \text{ sur } \partial D. \tag{1.48}$$

Il n'y a plus de condition initiale, puisque la résolution de (1.48) permet, par multiplication par le coefficient de dépendance en temps, de déterminer complètement la solution.

La différence avec le problème aux valeurs propres (1.38) tient en deux points :
— la valeur de ν est *connue*, alors qu'ω_k est une *inconnue* du problème (1.38) ;
— il y a un second membre g *a priori* non nul dans (1.48).

Chapitre 2

La méthode des différences finies

2.1 Introduction

Dans ce chapitre, nous présentons une méthode de discrétisation numérique. L'introduction de celle-ci nous permet de calculer des approximations des solutions d'EDP statiques ou instationnaires. On étudie le calcul du déplacement transversal d'un fil ou d'une poutre, puis le déplacement vertical d'une membrane élastique et enfin le potentiel électrostatique. Comme on le verra, à ces trois problèmes correspondent des approximations différentes, puisque les modèles dont ils sont issus sont respectivement posés dans $\mathbb{R}$, $\mathbb{R}^2$ et $\mathbb{R}^3$. Néanmoins, la technique d'approximation retenue, appelée **méthode des différences finies**, reste la même pour ces trois problèmes. En outre, nous expliquons comment on peut utiliser cette technique d'approximation pour des problèmes à coefficients variables. A la fin du chapitre, nous verrons comment cette méthode de discrétisation peut aussi être appliquée aux problèmes dépendant du temps.

2.2 Problèmes monodimensionnels

Dans cette section, nous considérons tout d'abord le problème modèle 1D, cf. §1.1. On considère ensuite un modèle plus général 1D, avec des coefficients variables.

2.2.1 Fil ou poutre

Soit donc un fil tendu entre ses extrémités, ou une poutre, appuyée en ses extrémités, situées en 0 et 1. On suppose qu'il ou elle est soumis à une force extérieure transverse (telle que son poids, dans le cas du fil pesant). On note $f : x \mapsto f(x)$ la densité linéique des forces appliquées, et $u : x \mapsto u(x)$ le déplacement transversal induit, que l'on cherche à approcher numériquement. Nous avons admis que les équations de l'élasticité linéaire monodimensionnelles (1D) normalisées permettent de modéliser correctement le phénomène. Rappelons-en la forme :

$$-u''(x) = f(x) \text{ sur }]0,1[, \quad u(0) = u(1) = 0, \tag{2.1}$$

composée d'une équation différentielle ordinaire et de conditions aux limites de Dirichlet. Dans le cas particulier où $f \equiv 1$, la solution est égale à

$$u_0(x) = \frac{1}{2}x(1-x). \tag{2.2}$$

Ce résultat élémentaire sera utile par la suite...

2.2.2 Schéma aux différences finies 1D

On suppose que la solution u est de classe $\mathcal{C}^4([0,1])$ ou, ce qui est *équivalent*, que la donnée f est de classe $\mathcal{C}^2([0,1])$. Pour déterminer une méthode d'approximation de l'équation aux dérivées partielles (2.1) (ça n'est pas la seule !), on utilise la

Proposition 2.1 *On suppose que $u \in \mathcal{C}^4([0,1])$. Soient $x \in]0,1[$ et h tel que $[x-h, x+h] \in [0,1]$. Alors*

$$\exists \theta \in]-1,1[\text{ tel que } -u''(x) = \frac{-u(x-h)+2u(x)-u(x+h)}{h^2} + \frac{h^2}{12}u^{(4)}(x+\theta\, h). \tag{2.3}$$

Démonstration : On utilise la formule de Taylor-Mac Laurin.

$$\begin{aligned} \exists \theta^- \in]-1,0[\quad | \quad u(x-h) &= u(x) - h\,u'(x) + \frac{h^2}{2}u''(x) - \frac{h^3}{6}u'''(x) + \frac{h^4}{24}u^{(4)}(x+\theta^- h) \\ \exists \theta^+ \in]0,1[\quad | \quad u(x+h) &= u(x) + h\,u'(x) + \frac{h^2}{2}u''(x) + \frac{h^3}{6}u'''(x) + \frac{h^4}{24}u^{(4)}(x+\theta^+ h). \end{aligned}$$

On somme les deux égalités, pour trouver

$$-u''(x) = \frac{-u(x-h)+2u(x)-u(x+h)}{h^2} + \frac{h^2}{24}(u^{(4)}(x+\theta^- h) + u^{(4)}(x+\theta^+ h)).$$

Pour arriver à l'expression annoncée, il faut se souvenir du théorème des valeurs intermédiaires. Il permet, puisque $u^{(4)}$ est continue, de remplacer les deux termes en $u^{(4)}$ par $2u^{(4)}(x+\theta\, h)$, mais avec un paramètre θ appartenant à $[\theta^-, \theta^+]$, donc à $]-1,1[$ comme annoncé. ◇

Remarque 2.2 *Le premier terme de (2.3) est une bonne approximation de $-u''(x)$, sous réserve que h est* petit. *En effet, comme on a la relation $-u^{(4)} = f''$, on sait que*

$$\left|\frac{h^2}{12}u^{(4)}(x+\theta\, h)\right| \leq \frac{C_{f,2}}{12}h^2, \text{ avec } C_{f,2} = \sup_{x\in[0,1]} |f''(x)|.$$

Ce résultat *simple* fournit une méthode de discrétisation et d'approximation de l'équation de départ (2.1) ; on parle souvent de **schéma numérique** de discrétisation. Le terme **différences finies** provient quant à lui de l'expression (2.3) : on remplace une dérivée, qui est par définition la *limite* d'un taux de variation, par un taux de variation, dont le dénominateur conserve une valeur finie *non nulle* (ici h^2 pour une dérivée seconde). En pratique, comment procède-t-on ?

Pour commencer, on choisit $N \in \mathbb{N}$, et on fixe $h = \dfrac{1}{N+1}$. Remarquons tout de suite que pour avoir une "bonne" approximation de $u''(x)$, il convient que h soit petit. Ceci signifie que N est un paramètre de discrétisation qui aura vocation à devenir "grand", lors de la réalisation des expériences numériques.

Nous allons construire une méthode qui permet d'approcher la valeur de u aux points $x_i = ih$, pour $i \in \{0, 1, \cdots, N+1\}$, par des nombres, notés $(u_i)_{0 \le i \le N+1}$. Puisque u est approchée en deux points consécutifs distants de h, on appelle h le **pas de discrétisation**.

FIGURE 2.1 – Le segment découpé ($N = 4$, $h = 0.2$)

Remarque 2.3 *Comme on sait que $u(0) = u(1) = 0$, on choisira toujours comme approximation $u_0 = u_{N+1} = 0$!*

On définit $f_i = f(x_i)$, pour $i \in \{1, \cdots, N\}$, et on considère l'ensemble des équations

$$\frac{-u_{i-1} + 2u_i - u_{i+1}}{h^2} = f_i, \ 1 \le i \le N, \text{ avec } u_0 = u_{N+1} = 0. \tag{2.4}$$

Chaque équation faisant intervenir trois nombres parmi $(u_i)_i$, on parle de **schéma à trois points**.

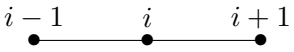


FIGURE 2.2 – Le schéma aux différences finies à trois points

NB. Noter la similitude entre (2.4) d'une part, et (2.3) et (2.1) en $x = x_i$, d'autre part.

2.2.3 Système linéaire pour les différences finies 1D

Si on appelle u (resp. f) le vecteur de $\mathbb{R}^N$ de composantes $(u_i)_{1 \le i \le N}$ (resp. $(f_i)_{1 \le i \le N}$), on peut réécrire le système (2.4) sous la *forme vectorielle équivalente*

$$\mathbb{A}_1 u = f, \text{ avec } \mathbb{A}_1 = \frac{1}{h^2} \begin{pmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -1 \\ 0 & \dots & 0 & -1 & 2 \end{pmatrix} \in \mathbb{R}^{N \times N}. \tag{2.5}$$

Par construction, la matrice [7] $\mathbb{A}_1$ est
- tridiagonale, c'est-à-dire que tous les termes non nuls sont regroupés sur trois diagonales ;
- symétrique, puisque $(\mathbb{A}_1)_{i,j} = (\mathbb{A}_1)_{j,i}$, pour $1 \leq i, j \leq N$.

Il convient maintenant de vérifier qu'il existe une solution u unique de (2.5). Qui plus est, est-il possible de calculer et majorer l'erreur commise ? C'est l'objet des résultats ci-dessous. Tout d'abord, nous allons vérifier que la matrice $\mathbb{A}_1$ est inversible. Outre l'obtention de l'existence et de l'unicité de u, ceci nous permettra de construire une formule explicite, exprimant l'erreur commise en fonction des données du problème. Par ailleurs, pour exploiter cette formule, c'est-à-dire pour majorer l'erreur, nous allons étudier les caractéristiques de l'inverse $\mathbb{A}_1^{-1}$.

Définition 2.4 *Un vecteur v de $\mathbb{R}^N$ est dit* **positif** *lorsque $v_i \geq 0$, $\forall 1 \leq i \leq N$.*
Une matrice A de $\mathbb{R}^{N\times N}$ est dite **positive** *lorsque $A_{i,j} \geq 0$, $\forall 1 \leq i, j \leq N$.*
Une matrice A de $\mathbb{R}^{N\times N}$ est dite **monotone** *lorsqu'elle est inversible, d'inverse positive.*

Avant de nous intéresser au cas particulier de la matrice issue du schéma à trois points, donnons une caractérisation simple des matrices monotones.

Proposition 2.5 *Une matrice A de $\mathbb{R}^{N\times N}$ est monotone si, et seulement si, on a l'inclusion*

$$\{v \in \mathbb{R}^N \ : \ Av \geq 0\} \subset \{v \in \mathbb{R}^N \ : \ v \geq 0\}. \tag{2.6}$$

Démonstration : Supposons que A est monotone.
Soit v tel que $Av \geq 0$, alors $v = A^{-1}(Av)$ et, pour $1 \leq i \leq N$, $v_i = \sum_j (A^{-1})_{i,j}(Av)_j \geq 0$, puisque $(A^{-1})_{i,j}$ et $(Av)_j$ sont positifs par hypothèse. Ainsi $v \geq 0$, et l'inclusion (2.6) est vérifiée.

Réciproquement, si l'inclusion est satisfaite, montrons tout d'abord que A est inversible.
Soit donc v tel que $Av = 0$: on a $Av \geq 0$ et $A(-v) \geq 0$, ce qui implique $v \geq 0$ et $(-v) \geq 0$, i.e. $v = 0$, d'où l'inversibilité.
Sachant que A^{-1} existe, étudions sa positivité...
On note $(\mathrm{e}_i)_i$ la base orthonormale canonique de $\mathbb{R}^N$. Alors les $f_i = A^{-1}\mathrm{e}_i$, pour i variant de 1 à N, sont les vecteurs colonnes de A^{-1}. On a bien sûr $\mathrm{e}_i = Af_i$, et l'inclusion (2.6) permet d'affirmer que f_i est positif, puisque e_i l'est. En d'autres termes, tous les éléments de A^{-1} sont positifs.
En conclusion, la matrice A est monotone. ◇

Proposition 2.6 *La matrice $\mathbb{A}_1$ correspondant à (2.5) est monotone.*

Démonstration : Pour prouver que $\mathbb{A}_1$ est monotone, on reprend la proposition 2.5. Soit v tel que $\mathbb{A}_1 v \geq 0$, et $v_k = \min_{1\leq i\leq N} v_i$ (ou, de façon équivalente, $v_k \leq v_i$, $\forall i$). Le but est d'arriver à l'inégalité $v_k \geq 0$. On a

$$\begin{cases} 2v_1 - v_2 \geq 0 \\ -v_{i-1} + 2v_i - v_{i+1} \geq 0, \quad 2 \leq i \leq N-1 \\ -v_{N-1} + 2v_N \geq 0 \end{cases} .$$

7. En règle générale, on écrira les matrices issues de la discrétisation par différences finies en caractères ombrés majuscules.

Si $v_k = v_1$, on trouve

$$v_k \geq v_2 - v_k \geq 0.$$

De même si $v_k = v_N$.

Si $k \in \{2, \cdots, N-1\}$, on trouve cette fois

$$(v_k - v_{k-1}) + (v_k - v_{k+1}) \geq 0.$$

Or, $v_k \leq v_{k-1}$ et, de même, $v_k \leq v_{k+1}$. On a donc $(v_k - v_{k-1}) + (v_k - v_{k+1}) \leq 0$, ce qui donne

$$v_{k-1} = v_{k+1} = v_k \,!$$

Par récurrence, on arrive facilement à $v_1 = \cdots = v_{k-1} = v_k = v_{k+1} = \cdots = v_N$. Et la première (ou la dernière) équation donne à nouveau $v_k \geq 0$. $\diamond$

NB. Dans le corps de la preuve, on a obtenu l'inclusion

$$\{v \in \mathbb{R}^N \ : \ v_i = \lambda, \ 1 \leq i \leq N, \ \lambda \in \mathbb{R}^+\} \subset \{v \in \mathbb{R}^N \ : \ \mathbb{A}_1 v \geq 0\}.$$

Exercice 2.1 *Déduire de la proposition précédente un* principe de positivité *pour le problème (2.5), homologue* discret *du principe général énoncé au théorème 1.8 (avec donnée nulle sur la frontière).*

2.2.4 Erreur pour les différences finies 1D

Comme $\mathbb{A}_1$ est monotone, elle est en particulier inversible : c'est cette propriété que nous allons utiliser maintenant, pour déterminer l'erreur commise. Soit e le vecteur de $\mathbb{R}^N$ dont les composantes sont égales à $e_i = u_i - u(x_i)$, pour i variant de 1 à N. Comme pour u, on adopte la convention $e_0 = e_{N+1} = 0$ (justifiée par le fait que $u(0) = u(1) = 0$, et $u_0 = u_{N+1} = 0\,!$). Sachant que u (resp. u) est solution de l'équation (2.1) (resp. (2.5)), on a alors, d'après (2.3), pour i compris entre 1 et N :

$$\begin{aligned}
(\mathbb{A}_1 e)_i &= \frac{-e_{i-1} + 2e_i - e_{i+1}}{h^2} = (\mathbb{A}_1 u)_i - \frac{-u(x_{i-1}) + 2u(x_i) - u(x_{i+1})}{h^2} \\
&= f(x_i) - \frac{-u(x_i - h) + 2u(x_i) - u(x_i + h)}{h^2} = f(x_i) + u''(x_i) + \frac{h^2}{12} u^{(4)}(x_i + \theta_i\, h) \\
&= \frac{h^2}{12} u^{(4)}(x_i + \theta_i\, h) = \frac{h^2}{12} f''(x_i + \theta_i\, h), \text{ avec } \theta_i \in]-1, 1[.
\end{aligned}$$

Dans l'esprit de la remarque 2.2, on introduit le vecteur ε de $\mathbb{R}^N$, dont les composantes sont égales à $\varepsilon_i = (\mathbb{A}_1 e)_i = \dfrac{h^2}{12} f''(x_i + \theta_i\, h)$, pour i variant de 1 à N, et dont la norme est telle que $\|\varepsilon\|_\infty \leq \dfrac{h^2}{12}\, C_{f,2}$ par construction. On en déduit alors l'expression de l'erreur commise

$$e = \mathbb{A}_1^{-1} \varepsilon. \tag{2.7}$$

Pour poursuivre, utilisons le fait que tous les éléments de $\mathbb{A}^{-1}$ sont positifs. En reprenant l'expression de l'erreur (2.7), on peut alors écrire

$$|e_i| = |\sum_j (\mathbb{A}_1^{-1})_{i,j} \varepsilon_j| \leq \sum_j (\mathbb{A}_1^{-1})_{i,j} |\varepsilon_j| \leq \sum_j (\mathbb{A}_1^{-1})_{i,j} \|\varepsilon\|_\infty \leq \frac{h^2}{12}\, C_{f,2} \sum_j (\mathbb{A}_1^{-1})_{i,j}. \tag{2.8}$$

Pour finalement arriver à une majoration de l'erreur commise, il suffit de majorer $\sum_j (\mathbb{A}_1^{-1})_{i,j}$ dans (2.8), pour $1 \leq i \leq N$.

Proposition 2.7 *La somme des éléments de chaque ligne de $\mathbb{A}_1^{-1}$ est inférieure ou égale à 1/8.*

Démonstration : On remarque que $\sum_j (\mathbb{A}_1^{-1})_{i,j} = \sum_j (\mathbb{A}_1^{-1})_{i,j}\delta_j$, sous réserve que $\delta_j = 1$, pour tout j.
A quoi correspond le vecteur δ ainsi construit ? On pose $u_0 = \mathbb{A}_1^{-1}\delta$, soit $\mathbb{A}_1 u_0 = \delta$.
δ joue le rôle d'un second membre de (2.5). Il correspond de fait à $f \equiv 1$, ce qui nous renvoie à la solution u_0 de (2.2). Or, dans ce cas *très particulier*,

$$-{u_0}''(x) = \frac{-u_0(x-h) + 2u_0(x) - u_0(x+h)}{h^2}, \ \forall x \in]0,1[, \ \forall h \text{ t. q. } [x-h, x+h] \subset [0,1].$$

Ainsi, u_0 tel que $(u_0)_i = u_0(x_i)$ vérifie

$$\mathbb{A}_1 u_0 = \delta, \text{ soit } u_0 = \mathbb{A}_1^{-1}\delta, \text{ ou } \sum_j (\mathbb{A}_1^{-1})_{i,j} = u_0(x_i), \ 1 \leq i \leq N.$$

Et $\sup_{x\in[0,1]} u_0(x) = u_0(1/2) = 1/8$, ce qui permet de conclure. ◇

On a donc démontré le

Théorème 2.8 *Lorsque la solution u est de classe $\mathcal{C}^4([0,1])$, l'erreur est telle que*

$$\|e\|_\infty \leq \frac{h^2}{96} \sup_{x\in[0,1]} |f''(x)|. \tag{2.9}$$

En conclusion, l'erreur "ponctuelle" tend **uniformément** vers 0 comme h^2.

NB. Lorsque h décroît, N croît en proportion inverse. Bref, l'erreur maximale décroît selon le carré de h, alors que le nombre d'inconnues croît en $1/h$...

Ceci étant, cette estimation dépend de la régularité de u (ou, ce qui revient au même dans le cas 1D, de celle de f). Que se passe-t-il si la solution u ou, ce qui est *équivalent*, la donnée f sont moins régulières ?

Théorème 2.9 *Lorsque la solution u est de classe $\mathcal{C}^2([0,1])$, l'erreur est telle que*

$$\lim_{h\to 0^+} \|e\|_\infty = 0. \tag{2.10}$$

Démonstration : On introduit à nouveau le vecteur ε, de composantes $\varepsilon_i = (\mathbb{A}_1 e)_i$. D'après les résultats portant sur la matrice $\mathbb{A}_1$ (propositions 2.6 et 2.7), il suffit de prouver que $\|\varepsilon\|_\infty \to 0$, lorsque h tend vers 0, c'est-à-dire :

$$\forall \eta > 0, \ \exists h_\eta > 0, \ 0 < h < h_\eta \ \Rightarrow \ \|\varepsilon\|_\infty < \eta.$$

(Bien évidemment, $\mathbb{A}_1$, e et ε dépendent de h, mais la dépendance est sous-entendue, notamment dans l'inégalité ci-dessus.)

Lorsque l'on sait simplement que f est continue, il n'est plus possible d'obtenir une expression des composantes $(\varepsilon_i)_i$ en fonction de la dérivée seconde f''... Mais tout n'est pas perdu ! Comme u est de classe $\mathcal{C}^2([0,1])$, on peut donc écrire, à l'aide de la formule de Taylor-Mac Laurin, les égalités

$$\begin{aligned} \exists \theta^- \in]-1,0[\quad \text{tel que} \quad u(x-h) &= u(x) - h\,u'(x) + \frac{h^2}{2} u''(x+\theta^- h) \\ \exists \theta^+ \in]0,1[\quad \text{tel que} \quad u(x+h) &= u(x) + h\,u'(x) + \frac{h^2}{2} u''(x+\theta^+ h). \end{aligned}$$

On en déduit que, pour i variant de 1 à N,

$$\varepsilon_i = f(x_i) + \frac{1}{2}\left(u''(x_i+\theta^- h) + u''(x_i+\theta^+ h)\right) = f(x_i) - \frac{1}{2}\left(f(x_i+\theta^- h) + f(x_i+\theta^+ h)\right).$$

Or, f étant continue sur le segment $[0,1]$, elle est uniformément continue. Ce que l'on peut exprimer mathématiquement sous la forme :

$$\forall \eta > 0,\ \exists h_\eta > 0,\ \forall x, y \in [0,1],\ |x-y| < h_\eta \ \Rightarrow\ |f(x) - f(y)| < \eta.$$

Or, si $h \in]0, h_\eta[$, on a $|x_i - (x_i + \theta^\pm h)| < h_\eta$, d'où $|\varepsilon_i| < \eta$, pour tout i : par voie de conséquence, $\|\varepsilon\|_\infty < \eta$. On en conclut finalement que, pour tout h dans $]0, h_\eta[$, on a l'inégalité

$$\|e\|_\infty < \frac{\eta}{8}.$$

$\diamond$

Sur cet exemple simple, on constate donc que la méthode des différences finies convergera *a priori* d'autant mieux que la solution du problème initial est régulière. Il est à noter, et c'est un point très important, que l'on retrouve effectivement ce type de comportement lorsque l'on réalise des expériences numériques...

Pour résoudre le système linéaire (2.5), nous allons établir une autre propriété concernant la matrice $\mathbb{A}_1$, qui nous permettra d'utiliser les algorithmes étudiés aux chapitres 5 et 6 :

— méthode de Cholesky (chapitre 5) ;
— méthodes de Jacobi ou de Gauss-Seidel, puisque $\mathbb{A}_1$ est tridiagonale (chapitre 6).

Définition 2.10 *Une matrice A de $\mathbb{R}^{N\times N}$ est dite* **définie-positive** *lorsque $(Av, v) > 0$, $\forall v \in \mathbb{R}^N \setminus \{0\}$.*

Notons tout de suite le résultat général ci-dessous.

Proposition 2.11 *Toute matrice définie-positive A de $\mathbb{R}^{N\times N}$ est inversible.*

Démonstration : En effet, $Av = 0 \Longrightarrow (Av, v) = 0 \Longrightarrow v = 0$. $\diamond$

Proposition 2.12 *La matrice $\mathbb{A}_1$ de (2.5) est symétrique définie-positive.*

Démonstration : Nous avons déjà remarqué que la matrice $\mathbb{A}_1$ est symétrique.
Il reste à vérifier qu'elle est définie-positive. On applique la définition, en formant le produit scalaire $h^2(\mathbb{A}_1 v, v)$, pour un vecteur v de $\mathbb{R}^N$ quelconque :

$$\begin{aligned} h^2(\mathbb{A}_1 v, v) &= h^2 \sum_{i=1}^{N} (\mathbb{A}_1 v)_i v_i \\ &= (2v_1 - v_2)v_1 + \sum_{i=2}^{N-1} (-v_{i-1} + 2v_i - v_{i+1})v_i + (-v_{N-1} + v_N)v_N \\ &= 2\sum_{i=1}^{N} {v_i}^2 - 2\sum_{i=1}^{N-1} v_i v_{i+1} = {v_1}^2 + {v_N}^2 + \sum_{i=1}^{N-1} \left({v_i}^2 - 2v_i v_{i+1} + {v_{i+1}}^2\right) \\ &= {v_1}^2 + {v_N}^2 + \sum_{i=1}^{N-1} (v_i - v_{i+1})^2 . \end{aligned}$$

Ainsi, $(\mathbb{A}_1 v, v) \geq 0$. De plus,$(\mathbb{A}_1 v, v) = 0$ entraîne que $v_1 = v_N = 0$, et $v_i = v_{i+1}$, pour $i = 1, \cdots, N-1$. On en déduit par récurrence que $v_i = 0$, pour $i = 2, \cdots, N-1$, et donc $v = 0$. ◇

2.2.5 Extension au Laplacien généralisé 1D

On s'intéresse ici à la généralisation de (2.1), à savoir

$$-(ku')'(x) + qu(x) = f(x) \text{ sur }]a,b[, \quad u(a) = u_a,\ u(b) = u_b. \tag{2.11}$$

Ci-dessus, k et q sont des coefficients qui peuvent dépendre de x. Nous expliquons ci-dessous les similitudes et différences avec le cas du Laplacien ($k = 1$, $q = 0$ et $u_a = u_b = 0$, avec $a = 0$, $b = 1$). Pour débuter, on construit le schéma d'approximation en deux étages, grâce à la proposition suivante qui repose encore une fois sur l'utilisation de la formule de Taylor-Mac Laurin.

Proposition 2.13 *On suppose que $v \in \mathcal{C}^3([0,1])$. Soient $x \in]0,1[$ et h' tel que $[x - h', x + h'] \in [0,1]$. Alors*

$$\exists \theta \in]-1,1[\text{ tel que } v'(x) = \frac{v(x+h') - v(x-h')}{2h'} + \frac{(h')^2}{3} v'''(x + \theta\, h'). \tag{2.12}$$

On applique ensuite deux fois ce résultat, à $v = ku'$, puis à $v = u$, pour en déduire un schéma aux différences finies pour (2.11). Choisissons $N \in \mathbb{N}$, et on fixe le pas de discrétisation $h = \dfrac{b-a}{N+1}$. On introduit les points $(x_i)_{0 \leq i \leq N+1}$ comme précédemment, ainsi que $x_{i+1/2} = x_i + \frac{1}{2}h$, pour $i \in \{0, 1, \cdots, N\}$.
A partir de là, on peut approcher (2.11) aux points $(x_i)_{0 \leq i \leq N+1}$. On a $u(x_0) = u_a$ et $u(x_{N+1}) = u_b$.
Ensuite, on choisit $v = ku'$, $x = x_i$ $(1 \leq i \leq N)$ et $h' = \frac{1}{2}h$ et on applique la proposition 2.13

$$\begin{aligned} (ku')'(x_i) &= \frac{(ku')(x_{i+1/2}) - (ku')(x_{i-1/2})}{h} + \frac{h^2}{12}(ku')'''(x_i + \theta\,\frac{h}{2}) \\ &= \frac{k_{i+1/2}u'(x_{i+1/2}) - k_{i-1/2}u'(x_{i-1/2})}{h} + O(h^2), \end{aligned}$$

où on a posé $k_{i+1/2} = k(x_{i+1/2})$ pour $i \in \{0, 1, \cdots, N\}$.
Si on choisit maintenant $v = u$, $x = x_{i+1/2}$ $(0 \leq i \leq N)$ et $h' = \frac{1}{2}h$, on trouve

$$u'(x_{i+1/2}) = \frac{u(x_{i+1}) - u(x_i)}{h} + O(h^2).$$

D'où finalement, pour $1 \leq i \leq N$,

$$-(ku')'(x_i) = \frac{-k_{i+1/2}u(x_{i+1}) + (k_{i+1/2} + k_{i-1/2})u(x_i) - k_{i-1/2}u(x_{i-1})}{h^2} + O(h).$$

Si on pose $q_i = q(x_i)$ et $f_i = f(x_i)$ pour $1 \leq i \leq N$, on est amené à considérer les équations

$$\begin{cases} \dfrac{-k_{i+1/2}u_{i+1} + (k_{i+1/2} + k_{i-1/2})u_i - k_{i-1/2}u_{i-1}}{h^2} + q_i u_i = f_i, \ 1 \leq i \leq N, \\ u_0 = u_a, \ u_{N+1} = u_b. \end{cases} \tag{2.13}$$

C'est le **schéma à trois points** pour le Laplacien généralisé 1D. En effet, du point de vue algorithmique, on retrouve exactement la structure de la figure 2.2.

On peut construire le système linéaire équivalent à (2.13), posé dans $\mathbb{R}^N$. Comme dans le cas du Laplacien, on connaît la valeur de u_0 et de u_N ; on appelle u (resp. f) le vecteur de $\mathbb{R}^N$ de composantes $(u_i)_{1\leq i\leq N}$ (resp. $(f_i)_{1\leq i\leq N}$). On introduit un troisième vecteur $g \in \mathbb{R}^N$ pour les conditions aux limites, de composantes : $g_1 = \frac{1}{h^2}k_{1/2}u_a$, $g_i = 0$ pour $2 \leq i \leq N-1$ et $g_N = \frac{1}{h^2}k_{N+1/2}u_b$. Le système (2.13) peut alors être écrit sous la *forme vectorielle équivalente*

$$\mathbb{A}_1' \, u = f + g, \text{ avec } \mathbb{A}_1' = \mathbb{K}_1 + \mathbb{Q}_1 \in \mathbb{R}^{N\times N} \text{ et} \tag{2.14}$$

$$\mathbb{K}_1 = \frac{1}{h^2}\begin{pmatrix} k_{1/2}+k_{3/2} & -k_{3/2} & 0 & \dots & 0 \\ -k_{3/2} & \ddots & \ddots & \ddots & \vdots \\ 0 & -k_{i-1/2} & k_{i-1/2}+k_{i+1/2} & -k_{i+1/2} & 0 \\ \vdots & \ddots & \ddots & \ddots & -k_{N-1/2} \\ 0 & \dots & 0 & -k_{N-1/2} & k_{N-1/2}+k_{N+1/2} \end{pmatrix},$$

$$\mathbb{Q}_1 = \begin{pmatrix} q_1 & 0 & 0 & \dots & 0 \\ 0 & q_2 & 0 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 0 & q_N \end{pmatrix}. \tag{2.15}$$

Par construction, les matrices $\mathbb{K}_1$ et $\mathbb{A}_1'$ sont tridiagonales et symétriques. Pour obtenir plus de propriétés, et notamment que $\mathbb{A}_1'$ est inversible, nous faisons des hypothèses classiques sur les coefficients du modèle, à savoir que ceux-ci sont de même signe constant :

$$\exists k_0 > 0, \ \forall x \in]a,b[, \ k(x) \geq k_0\,; \quad \forall x \in]a,b[, \ q(x) \geq 0. \tag{2.16}$$

Proposition 2.14 *Sous l'hypothèse (2.16), les matrices* $\mathbb{K}_1$ *et* $\mathbb{A}_1'$ *sont monotones.*

Démonstration : On reprend la proposition 2.5 pour la condition équivalente sur la monotonie. Soit v tel que $\mathbb{K}_1 v \geq 0$, et $v_k = \min_{1\leq i\leq N} v_i$. On a

$$\left\{\begin{array}{l} (k_{1/2}+k_{3/2})v_1 - k_{3/2}v_2 \geq 0 \\ -k_{i-1/2}v_{i-1} + (k_{i-1/2}+k_{i+1/2})v_i - k_{i+1/2}v_{i+1} \geq 0, \quad 2\leq i\leq N-1 \\ -k_{N-1/2}v_{N-1} + (k_{N-1/2}+k_{N+1/2})v_N \geq 0 \end{array}\right. .$$

Par hypothèse, on a $k_{i+1/2} = k(x_{i+1/2}) > 0$ pour tout $i \in \{0,1,\cdots,N\}$.
Si $v_k = v_1$, on trouve, puisque $k_{1/2}, k_{3/2} > 0$,

$$(k_{1/2}+k_{3/2})v_k \geq k_{3/2}v_2 \geq k_{3/2}v_k \;\Rightarrow\; k_{1/2}v_k \geq 0 \;\Rightarrow\; v_k \geq 0.$$

De même si $v_k = v_N$ $(k_{N-1/2}, k_{N+1/2} > 0)$.

Si $k \in \{2,\cdots,N-1\}$, on trouve cette fois

$$k_{k-1/2}(v_k - v_{k-1}) + k_{k+1/2}(v_k - v_{k+1}) \geq 0.$$

Or, $v_k \leq v_{k-1}$ et $v_k \leq v_{k+1}$. On a donc $k_{k-1/2}(v_k - v_{k-1}) + k_{k+1/2}(v_k - v_{k+1}) \leq 0$ puisque $k_{k-1/2}, k_{k+1/2} > 0$. On en déduit que $(v_k - v_{k-1}) = (v_k - v_{k+1}) = 0$, c'est-à-dire $v_{k-1} = v_k = v_{k+1}$. Par récurrence, on arrive à $v_1 = \cdots = v_{k-1} = v_k = v_{k+1} = \cdots = v_N$, et la première (ou la dernière) équation donne à nouveau $v_k \geq 0$.
Pour la matrice $\mathbb{A}_1'$, la démonstration est élémentaire car on ajoute des termes strictement positifs sur la diagonale. Si $k \in \{2,\cdots,N-1\}$, on trouve directement

$$q_k v_k \geq k_{k-1/2}(v_{k-1} - v_k) + k_{k+1/2}(v_{k+1} - v_k) \geq 0.$$

Idem si $k = 1$ ou $k = N$. ◇

Proposition 2.15 *Sous l'hypothèse (2.16), les matrices* $\mathbb{K}_1$ *et* $\mathbb{A}_1'$ *sont définies-positives.*

Démonstration : On applique la définition, en formant le produit scalaire $h^2(\mathbb{K}_1 v, v)$, pour un vecteur v de $\mathbb{R}^N$ quelconque :

$$\begin{aligned} h^2(\mathbb{K}_1 v, v) &= h^2 \sum_{i=1}^{N} (\mathbb{K}_1 v)_i v_i \\ &= ((k_{1/2}+k_{3/2})v_1 - k_{3/2}v_2)v_1 \\ &\quad + \sum_{i=2}^{N-1} (-k_{i-1/2}v_{i-1} + (k_{i-1/2}+k_{i+1/2})v_i - k_{i+1/2}v_{i+1})v_i \\ &\quad + (-k_{N-1/2}v_{N-1} + (k_{N-1/2}+k_{N+1/2})v_N)v_N \\ &= k_{1/2}{v_1}^2 + \sum_{i=2}^{N} k_{i-1/2}({v_{i-1}}^2 - 2v_{i-1}v_i + {v_i}^2) + k_{N+1/2}{v_N}^2 \\ &= k_{1/2}{v_1}^2 + k_{N+1/2}{v_N}^2 + \sum_{i=2}^{N} k_{i-1/2}(v_{i-1} - v_i)^2. \end{aligned}$$

Puisque $k_{i-1/2} > 0$ pour tout $i \in \{1, 2, \cdots, N+1\}$, on a $(\mathbb{K}_1 v, v) \geq 0$. De plus,$(\mathbb{K}_1 v, v) = 0$ entraîne que $v_1 = v_N = 0$, et $v_i = v_{i+1}$, pour $i = 1, \cdots, N-1$. Par récurrence on a $v_i = 0$, pour $i = 2, \cdots, N-1$, et donc $v = 0$
La matrice $\mathbb{A}'_1$ est définie-positive puisque $\mathbb{K}_1$ et $\mathbb{Q}_1$ le sont. $\diamond$

On en déduit le résultat suivant, qui découle de l'une ou l'autre des propositions 2.14 ou 2.15.

Corollaire 2.16 *Sous l'hypothèse (2.16), la matrice $\mathbb{A}'_1$ est inversible.*

Exercice 2.2 *Déduire des résultats précédents un principe de positivité pour l'approximation (2.14) du Laplacien généralisé 1D.*

2.3 Problème bidimensionnels

Dans cette section, on généralise la méthode des différences finies en 2D. On commence par le calcul des (petits) déplacements de la membrane élastique, soumise à des forces verticales, puis on examine le cas de la diffusion neutronique.
Dans $\mathbb{R}^2$, on considère une membrane carrée, $D =]0,1[^2$, soumise à une force verticale, de densité surfacique $f(x,y)$, et fixée en sa **frontière** ∂D. On note $u : (x,y) \mapsto u(x,y)$ le déplacement transversal induit par l'application de la force, dont on rappelle qu'il est solution du problème bidimensionnel (2D) *normalisé*

$$-\Delta_2 u = f \text{ sur } D, \quad u = 0 \text{ sur } \partial D. \tag{2.17}$$

Pour discrétiser le problème à l'aide de la méthode des différences finies, on s'inspire très fortement de la méthode employée pour le problème 1D (2.1). En effet, on remarque qu'en 1D on a les égalités $-u'' = -\dfrac{d^2}{dx^2}u = -\Delta_1 u$, i.e. (2.1) est un Laplacien 1D à résoudre !

Si la solution u est de classe $\mathcal{C}^4([0,1]^2)$, on peut écrire l'équivalent de la proposition 2.1.

NB. Malheureusement, et contrairement au cas 1D, on n'a plus l'équivalence entre u de classe $\mathcal{C}^4([0,1]^2)$ et f de classe $\mathcal{C}^2([0,1]^2)$...

Proposition 2.17 *Soient $(x,y) \in]0,1[^2$ et (h_1,h_2) tels que*
$[x-h_1, x+h_1] \in [0,1]$, *et* $[y-h_2, y+h_2] \in [0,1]$. *Alors*

$$\begin{aligned}
\exists(\theta_1,\theta_2) \quad \in]-1,1[^2 \text{ tels que} \\
-\frac{\partial^2 u}{\partial x^2}(x,y) &= \frac{-u(x-h_1,y) + 2u(x,y) - u(x+h_1,y)}{{h_1}^2} + \frac{{h_1}^2}{12}\frac{\partial^4 u}{\partial x^4}(x+\theta_1 h_1, y) \\
-\frac{\partial^2 u}{\partial y^2}(x,y) &= \frac{-u(x,y-h_2) + 2u(x,y) - u(x,y+h_2)}{{h_2}^2} + \frac{{h_2}^2}{12}\frac{\partial^4 u}{\partial y^4}(x, y+\theta_2 h_2).
\end{aligned}$$

On en déduit que

$$\begin{aligned} -\Delta_2 u(x,y) &= \frac{-u(x-h_1,y)+2u(x,y)-u(x+h_1,y)}{{h_1}^2} \\ &\quad + \frac{-u(x,y-h_2)+2u(x,y)-u(x,y+h_2)}{{h_2}^2} \\ &\quad + \frac{{h_1}^2}{12}\frac{\partial^4 u}{\partial x^4}(x+\theta_1 h_1, y) + \frac{{h_2}^2}{12}\frac{\partial^4 u}{\partial y^4}(x, y+\theta_2 h_2). \end{aligned} \tag{2.18}$$

Remarque 2.18 *Les deux premiers termes de (2.18) sont une bonne approximation de* $-\Delta_2 u(x,y)$*, sous réserve que* h_1 *et* h_2 *soient* petits*. Le reste est en effet borné par*

$$\frac{1}{12}\left({h_1}^2+{h_2}^2\right) C_{u,4}, \text{ avec } C_{u,4} = \max\left(\sup_{(x,y)\in[0,1]^2} |\frac{\partial^4 u}{\partial x^4}(x,y)|, \sup_{(x,y)\in[0,1]^2} |\frac{\partial^4 u}{\partial y^4}(x,y)| \right).$$

Dans la suite, on va considérer un pas de discrétisation *identique*[8] selon la direction des x, et celle des y, c'est-à-dire $h = h_1 = h_2$. Chacun des intervalles $[0,1]$ est découpé en $n+1$ intervalles égaux de longueur $h = 1/(n+1)$. Puis on calcule les nombres $(u_{i,j})_{0\le i,j\le n+1}$, qui sont les *valeurs approchées* de la solution u aux points d'abscisse $x_i = ih$ et d'ordonnée $y_j = jh$, $0 \le i,j \le n+1$. On note $(f_{i,j})_{1\le i,j\le n}$ les valeurs $f_{i,j} = f(x_i, y_j)$, pour i et j variant de 1 à n.

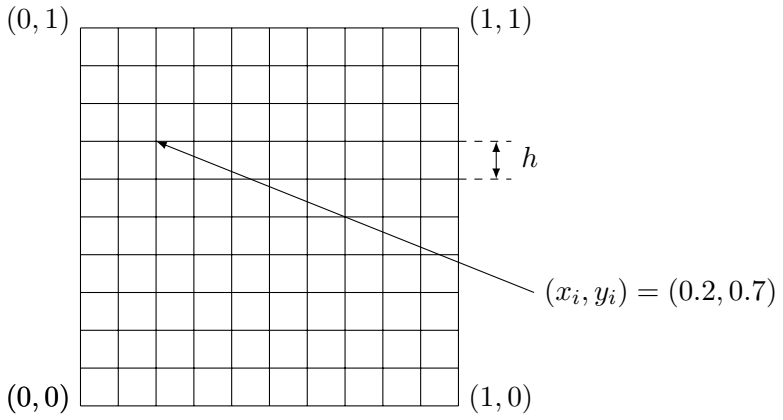


FIGURE 2.3 – D et les points de discrétisation ($n = 9$, $h = 0.1$)

Le Laplacien 2D est *approché* par une combinaison linéaire de valeurs $u_{i,j}$, selon le **schéma à cinq points**

$$-\Delta_2 u(x_i,y_j) \approx \frac{-u_{i,j-1}-u_{i-1,j}+4u_{i,j}-u_{i+1,j}-u_{i,j+1}}{h^2}, \quad 1\le i,j\le n. \tag{2.19}$$

8. En pratique, il est tout à fait possible de considérer des pas différents selon les directions x et y, voire variables dans chaque direction...

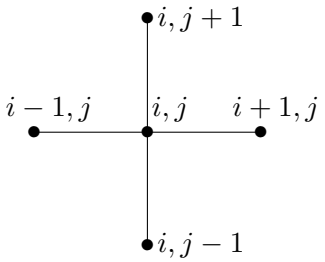


FIGURE 2.4 – Le schéma aux différences finies à cinq points

Le problème (2.17) est donc approché de la manière suivante : on remplace la recherche de la fonction u, par la recherche des n^2 valeurs $u_{i,j} \in \mathbb{R}$ qui vérifient les relations

$$\frac{-u_{i,j-1} - u_{i-1,j} + 4u_{i,j} - u_{i+1,j} - u_{i,j+1}}{h^2} = f_{i,j}, \quad 1 \leq i, j \leq n. \tag{2.20}$$

Les valeurs de u sur la frontière ∂D, ici $u(0, \cdot)$ et $u(1, \cdot)$, ainsi que $u(\cdot, 0)$ et $u(\cdot, 1)$, sont connues (et égales à zéro). Il en est donc de même pour $u_{0,j}$, $u_{n+1,j}$, $u_{i,0}$ et $u_{i,n+1}$, pour i et j variant de 1 à n. Il reste donc au total $N = n^2$ valeurs à calculer.

On les regroupe n par n, ainsi que les $(f_{i,j})_{i,j}$, en opérant l'*identification* $u_{\cdot,j} = (u_{i,j})_{1 \leq i \leq n}$. Le bloc $u_{\cdot,j}$ appartient à $\mathbb{R}^n$, avec

$$u_{\cdot,j} = \begin{pmatrix} u_{1,j} \\ \vdots \\ u_{n,j} \end{pmatrix}.$$

Il en est de même pour $f_{\cdot,j} \in \mathbb{R}^n$. Ensuite, on pose

$$u = \begin{pmatrix} u_{\cdot,1} \\ \vdots \\ u_{\cdot,n} \end{pmatrix} \in \mathbb{R}^N \text{ et } f = \begin{pmatrix} f_{\cdot,1} \\ \vdots \\ f_{\cdot,n} \end{pmatrix} \in \mathbb{R}^N.$$

Ainsi, on a *numéroté* les inconnues ligne par ligne, dans le sens croissant, pour les indices i (au sein d'une ligne) et j (numéro de ligne). Les inconnues $(u_{i,j})_{1 \leq i,j \leq n}$ sont solutions du système linéaire formé des N relations (2.20). Ce système linéaire peut être écrit sous la forme

$$\mathbb{A}_2 u = f, \tag{2.21}$$

avec $\mathbb{A}_2$ une matrice de $\mathbb{R}^{N \times N}$. Si l'on s'intéresse à sa structure interne, on vérifie facilement que l'on peut écrire

$$\mathbb{A}_2 = \frac{1}{h^2} \begin{pmatrix} \mathbb{B}_1 & T & & & \\ T & \mathbb{B}_1 & T & & \\ & \ddots & \ddots & \ddots & \\ & & T & \mathbb{B}_1 & T \\ & & & T & \mathbb{B}_1 \end{pmatrix}. \tag{2.22}$$

Ci-dessus, les blocs autres que $\mathbb{B}_1$ et T sont nuls et, par ailleurs, $T = -I_n$, avec I_n la matrice identité d'ordre n, et $\mathbb{B}_1 \in \mathbb{R}^{n\times n}$ est la matrice tridiagonale définie par

$$\mathbb{B}_1 = \begin{pmatrix} 4 & -1 & & & \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 4 & -1 \\ & & & -1 & 4 \end{pmatrix} = 2I_n + h^2\mathbb{A}_1.$$

La matrice $\mathbb{A}_2$ est donc pentadiagonale par points (i. e. avec tous les éléments non nuls regroupés sur cinq diagonales), et tridiagonale par blocs, *lorsque* la numérotation est celle indiquée ci-dessus : ligne par ligne (j croissant), et i croissant au sein d'une ligne.

Récapitulons. Si on note avec un seul indice les composantes de u, c'est-à-dire $(u_I)_{1\le I\le N}$, on a les correspondances :

$$\text{composante } _I \equiv \ _{i,j} \quad \Longleftrightarrow \quad I = i + (j-1)n \text{ ou } \begin{cases} i = (I-1) \mod n + 1 \\ j = \lfloor (I-1)/n \rfloor + 1 \end{cases} . \tag{2.23}$$

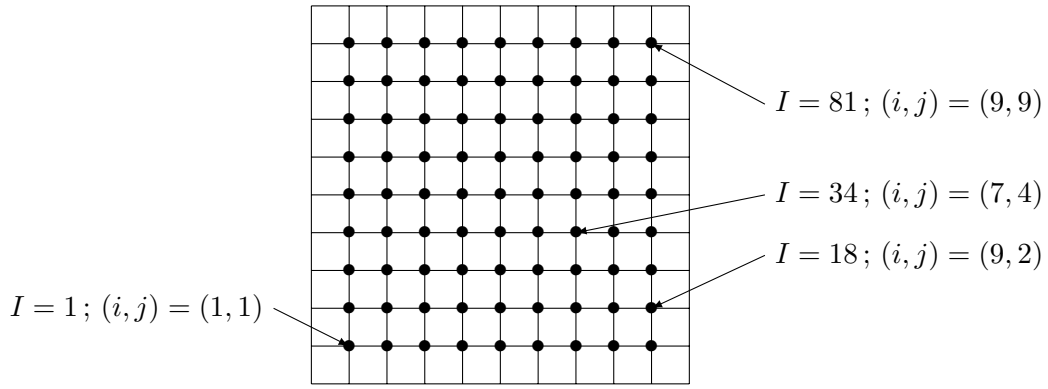


FIGURE 2.5 – Les deux numérotations : $I \in \{1, \cdots, 81\}$ et $i, j \in \{1, \cdots, 9\}$

Nous allons maintenant étudier les propriétés de la matrice $\mathbb{A}_2$. Nous allons d'abord établir qu'elle est monotone, puis qu'elle est symétrique définie-positive, comme $\mathbb{A}_1$.

Théorème 2.19 *La matrice $\mathbb{A}_2$ des systèmes linéaires (2.21) et (2.22) est monotone.*

Démonstration : Soit donc $v \in \mathbb{R}^N$ tel que $\mathbb{A}_2 v \ge 0$. Composante par composante (avec

le double indiçage $_{i,j}$), ceci signifie

$$4v_{1,1} - v_{2,1} - v_{1,2} \geq 0 \tag{2.24}$$
$$4v_{n,1} - v_{n-1,1} - v_{n,2} \geq 0 \tag{2.25}$$
$$4v_{1,n} - v_{1,n-1} - v_{2,n} \geq 0 \tag{2.26}$$
$$4v_{n,n} - v_{n,n-1} - v_{n-1,n} \geq 0 \tag{2.27}$$
$$4v_{i,1} - v_{i-1,1} - v_{i+1,1} - v_{i,2} \geq 0 \qquad 2 \leq i \leq n-1 \tag{2.28}$$
$$4v_{1,j} - v_{1,j-1} - v_{2,j} - v_{1,j+1} \geq 0 \qquad 2 \leq j \leq n-1 \tag{2.29}$$
$$4v_{n,j} - v_{n,j-1} - v_{n-1,j} - v_{n,j+1} \geq 0 \qquad 2 \leq j \leq n-1 \tag{2.30}$$
$$4v_{i,n} - v_{i,n-1} - v_{i-1,n} - v_{i+1,n} \geq 0 \qquad 2 \leq i \leq n-1 \tag{2.31}$$
$$4v_{i,j} - v_{i,j-1} - v_{i-1,j} - v_{i+1,j} - v_{i,j+1} \geq 0 \qquad 2 \leq i,j \leq n-1. \tag{2.32}$$

Ci-dessus, on a isolé des lignes de la matrice $\mathbb{A}_2$ correspondant respectivement

1. En (2.24)-(2.27) :
   aux *coins* de la grille, d'indices (i,j) parmi $\{(1,1),(n,1),(1,n),(n,n)\}$.
2. En (2.28)-(2.31) :
   au *bord* de la grille, coins exclus, d'indices (i,j) avec $i \in \{1,n\}$ ou (exclusif) $j \in \{1,n\}$.
3. En (2.32) :
   aux *points internes* de la grille, d'indices (i,j) avec $i,j \in \{2,\cdots,n-1\}$.

Soit maintenant $v_{min} = \min_{i,j} v_{i,j}$. On veut prouver que $v_{min} \geq 0$, pour pouvoir conclure grâce à la proposition 2.5. Comme la grille comprend des points aux coins, sur le bord (coins exclus) et intérieurs, on traite les trois cas correspondants.

1. Si v_{min} correspond à un coin (par exemple $v_{min} = v_{1,1}$), (2.24) fournit l'inégalité

   $$2v_{min} \geq (v_{2,1} - v_{min}) + (v_{1,2} - v_{min}) \geq 0.$$

2. Si v_{min} correspond à un point du bord différent d'un coin (par exemple $v_{min} = v_{i,1}$, $i \in \{2,\cdots,n-1\}$ donné), (2.28) fournit l'inégalité

   $$v_{min} \geq (v_{i-1,1} - v_{min}) + (v_{i+1,1} - v_{min}) + (v_{i,2} - v_{min}) \geq 0.$$

3. Si v_{min} correspond à un point intérieur, (2.32) fournit l'inégalité

   $$(v_{min} - v_{i,j-1}) + (v_{min} - v_{i-1,j}) + (v_{min} - v_{i+1,j}) + (v_{min} - v_{i,j+1}) \geq 0.$$

   Pour conclure dans le cas 3., il faut se ramener aux cas 1. ou 2. Or, d'après la définition de v_{min}, chacun des quatre termes entre parenthèses est négatif ou nul. Il sont donc tous nuls, c'est-à-dire

   $$v_{min} = v_{i,j-1} = v_{i-1,j} = v_{i+1,j} = v_{i,j+1}.$$

   Comme dans le cas 1D (cf. la preuve de la proposition 2.6), on peut raisonner par récurrence (sur i et sur j), pour trouver finalement que $v_{i,j} = v_{min}$ en tous les points, coins *exceptés*. Mais ceci est suffisant pour conclure, car les points du bord (hors coins) sont compris, et le cas 2. s'applique.

$\diamond$

Pour établir le caractère défini-positif de $\mathbb{A}_2$, nous allons reprendre l'écriture par blocs $n \times n$ de celle-ci (2.22).

Proposition 2.20 *La matrice $\mathbb{A}_2$ du système linéaire (2.21) est symétrique définie-positive.*

Démonstration : Pour commencer, la matrice $\mathbb{A}_2$ est symétrique par construction.
On forme, pour $v \in \mathbb{R}^N$, $h^2(\mathbb{A}_2 v, v)$:

$$\begin{aligned}
h^2(\mathbb{A}_2 v, v) &= \left(\begin{pmatrix} \mathbb{B}_1 & -I_n & & & \\ -I_n & \mathbb{B}_1 & -I_n & & \\ & \ddots & \ddots & \ddots & \\ & & -I_n & \mathbb{B}_1 & -I_n \\ & & & -I_n & \mathbb{B}_1 \end{pmatrix} \begin{pmatrix} v_{\cdot,1} \\ \vdots \\ v_{\cdot,n} \end{pmatrix}, \begin{pmatrix} v_{\cdot,1} \\ \vdots \\ v_{\cdot,n} \end{pmatrix} \right) \\
&= (\mathbb{B}_1 v_{\cdot,1} - v_{\cdot,2}, v_{\cdot,1})_n + \sum_{2 \le j \le n-1} (-v_{\cdot,j-1} + \mathbb{B}_1 v_{\cdot,j} - v_{\cdot,j+1}, v_{\cdot,j})_n \\
&\quad + (-v_{\cdot,n-1} + \mathbb{B}_1 v_{\cdot,n}, v_{\cdot,n})_n \\
&= \sum_{1 \le j \le n} (\mathbb{B}_1 v_{\cdot,j}, v_{\cdot,j})_n - 2 \sum_{1 \le j \le n-1} (v_{\cdot,j}, v_{\cdot,j+1})_n \\
&= \sum_{1 \le j \le n} \left(2\|v_{\cdot,j}\|_n^2 + h^2 (\mathbb{A}_1 v_{\cdot,j}, v_{\cdot,j})_n \right) - 2 \sum_{1 \le j \le n-1} (v_{\cdot,j}, v_{\cdot,j+1})_n \\
&= \|v_{\cdot,1}\|_n^2 + \|v_{\cdot,n}\|_n^2 + h^2 \sum_{1 \le j \le n} (\mathbb{A}_1 v_{\cdot,j}, v_{\cdot,j})_n + \sum_{1 \le j \le n-1} \|v_{\cdot,j} - v_{\cdot,j+1}\|_n^2 .
\end{aligned}$$

Ci-dessus, on a noté $(\cdot,\cdot)_n$ le produit scalaire usuel de $\mathbb{R}^n$, et $\|\cdot\|_n$ la norme associée. Comme $\mathbb{A}_1$ est définie-positive, on en déduit immédiatement que $h^2(\mathbb{A}_2 v, v)$ est positif ou nul, puisque c'est une somme de termes positifs. Si on suppose que $h^2(\mathbb{A}_2 v, v) = 0$, on trouve $\sum_{1 \le j \le n} (\mathbb{A}_1 v_{\cdot,j}, v_{\cdot,j})_n = 0$, ce qui implique $v_{\cdot,j} = 0$ pour $1 \le j \le n$, soit finalement $v = 0$. $\diamond$

Malheureusement, on ne peut pas aller plus loin, c'est-à-dire estimer la précision de l'approximation obtenue grâce au schéma à cinq points, en continuant à suivre la méthode pour le cas 1D([9]). Cependant, à l'aide d'autres techniques relativement lourdes à mettre en œuvre (cf. [13]), on peut malgré tout prouver le

Théorème 2.21 *Lorsque la solution u est de classe $\mathcal{C}^4([0,1]^2)$, l'erreur est telle que*

$$\|e\|_\infty \le C\, h^2 (C_{u,4} + h\, C_{u,3}), \ \textit{avec}\ C_{u,3} = \max \left(\sup_{(x,y) \in [0,1]^2} |\frac{\partial^3 u}{\partial x^3}(x,y)|, \sup_{(x,y) \in [0,1]^2} |\frac{\partial^3 u}{\partial y^3}(x,y)| \right),$$

où C est une constante qui est indépendante de u et de h.

Bref, *"tout est bien qui finit bien"* (librement traduit de [35]...).

9. En effet, si $f \equiv 1$, quelle est la solution du problème (2.17), quelle est sa régularité ?

Notons au passage que comme $\mathbb{A}_2$ est symétrique définie-positive, on peut utiliser, pour résoudre (2.21), la méthode de Cholesky (chapitre 5), ou bien les méthodes de Jacobi ou de Gauss-Seidel, puisque $\mathbb{A}_2$ est tridiagonale par blocs $n \times n$ (chapitre 6).

Pour conclure cette section, considérons un problème 2D plus général, à savoir la diffusion neutronique (ou Laplacien généralisé), posé dans le domaine $D' =]a_1, b_1[\times]a_2, b_2[$. On doit alors calculer la solution de

$$-\operatorname{div}(k\,\mathbf{grad}\,u) + q\,u = f \text{ sur } D', \quad u = g \text{ sur } \partial D'. \tag{2.33}$$

Ci-dessus, k et q sont des coefficients qui peuvent dépendre de (x, y). Notons que

$$-\operatorname{div}(k\,\mathbf{grad}\,u) = -\frac{\partial}{\partial x}\left(k\,\frac{\partial u}{\partial x}\right) - \frac{\partial}{\partial y}\left(k\,\frac{\partial u}{\partial y}\right). \tag{2.34}$$

On peut donc utiliser une approximation du type de celle de la proposition 2.13 en x, puis en y. Pour cela, on introduit

$$x_i = a_1 + ih_1,\ 0 \le i \le n_1, \quad y_j = a_2 + jh_2,\ 0 \le j \le n_2,$$

avec $h_d = (b_d - a_d)/(n_d + 1)$, $d = 1, 2$ les pas de discrétisation. Dans la suite, on introduit les nombres $(u_{i,j})_{0\le i\le n_1+1, 0\le j\le n_2+1}$, qui sont les *valeurs approchées* de la solution u aux points $(x_i, y_j)_{0\le i\le n_1+1, 0\le j\le n_2+1}$, et on note $(f_{i,j})_{1\le i\le n_1, 1\le j\le n_2}$ les valeurs $f_{i,j} = f(x_i, y_j)$, resp. $(q_{i,j})_{1\le i\le n_1, 1\le j\le n_2}$ les valeurs $q_{i,j} = q(x_i, y_j)$, pour les points (x_i, y_j) intérieurs à D'. On utilise une approximation 1D pour chacun des deux termes de (2.34). Sous réserve de régularité suffisante, on trouve tout d'abord, pour y_j fixé,

$$\begin{aligned} -\frac{\partial}{\partial x}\left(k\,\frac{\partial u}{\partial x}\right)(x_i, y_j) \;\approx\; \frac{1}{h_1^2}\Big(& -k(x_{i-1/2}, y_j)u(x_{i-1}, y_j) \\ & +(k(x_{i-1/2}, y_j) + k(x_{i+1/2}, y_j))u(x_i, y_j) \\ & -k(x_{i+1/2}, y_j)u(x_{i+1}, y_j)\Big)\,; \end{aligned}$$

et, pour x_i fixé

$$\begin{aligned} -\frac{\partial}{\partial y}\left(k\,\frac{\partial u}{\partial y}\right)(x_i, y_j) \;\approx\; \frac{1}{h_2^2}\Big(& -k(x_i, y_{j-1/2})u(x_i, y_{j-1}) \\ & +(k(x_i, y_{j-1/2}) + k(x_i, y_{j+1/2}))u(x_i, y_j) \\ & -k(x_i, y_{j+1/2})u(x_i, y_{j+1})\Big). \end{aligned}$$

Ci-dessus on a noté $x_{i+1/2} = x_i + \frac{1}{2}h_1$, $0 \le i \le n_1$ et $y_{j+1/2} = y_j + \frac{1}{2}h_2$, $0 \le j \le n_2$. On introduit finalement

$$\begin{cases} k_{i+1/2,j} = k(x_{i+1/2}, y_j),\ 0 \le i \le n_1, 1 \le j \le n_2, \\ k_{i,j+1/2} = k(x_i, y_{j+1/2}),\ 1 \le i \le n_1, 0 \le j \le n_2. \end{cases}$$

Comme précédemment, on approche le problème (2.33) en deux temps. Tout d'abord aux points (intérieurs) de D' :

$$\begin{aligned}&\frac{1}{h_1^2}\Big(-k_{i-1/2,j}u_{i-1,j}+(k_{i-1/2,j}+k_{i+1/2,j})u_{i,j}-k_{i+1/2,j}u_{i+1,j}\Big)\\+&\frac{1}{h_2^2}\Big(-k_{i,j-1/2}u_{i,j-1}+(k_{i,j-1/2}+k_{i,j+1/2})u_{i,j}-k_{i,j+1/2}u_{i,j+1}\Big)\\&+q_{i,j}u_{i,j}=f_{i,j},\quad 1\le i\le n_1, 1\le j\le n_2.\end{aligned}\tag{2.35}$$

Puis aux points sur $\partial D'$:

$$u_{i,j}=g(x_i,y_j),\quad (i,j)\in\{0,n_1+1\}\times\{0,\cdots,n_2+1\}\cup\{0,\cdots,n_1+1\}\times\{0,n_2+1\}.\tag{2.36}$$

Du point de vue algorithmique, on retrouve la structure de la figure 2.4, c'est donc bien un **schéma à cinq points**. Les valeurs sur la frontière étant connues (2.36), il reste à calculer les $N=n_1n_2$ valeurs $(u_{i,j})_{1\le i\le n_1,1\le j\le n_2}$. On peut exprimer (2.35)-(2.36) sous la forme d'un système linéaire équivalent dont les inconnues sont ces N valeurs. On choisit de regrouper les inconnues $(u_{i,j})_{1\le i\le n_1,1\le j\le n_2}$ par blocs de n_1 valeurs [10], en opérant l'*identification* $u_{\cdot,j}=(u_{i,j})_{1\le i\le n_1}$. Chaque bloc $u_{\cdot,j}$ appartient à $\mathbb{R}^{n_1}$:

$$u_{\cdot,j}=\begin{pmatrix}u_{1,j}\\ \vdots\\ u_{n_1,j}\end{pmatrix}\in\mathbb{R}^{n_1},\quad 1\le j\le n_2.$$

On a donc n_2 blocs. On procède de même pour les $(f_{i,j})_{i,j}$, en introduisant $f_{\cdot,j}\in\mathbb{R}^{n_1}$, $1\le j\le n_2$. Ensuite, on pose

$$u=\begin{pmatrix}u_{\cdot,1}\\ \vdots\\ u_{\cdot,n_2}\end{pmatrix}\in\mathbb{R}^N \text{ et } f=\begin{pmatrix}f_{\cdot,1}\\ \vdots\\ f_{\cdot,n_2}\end{pmatrix}\in\mathbb{R}^N.$$

Il reste à définir le vecteur $g\in\mathbb{R}^N$ regroupant les valeurs sur la frontière $\partial D'$. On introduit $(g_{i,j})_{1\le i\le n_1,1\le j\le n_2}$ avec

$$\begin{array}{lll}g_{1,j}=\frac{1}{h_1^2}k_{1/2,j}\,g(a_1,y_j), & g_{n_1,j}=\frac{1}{h_1^2}k_{n_1+1/2,j}\,g(b_1,y_j), & 1\le j\le n_2,\\ g_{i,1}=\frac{1}{h_2^2}k_{i,1/2}\,g(x_i,a_2), & g_{i,n_2}=\frac{1}{h_2^2}k_{i,n_2+1/2}\,g(x_i,b_2), & 1\le i\le n_1,\\ g_{i,j}=0 & \text{sinon.} & \end{array}$$

On regroupe les inconnues $(g_{i,j})_{1\le i\le n_1,1\le j\le n_2}$ par blocs de n_1 valeurs, comme pour $(u_{i,j})$ et $(f_{i,j})$, c'est-à-dire qu'on introduit $g_{\cdot,j}\in\mathbb{R}^{n_1}$ pour $1\le j\le n_2$, puis

$$g=\begin{pmatrix}g_{\cdot,1}\\ \vdots\\ g_{\cdot,n_2}\end{pmatrix}\in\mathbb{R}^N.$$

10. On pourrait choisir de les regrouper par blocs de n_2 valeurs :

$$u_{i,\cdot}=\begin{pmatrix}u_{i,1}\\ \vdots\\ u_{i,n_2}\end{pmatrix}\in\mathbb{R}^{n_2},\quad 1\le i\le n_1,\ \text{etc.}$$

Le système (2.35)-(2.36) peut finalement être écrit sous la *forme vectorielle équivalente*

$$\mathbb{A}_2' u = f + g, \text{ avec } \mathbb{A}_2' = \mathbb{K}_2 + \mathbb{Q}_2 \in \mathbb{R}^{N\times N}. \tag{2.37}$$

Il reste à préciser la structure des matrices $\mathbb{K}_2$ et $\mathbb{Q}_2$. La matrice $\mathbb{Q}_2$ est diagonale, et en particulier on peut l'écrire par blocs (eux-mêmes diagonaux par points) selon

$$\mathbb{Q}_2 = \begin{pmatrix} \mathbb{Q}_{1,1} & & & & \\ & \ddots & & & \\ & & \mathbb{Q}_{j,j} & & \\ & & & \ddots & \\ & & & & \mathbb{Q}_{n_2,n_2} \end{pmatrix}, \quad \mathbb{Q}_{j,j} \in \mathbb{R}^{n_1\times n_1},\ 1 \le j \le n_2. \tag{2.38}$$

Pour $1 \le j \le n_2$, $\mathbb{Q}_{j,j}$ agit de $u_{\cdot,j}$ vers $(f+g)_{\cdot,j}$, et on a

$$\mathbb{Q}_{j,j} = \begin{pmatrix} q_{1,j} & & & & \\ & \ddots & & & \\ & & q_{i,j} & & \\ & & & \ddots & \\ & & & & q_{n_1,j} \end{pmatrix}.$$

Si l'on s'intéresse à la structure interne de $\mathbb{K}_2$, on vérifie facilement que l'on peut écrire

$$\mathbb{K}_2 = \begin{pmatrix} \mathbb{K}_{1,1} & \mathbb{K}_{1,2} & & & & \\ \ddots & \ddots & \ddots & & & \\ & \mathbb{K}_{j,j-1} & \mathbb{K}_{j,j} & \mathbb{K}_{j,j+1} & & \\ & & \ddots & \ddots & \ddots & \\ & & & \mathbb{K}_{n_2,n_2-1} & \mathbb{K}_{n_2,n_2} \end{pmatrix}, \quad \mathbb{K}_{j,j'} \in \mathbb{R}^{n_1\times n_1},\ 1 \le j, j' \le n_2. \tag{2.39}$$

En effet, si on examine (2.35), le bloc $u_{\cdot,j}$ est lié *uniquement* aux blocs $u_{\cdot,j-1}$ (si $j \ge 2$) et $u_{\cdot,j+1}$ (si $j \le n_2 - 1$). En d'autres termes, $\mathbb{K}_2$ est bien tridiagonale par blocs.
Il reste à détailler la structure des blocs $\mathbb{K}_{j,j'}$. Pour $1 \le j \le n_2$, un bloc diagonal $\mathbb{K}_{j,j} \in \mathbb{R}^{n_1\times n_1}$ agit de $u_{\cdot,j}$ vers $(f+g)_{\cdot,j}$, et par inspection on trouve que c'est une matrice tridiagonale et symétrique par points, avec :

$$\begin{aligned} (\mathbb{K}_{j,j})_{i,i-1} &= -\tfrac{1}{h_1^2}\, k_{i-1/2,j} && 2 \le i \le n_1, \\ (\mathbb{K}_{j,j})_{i,i} &= \tfrac{1}{h_1^2}(k_{i-1/2,j} + k_{i+1/2,j}) + \tfrac{1}{h_2^2}(k_{i,j-1/2} + k_{i,j+1/2}) && 1 \le i \le n_1, \\ (\mathbb{K}_{j,j})_{i,i+1} &= -\tfrac{1}{h_1^2}\, k_{i+1/2,j} && 1 \le i \le n_1 - 1. \end{aligned}$$

Pour les blocs non-diagonaux, on note que $\mathbb{K}_{j,j-1} \in \mathbb{R}^{n_1\times n_1}$ agit de $u_{\cdot,j-1}$ vers $(f+g)_{\cdot,j}$ pour $2 \le j \le n_2$, alors que $\mathbb{K}_{j,j+1} \in \mathbb{R}^{n_1\times n_1}$ agit de $u_{\cdot,j+1}$ vers $(f+g)_{\cdot,j}$ pour $1 \le j \le n_2-1$. En outre, on vérifie facilement [11] que tous ces blocs sont des matrices diagonales, avec :

$$\begin{aligned} (\mathbb{K}_{j,j-1})_{i,i} &= -\tfrac{1}{h_2^2}\, k_{i,j-1/2} \quad 1 \le i \le n_1, \\ (\mathbb{K}_{j,j+1})_{i,i} &= -\tfrac{1}{h_2^2}\, k_{i,j+1/2} \quad 1 \le i \le n_1. \end{aligned}$$

11. $(\mathbb{K}_{j,j-1})_{i,i}$ correspond au terme d'indice $j-1$ dans le schéma à cinq points, resp. $(\mathbb{K}_{j,j+1})_{i,i}$ au terme d'indice $j+1$.

En particulier, comme toutes ces matrices sont diagonales, on a les relations $\mathbb{K}_{j,j+1} = (\mathbb{K}_{j+1,j})^T$ pour $1 \leq j \leq n_2 - 1$. Conclusion, la matrice $\mathbb{K}_2$ est symétrique. On aurait pu retrouver ce résultat en raisonnant directement sur la structure pentadiagonale par points de $\mathbb{K}_2$!

Remarque 2.22 *Si l'on revient sur les blocs diagonaux $\mathbb{K}_{j,j} \in \mathbb{R}^{n_1 \times n_1}$, on remarque qu'ils possèdent la même structure que les matrices $\mathbb{K}_1$ des différences finies 1D, où $k_{1D}(x) = k(x, y_j)$ pour x parcourant $]a_1, b_1[$, avec des contributions diagonales supplémentaires, du type $\frac{1}{h_2^2}(k_{i,j-1/2} + k_{i,j+1/2})$ pour $1 \leq i \leq n_1$.*

Par construction, les matrices $\mathbb{K}_2$ et $\mathbb{A}'_2$ sont tridiagonales par blocs, pentadiagonales par points, et symétriques. Pour obtenir plus de propriétés, et notamment que $\mathbb{A}'_2$ est inversible, nous supposons que les coefficients du modèle sont de même signe constant :

$$\exists k_0 > 0,\ \forall (x,y) \in D,\ k(x,y) \geq k_0\,; \quad \forall (x,y) \in D,\ q(x,y) \geq 0. \tag{2.40}$$

Proposition 2.23 *Sous l'hypothèse (2.40), les matrices $\mathbb{K}_2$ et $\mathbb{A}'_2$ sont monotones.*

Démonstration : Laissée en exercice. Il suffit de reprendre la démonstration du théorème 2.19 avec le double indiçage $_{i,j}$. ◇

Proposition 2.24 *Sous l'hypothèse (2.40), les matrices $\mathbb{K}_2$ et $\mathbb{A}'_2$ sont définies-positives.*

Démonstration : Laissée en exercice. L'idée est de s'appuyer sur la structure tridiagonale par blocs de $\mathbb{K}_2$, et d'utiliser le fait que ses blocs diagonaux $\mathbb{K}_{j,j}$ pour $1 \leq j \leq n_2$ sont (symétriques) définis-positifs d'après la remarque 2.22. ◇

Par conséquent, on a le résultat suivant.

Corollaire 2.25 *Sous l'hypothèse (2.40), la matrice $\mathbb{A}'_2$ est inversible.*

Exercice 2.3 *Établir un* principe de positivité *pour l'approximation (2.37) de la diffusion neutronique.*

2.4 Problème tridimensionnels, ou multi-dimensionnels

Dans cette section, on propose une méthode de calcul numérique du potentiel électrostatique, toujours basée sur les différences finies, sous la forme d'un exercice. Précisons tout d'abord le modèle. Dans $\mathbb{R}^3$, on considère une cavité cubique, $C =]0,1[^3$, dans laquelle on a fait le vide. On suppose qu'elle est entourée d'un conducteur parfait, et que le potentiel électrostatique, sur sa frontière ∂C, est nul. Enfin, on place des charges dans la cavité. On rappelle que si ρ est la densité de charge et ϵ_0 la permittivité électrique, le potentiel électrostatique u généré par les charges dans la cavité est solution de l'équation dite de Coulomb, qui s'écrit

$$-\Delta_3 u = \frac{\rho}{\epsilon_0} \text{ sur } C, \quad u = 0 \text{ sur } \partial C. \tag{2.41}$$

Exercice 2.4 *Suggérer une généralisation de l'approche du problème 2D au cas 3D, pour résoudre le problème du potentiel électrostatique (2.41).*
1. Quel sera a priori *le nombre de points du schéma aux différences finies, dans ce cas ? Le construire, en le justifiant.*
2. Examiner en détail les propriétés de la matrice $\mathbb{A}_3$ ainsi obtenue. Etablir successivement que :

- *$\mathbb{A}_3$ est monotone ;*
- *$\mathbb{A}_3$ est symétrique ;*
- *$\mathbb{A}_3$ est définie-positive.*

3. En déduire un principe de positivité discret associé à $\mathbb{A}_3$.

Nous proposons pour finir les exercices ci-dessous.

Exercice 2.5 *Suggérer une généralisation au cas de la diffusion neutronique en 3D, c'est-à-dire la résolution de*

$$-\operatorname{div}(k\,\mathbf{grad}\,u) + q\,u = f \text{ sur } C', \quad u = g \text{ sur } \partial C', \tag{2.42}$$

posé dans $C' =]a_1, b_1[\times]a_2, b_2[\times]a_3, b_3[$.

Exercice 2.6 *Suggérer une généralisation au cas d-dimensionnel ($d \geq 4$), pour résoudre le problème ci-dessous, posé dans $\Omega_d =]0,1[^d$:*

$$-\Delta_d u = f \text{ sur } \Omega_d, \quad u = g \text{ sur } \partial\Omega_d, \qquad \text{où } \Delta_d = \sum_{k=1}^{d} \frac{\partial^2}{\partial {x_k}^2}. \tag{2.43}$$

1. Quel sera a priori *le nombre de points du schéma aux différences finies, dans ce cas ? Le construire, en le justifiant.*
2. Examiner en détail les propriétés de la matrice $\mathbb{A}_d$ ainsi obtenue. Etablir successivement que :

- *$\mathbb{A}_d$ est monotone ;*
- *$\mathbb{A}_d$ est symétrique ;*
- *$\mathbb{A}_d$ est définie-positive.*

(On pourra raisonner par récurrence.)
3. En déduire un principe de positivité discret associé à $\mathbb{A}_d$.

2.5 Problèmes dépendant du temps

Dans cette section, nous considérons tout d'abord un problème instationnaire 1D, à savoir celui de la corde vibrante, fixée en ses deux extrémités. On souhaite calculer numériquement les déplacements verticaux, à partir d'une position initiale connue (à $t = 0$), jusqu'à un instant final $T > 0$. Si on suppose que la corde est de longueur L, on a vu que u vérifie

$$\frac{\partial^2 u}{\partial t^2} - c^2 \frac{\partial^2 u}{\partial x^2} = 0 \text{ pour } x \in]0, L[\text{ et } t \in]0, T[, \tag{2.44}$$

$$u(0,t) = u(L,t) = 0 \text{ pour } t \in]0, T[, \tag{2.45}$$

$$u(x,0) = u^0(x) \text{ et } \frac{\partial u}{\partial t}(x,0) = u^1(x) \text{ pour } x \in]0, L[. \tag{2.46}$$

La solution dépend de deux variables, l'une spatiale x, et l'autre temporelle t, et l'équation (2.44) met en jeu deux dérivées partielles secondes, l'une par rapport à x, et l'autre par rapport à t. Il est donc naturel de considérer une approximation par différences finies, pour chacune des deux dérivées. Par ailleurs, la seconde condition initiale comprend une dérivée première par rapport au temps, que nous allons également approcher par différences finies. Enfin, pour avoir accès aux valeurs en $(x,t) \in \{(0,0),(L,0),(0,T),(L,T)\}$, nous allons supposer que d'une part (2.45) est valable pour $t = T$, et d'autre part que la première partie de (2.46) est valable en $x = 0$ et en $x = L$, avec $u^0(0) = u^0(L) = 0$.
Dans la suite, on introduit un pas de discrétisation spatial, à savoir $h_x = L/(n_x+1)$, et un pas de discrétisation temporel *a priori* différent, $h_t = T/(n_t+1)$. L'intervalle $[0,L]$ est donc découpé en n_x+1 intervalles, et $[0,T]$ est découpé en n_t+1 intervalles. Pour bien différencier les discrétisations en espace et en temps, on note les *valeurs approchées* de la solution u aux "points" d'abscisse $x_i = ih_x$ et d'ordonnée $t_m = mh_t$ par $(u_i^m)_{0\leq i\leq n_x+1, 0\leq m\leq n_t+1}$. On arrive donc aux approximations :

$$\frac{\partial^2 u}{\partial t^2}(x_i,t_m) \approx \frac{u_i^{m+1} - 2u_i^m + u_i^{m-1}}{h_t^2},\ 1 \leq i \leq n_x,\ 1 \leq m \leq n_t, \tag{2.47}$$

$$\frac{\partial^2 u}{\partial x^2}(x_i,t_m) \approx \frac{u_{i+1}^m - 2u_i^m + u_{i-1}^m}{h_x^2},\ 1 \leq i \leq n_x,\ 1 \leq m \leq n_t, \tag{2.48}$$

$$\frac{\partial u}{\partial t}(x_i,0) \approx \frac{u_i^1 - u_i^0}{h_t},\ 1 \leq i \leq n_x. \tag{2.49}$$

A partir de là, il est aisé de construire l'ensemble des équations vérifiées par les inconnues $(u_i^m)_{0\leq i\leq n_x+1, 0\leq m\leq n_t+1}$. Commençons par l'approximation de (2.44) aux "points" *intérieurs* de discrétisation $(x_i,t_m)_{1\leq i\leq n_x, 1\leq m\leq n_t}$, c'est-à-dire dans l'*ouvert* $]0,L[\times]0,T[$. Ainsi, on trouve :

$$\frac{u_i^{m+1} - 2u_i^m + u_i^{m-1}}{h_t^2} - c^2\frac{u_{i+1}^m - 2u_i^m + u_{i-1}^m}{h_x^2} = 0,\ 1 \leq i \leq n_x,\ 1 \leq m \leq n_t. \tag{2.50}$$

Les conditions aux limites (2.45) permettent de déterminer les valeurs extrémales *en espace*, c'est-à-dire pour $i = 0$ et $i = n_x + 1$, ce qui correspond aux extrémités du domaine de calcul spatial $x_0 = 0$ et $x_{n_x+1} = L$. Par identification, on en déduit

$$u_0^m = u_{n_x+1}^m = 0,\ 1 \leq m \leq n_t + 1. \tag{2.51}$$

Enfin, les deux conditions initiales permettent de calculer les valeurs approchées aux deux premiers instant, à savoir en t_0 et t_1, par l'intermédiaire de

$$u_i^0 = u^0(x_i),\ 0 \leq i \leq n_x + 1, \tag{2.52}$$

$$\frac{u_i^1 - u_i^0}{h_t} = u^1(x_i),\ 1 \leq i \leq n_x. \tag{2.53}$$

Si on regroupe (2.50-2.53), on aboutit après réorganisation à

$$m = 0 \ : \quad \left\{ \begin{array}{l} u_i^0 = u^0(x_i), \text{ pour } 1 \leq i \leq n_x \\ u_0^0 = u_{n_x+1}^0 = 0 \end{array} \right. ; \tag{2.54}$$

$$m = 1 \ : \quad \left\{ \begin{array}{l} u_i^1 = u^0(x_i) + h_t u^1(x_i), \text{ pour } 1 \leq i \leq n_x \\ u_0^1 = u_{n_x+1}^1 = 0 \end{array} \right. ; \tag{2.55}$$

Pour $m = 1, \cdots, n_t$

$$\left\{ \begin{array}{l} u_i^{m+1} = \dfrac{c^2 h_t^2}{h_x^2}(u_{i+1}^m - 2u_i^m + u_{i-1}^m) + 2u_i^m - u_i^{m-1}, \text{ pour } 1 \leq i \leq n_x \\ u_0^{m+1} = u_{n_x+1}^{m+1} = 0 \end{array} \right. \tag{2.56}$$

Que constate-t-on ?
Tout d'abord, on calcule les valeurs de la solution approchée en *incrémentant en temps*, en commençant par l'instant $t_0 = 0$, en poursuivant par t_1, puis t_2, $\cdots$, jusqu'à l'instant final $t_{n_t+1} = T$. La valeur finale est bien une inconnue du problème, comme suggéré par la remarque 1.6 sur la solution exacte.
Par ailleurs, si on s'intéresse à la dépendance entre les données, on constate que u_i^{m+1} dépend de $(u_{i_1}^m)_{i-1 \leq i_1 \leq i+1}$ et de u_i^{m-1}. Ces valeurs dépendent elles-mêmes de $(u_{i_2}^{m-1})_{i-2 \leq i_2 \leq i+2}$, de $(u_{i_1}^{m-2})_{i-1 \leq i_1 \leq i+1}$ et de u_i^{m-3}, et ainsi de suite... Le schéma ci-dessous résume la situation, où l'on a représenté le **cône de dépendance discret** de la solution calculée.

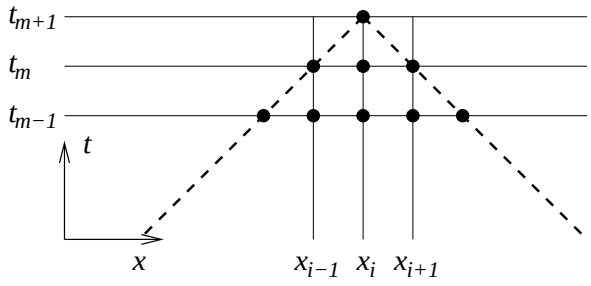


FIGURE 2.6 – Cône de dépendance discret

D'un point de vue algorithmique, il est intéressant d'introduire la suite de vecteurs $(v^m)_{0 \leq m \leq n_t+1}$ de $\mathbb{R}^{n_x}$, contenant les approximations spatiales successives sur $]0, L[$, calculées à l'aide de (2.54-2.56) : $(v^m)_i = u_i^m$, pour $1 \leq i \leq n_x$, et $0 \leq m \leq n_t + 1$. Comme dans le cas statique, il n'est pas nécessaire de stocker explicitement les valeurs aux extrémités $x = 0$ et $x = L$, qui sont connues à tout instant. On peut alors réécrire (2.56) sous la *forme vectorielle équivalente*, pour $1 \leq m \leq n_t$:

$$v^{m+1} = (2I_{n_x} + (ch_t)^2 \mathbb{A}_1')v^m - v^{m-1}, \ \mathbb{A}_1' = \frac{1}{(h_x)^2} \begin{pmatrix} 2 & -1 & \dots & \dots & 0 \\ -1 & 2 & -1 & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & -1 \\ 0 & \dots & \dots & -1 & 2 \end{pmatrix} \in \mathbb{R}^{n_x \times n_x}. \tag{2.57}$$

De cette façon, on retrouve bien une forme vectorielle de type 1D, semblable à la formulation vectorielle statique 1D (2.5). Par contre, on obtient un **schéma explicite**, au sens où il n'est pas nécessaire de résoudre un système linéaire pour déterminer la solution approchée. Enfin, les matrices $\mathbb{A}_1$ et $\mathbb{A}'_1$ diffèrent par un coefficient multiplicatif...

Dans la suite, nous n'étudions pas dans les détails la convergence de la solution discrète (notée u_{app}) vers la solution exacte u. Nous nous contentons, pour fixer les idées, d'une définition, avec une mesure de l'erreur par l'intermédiaire d'une norme *abstraite* ($\|\cdot\|$.)

Définition 2.26 *Un schéma est dit* **convergent** *pour la norme $\|\cdot\|$ si, et seulement si, pour toute donnée initiale (u^0, u^1), on a la propriété*

$$\|u_{app} - u\| \to 0, \text{ lorsque } h_x, h_t \to 0,$$

le rapport h_t/h_x étant fixé.

NB. Dans la définition ci-dessus, on remarque que les pas de discrétisation h_x et h_t sont liés entre eux.

Cette étude de la convergence repose usuellement (cf. [2]) sur deux ingrédients : la **consistance**, et la **stabilité** du schéma. Nous allons uniquement évoquer la notion de stabilité, à l'aide d'un calcul simple. Si on suppose que le domaine spatial est égal à $\mathbb{R}$, alors on peut vérifier que la solution exacte est de la forme

$$u(x,t) = \frac{1}{2}(u^0(x+ct)+u^0(x-ct))+\frac{1}{2}(v^1(x+ct)-v^1(x-ct)),\ v^1(t) = \int_0^t u^1(s)ds. \quad (2.58)$$

Remarque 2.27 *La vitesse de propagation de l'information, associée à (2.58), est égale à $\pm c$. Les ondes se propagent donc à la vitesse c (en module) sur la corde. Cette propriété reste valable pour les problèmes hyperboliques en dimension 2 ou 3 d'espace (cas 2D, 3D instationnaires hyperboliques.)*

En particulier, la valeur de u au "point" (x,t) dépend de
- la valeur de u^0 en $x \pm ct$;
- la valeur de v^1 en $x \pm ct$, c'est-à-dire la valeur de u^1 sur $[x-ct, x+ct]$.

On peut donc, à l'instar du cas discret, définir le **cône de dépendance** de la solution exacte.

Si maintenant on revient à la question de la stabilité, il est clair qu'une *condition nécessaire* de stabilité du schéma est que le cône de dépendance discret contienne *suffisamment* d'informations : en d'autres termes, il doit contenir le cône de convergence associé à la solution exacte !

Si l'on compare les pentes, à savoir h_t/h_x pour le schéma discret et $1/c$ pour la solution exacte, la condition nécessaire de stabilité du schéma explicite se résume à

$$ch_t \leq h_x. \quad (2.59)$$

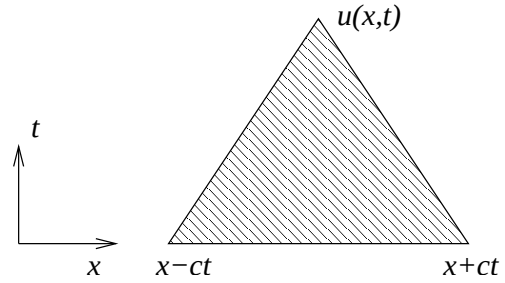


FIGURE 2.7 – Cône de dépendance

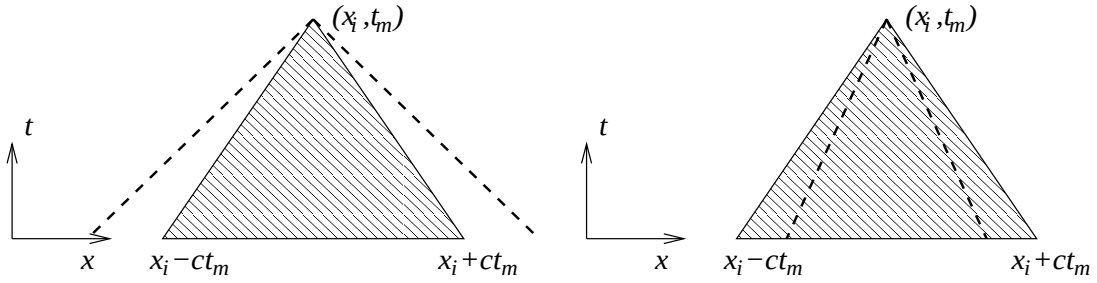


FIGURE 2.8 – Stabilité ou instabilité

On parle habituellement de **condition CFL**, pour Courant, Lax et Friedrichs.

Il est possible (cf. [2]) de prouver que la condition CFL (2.59) est en fait une *condition nécessaire et suffisante* de stabilité du schéma explicite. Pour s'affranchir d'une condition de stabilité, on peut introduire des **schémas implicites**, c'est-à-dire que l'on va déterminer $(u_i^{m+1})_i$ en résolvant un système linéaire dont la solution n'est plus explicite contrairement à (2.56). Pour cela, on réécrit l'équation (2.48) sous la forme suivante :

$$\frac{\partial^2 u}{\partial x^2}(x_i, t_m) \approx \frac{u_{i+1}^{m+1} - 2u_i^{m+1} + u_{i-1}^{m+1}}{h_x^2},\ 1 \leq i \leq n_x,\ 1 \leq m \leq n_t, \tag{2.60}$$

Ainsi, à la place de l'équation(2.50), on obtient :

$$\frac{u_i^{m+1} - 2u_i^m + u_i^{m-1}}{h_t^2} - c^2 \frac{u_{i+1}^{m+1} - 2u_i^{m+1} + u_{i-1}^{m+1}}{h_x^2} = 0,\ 1 \leq i \leq n_x,\ 1 \leq m \leq n_t. \tag{2.61}$$

Le calcul de $(u_i^{m+1})_i$ est alors effectué en résolvant le système linéaire suivant :

$$m = 0 \ : \quad \left\{ \begin{array}{l} u_i^0 = u^0(x_i), \text{ pour } 1 \leq i \leq n_x \\ u_0^0 = u_{n_x+1}^0 = 0 \end{array} \right. ; \tag{2.62}$$

$$m = 1 \ : \quad \left\{ \begin{array}{l} u_i^1 = u^0(x_i) + h_t u^1(x_i), \text{ pour } 1 \leq i \leq n_x \\ u_0^1 = u_{n_x+1}^1 = 0 \end{array} \right. ; \tag{2.63}$$

$$\text{Pour } m = 1, \cdots, n_t \text{ et pour } 1 \leq i \leq n_x \quad \left\{ \begin{array}{l} \left(1 + 2\dfrac{c^2 h_t^2}{h_x^2}\right) u_i^{m+1} - \dfrac{c^2 h_t^2}{h_x^2}(u_{i+1}^{m+1} + u_{i-1}^{m+1}) = 2u_i^m - u_i^{m-1} \\ u_0^{m+1} = u_{n_x+1}^{m+1} = 0 \end{array} \right. . \tag{2.64}$$

De la même façon que dans le cas du schéma explicite (2.57), en utilisant la suite de vecteurs $(v^m)_{0 \leq m \leq n_t+1}$, on peut réécrire (2.64) sous la *forme vectorielle équivalente*, pour $1 \leq m \leq n_t$:

$$(I_{n_x} + (ch_t)^2 \mathbb{A}_1) v^{m+1} = 2\, v^m - v^{m-1}. \tag{2.65}$$

Ainsi, dans le cas implicite, pour évaluer la solution v^{m+1}, il faut résoudre un système linéaire, alors que dans le cas explicite, la calcul de v_{m+1} est direct. Cependant, dans le cas implicite, le pas en temps h_t peut être choisi indépendamment du pas en espace h_x. Notons pour finir que la matrice $I_{n_x} + (ch_t)^2 \mathbb{A}_1$ est, comme la matrice $\mathbb{A}_1$, tridiagonale, mais elle est de plus à diagonale strictement dominante (voir la définition 6.17) : on peut, par exemple, utiliser la méthode de Jacobi ou de Gauss-Seidel par points (voir le §6.5 et le §6.6) pour résoudre le système (2.65) (voir la proposition 6.19).

Bien sûr, ce type d'approximation à l'aide des différences finies, résultant en des schémas explicites ou implicites, est utilisable *a priori* pour les problèmes instationnaires 2D et 3D, hyperboliques ou paraboliques, présentés au chapitre 1.

Pour conclure cette section, nous abordons brièvement la discrétisation des problèmes aux valeurs propres, puis celle des problèmes stationnaires. Reprenons le cas de la membrane élastique $\Omega_2 =]0,1[^2$.

Le problème aux valeurs propres s'écrit : trouver les couples modes propres [12] - valeurs propres $(x,y) \mapsto u_k(x,y)$ et $\lambda_k > 0$ tels que

$$-c^2 \Delta_2 u_k = \lambda_k u_k \text{ sur } \Omega_2, \quad u_k = 0 \text{ sur } \partial\Omega_2. \tag{2.66}$$

(Ci-dessus, k est un entier naturel non nul par convention.)
Une fois le pas de discrétisation h fixé ($h = 1/(n+1)$, $N = n^2$), on en déduit l'approximation par différences finies : trouver les couples modes propres discrets - valeurs propres discrètes $u_l \in \mathbb{R}^N$ et λ_l tels que

$$c^2 \mathbb{A}_2 u_l = \lambda_l u_l. \tag{2.67}$$

Comme $\mathbb{A}_2$ est définie-positive (proposition 2.20), on en déduit immédiatement que toute valeur propre discrète λ_l est strictement positive. En effet :

$$\lambda_l \|u_l\|_2^2 = \lambda_l (u_l, u_l) = c^2 (\mathbb{A}_2 u_l, u_l) > 0.$$

12. On rappelle qu'un mode propre n'est pas identiquement nul.

Des méthodes de calcul des valeurs propres et vecteurs propres d'une matrice sont présentées au chapitre 8. Ceci répond à la question de la résolution du problème discret. Comment peut-on vérifier que celui-ci est une bonne approximation du problème initial ? La question étant complexe, nous nous contentons de mettre en avant la problématique associée... Les valeurs propres discrètes étant en nombre fini d'après un résultat classique (proposition A.2), il est clair qu'on ne peut espérer approcher tous les couples $(u_k, \lambda_k)_{k\in\mathbb{N}}$, pour h (et donc N) fixés ! Néanmoins, le nombre de modes propres discrets croît comme N, et tend bien vers l'infini lorsque $h \to 0$... Ensuite, concernant la convergence des modes discrets vers les modes du problème initial, nous sommes confrontés à deux difficultés : identifier vers quel mode propre une suite de modes propres discrets (indexée par N) converge, et s'assurer qu'à la limite $N \to +\infty$, on atteint tous les modes propres.

Nous finissons par les problèmes stationnaires du type (pour $\nu > 0$ donné) : trouver $(x, y) \mapsto u(x, y)$ tel que

$$-\nu^2 u - c^2 \Delta_2 u = g \text{ sur } \Omega_2, \quad u = 0 \text{ sur } \partial\Omega_2. \tag{2.68}$$

Le pas h étant fixé ($h = 1/(n+1)$, $N = n^2$), l'approximation par différences finies consiste en le système linéaire

$$\mathbb{A}_{2,\nu} u = u, \text{ avec } \mathbb{A}_{2,\nu} = c^2 \mathbb{A}_2 - \nu^2 I_N. \tag{2.69}$$

On retrouve là un système linéaire à résoudre, dont la matrice $\mathbb{A}_{2,\nu}$ est symétrique. Malheureusement, lorsque ν est suffisamment grand, cette matrice n'est plus défnie-positive. Il est donc nécessaire d'utiliser une méthode de type factorisation de Gauss (chapitre 5), pour laquelle des problèmes de stabilité numérique peuvent se produirent. Quant à la question de la convergence, elle est également complexe, étant étroitement reliée à celle de la convergence du problème aux valeurs propres.

Chapitre 3

La méthode des éléments finis

3.1 Introduction

Dans ce chapitre, nous présentons une seconde méthode de discrétisation numérique, plus flexible que la méthode des différences finies car elle permet de considérer des domaines de calculs dont la géométrie est moins contrainte (non-tensorielle). En outre, on peut mener à bien l'analyse numérique de la méthode, c'est-à-dire estimer l'erreur entre la solution exacte et la solution calculée, en fonction de certains paramètres. On suppose que le lecteur est familier avec les notions d'espaces de Hilbert et les formulations variationnelles (voir l'Annexe C), ainsi qu'avec celles des espaces de Lebesgue L^p et la théorie des distributions. Comme précédemment, le but est de calculer des approximations des solutions d'EDP statiques. Nous considérons la diffusion neutronique (ou Laplacien généralisé), posé dans un domaine Ω de $\mathbb{R}^d$, pour $d = 1, 2, 3$. Le but est de calculer la solution de

$$-\operatorname{div}(k\,\mathbf{grad}\,u) + q\,u = f \text{ sur } \Omega, \quad u = 0 \text{ sur } \partial\Omega. \tag{3.1}$$

Ci-dessus, k et q sont des coefficients qui peuvent dépendre de la variable spatiale. En outre, on peut également considérer une condition aux limites de Dirichlet non-homogène, ou de Neumann ou Robin.

On renvoie à l'Annexe D pour la notion de domaine de $\mathbb{R}^d$, ainsi que sur les diverses propriétées liées aux espaces fonctionnels définis sur Ω, et notamment $H^1(\Omega)$, $H^1_0(\Omega)$, $\mathbf{L}^2(\Omega)$ et $\mathbf{H}(\operatorname{div}, \Omega)$. Ici et dans la suite, on note en gras les espaces fonctionnels dont les éléments sont à valeurs vectorielles.
Concernant les coefficients k et q, on les choisit a priori dans $L^\infty(\Omega)$, et on fait les hypothèses suivantes :

$$\begin{cases} \exists k_{min}, k_{max} > 0 \ \text{ tels que } k_{min} \le k(\boldsymbol{x}) \le k_{max} \text{ p.p. tout } \boldsymbol{x} \in \Omega, \\ \exists q_{min}, q_{max} \ge 0 \ \text{ tels que } q_{min} \le q(\boldsymbol{x}) \le q_{max} \text{ p.p. tout } \boldsymbol{x} \in \Omega. \end{cases} \tag{3.2}$$

On considère une donnée :

$$f \in L^2(\Omega). \tag{3.3}$$

Enfin, on cherche une solution u d'énergie finie sur Ω, à savoir

$$\int_\Omega (|u|^2 + |\,\mathbf{grad}\,u|^2) d\boldsymbol{x} < \infty.$$

3.2 Formulations variationnelles, existence de solutions

Par hypothèse, une solution u de (3.1) d'énergie finie est telle que $u \in H^1_0(\Omega)$. Dans la suite, nous proposons deux approches pour résoudre (3.1) et prouver l'existence d'une solution dans $H^1_0(\Omega)$, ces deux approches conduisant à des techniques de résolution numériques différentes, mais complémentaires.

3.2.1 Problème à une inconnue

On rappelle qu'on a la formule d'intégration par parties suivante

$$\forall v \in H^1_0(\Omega),\ \forall \boldsymbol{q} \in \mathbf{H}(\mathrm{div},\Omega), \quad \int_\Omega (\mathbf{grad}\, v \cdot \boldsymbol{q} + v \,\mathrm{div}\, \boldsymbol{q}) d\boldsymbol{x} = 0. \tag{3.4}$$

Pour l'utiliser pour le problème (3.1), on introduit la variable auxiliaire $\boldsymbol{p} := -k\,\mathbf{grad}\, u$. Puisque $k \in L^\infty(\Omega)$ et $\mathbf{grad}\, u \in \mathbf{L}^2(\Omega)$, on a $\boldsymbol{p} \in \mathbf{L}^2(\Omega)$. En outre, puisque la donnée $f \in L^2(\Omega)$, on a aussi $\mathrm{div}\, \boldsymbol{p} = f - q\,u \in L^2(\Omega)$ et ainsi $\boldsymbol{p} \in \mathbf{H}(\mathrm{div},\Omega)$. A partir de là, on choisit $v \in H^1_0(\Omega)$, et on remarque que si $u \in H^1_0(\Omega)$ est solution du problème (3.1), on a $\mathrm{div}\, \boldsymbol{p} + q\,u - f = 0$ dans $L^2(\Omega)$, et en prenant le produit scalaire par v dans $L^2(\Omega)$, on obtient :

$$\begin{aligned} \forall v \in H^1_0(\Omega), \quad 0 &= \int_\Omega (\mathrm{div}\, \boldsymbol{p} + q\,u - f)\, v \, d\boldsymbol{x} \\ &\overset{\text{i.p.p.}}{=} \int_\Omega (-\boldsymbol{p} \cdot \mathbf{grad}\, v + q\,uv - fv)\, d\boldsymbol{x} \\ &= \int_\Omega (k\,\mathbf{grad}\, u \cdot \mathbf{grad}\, v + q\,uv - fv)\, d\boldsymbol{x} \end{aligned}$$

On en déduit donc que si $u \in H^1_0(\Omega)$ est solution de (3.1), alors u est également solution de la **formulation variationnelle** ci-dessous :

$$\left\{ \begin{array}{l} \text{Trouver } u \in H^1_0(\Omega) \text{ tel que} \\ \forall v \in H^1_0(\Omega), \quad \displaystyle\int_\Omega (k\,\mathbf{grad}\, u \cdot \mathbf{grad}\, v + q\,uv)\, d\boldsymbol{x} = \int_\Omega fv\, d\boldsymbol{x}. \end{array} \right. \tag{3.5}$$

Exercice 3.1 *En raisonnant au sens des distributions, montrer que si u est solution de (3.5), alors u est solution de (3.1) d'énergie finie.*

A partir de là, on se concentre sur la résolution de la formulation variationnelle (3.5) sous les hypothèses (3.2). On remarque qu'elle peut s'écrire sous la forme abstraite (C.3), avec $V = H^1_0(\Omega)$ muni de la norme [13] $\|\cdot\|_{H^1(\Omega)}$, la forme bilinéaire a_1 définie sur $V \times V$:

$$(v, w) \mapsto \int_\Omega (k\,\mathbf{grad}\, v \cdot \mathbf{grad}\, w + q\,vw)\, d\boldsymbol{x}, \tag{3.6}$$

et $\mathtt{f} \in V'$:

$$v \mapsto \int_\Omega fv\, d\boldsymbol{x}. \tag{3.7}$$

13. On rappelle que $\|v\|_{H^1(\Omega)} = (\|v\|^2_{L^2(\Omega)} + \|\,\mathbf{grad}\, v\|^2_{\mathbf{L}^2(\Omega)})^{1/2} = (\int_\Omega (|v|^2 + |\,\mathbf{grad}\, v|^2) d\boldsymbol{x})^{1/2}$.

Le fait que $\mathtt{f}$ est une forme continue sur $H^1_0(\Omega)$ est une conséquence de l'inégalité de Cauchy-Schwarz dans $L^2(\Omega)$. En effet, soit $v \in H^1_0(\Omega)$, on a :

$$\begin{aligned} |\langle \mathtt{f}, v\rangle| &= \left|\int_\Omega f v \, d\boldsymbol{x}\right| = \left|(f,v)_{L^2(\Omega)}\right| \leq \|f\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} \leq \|f\|_{L^2(\Omega)}\|v\|_{H^1(\Omega)}, \\ &\Longrightarrow \quad \sup_{v\in H^1_0(\Omega)\setminus\{0\}} \frac{|\langle \mathtt{f}, v\rangle|}{\|v\|_{H^1(\Omega)}} \leq \|f\|_{L^2(\Omega)}. \end{aligned}$$

Ainsi, on a bien $\mathtt{f} \in (H^1_0(\Omega))'$, et de plus $\|\mathtt{f}\|_{(H^1_0(\Omega))'} \leq \|f\|_{L^2(\Omega)}$.
De la même façon, on prouve que la forme a_1 est continue sur $H^1_0(\Omega) \times H^1_0(\Omega)$. Soient $v, w \in H^1_0(\Omega)$, on a :

$$\begin{aligned} |a_1(v,w)| &= \left|\int_\Omega (k \,\mathbf{grad}\, v \cdot \mathbf{grad}\, w + q\, vw)\, d\boldsymbol{x}\right| \\ &\leq \left|\int_\Omega k \,\mathbf{grad}\, v \cdot \mathbf{grad}\, w \, d\boldsymbol{x}\right| + \left|\int_\Omega q\, vw \, d\boldsymbol{x}\right| \\ &\leq k_{max}\int_\Omega |\,\mathbf{grad}\, v \cdot \mathbf{grad}\, w|\, d\boldsymbol{x} + q_{max}\int_\Omega |vw|\, d\boldsymbol{x} \\ &\leq k_{max}\,\|\,\mathbf{grad}\, v\|_{\mathbf{L}^2(\Omega)}\|\,\mathbf{grad}\, w\|_{\mathbf{L}^2(\Omega)} + q_{max}\|v\|_{L^2(\Omega)}\|w\|_{L^2(\Omega)} \\ &\leq \max(k_{max}, q_{max})\left(\|\,\mathbf{grad}\, v\|_{\mathbf{L}^2(\Omega)}\|\,\mathbf{grad}\, w\|_{\mathbf{L}^2(\Omega)} + \|v\|_{L^2(\Omega)}\|w\|_{L^2(\Omega)}\right) \\ &\leq \max(k_{max}, q_{max})\,\|v\|_{H^1(\Omega)}\|w\|_{H^1(\Omega)}. \end{aligned}$$

Pour aboutir à la dernière inégalité, on a simplement utilisé l'inégalité de Cauchy-Schwarz dans $\mathbb{R}^2$: pour tous $(x_1,y_1),(x_2,y_2) \in \mathbb{R}^2$, on a $x_1x_2 + y_1y_2 \leq (x_1^2+y_1^2)^{1/2}(x_2^2+y_2^2)^{1/2}$. Ainsi,

$$\sup_{v,w\in H^1_0(\Omega)\setminus\{0\}} \frac{|a_1(v,w)|}{\|v\|_{H^1(\Omega)}\|w\|_{H^1(\Omega)}} \leq \max(k_{max}, q_{max}).$$

On veut maintenant utiliser le formalisme du §C.2.

Théorème 3.1 *Sous les hypothèses (3.2), la formulation variationnelle (3.5) est bien posée. En particulier*

$$\exists C > 0,\ \forall f \in L^2(\Omega),\ \exists! u \text{ solution de (3.5) tel que } \|u\|_{H^1(\Omega)} \leq C\,\|f\|_{L^2(\Omega)}.$$

Démonstration : Pour pouvoir appliquer le théorème de Lax-Milgram, il faut démontrer la coercivité de la forme a_1, ce que l'on fait grâce à l'inégalité de Poincaré (théorème D.2). En effet, on a, pour $v \in H^1_0(\Omega)$:

$$\begin{aligned} a_1(v,v) &= \int_\Omega \left(k|\,\mathbf{grad}\, v|^2 + q|v|^2\right) d\boldsymbol{x} \\ &\geq k_{min}\int_\Omega |\,\mathbf{grad}\, v|^2\, d\boldsymbol{x} + q_{min}\int_\Omega |v|^2\, d\boldsymbol{x} \\ &\geq k_{min}\,\|\,\mathbf{grad}\, v\|^2_{\mathbf{L}^2(\Omega)} = k_{min}\frac{1+C_P^2}{1+C_P^2}\,\|\,\mathbf{grad}\, v\|^2_{\mathbf{L}^2(\Omega)} \\ &\geq \frac{k_{min}}{1+C_P^2}\left(\|\,\mathbf{grad}\, v\|^2_{\mathbf{L}^2(\Omega)} + \|v\|^2_{L^2(\Omega)}\right) = \frac{k_{min}}{1+C_P^2}\,\|v\|^2_{H^1(\Omega)}. \end{aligned}$$

Ci-dessus, $C_P > 0$ est la constante apparaissant dans l'inégalité de Poincaré. Ainsi, la forme a_1 est coercive (voir la définition C.4) avec une constante de coercivité égale à $\alpha = k_{min}(1 + C_P^2)^{-1} > 0$. On peut donc appliquer le théorème de Lax-Milgram, et en conclure que la formulation variationnelle (3.5) est bien posée, et qu'en particulier pour toute donnée $f \in L^2(\Omega)$, il existe une solution et une seule $u \in H_0^1(\Omega)$ de (3.5) ou de (3.1), qui dépend continûment de $\mathtt{f} \in (H_0^1(\Omega))'$ (et $\mathtt{f}$ est telle que $\|\mathtt{f}\|_{(H_0^1(\Omega))'} \leq \|f\|_{L^2(\Omega)}$). Plus directement, en prenant $v = u$ dans la formulation variationnelle (3.5), on trouve que

$$\|u\|_{H^1(\Omega)} \leq \left(\frac{1 + C_P^2}{k_{min}}\right)^{1/2} \|f\|_{L^2(\Omega)}.$$

$\diamond$

3.2.2 Problème à deux inconnues

On choisit ici de conserver explicitement la variable auxiliaire $\boldsymbol{p} = -k\,\mathbf{grad}\,u$ introduite précédemment. En d'autres termes, on s'intéresse aussi au gradient de u (multiplié par k). Comme on cherche des solutions u d'énergie finie, on a $\boldsymbol{p} \in \mathbf{L}^2(\Omega)$; et, puisque la donnée f appartient à $L^2(\Omega)$, on a $\boldsymbol{p} \in \mathbf{H}(\mathrm{div}, \Omega)$. On réécrit donc le problème de départ (3.1) avec une solution d'énergie finie sous la forme

$$\left\{ \begin{array}{l} \text{Trouver } u \in H_0^1(\Omega),\ \boldsymbol{p} \in \mathbf{H}(\mathrm{div}, \Omega) \text{ tels que} \\ \mathrm{div}\,\boldsymbol{p} + q\,u = f \text{ et } k^{-1}\boldsymbol{p} + \mathbf{grad}\,u = 0 \text{ sur } \Omega. \end{array} \right. \tag{3.8}$$

Pour résoudre ce problème à deux inconnues, nous supposons que la borne inférieure q_{min} dans (3.2) est *strictement positive* :

$$\left\{ \begin{array}{l} \exists k_{min}, k_{max} > 0 \ \text{ tels que } k_{min} \leq k(\boldsymbol{x}) \leq k_{max} \text{ p.p. tout } \boldsymbol{x} \in \Omega, \\ \exists q_{min}, q_{max} > 0 \ \text{ tels que } q_{min} \leq q(\boldsymbol{x}) \leq q_{max} \text{ p.p. tout } \boldsymbol{x} \in \Omega. \end{array} \right. \tag{3.9}$$

Pour construire une formulation variationnelle équivalente, nous procédons en deux étapes (il y a maintenant deux équations à prendre en compte). Tout d'abord, on choisit $\boldsymbol{q} \in \mathbf{H}(\mathrm{div}, \Omega)$, et on effectue le produit scalaire dans $\mathbf{L}^2(\Omega)$ de la seconde équation de (3.8). Puisque $(u, \boldsymbol{q})$ appartient à $H_0^1(\Omega) \times \mathbf{H}(\mathrm{div}, \Omega)$, on peut intégrer par parties, cf. (3.4). On a donc :

$$\forall \boldsymbol{q} \in \mathbf{H}(\mathrm{div}, \Omega), \quad 0 = \int_\Omega (k^{-1}\boldsymbol{p} + \mathbf{grad}\,u) \cdot \boldsymbol{q}\, d\boldsymbol{x} \overset{\text{i.p.p.}}{=} \int_\Omega (k^{-1}\boldsymbol{p} \cdot \boldsymbol{q} - u\,\mathrm{div}\,\boldsymbol{q}) d\boldsymbol{x}.$$

A l'issue de cette intégration par parties, on remarque qu'il suffit maintenant que u appartienne à $L^2(\Omega)$ pour donner un sens à la dernière intégrale. Le fait que $u \in H^1(\Omega)$ et que la trace de u s'annule sur $\partial\Omega$ est contenu dans l'égalité variationnelle ci-dessus, valable pour toute fonction-test $\boldsymbol{q}$ dans $\mathbf{H}(\mathrm{div}, \Omega)$. par exemple, si on choisit $\boldsymbol{q} \in \boldsymbol{D}(\Omega)$, on retrouve sans peine que $\mathbf{grad}\,u = -k^{-1}\boldsymbol{p}$ au sens des distributions dans Ω, et donc que $u \in H^1(\Omega)$. Pour retrouver la condition aux limites homogène, il faut utiliser la surjectivité de l'application trace normale, voir le théorème D.3.
On cherche dorénavant une solution $(u, \boldsymbol{p}) \in L^2(\Omega) \times \mathbf{H}(\mathrm{div}, \Omega)$.
Pour prendre en compte la première équation de (3.8), on utilise simplement l'égalité variationnelle équivalente :

$$\forall v \in L^2(\Omega), \quad 0 = \int_\Omega (\mathrm{div}\,\boldsymbol{p} + q\,u - f) v\, d\boldsymbol{x}.$$

Par soustraction des égalités variationnelles, on en conclut que si $(u, \boldsymbol{p})$ est solution de (3.8), alors $(u, \boldsymbol{p})$ est solution de la **formulation variationnelle** ci-dessous :

$$\left\{ \begin{array}{l} \text{Trouver } (u, \boldsymbol{p}) \in L^2(\Omega) \times \mathbf{H}(\mathrm{div}, \Omega) \text{ tel que} \\ \forall (v, \boldsymbol{q}) \in L^2(\Omega) \times \mathbf{H}(\mathrm{div}, \Omega), \\ \qquad \displaystyle\int_\Omega \left(-k^{-1} \boldsymbol{p} \cdot \boldsymbol{q} + u \,\mathrm{div}\, \boldsymbol{q} + v \,\mathrm{div}\, \boldsymbol{p} + q\, uv \right) d\boldsymbol{x} = \int_\Omega f v \, d\boldsymbol{x}. \end{array} \right. \tag{3.10}$$

Comme on l'a vu, si $(u, \boldsymbol{p})$ est solution de (3.10), alors $(u, \boldsymbol{p})$ est solution de (3.8).
Sous la forme abstraite (C.3), on se place dans l'espace produit $\mathcal{V} = L^2(\Omega) \times \mathbf{H}(\mathrm{div}, \Omega)$ muni de la norme produit $\|(v, \boldsymbol{q})\|_{\mathcal{V}} = (\|v\|^2_{L^2(\Omega)} + \|\boldsymbol{q}\|^2_{\mathbf{H}(\mathrm{div},\Omega)})^{1/2}$, avec la forme bilinéaire définie sur $\mathcal{V} \times \mathcal{V}$:

$$a_2 \ : \ ((v, \boldsymbol{q}), (w, \boldsymbol{r})) \mapsto \int_\Omega \left(-k^{-1} \boldsymbol{q} \cdot \boldsymbol{r} + v \,\mathrm{div}\, \boldsymbol{r} + w \,\mathrm{div}\, \boldsymbol{q} + q\, vw \right) d\boldsymbol{x}, \tag{3.11}$$

et $\mathtt{f} \in \mathcal{V}' : (v, \boldsymbol{q}) \mapsto \int_\Omega f v \, d\boldsymbol{x}$.
La continuité de la forme $\mathtt{f}$ est immédiate, avec $\|\mathtt{f}\|_{\mathcal{V}'} \leq \|f\|_{L^2(\Omega)}$. En outre, on peut vérifier sans peine que la forme a_2 est continue sur $\mathcal{V} \times \mathcal{V}$. Qu'en est-il de la coercivité de la forme a_2 ? On a :

$$\begin{array}{rl} a_2((0, \boldsymbol{q}), (0, \boldsymbol{q})) = -\int_\Omega k^{-1} |\boldsymbol{q}|^2 d\boldsymbol{x} < 0 & \text{si } \boldsymbol{q} \neq 0, \\ a_2((v, 0), (v, 0)) = \int_\Omega q |v|^2 d\boldsymbol{x} > 0 & \text{si } v \neq 0. \end{array}$$

D'après la Remarque C.7, comme la forme a_2 n'a pas de signe, elle n'est donc pas coercive ! ? En conséquence, il faut appliquer le théorie de Banach-Necas-Babuska pour établir le caractère bien posé de (3.10).

Théorème 3.2 *Sous les hypothèses (3.9), la formulation variationnelle (3.10) est bien posée. En particulier*

$$\exists C > 0, \ \forall f \in L^2(\Omega), \ \exists ! (u, \boldsymbol{p}) \text{ solution de (3.5) tel que } \|u\|_{L^2(\Omega)} + \|\boldsymbol{p}\|_{\mathbf{H}(\mathrm{div},\Omega)} \leq C \, \|f\|_{L^2(\Omega)}.$$

Démonstration : On remarque tout d'abord que la forme a_2 est symétrique, puisque

$$\forall (v, \boldsymbol{q}), (w, \boldsymbol{r}) \in \mathcal{V}, \quad a_2((v, \boldsymbol{q}), (w, \boldsymbol{r})) = a_2((w, \boldsymbol{r}), (v, \boldsymbol{q})).$$

On peut donc utiliser le Corollaire C.12 pour conclure : il reste à établir une condition de stabilité (C.7) que nous rappelons ici :

$$\exists \alpha' > 0, \ \forall (v, \boldsymbol{q}) \in \mathcal{V}, \quad \sup_{(w,\boldsymbol{r}) \in \mathcal{V} \setminus \{0\}} \frac{|a_2((v, \boldsymbol{q}), (w, \boldsymbol{r}))|}{\|(w, \boldsymbol{r})\|_{\mathcal{V}}} \geq \alpha' \, \|(v, \boldsymbol{q})\|_{\mathcal{V}}.$$

Pour tout $(v, \boldsymbol{q}) \in \mathcal{V}$ non-nul (c'est évident pour $(0, \boldsymbol{0})$!), on doit donc trouver un représentant $(w^*, \boldsymbol{r}^*)$ permettant de vérifier cette condition.
On commence par le cas où $\mathrm{div}\, \boldsymbol{q} = 0$. On a :

$$a_2((v, \boldsymbol{q}), (w^*, \boldsymbol{r}^*)) = \int_\Omega \left(-k^{-1} \boldsymbol{q} \cdot \boldsymbol{r}^* + v \,\mathrm{div}\, \boldsymbol{r}^* + q\, vw^* \right) d\boldsymbol{x}.$$

On remarque que si on choisit $\boldsymbol{r}^* = -\boldsymbol{q}$, le premier terme de l'intégrale est strictement positif, et le second est nul :

$$a_2((v,\boldsymbol{q}),(w^*,-\boldsymbol{q})) = \int_\Omega \left(k^{-1}|\boldsymbol{q}|^2 + q\,vw^*\right) d\boldsymbol{x}.$$

Ensuite, choisir $w^* = v$ donne

$$a_2((v,\boldsymbol{q}),(v,-\boldsymbol{q})) = \int_\Omega \left(k^{-1}|\boldsymbol{q}|^2 + q|v|^2\right) d\boldsymbol{x}.$$

Avec ce choix, on a $\|(w^*,\boldsymbol{r}^*)\|_{\mathcal{V}} = (\|v\|^2_{L^2(\Omega)} + \|\boldsymbol{q}\|^2_{\mathbf{H}(\mathrm{div},\Omega)})^{1/2} = \|(v,\boldsymbol{q})\|_{\mathcal{V}}$, et de plus

$$\begin{aligned}
a_2((v,\boldsymbol{q}),(w^*,\boldsymbol{r}^*)) &= \int_\Omega \left(k^{-1}|\boldsymbol{q}|^2 + q|v|^2\right) d\boldsymbol{x} \\
&\geq (k_{max})^{-1}\int_\Omega |\boldsymbol{q}|^2\,d\boldsymbol{x} + q_{min}\int_\Omega |v|^2\,d\boldsymbol{x} \\
&\geq \min((k_{max})^{-1}, q_{min})\,\|(v,\boldsymbol{q})\|^2_{\mathcal{V}} \\
&= \min((k_{max})^{-1}, q_{min})\,\|(v,\boldsymbol{q})\|_{\mathcal{V}}\,\|(w^*,\boldsymbol{r}^*)\|_{\mathcal{V}}, \\
\Longrightarrow \quad & \sup_{(w,\boldsymbol{r})\in\mathcal{V}\setminus\{0\}} \frac{|a_2((v,\boldsymbol{q}),(w,\boldsymbol{r}))|}{\|(w,\boldsymbol{r})\|_{\mathcal{V}}} \geq \min((k_{max})^{-1}, q_{min})\,\|(v,\boldsymbol{q})\|_{\mathcal{V}}.
\end{aligned}$$

Poursuivons par le cas général ($\mathrm{div}\,\boldsymbol{q} \neq 0$ est possible), en choisissant *a priori* $\boldsymbol{r}^* = -\boldsymbol{q}$:

$$a_2((v,\boldsymbol{q}),(w^*,-\boldsymbol{q})) = \int_\Omega \left(k^{-1}|\boldsymbol{q}|^2 + (w^*-v)\,\mathrm{div}\,\boldsymbol{q} + q\,vw^*\right) d\boldsymbol{x}.$$

Cette fois, on choisit $w^* = \lambda(v + q^{-1}\,\mathrm{div}\,\boldsymbol{q})$, avec λ à déterminer [14] :

$$a_2((v,\boldsymbol{q}),(\lambda(v+q^{-1}\,\mathrm{div}\,\boldsymbol{q}),-\boldsymbol{q})) = \int_\Omega \left(k^{-1}|\boldsymbol{q}|^2 + (2\lambda-1)v\,\mathrm{div}\,\boldsymbol{q} + \lambda\,q^{-1}|\,\mathrm{div}\,\boldsymbol{q}|^2 + \lambda\,q\,|v|^2\right) d\boldsymbol{x}.$$

Prendre $\lambda = \frac{1}{2}$, i.e. $w^* = \frac{1}{2}(v + q^{-1}\,\mathrm{div}\,\boldsymbol{q})$, permet d'éliminer le terme croisé :

$$a_2((v,\boldsymbol{q}),(\frac{1}{2}(v+q^{-1}\,\mathrm{div}\,\boldsymbol{q}),-\boldsymbol{q})) = \int_\Omega \left(k^{-1}|\boldsymbol{q}|^2 + \frac{1}{2}q^{-1}|\,\mathrm{div}\,\boldsymbol{q}|^2 + \frac{1}{2}\,q\,|v|^2\right) d\boldsymbol{x}.$$

On en déduit que

$$\begin{aligned}
a_2((v,\boldsymbol{q}),(w^*,\boldsymbol{r}^*)) &\geq \min((k_{max})^{-1}, \frac{1}{2}(q_{max})^{-1}, \frac{1}{2}q_{min}) \int_\Omega \left(|\boldsymbol{q}|^2 + |\,\mathrm{div}\,\boldsymbol{q}|^2 + |v|^2\right) d\boldsymbol{x} \\
&\geq \min((k_{max})^{-1}, \frac{1}{2}(q_{max})^{-1}, \frac{1}{2}q_{min})\,\|(v,\boldsymbol{q})\|^2_{\mathcal{V}}.
\end{aligned}$$

Pour conclure la démonstration, il reste à majorer $\|(w^*,\boldsymbol{r}^*)\|_{\mathcal{V}}$ par $\|(v,\boldsymbol{q})\|_{\mathcal{V}}$. On obtient :

$$\begin{aligned}
\|(w^*,\boldsymbol{r}^*)\|^2_{\mathcal{V}} &= \|\frac{1}{2}(v+q^{-1}\,\mathrm{div}\,\boldsymbol{q})\|^2_{L^2(\Omega)} + \|\boldsymbol{q}\|^2_{\mathbf{L}^2(\Omega)} + \|\,\mathrm{div}\,\boldsymbol{q}\|^2_{L^2(\Omega)} \\
&\leq \frac{1}{2}\|v\|^2_{L^2(\Omega)} + \|\boldsymbol{q}\|^2_{\mathbf{L}^2(\Omega)} + (1+\frac{1}{2}(q_{min})^{-2})\|\,\mathrm{div}\,\boldsymbol{q}\|^2_{L^2(\Omega)} \\
&\leq (1+\frac{1}{2}(q_{min})^{-2})\left(\|v\|^2_{L^2(\Omega)} + \|\boldsymbol{q}\|^2_{\mathbf{L}^2(\Omega)} + \|\,\mathrm{div}\,\boldsymbol{q}\|^2_{L^2(\Omega)}\right), \\
\Longrightarrow \quad & \|(w^*,\boldsymbol{r}^*)\|_{\mathcal{V}} \leq (1+\frac{1}{2}(q_{min})^{-2})^{1/2}\,\|(v,\boldsymbol{q})\|_{\mathcal{V}}.
\end{aligned}$$

14. Le choix de la "correction" $q^{-1}\,\mathrm{div}\,\boldsymbol{q}$ provient de la première équation de (3.8), où l'on voit que le champ scalaire u est homogène à $q^{-1}\,\mathrm{div}\,\boldsymbol{p}$...

Et finalement, on a dans le cas général

$$\sup_{(w,\boldsymbol{r})\in\mathcal{V}\setminus\{0\}} \frac{|a_2((v,\boldsymbol{q}),(w,\boldsymbol{r}))|}{\|(w,\boldsymbol{r})\|_{\mathcal{V}}} \geq \frac{\min((k_{max})^{-1},\frac{1}{2}(q_{max})^{-1},\frac{1}{2}q_{min})}{(1+\frac{1}{2}(q_{min})^{-2})^{1/2}}\|(v,\boldsymbol{q})\|_{\mathcal{V}},$$

qui est la condition de stabilité cherchée. La formulation variationnelle (3.10) est bien posée, et en particulier $\|(u,\boldsymbol{p})\|_{\mathcal{V}}$ dépend continûment de $\|\mathtt{f}\|_{\mathcal{V}'}$ (la donnée). Comme $\|\mathtt{f}\|_{\mathcal{V}'} \leq \|f\|_{L^2(\Omega)}$, on a bien la conclusion cherchée, si on se souvient que :

$$\|u\|_{L^2(\Omega)} + \|\boldsymbol{p}\|_{\mathbf{H}(\mathrm{div},\Omega)} \leq \sqrt{2}\left(\|u\|^2_{L^2(\Omega)} + \|\boldsymbol{p}\|^2_{\mathbf{H}(\mathrm{div},\Omega)}\right)^{1/2} = \sqrt{2}\,\|(u,\boldsymbol{p})\|_{\mathcal{V}}.$$

$\diamond$

Plutôt que de démontrer la condition de stabilité (ou condition inf-sup) de la forme a_2, on peut passer par la théorie de la $\mathbb{T}$-coercivité, voir le théorème C.14. La démarche est très similaire à celle proposée précédemment. Pour tout $(v,\boldsymbol{q}) \in \mathcal{V}\setminus\{(0,0)\}$, on cherche $(w^*,\boldsymbol{r}^*)$ *dépendant linéairement de* $(v,\boldsymbol{q})$ et tel que

$$a_2((v,\boldsymbol{q}),(w^*,\boldsymbol{r}^*)) \geq \underline{\alpha}\|(v,\boldsymbol{q})\|^2_{\mathcal{V}},$$

avec $\underline{\alpha} > 0$ indépendant de $(v,\boldsymbol{q})$. A partir de là, on peut choisir $\mathbb{T} \in \mathcal{L}(\mathcal{V})$ défini par $\mathbb{T}((v,\boldsymbol{q})) = (w^*,\boldsymbol{r}^*)$. Dans notre cas, on choisit donc

$$\mathbb{T}((v,\boldsymbol{q})) = (\frac{1}{2}(v + q^{-1}\,\mathrm{div}\,\boldsymbol{q}), -\boldsymbol{q}), \tag{3.12}$$

et $\underline{\alpha} = \min((k_{max})^{-1},\frac{1}{2}(q_{max})^{-1},\frac{1}{2}q_{min})$. Et on sait que $|||\mathbb{T}||| \leq (1+\frac{1}{2}(q_{min})^{-2})^{1/2}$. D'après la définition C.13, la forme a_2 est $\mathbb{T}$-coercive. Le théorème C.14 s'applique, et on a établi le théorème 3.2 par ce biais.

3.3 Discrétisation par éléments finis

Après la méthode des différences finies, nous introduisons une seconde méthode de discrétisation numérique, basée sur l'approximation des formulations variationnelles, et appelée **méthode des éléments finis**. On obtient ainsi une nouvelle catégorie de schémas numériques de discrétisation. La première différence par rapport à la méthode des différences finies est qu'on n'approche plus les EDP plus les conditions aux limites, mais une forme variationnelle équivalente. En outre, on peut considérer des domaines de forme plus générale lorsqu'on utilise la méthode des éléments finis.
Pour commencer, nous la présentons pour le problème à une inconnue, posé dans des domaines Ω *polygonaux* ($d = 2$) ou *polyédriques* ($d = 3$). Plus précisément, l'approximation que nous choisissons est de type conforme, ou Galerkin, cf. §C.4.2, c'est-à-dire que les espaces discrets $(V_h)_h$ que nous construisons sont des sous-espaces vectoriels de dimension finie de l'espace V.
Nous poursuivons ensuite par une étude similaire pour le problème à deux inconnues, toujours posé des domaines Ω *polygonaux* ($d = 2$) ou *polyédriques (*$d = 3$*)*, en choisissant encore une approximation de type conforme : cette fois les espaces discrets $(\mathcal{V}_h)_h$ que nous construisons sont des sous-espaces vectoriels de dimension finie de l'espace $\mathcal{V}$.
Dans les deux cas, le paramètre positif h est destiné à tendre vers 0, et $\lim_{h\to 0} dim(V_h) = +\infty$, respectivement $\lim_{h\to 0} dim(\mathcal{V}_h) = +\infty$.

3.3.1 Problème à une inconnue

On suppose que l'on dispose d'une famille $(V_h)_h$ de sous-espaces vectoriels de dimension finie de $V = H_0^1(\Omega)$. On munit les sous-espaces V_h de la norme $\|\cdot\|_V = \|\cdot\|_{H^1(\Omega)}$.

Soit h donné. Dans l'espace V_h correspondant, la **formulation variationnelle discrète** associée à (3.5) est :

$$\left\{ \begin{array}{l} \text{Trouver } u_h \in V_h \text{ tel que} \\ \forall v_h \in V_h, \quad a_1(u_h, v_h) = \langle \mathtt{f}, v_h \rangle, \end{array} \right. \tag{3.13}$$

où la forme bilinéaire a_1 est donnée par (3.6), et la forme linéaire $\mathtt{f}$ est donnée par (3.7).

Remarque 3.3 *Comme on cherche la solution discrète u_h dans $V_h \subset H_0^1(\Omega)$, celle-ci est définie presque partout, ce qui constitue une autre différence par rapport à la méthode des différences finies, où la solution exacte n'est approchée qu'aux points de discrétisation.*

Comme la solution discrète u_h dépend linéairement de la donnée f, on peut réexprimer (3.13) sous la forme d'un système linéaire équivalent [15].

Lemme 3.4 *Soit $(\varphi_j)_{1\leq j\leq N}$ une base de l'espace vectoriel V_h. Résoudre la formulation variationnelle discrète (3.13) est équivalent à résoudre le système linéaire posé dans $\mathbb{R}^N$:*

$$\left\{ \begin{array}{l} \textit{Trouver } X \in \mathbb{R}^N \textit{ tel que} \\ \mathbb{A}X = F \end{array} \right. , \tag{3.14}$$

avec

$$\mathbb{A}_{i,j} = a_1(\varphi_j, \varphi_i),\ 1 \leq i, j \leq N, \quad F_i = \langle \mathtt{f}, \varphi_i \rangle,\ 1 \leq i \leq N, \tag{3.15}$$

et l'on a la correspondance

$$u_h = \sum_{j=1}^{N} X_j \varphi_j. \tag{3.16}$$

Remarque 3.5 *Sauf exception, on omet la dépendance de N, $\mathbb{A}$, X et F par rapport à h.*

Démonstration : Soit u_h résolvant (3.13) : on note $(X_j)_{1\leq j\leq N}$ ses composantes dans la base $(\varphi_j)_{1\leq j\leq N}$. On choisit $v_h = \varphi_i$ dans la formulation variationnelle (3.13), ce qui nous donne avec la définition (3.16) du vecteur X, et la définition (3.15) de $\mathbb{A}$ et F :

$$\forall i, \quad \sum_{j=1}^{N} \mathbb{A}_{i,j} X_j = \sum_{j=1}^{N} X_j a_1(\varphi_j, \varphi_i) = a_1(\sum_{j=1}^{N} X_j \varphi_j, \varphi_i) \overset{(3.13)}{=} \langle \mathtt{f}, \varphi_i \rangle = F_i,$$

c'est-à-dire le système linéaire (3.14), écrit ligne par ligne.
Réciproquement, à partir de X une solution de (3.14), on définit u_h selon (3.16), et on

15. En règle générale, on écrira les matrices issues de la discrétisation par éléments finis en caractères ombrés majuscules.

choisit $v_h = \sum_{i=1}^N Y_i\varphi_i$ un élément quelconque de V_h. On trouve

$$\begin{aligned} a_1(u_h, v_h) &= a\left(\sum_{j=1}^N X_j\varphi_j, \sum_{i=1}^N Y_i\varphi_i\right) = \sum_{i=1}^N Y_i\left(\sum_{j=1}^N X_j\, a_1(\varphi_j,\varphi_i)\right) \\ &\overset{(3.15)}{=} \sum_{i=1}^N Y_i\left(\sum_{j=1}^N \mathbb{A}_{i,j}X_j\right) \\ &\overset{(3.14)}{=} \sum_{i=1}^N Y_iF_i \overset{(3.15)}{=} \sum_{i=1}^N Y_i\,\langle \mathtt{f},\varphi_i\rangle = \langle \mathtt{f}, \sum_{i=1}^N Y_i\varphi_i\rangle = \langle \mathtt{f}, v_h\rangle. \end{aligned}$$

◇

Dans la suite, on suppose qu'une base $(\varphi_j)_{1\le j\le N}$ de V_h est donnée, et que $\mathbb{A}$ et F sont définis par (3.15). Notons en passant qu'on a établi (voir la démonstration de la réciproque) une égalité fondamentale reliant les fonctions discrètes aux vecteurs de $\mathbb{R}^N$ les représentant dans cette base $(\varphi_j)_{1\le j\le N}$:

$$\forall v_h = \sum_{j=1}^N Y_j\varphi_j,\ \forall w_h = \sum_{j=1}^N Z_j\varphi_j, \quad a_1(v_h, w_h) = (\mathbb{A}Y, Z). \tag{3.17}$$

Proposition 3.6 *La matrice $\mathbb{A}$ est symétrique par construction. Sous les hypothèses (3.2), elle est en outre définie-positive.*

Démonstration : La matrice $\mathbb{A}$ est symétrique, puisque la forme a_1 l'est. Sous les hypothèses (3.2), la forme a_1 est coercive, et on en déduit que,

$$\forall Y \in \mathbb{R}^N \setminus \{0\}, \quad (\mathbb{A}Y, Y) \overset{(3.17)}{=} a_1(v_h, v_h) > 0.$$

◇

D'après la proposition 2.11, la matrice $\mathbb{A}$ est inversible, et le système linéaire (3.14) admet une solution X et une seule. On en déduit finalement que la formulation variationnelle discrète (3.13) admet une solution u_h et une seule.
Pour comparer cette solution discrète à la solution exacte u, on utilise le lemme de Céa (cf. théorème C.20). On trouve dans notre cas :

$$\|u - u_h\|_{H^1(\Omega)} \le \frac{\max(k_{max}, q_{max})}{k_{min}}(1 + C_P^2) \inf_{v_h\in V_h} \|u - v_h\|_{H^1(\Omega)}, \tag{3.18}$$

où C_P est la constante qui apparaît dans l'inégalité de Poincaré. En particulier, l'erreur $\|u-u_h\|_{H^1(\Omega)}$ et $\inf_{v_h\in V_h}\|u-v_h\|_{H^1(\Omega)}$ sont du même ordre (cf. remarque C.21). En outre, comme V_h est un sous-espace vectoriel fermé de $H^1(\Omega)$, on a $\inf_{v_h\in V_h}\|u - v_h\|_{H^1(\Omega)} = \min_{v_h\in V_h}\|u - v_h\|_{H^1(\Omega)} = \|u - P_hu\|_{H^1(\Omega)}$, où P_h est la projection orthogonale de $H^1(\Omega)$ sur V_h.

On laisse maintenant h *varier*. On dispose donc des solutions approchées $(u_h)_h$. D'après le théorème C.23, pour que l'erreur tende vers 0 quand h tend vers 0, il suffit d'avoir la propriété d'approximabilité minimale (C.26). On réécrit ci-dessous le résultat de convergence pour le problème à une inconnue.

Théorème 3.7 *Sous les hypothèses (3.2) et si la propriété d'approximabilité minimale (C.26) est vraie pour* $(V_h)_h$ *dans* $H^1(\Omega)$*, alors l'erreur tend vers* 0 *quand* h *tend vers* 0 *:*

$$\lim_{h\to 0} \|u - u_h\|_{H^1(\Omega)} = 0. \tag{3.19}$$

Concernant la qualité des solutions approchées $(u_h)_h$, on introduit habituellement la notion de **vitesse de convergence**.

Définition 3.8 *Les solutions approchées* $(u_h)_h$ *convergent vers* u *à l'ordre* $\alpha_{cv} > 0$ *si, et seulement si,*

$$\exists C_{cv} > 0,\ \forall f,\ \forall h,\ \|u - u_h\|_{H^1(\Omega)} \leq C_{cv}\, h^{\alpha_{cv}}\, \|f\|. \tag{3.20}$$

Comment définir les espaces discrets $(V_h)_h$ pour le problème à une inconnue ? Une idée naturelle consiste à construire un espace d'approximation à partir de fonctions continues, et affines par morceaux. En effet, les fonctions globalement continues, et régulières par morceaux, appartiennent à $H^1(\Omega)$ [11]. Il faut enfin imposer une condition de trace nulle sur $\partial\Omega$ pour définir des fonctions de $H^1_0(\Omega)$. Afin de construire de tels espaces, on suppose maintenant que l'on résout le problème posé dans un domaine Ω *polygonal* (si $d = 2$), ou *polyédrique* (si $d = 3$). On choisit un **maillage** de Ω, ici une partition en simplexes fermés, c'est-à-dire des *triangles fermés* pour $d = 2$ et des *tétraèdres fermés* pour $d = 3$, $\overline{\Omega} = \cup_{1\leq\ell\leq L} T_\ell$, qui vérifie les propriétés suivantes :

$$\left\{\begin{array}{l} int(T_\ell) \neq \emptyset \text{ pour tout } \ell\ ; \\ int(T_\ell) \cap int(T_m) = \emptyset \text{ si } \ell \neq m\ ; \\ \text{toute facette d'un simplexe est :} \\ \quad \text{soit facette d'un autre simplexe, soit incluse dans } \partial\Omega. \end{array}\right. \tag{3.21}$$

Par définition, pour $d = 2$, une facette est une arête du bord du triangle et, pour $d = 3$, une facette est une face du bord du tétraèdre. Il y a donc $d + 1$ facettes par simplexe. On appelle $(M_j)_{j=1,N^+}$ l'ensemble de tous les sommets du maillage, numérotés de sorte que $(M_j)_{j=1,N}$ est l'ensemble des sommets intérieurs, et $(M_j)_{j=N+1,N^+}$ est l'ensemble des sommets situés sur la frontière. On appelle **pas du maillage** le nombre strictement positif $h = \max_{1\leq\ell\leq L} h_\ell$ où, pour chaque ℓ, h_ℓ est le diamètre du plus petit cercle contenant T_ℓ. On note également, pour chaque ℓ, ρ_ℓ le diamètre du plus grand cercle inclus dans T_ℓ. Enfin, on appelle $\mathcal{T}_h$ le maillage obtenu. Dans la suite, on s'appuie sur une famille de maillages $(\mathcal{T}_h)_h$ pour construire les espaces discrets $(V_h)_h$.

Remarque 3.9 *Pour que* h *tende vers* 0*, on doit construire des partitions telles que le diamètre maximal des simplexes tende vers* 0*. Ceci implique en particulier que* $\max_{1\leq\ell\leq L} |T_\ell|$ *tende vers* 0*, et donc que le nombre de simplexes tende vers* $+\infty$*.*

On raisonne pour *une partition, et donc pour un maillage, donnés* : $\overline{\Omega} = \cup_{1\leq\ell\leq L} T_\ell$. On peut maintenant définir V_h, sous-espace vectoriel de $H^1_0(\Omega)$ de dimension finie, en deux temps. On introduit, pour $k \in \mathbb{N}$,

$$P_k(T) := \{q \in P(T)\ :\ d^\circ(q) \leq k\}.$$

Lorsque $k = 1$, on peut également utiliser la définition équivalente

$$P_1(T) := \{q \in P(T) \ : \ \exists a \in P_0(T),\, \exists \boldsymbol{b} \in (P_0(T))^d,\ \forall x \in T,\ q(\boldsymbol{x}) = a + \boldsymbol{b} \cdot \boldsymbol{x}\}.$$

On définit, pour $k \in \mathbb{N} \setminus \{0\}$, l'espace discret

$$V_{h,k}^+ := \{v_h \in C^0(\overline{\Omega}) \ : \ \forall \ell,\ v_{h|T_\ell} \in P_k(T_\ell)\}.$$

On parle [16] d'**élement fini de Lagrange d'ordre** k. Par construction, $V_{h,k}^+ \subset H^1(\Omega)$. Ensuite, on définit, pour $k \in \mathbb{N} \setminus \{0\}$, l'espace discret

$$V_{h,k} := \{v_h \in V_{h,k}^+ \ : \ v_{h|\partial\Omega} = 0\}. \tag{3.22}$$

Par construction, $V_{h,k} \subset H_0^1(\Omega)$. Dans la suite, on se concentre sur le cas $k = 1$, et on pose $V_h = V_{h,1}$. Les résultats obtenus ci-dessous se généralisent sans peine au cas $k \geq 2$ [11]. Pour construire une base de V_h, on choisit une famille de N fonctions $(w_j)_{1 \leq j \leq N}$:

$$\forall j,\ w_j \in V_h,\ \text{et } w_j(M_i) = \delta_{ij} \ \text{ pour } 1 \leq i \leq N,$$

où δ_{ij} est le **symbole de Kronecker**, c'est-à-dire que $\delta_{ij} = 1$ si $i = j$, et $\delta_{ij} = 0$ si $i \neq j$. On appelle $(w_j)_{j=1,N}$ les **fonctions "chapeau"**.

Exercice 3.2 *Montrer que si la deuxième ou la troisième propriété (3.21) est violée, alors on peut toujours trouver un sommet M_j pour lequel il n'existe pas de fonction "chapeau" associée.*

La propriété fondamentale de ces fonctions "chapeau" est que leur support est local, ce qui a une grande importance lorsqu'on construit la matrice $\mathbb{A}$ représentant la forme a_1 dans cette base.

Proposition 3.10 *On suppose les propriétés (3.21) vérifiées. Alors*

$$supp(w_i) = \cup_{\ell \ t.q. \ M_i \in T_\ell} T_\ell \quad pour\ 1 \leq i \leq N.$$

Démonstration : Soit un indice i donné. Considérons un simplexe T_ℓ. Deux cas peuvent se présenter :

- soit $M_i \notin T_\ell$, et alors w_i s'annule aux $d+1$ sommets de T_ℓ ; comme $w_{i|T_\ell}$ est affine dans le simplexe et que les $d+1$ sommets ne sont pas alignés (c'est la première propriété (3.21)), on en déduit que $w_{i|T_\ell} = 0$;
- soit $M_i \in T_\ell$, et alors w_i vaut 1 en M_i, et s'annule aux d autres sommets de T_ℓ ; si on note a la facette contenant ces autres sommets, on en déduit que $w_i(\boldsymbol{x}) \neq 0$ pour $\boldsymbol{x} \in T_\ell \setminus a$ puisque $w_{i|T_\ell}$ est affine dans le simplexe.

Le résultat suit. ◇

Remarque 3.11 *Soit w_i un fonction "chapeau". Si on reprend la démonstration ci-dessus (2ème item), on note que, pour tout simplexe T_ℓ tel que $M_i \in T_\ell$, la restriction $w_{i|T_\ell}$ coïncide avec la coordonnée barycentrique associée au sommet M_i dans T_ℓ. En effet, par définition, les $d+1$ coordonnées barycentriques $(\lambda_i^{T_\ell})_{i=1,d+1}$ dans T_ℓ sont affines, valent 1 en un sommet, et 0 en tous les autres.*

16. Lorsque $k = 0$, l'espace discret $V_{h,0}^+$ est réduit à l'ensemble des fonctions constantes sur Ω.

Ensuite, on prouve que ces fonctions engendrent l'espace discret V_h.

Proposition 3.12 *On suppose les propriétés (3.21) vérifiées. Alors la famille $(w_j)_{1\leq j\leq N}$ constitue une base de V_h. De plus, on a l'identité :*

$$\forall v_h \in V_h, \quad v_h(\boldsymbol{x}) = \sum_{j=1,N} v_h(M_j)\, w_j(\boldsymbol{x}) \quad \textit{pour tout } \boldsymbol{x} \in \overline{\Omega}. \tag{3.23}$$

Démonstration : Montrons que la famille $(w_j)_{1\leq j\leq N}$ est libre. Soit $(\alpha_j)_{j=\underline{1,N}} \in \mathbb{R}^N$ telle que $\sum_{j=1,N} \alpha_j\, w_j = 0$. On a donc $\sum_{j=1,N} \alpha_j\, w_j(\boldsymbol{x}) = 0$ pour tout $\boldsymbol{x} \in \overline{\Omega}$. On choisit $\boldsymbol{x} = M_i$, pour $i = 1, N$, ce qui donne

$$0 = \sum_{j=1,N} \alpha_j\, w_j(M_i) = \alpha_i.$$

Montrons maintenant que la famille $(w_j)_{1\leq j\leq N}$ est génératrice.
On remarque que si cette famille est génératrice, alors pour tout $v_h \in V_h$ il existe $(\alpha_j)_{j=1,N} \in \mathbb{R}^N$ tels que $v_h = \sum_{j=1,N} \alpha_j\, w_j$. Si on choisit $\boldsymbol{x} = M_i$, pour $i = 1, N$, on trouve $v_h(M_i) = \alpha_i$, et ainsi on a nécessairement $v_h = \sum_{j=1,N} v_h(M_j)\, w_j$ (c'est-à-dire (3.23)).
Ainsi pour prouver que $(w_j)_{1\leq j\leq N}$ est génératrice, pour tout $v_h \in V_h$, on doit prouver que $v_h - \sum_{j=1,N} v_h(M_j)\, w_j = 0$. Comme les fonctions sont continues sur $\overline{\Omega}$, ceci revient à prouver que pour tout simplexe T_ℓ, on a $v_h(\boldsymbol{x}) - \sum_{j=1,N} v_h(M_j)\, w_j(\boldsymbol{x}) = 0$ pour tout $\boldsymbol{x}$ dans T_ℓ. Par construction, $(v_h - \sum_{j=1,N} v_h(M_j)\, w_j)_{|T_\ell}$ est affine, et de plus elle s'annule aux $d+1$ sommets du simplexe qui sont non-alignés d'après la première propriété (3.21) : elle est donc nulle sur T_ℓ. Comme le résultat est valable pour tout simplexe, on a bien $v_h - \sum_{j=1,N} v_h(M_j)\, w_j = 0$ sur $\overline{\Omega}$. ◇

Les fonctions discrètes $v_h \in V_h$ sont donc caractérisées par les valeurs $(v_h(M_j))_{j=1,N}$. On appelle l'application $v_h \mapsto (v_h(M_j))_{j=1,N}$ les **degrés de liberté**.
On retient dorénavant $(w_j)_{j=1,N}$ comme base de V_h. Grâce à la proposition 3.10, on peut estimer le nombre d'éléments non-nuls de la matrice $\mathbb{A}_h$, noté $nnz(\mathbb{A}_h)$. Nous proposons des estimations "optimales" lorsque $d = 2$. Si $d = 3$, on peut obtenir des résultats similaires, voir la remarque 3.16 ci-après.

Proposition 3.13 *On suppose les propriétés (3.21) vérifiées et que $\Omega \subset \mathbb{R}^2$. Alors le nombre d'éléments non-nuls de la matrice $\mathbb{A}_h \in \mathbb{R}^{N_h \times N_h}$ est asymptotiquement au plus de l'ordre de $7N_h$:*

$$\forall h, \quad \frac{nnz(\mathbb{A}_h)}{N_h} < 7.$$

Remarque 3.14 *En d'autres termes, la matrice $\mathbb{A}_h$ associée possède, en moyenne, moins de 7 éléments non-nuls par ligne. Pour rappel, en 2D et pour tout h, les matrices type $\mathbb{A}'_2$ issues de la discrétisation par différences finies, voir (2.37), possèdent au plus 5 éléments non-nuls par ligne.*

Démonstration : On introduit A^+ le nombre d'arêtes du maillage : on décompose A^+ en $A+A_b$, où A_b est le nombre d'arêtes situées sur la frontière. Pour rappel, il y a $N^+ = N+N_b$ sommets, avec N_b sommets situés sur la frontière. Si on compte les arêtes frontière et les sommets frontière, on trouve $A_b = N_b$. Ensuite, on note A_{int} le nombre d'arêtes dont

aucune extrémité n'est située sur la frontière. Si on fait l'hypothèse (raisonnable) que de chaque sommet situé sur la frontière, il part au moins une arête vers l'intérieur, on a $A_{int} \leq A - N_b$. Et enfin, il y a L triangles.
Pour $1 \leq i, j \leq N$, on a :

$$\begin{aligned}
\mathbb{A}_{i,j} &= \int_\Omega (k\,\mathbf{grad}\, w_j \cdot \mathbf{grad}\, w_i + q\, w_j w_i)\, d\boldsymbol{x} \\
&= \sum_{\ell=1,L} \int_{int(T_\ell)} \left(k\,\mathbf{grad}\, w_{j|T_\ell} \cdot \mathbf{grad}\, w_{i|T_\ell} + q\, w_{j|T_\ell} w_{i|T_\ell}\right) d\boldsymbol{x} \qquad (3.24) \\
&= \sum_{\ell=1,L \ t.q.\ M_i, M_j \in T_\ell} \int_{int(T_\ell)} \left(k\,\mathbf{grad}\, w_{j|T_\ell} \cdot \mathbf{grad}\, w_{i|T_\ell} + q\, w_{j|T_\ell} w_{i|T_\ell}\right) d\boldsymbol{x},
\end{aligned}$$

où on a utilisé la proposition 3.10 pour la dernière égalité. En effet, si $M_i \notin T_\ell$ ou si $M_j \notin T_\ell$, l'intégrale sur $int(T_\ell)$ est nulle. Ainsi, pour avoir $\mathbb{A}_{i,j} \neq 0$, il est nécessaire que $M_i = M_j$, ou que $[M_i, M_j]$ soit un côté d'un des triangles $(T_\ell)_{1 \leq \ell \leq L}$. Il y a N occurences du premier cas, et $2A_{int}$ occurences du second cas. On a donc :

$$nnz(\mathbb{A}) \leq N + 2A_{int}.$$

Pour obtenir une borne qui ne dépende que de N, on doit maintenant borner A par N. Pour simplifier, on fait l'hypothèse que le domaine Ω est sans trous. Alors, si on décompte les arêtes, on trouve tout d'abord que :

$$3L = 2A + A_b.$$

Ensuite, si on compte les angles aux sommets, et si on se souvient que la somme des angles aux sommets d'un polygone à n côtés est égale à $(n-2)\pi$, on trouve :

$$L = 2N + N_b - 2.$$

Ainsi, $2A = 6N + 2N_b - 6$, et donc

$$2A_{int} = 2A_{int} - 2A + 6N + 2N_b - 6 = 6N + 2(A_{int} + N_b - A) - 6 < 6N,$$

et on en conclut que

$$nnz(\mathbb{A}) < 7N.$$

On en déduit le résultat annoncé. ◇

Remarque 3.15 *Plus généralement, on peut prouver que cette propriété est valable quel que soit l'ordre $k > 0$ de l'élément fini de Lagrange [11]. Si on note $N_{h,k}$ la dimension de $V_{h,k}$, et $\mathbb{A}_{h,k} \in \mathbb{R}^{N_{h,k} \times N_{h,k}}$ la matrice associée à la forme a_1 dans une base "bien choisie", alors :*

$$\exists C_k > 0, \quad \forall h, \quad \frac{nnz(\mathbb{A}_{h,k})}{N_{h,k}} \leq C_k.$$

Exercice 3.3 *On peut également résoudre numériquement le problème (3.1), où on a remplacé la condition aux limites de Dirichlet par une condition aux limites de type Neumann ou Robin. Dans ce cas, si on utilise un élément fini de Lagrange d'ordre 1, on se place dans* $V_h^+ = V_{h,1}^+$. *La matrice* $\mathbb{A}'_h$ *correspondante appartient alors à* $\mathbb{R}^{N_h^+ \times N_h^+}$, *où* $N_h^+ = dim(V_h^+)$. *Pour* $d = 2$, *montrer que l'on a* $nnz(\mathbb{A}'_h) \leq N_h^+ + 2A_h^+$, *où* A_h^+ *est le nombre d'arêtes du maillage ; en utilisant les bornes établies dans la proposition 3.13, en conclure que :*

$$\forall h, \quad \frac{nnz(\mathbb{A}'_h)}{N_h^+} < 7.$$

Remarque 3.16 *Lorsque* $d = 3$, *on peut arriver à des estimations similaires, afin de garantir un nombre d'éléments non-nuls de la matrice* $\mathbb{A}_h$ *croissant linéairement par rapport à sa dimension. Pour cela, on a besoin d'une relation supplémentaire de dénombrement, puisque la somme des angles solides au sommet d'un tétraèdre n'est pas constante. Elle s'obtient sous réserve que la famille de maillages soit régulière (voir (3.25) ci-après) : cette condition fournit une borne uniforme (indépendante de h) sur le nombre maximal de tétraèdres auxquels un sommet peut appartenir, cf. [11] pour les détails.*

Pour établir la propriété d'approximabilité minimale (C.26), on choisit $V_+ = \mathcal{D}(\Omega)$ comme sous-espace dense de $V = H_0^1(\Omega)$. Pour cela, on utilise un opérateur d'**interpolation** $\pi_h : V_+ \to V_h$ et le passage à l'élément fini de référence (voir par exemple [11, §2.3] pour les calculs détaillés).
La démarche est la suivante. Pour $v_+ \in V_+$, on écrit tout d'abord :

$$\|v_+ - \pi_h v_+\|_{H^1(\Omega)} = \left(\sum_\ell \|(v_+)_{|T_\ell} - (\pi_h v_+)_{|T_\ell}\|^2_{H^1(T_\ell)} \right)^{1/2}.$$

Puis on estime, pour chaque ℓ, l'écart $\|(v_+)_{|T_\ell} - (\pi_h v_+)_{|T_\ell}\|_{H^1(T_\ell)}$. Pour cela, on passe de T_ℓ au *simplexe de référence*, voir (3.29) ci-après ; on utilise les propriétés d'approximabilité de l'opérateur d'interpolation sur ce simplexe, en utilisant une mesure dans un espace de l'échelle de Sobolev $(H^{1+\beta})_{\beta \geq 0}$ (ces propriétés *ne dépendent donc pas* du simplexe T_ℓ) ; et enfin on revient au simplexe T_ℓ.
Pour obtenir une estimation exploitable, on se place dans la situation où

$$\exists \sigma > 0, \ \forall h, \ \forall \ell, \ h_\ell \leq \sigma \rho_\ell. \tag{3.25}$$

On dit que la famille de maillages $(\mathcal{T}_h)_h$ est **régulière**.
Donnons quelques détails. On choisit l'opérateur d'interpolation de Scott-Zhang [18, §1.6.2], utilisable pour tout élément de $H^1(\Omega)$, et donc en particulier pour tout élément de $H_0^1(\Omega)$. Si pour chaque T_ℓ, on note S_{T_ℓ} l'union des simplexes ayant une intersection non-vide avec T_ℓ, on obtient, tous calculs faits, une estimation du type

$$\begin{aligned} &\exists C > 0, \ \forall \beta \in [0,1], \ \forall h, \ \forall T_\ell \in \mathcal{T}_h, \\ &\|(v_+)_{|T_\ell} - (\pi_h v_+)_{|T_\ell}\|_{H^1(T_\ell)} \leq C h_\ell^\beta \, |(v_+)_{|S_{T_\ell}}|_{H^{1+\beta}(S_{T_\ell})}. \end{aligned} \tag{3.26}$$

Remarque 3.17 *Lorsque $\beta \in]1/2, 1]$ et puisque $d \leq 3$, on sait d'après les injections de Sobolev que $H^{1+\beta}(\Omega) \subset \mathcal{C}^0(\overline{\Omega})$ (voir par exemple [3, Chapitre 2]) et on peut "simplifier" l'estimation (3.26) en se servant de l'opérateur "classique" d'interpolation de Lagrange. Dans ce cas, on peut remplacer à droite $|(v_+)_{|S_{T_\ell}}|_{H^{1+\beta}(S_{T_\ell})}$ par $|(v_+)_{|T_\ell}|_{H^{1+\beta}(T_\ell)}$.*

Pour conclure que l'hypothèse d'approximabilité minimale (C.26) est vérifiée, il reste à sommer (3.26) sur tous les simplexes. Mais, dans l'estimation (3.26), la contribution sur un simplexe T' donné apparaît plusieurs fois à droite, puisque par définition ce simplexe est inclus dans tous les S_{T_ℓ} tels que $T' \cap T_\ell \neq \emptyset$.
Lorsque $d = 3$, on peut vérifier que l'hypothèse de régularité de la famille de maillages implique que la valeur minimale (prise sur tous les simplexes de tous les maillages) des angles solides aux sommets des simplexes est strictement positive, et de même pour la valeur minimale des angles dièdriques aux arêtes des simplexes. Et, lorsque $d = 2$, la valeur minimale (prise sur tous les triangles de tous les maillages) des angles aux sommets est strictement positive. Ceci implique à son tour que

$$\exists C_{reg} > 0,\ \forall h,\ \forall \ell, \quad \max_{T' \in \mathcal{T}_h} card(\{T'\, t.q.\, T \subset S_{T_\ell}\}) \leq C_{reg}, \tag{3.27}$$

c'est-à-dire que le nombre d'occurences de T' est uniformément borné.
Par sommation sur tous les simplexes de $\mathcal{T}_h$, on trouve donc que

$$\begin{aligned}
\|v_+ - \pi_h v_+\|_{H^1(\Omega)} &\leq \left(\sum_\ell C^2 h_\ell^{2\beta}\, |(v_+)_{|S_{T_\ell}}|^2_{H^{1+\beta}(S_{T_\ell})} \right)^{1/2} \\
\big\{\text{Pour tout } \ell,\, h_\ell \leq h\big\} &\leq Ch^\beta \left(\sum_\ell |(v_+)_{|S_{T_\ell}}|^2_{H^{1+\beta}(S_{T_\ell})} \right)^{1/2} \\
\big\{\text{Au plus } C_{reg} \text{ occurences}\big\} &\leq CC_{reg} h^\beta \left(\sum_\ell |(v_+)_{|T_\ell}|^2_{H^{1+\beta}(T_\ell)} \right)^{1/2}.
\end{aligned}$$

Si on a choisi $\beta \in \{0, 1\}$, on a $(\sum_\ell |(v_+)_{|T_\ell}|^2_{H^{1+\beta}(T_\ell)})^{1/2} = |v_+|_{H^{1+\beta}(\Omega)}$. Pour $\beta \in]0, 1[$, d'après la propriété de sous-additivité des semi-normes $|\cdot|_{H^{1+\beta}(\Omega)}$ (cf. [3, Chapitre 2]), on a $(\sum_\ell |(v_+)_{|T_\ell}|^2_{H^{1+\beta}(T_\ell)})^{1/2} \leq |v_+|_{H^{1+\beta}(\Omega)}$. Dans tous les cas, on en conclut que

$$\forall v_+ \in V_+, \quad \|v_+ - \pi_h v_+\|_{H^1(\Omega)} \leq CC_{reg} h^\beta\, |v_+|_{H^{1+\beta}(\Omega)}, \tag{3.28}$$

et finalement (C.26) est vérifiée. En conclusion, l'erreur $\|u - u_h\|_{H^1(\Omega)}$ tend vers 0 quand h tend vers 0, cf. (3.19).

Remarque 3.18 *La propriété (3.28) se généralise sans difficulté aux éléments de $C^\infty(\overline{\Omega})$, avec des opérateurs à valeurs dans $(V^+_{h,k})_h$. D'après la densité de $C^\infty(\overline{\Omega})$ dans $H^1(\Omega)$ et dans $L^2(\Omega)$, cf. Annexe D, on dispose d'une propriété d'approximabilité (C.26) permettant d'approcher tout élement de $H^1(\Omega)$; et on dispose également d'une propriété similaire pour tout élement de $L^2(\Omega)$ (puisque $\|v_+ - \pi_h v_+\|_{L^2(\Omega)} \leq \|v_+ - \pi_h v_+\|_{H^1(\Omega)}$).*

Nous donnons quelques précisions ci-dessous, concernant le passage à l'élément fini de référence, on note habituellement $\hat{T}$ le *simplexe de référence* dont :

— (cas $d = 2$) les sommets ont pour coordonnées $(0, 0)$, $(0, 1)$ et $(1, 0)$ dans le répère de référence $(0, \hat{x}, \hat{y})$;
— (cas $d = 3$) les sommets ont pour coordonnées $(0, 0, 0)$, $(0, 0, 1)$, $(0, 1, 0)$ et $(1, 0, 0)$ dans le répère de référence $(0, \hat{x}, \hat{y}, \hat{z})$.

On passe du simplexe $\hat{T}$ à un simplexe T_ℓ donné à l'aide d'une transformation affine

$$F_\ell \ : \ \left\{ \begin{array}{l} \hat{T} \to T_\ell \\ \hat{\boldsymbol{x}} \mapsto \boldsymbol{x} = \mathbb{A}_\ell \hat{\boldsymbol{x}} + \boldsymbol{b}_\ell \end{array} \right. , \quad \text{avec } \mathbb{A}_\ell \in \mathbb{R}^{d\times d} \text{ inversible, et } \boldsymbol{b}_\ell \in \mathbb{R}^d. \tag{3.29}$$

Bien sûr, on a la transformation affine inverse F_ℓ^{-1} de T_ℓ dans $\hat{T}$: $\hat{\boldsymbol{x}} = \mathbb{A}_\ell^{-1}\boldsymbol{x} - \mathbb{A}_\ell^{-1}\boldsymbol{b}_\ell$. Dans la suite, le passage à l'élément fini de référence est équivalent à ces correspondances entre $\hat{\boldsymbol{x}} \in \hat{T}$ et $\boldsymbol{x}_\ell \in T_\ell$. Puisque les transformations sont affines, les formules de changement de variables s'écrivent pour toute fonction $w \in L^1(T_\ell)$, $\hat{w} \in L^1(\hat{T})$:

$$\begin{array}{l} \displaystyle\int_{T_\ell} w(\boldsymbol{x})\, d\boldsymbol{x} = \int_{\hat{T}} (w \circ F_\ell)(\hat{\boldsymbol{x}})\, |det(\mathbb{A}_\ell)|\, d\hat{\boldsymbol{x}}\,; \\ \displaystyle\int_{\hat{T}} \hat{w}(\hat{\boldsymbol{x}})\, d\hat{\boldsymbol{x}} = \int_{T_\ell} (\hat{w} \circ F_\ell^{-1})(\boldsymbol{x})\, |det(\mathbb{A}_\ell)|^{-1}\, d\boldsymbol{x}. \end{array} \tag{3.30}$$

On transforme une fonction à valeurs scalaires $\hat{v}$ définie sur $\hat{T}$ en v à valeurs scalaires définie sur T_ℓ, et réciproquement, selon :

$$v = \hat{v} \circ F_\ell^{-1} \qquad \Longleftrightarrow \qquad \hat{v} = v \circ F_\ell. \tag{3.31}$$

On écrit $v(\boldsymbol{x}) = \hat{v}(\hat{\boldsymbol{x}})$ pour tout $\boldsymbol{x} \in T_\ell$ (ou tout $\hat{\boldsymbol{x}} \in \hat{T}$). On peut facilement établir les identités suivantes sur les gradients :

$$\forall \hat{\boldsymbol{x}} \in \hat{T}, \quad \nabla v(\boldsymbol{x}) = (\mathbb{A}_\ell^T)^{-1} \hat{\nabla} \hat{v}(\hat{\boldsymbol{x}})\,; \quad \forall \boldsymbol{x} \in T_\ell, \quad \hat{\nabla} \hat{v}(\hat{\boldsymbol{x}}) = \mathbb{A}_\ell^T \nabla v(\boldsymbol{x}). \tag{3.32}$$

Pour établir la propriété d'approximabilité uniforme (C.29) et estimer la vitesse de convergence, on choisit

$$\tilde{V} = \{\tilde{v} \in H_0^1(\Omega) \ : \ \mathrm{div}(k\ \mathbf{grad}\, \tilde{v}) \in L^2(\Omega)\},$$

et on utilise le résultat technique ci-dessous (voir [9]), sur la régularité *a priori* des solutions [17] du problème (3.1).

Proposition 3.19 *On suppose que les coefficients k et q appartiennent à $\mathcal{P}W^{1,\infty}(\Omega)$, et qu'ils vérifient les hypothèses (3.2). Alors il existe $r_{\max} \in]0, 1]$, appelé exposant de régularité, tel que pour tout $f \in L^2(\Omega)$, la solution $u \in H_0^1(\Omega)$ de (3.1) appartient à $\bigcap_{0\le r<r_{\max}} \mathcal{P}H^{1+r}(\Omega)$ ($r_{\max} < 1$), ou $\mathcal{P}H^2(\Omega)$ ($r_{\max} = 1$), avec une dépendance continue.*

— *Si $r_{\max} < 1$:*
$\forall r \in [0, r_{\max}[$, $\exists C_r > 0$, $\forall f \in L^2(\Omega)$, $\|u\|_{PH^{1+r}(\Omega)} \le C_r \|f\|_{L^2(\Omega)}$;
— *Si $r_{\max} = 1$:*
$\exists C_1 > 0$, $\forall f \in L^2(\Omega)$, $\|u\|_{PH^2(\Omega)} \le C_1 \|f\|_{L^2(\Omega)}$.

17. Le préfixe $\mathcal{P}$ renvoie à une partition $\mathcal{P} := \{\Omega_p\}_{p=1,\cdots,P}$ de Ω : $(\Omega_p)_{p=1,\cdots,P}$ sont des domaines disjoints, et $\overline{\Omega} = \cup_{p=1,P}\overline{\Omega_p}$. Dans ce cas, $k \in \mathcal{P}W^{1,\infty}(\Omega)$ signifie que $k_{|\Omega_p} \in W^{1,\infty}(\Omega_p)$, pour $1 \le p \le P$. De même, $u \in \mathcal{P}H^{1+r}(\Omega)$ signifie que $u_{|\Omega_p} \in H^{1+r}(\Omega_p)$, pour $1 \le p \le P$.

Définition 3.20 *On dit qu'une famille de maillages $(\mathcal{T}_h)_h$ de Ω est conforme à une partition $(\Omega_p)_{1\leq p\leq P}$ de Ω si*

$$\forall h,\ \forall \ell,\ \exists p \in \{1,\dots,P\} \text{ tel que } T_\ell \subset \overline{\Omega_p}.$$

Encore une fois par passage à l'élément fini de référence et toujours à l'aide d'un opérateur d'interpolation appliqué cette fois aux éléments de $\tilde{V}$, on en déduit le résultat suivant sur l'ordre de convergence (voir la définition 3.8).

Théorème 3.21 *(**problème à une inconnue**) Sous les hypothèses de la proposition 3.19 et si la famille de maillage est régulière et conforme, alors l'ordre de convergence α_{cv} est égal à :*

- *$\alpha_{cv} = r_{\max} - \epsilon$ pour tout $\epsilon \in]0, r_{\max}[$ si $r_{\max} < 1$, ou*
- *$\alpha_{cv} = 1$ si $r_{\max} = 1$.*

Qui plus est, la constante C_{cv} ne dépend ni de la solution u, ni de la donnée f :

$$\exists C_{cv} > 0,\ \forall h,\ \forall f \in L^2(\Omega), \quad \|u - u_h\|_{H^1(\Omega)} \leq C_{cv}\, h^{\alpha_{cv}}\, \|f\|_{L^2(\Omega)}.$$

Démonstration : On choisit $r \in]0,1]$ comme à la proposition 3.19, où le cas $r = 1$ correspond à $r_{\max} = 1$.
Alors, d'après (3.28), [18] on a la première estimation

$$\inf_{v_h \in V_h} \|u - v_h\|_{H^1(\Omega)} \leq \|u - \pi_h u\|_{H^1(\Omega)} \leq C h^r\, |u|_{PH^{1+r}(\Omega)},$$

avec $C > 0$ indépendante de h et de u. Puis on applique l'estimation de la proposition 3.19, à savoir $|u|_{PH^{1+r}(\Omega)} \leq \|u\|_{PH^{1+r}(\Omega)} \leq C_r\, \|f\|_{L^2(\Omega)}$. L'utilisation du lemme de Céa, voir (3.18), donne finalement le résultat avec $\alpha_{cv} = r$. ◇

3.3.2 Problème à deux inconnues

On suppose ici que l'on dispose d'une famille $(\mathcal{V}_h)_h$ de sous-espaces vectoriels de dimension finie de $\mathcal{V} = L^2(\Omega) \times \mathbf{H}(\mathrm{div}, \Omega)$. Comme $\mathcal{V}$ est un espace produit, on choisit une discrétisation produit, c'est-à-dire de la forme $\mathcal{V}_h = M_h \times \boldsymbol{Q}_h$. On munit les sous-espaces M_h de la norme $\|\cdot\|_{L^2(\Omega)}$, les sous-espaces $\boldsymbol{Q}_h$ de la norme $\|\cdot\|_{\mathbf{H}(\mathrm{div},\Omega)}$, et enfin les espaces $\mathcal{V}_h$ de la norme produit $\|\cdot\|_{\mathcal{V}}$.

Soit h donné. Dans $M_h \times \boldsymbol{Q}_h$, la **formulation variationnelle discrète** associée à (3.10) est :

$$\begin{cases} \text{Trouver } (u_h, \boldsymbol{p}_h) \in M_h \times \boldsymbol{Q}_h \text{ tel que} \\ \forall (v_h, \boldsymbol{q}_h) \in M_h \times \boldsymbol{Q}_h, \quad a_2((u_h, \boldsymbol{p}_h), (v_h, \boldsymbol{q}_h)) = \displaystyle\int_\Omega f v_h \, d\boldsymbol{x}, \end{cases} \tag{3.33}$$

où la forme bilinéaire a_2 est définie en (3.11). La solution discrète $(u_h, \boldsymbol{p}_h)$ est définie presque partout, comme pour la discrétisation du problème à une inconnue (cf. remarque 3.3). Et puisqu'elle dépend linéairement de la donnée f, on peut réexprimer (3.33) sous la forme d'un système linéaire.

18. A l'aide, lorsque $r > 1/2$, de la remarque 3.17 qui permet de découpler les contributions sur les différents Ω_p.

Lemme 3.22 *Soient $(\phi_j)_{1\leq j\leq N_u}$ une base de l'espace vectoriel M_h et $(\boldsymbol{\kappa}_m)_{1\leq m\leq N_{\boldsymbol{p}}}$ une base de l'espace vectoriel $\boldsymbol{Q}_h$. Résoudre la formulation variationnelle discrète (3.33) est équivalent à résoudre le système linéaire posé dans $\mathbb{R}^{N_u+N_{\boldsymbol{p}}}$:*

$$\left\{\begin{array}{l} \text{Trouver } X\in\mathbb{R}^{N_u+N_{\boldsymbol{p}}} \text{ tel que} \\ \mathbb{A}X=F \end{array}\right. , \tag{3.34}$$

avec

$$\left\{\begin{array}{l} \mathbb{A}=\begin{pmatrix}\mathbb{M}_u & \mathbb{B}^T\\ \mathbb{B} & -\mathbb{M}_{\boldsymbol{p}}\end{pmatrix}\in\mathbb{R}^{(N_u+N_{\boldsymbol{p}})\times(N_u+N_{\boldsymbol{p}})},\ X=\begin{pmatrix}U\\P\end{pmatrix},\ F=\begin{pmatrix}F_u\\0\end{pmatrix}\in\mathbb{R}^{N_u+N_{\boldsymbol{p}}},\\ (\mathbb{M}_u)_{i,j}=\displaystyle\int_\Omega q\phi_j\phi_i\,d\boldsymbol{x},\ 1\leq i,j\leq N_u,\quad (F_u)_i=\int_\Omega f\phi_i\,d\boldsymbol{x},\ 1\leq i\leq N_u,\\ (\mathbb{M}_{\boldsymbol{p}})_{l,m}=\displaystyle\int_\Omega k^{-1}\boldsymbol{\kappa}_m\cdot\boldsymbol{\kappa}_l\,d\boldsymbol{x},\ 1\leq l,m\leq N_{\boldsymbol{p}},\\ \mathbb{B}_{l,j}=\displaystyle\int_\Omega \phi_j\,\mathrm{div}\,\boldsymbol{\kappa}_l\,d\boldsymbol{x},\ 1\leq l\leq N_{\boldsymbol{p}},\ 1\leq j\leq N_u, \end{array}\right. \tag{3.35}$$

et l'on a la correspondance

$$u_h=\sum_{j=1}^{N_u}U_j\phi_j\qquad \boldsymbol{p}_h=\sum_{m=1}^{N_{\boldsymbol{p}}}P_m\boldsymbol{\kappa}_m. \tag{3.36}$$

Remarque 3.23 *Encore une fois et sauf exception, on omet les dépendances en h.*

Démonstration : Soit $(u_h,\boldsymbol{p}_h)$ résolvant (3.33) : on note $(X_j)_{1\leq j\leq N_u}$ les composantes de u_h dans la base $(\phi_j)_{1\leq j\leq N_u}$, et $(P_m)_{1\leq j\leq N_{\boldsymbol{p}}}$ les composantes de $\boldsymbol{p}_h$ dans la base $(\boldsymbol{\kappa}_m)_{1\leq m\leq N_{\boldsymbol{p}}}$, c'est-à-dire (3.36). On choisit $(v_h,\boldsymbol{q}_h)=(\phi_i,0)$ dans la formulation variationnelle (3.33), ce qui nous donne avec la définition (3.36) des vecteurs U et P, et la définition (3.35) de $\mathbb{M}_u$, $\mathbb{B}$ et F_u :

$$\begin{aligned}\forall i,\quad \sum_{j=1}^{N_u}(\mathbb{M}_u)_{i,j}U_j+\sum_{m=1}^{N_{\boldsymbol{p}}}(\mathbb{B}^T)_{i,m}P_m&=\sum_{j=1}^{N_u}(\mathbb{M}_u)_{i,j}U_j+\sum_{m=1}^{N_{\boldsymbol{p}}}\mathbb{B}_{m,i}P_m\\ &=\sum_{j=1}^{N_u}U_j\int_\Omega q\phi_j\phi_i\,d\boldsymbol{x}+\sum_{m=1}^{N_{\boldsymbol{p}}}P_m\int_\Omega\phi_i\,\mathrm{div}\,\boldsymbol{\kappa}_m\,d\boldsymbol{x}\\ &=\int_\Omega q(\sum_{j=1}^{N_u}U_j\phi_j)\phi_i\,d\boldsymbol{x}+\int_\Omega\phi_i\,\mathrm{div}(\sum_{m=1}^{N_{\boldsymbol{p}}}P_m\boldsymbol{\kappa}_m)\,d\boldsymbol{x}\\ &=\int_\Omega qu_h\phi_i\,d\boldsymbol{x}+\int_\Omega\phi_i\,\mathrm{div}\,\boldsymbol{p}_h\,d\boldsymbol{x}\overset{(3.33)_{(\phi_i,0)}}{=}\int_\Omega f\phi_i\,d\boldsymbol{x}=(F_u)_i,\end{aligned}$$

soit les N_u premières lignes de (3.34), $\mathbb{M}_uU+\mathbb{B}^TP=F_u$, écrit ligne par ligne.
On choisit ensuite $(v_h,\boldsymbol{q}_h)=(0,\boldsymbol{\kappa}_l)$ dans la formulation variationnelle (3.33), ce qui nous

donne avec la définition (3.36) des vecteurs U et P, et la définition (3.35) de $\mathbb{M}_{\boldsymbol{p}}$ et $\mathbb{B}$:

$$\begin{aligned}\forall l, \quad & \sum_{j=1}^{N_u} \mathbb{B}_{l,j} U_j - \sum_{m=1}^{N_{\boldsymbol{p}}} (\mathbb{M}_{\boldsymbol{p}})_{l,m} P_m \\ &= \sum_{j=1}^{N_u} U_j \int_\Omega \phi_j \operatorname{div} \boldsymbol{\kappa}_l \, d\boldsymbol{x} + \sum_{m=1}^{N_{\boldsymbol{p}}} P_m \int_\Omega -k^{-1} \boldsymbol{\kappa}_m \cdot \boldsymbol{\kappa}_l \, d\boldsymbol{x} \\ &= \int_\Omega (\sum_{j=1}^{N_u} U_j \phi_j) \operatorname{div} \boldsymbol{\kappa}_l \, d\boldsymbol{x} + \int_\Omega -k^{-1} (\sum_{m=1}^{N_{\boldsymbol{p}}} P_m \boldsymbol{\kappa}_m) \cdot \boldsymbol{\kappa}_l \, d\boldsymbol{x} \\ &= \int_\Omega u_h \operatorname{div} \boldsymbol{\kappa}_l \, d\boldsymbol{x} + \int_\Omega -k^{-1} \boldsymbol{p}_h \cdot \boldsymbol{\kappa}_l \, d\boldsymbol{x} \overset{(3.33)_{(0,\boldsymbol{\kappa}_l)}}{=} 0,\end{aligned}$$

soit les $N_{\boldsymbol{p}}$ dernières lignes de (3.34), $\mathbb{B}U - \mathbb{M}_{\boldsymbol{p}} P = 0$, écrit ligne par ligne.
Réciproquement, à partir de X une solution de (3.34), on définit $u_h \in M_h$ et $\boldsymbol{p}_h \in \boldsymbol{Q}_h$ selon (3.36), et on choisit $(v_h, \boldsymbol{q}_h) = (\sum_{i=1}^{N_u} Y_i \phi_i, \sum_{l=1}^{N_{\boldsymbol{p}}} Q_l \boldsymbol{\kappa}_l)$ un élément quelconque de $M_h \times \boldsymbol{Q}_h$. On trouve successivement :

$$\begin{aligned}& a_2((u_h, \boldsymbol{p}_h), (v_h, 0)) \\ &= \int_\Omega \Big((\sum_{i=1}^{N_u} Y_i \phi_i) \operatorname{div} (\sum_{m=1}^{N_{\boldsymbol{p}}} P_m \boldsymbol{\kappa}_m) + q \, (\sum_{j=1}^{N_u} U_j \phi_j)(\sum_{i=1}^{N_u} Y_i \phi_i) \Big) d\boldsymbol{x} \\ &= \sum_{i=1}^{N_u} Y_i \left(\int_\Omega \Big(\phi_i \operatorname{div} (\sum_{m=1}^{N_{\boldsymbol{p}}} P_m \boldsymbol{\kappa}_m) + q \, (\sum_{j=1}^{N_u} U_j \phi_j) \phi_i \Big) d\boldsymbol{x} \right) \\ &\overset{(3.35)}{=} \sum_{i=1}^{N_u} Y_i \Big(\sum_{m=1}^{N_{\boldsymbol{p}}} (\mathbb{B}^T)_{i,m} P_m + \sum_{j=1}^{N_u} (\mathbb{M}_u)_{i,j} U_j \Big) = \sum_{i=1}^{N_u} Y_i (\mathbb{B}^T P + \mathbb{M}_u U)_i \\ &\overset{(3.34)}{=} \sum_{i=1}^{N_u} Y_i (F_u)_i \overset{(3.35)}{=} \sum_{i=1}^{N_u} Y_i \int_\Omega f \phi_i \, d\boldsymbol{x} = \int_\Omega f \Big(\sum_{i=1}^{N_u} Y_i \phi_i \Big) \, d\boldsymbol{x} = \int_\Omega f v_h \, d\boldsymbol{x} \, ;\end{aligned}$$

$$\begin{aligned}& a_2((u_h, \boldsymbol{p}_h), (0, \boldsymbol{q}_h)) \\ &= \int_\Omega \Big(- k^{-1} (\sum_{m=1}^{N_{\boldsymbol{p}}} P_m \boldsymbol{\kappa}_m) \cdot (\sum_{l=1}^{N_{\boldsymbol{p}}} Q_l \boldsymbol{\kappa}_l) + (\sum_{j=1}^{N_u} U_j \phi_j) \operatorname{div} (\sum_{l=1}^{N_{\boldsymbol{p}}} Q_l \boldsymbol{\kappa}_l) \Big) d\boldsymbol{x} \\ &= \sum_{l=1}^{N_{\boldsymbol{p}}} Q_l \left(\int_\Omega \Big(- k^{-1} (\sum_{m=1}^{N_{\boldsymbol{p}}} P_m \boldsymbol{\kappa}_m) \cdot \boldsymbol{\kappa}_l + (\sum_{j=1}^{N_u} U_j \phi_j) \operatorname{div} \boldsymbol{\kappa}_l \Big) d\boldsymbol{x} \right) \\ &\overset{(3.35)}{=} \sum_{l=1}^{N_{\boldsymbol{p}}} Q_l \Big(\sum_{m=1}^{N_{\boldsymbol{p}}} (-\mathbb{M}_{\boldsymbol{p}})_{l,m} P_m + \sum_{j=1}^{N_u} \mathbb{B}_{l,j} U_j \Big) = \sum_{l=1}^{N_{\boldsymbol{p}}} Q_l (-\mathbb{M}_{\boldsymbol{p}} P + \mathbb{B} U)_l \overset{(3.34)}{=} 0.\end{aligned}$$

On conclut par linéarité par rapport au second argument de $a_2(\cdot, \cdot)$:

$$a_2((u_h, \boldsymbol{p}_h), (v_h, \boldsymbol{q}_h)) = a_2((u_h, \boldsymbol{p}_h), (v_h, 0)) + a_2((u_h, \boldsymbol{p}_h), (0, \boldsymbol{q}_h)) = \int_\Omega f v_h \, d\boldsymbol{x},$$

c'est-à-dire la formulation variationnelle discrète (3.33). ◇

Supposons à partir de maintenant qu'une base $(\phi_j)_{1\leq j\leq N_u}$ de M_h, respectivement une base $(\boldsymbol{\kappa}_m)_{1\leq m\leq N_{\boldsymbol{p}}}$ de $\boldsymbol{Q}_h$, sont données. On définit $\mathbb{M}_u$, $\mathbb{B}$, $\mathbb{M}_{\boldsymbol{p}}$ et $\mathbb{A}$, F_u et F à l'aide de (3.35). Comme dans le cas du problème à une inconnue, on vérifie successivement que :

$$\forall v_h=\sum_{j=1}^{N_u} Y_j\phi_j,\ \forall w_h=\sum_{j=1}^{N_u} Z_j\phi_j,\qquad \int_\Omega q\, v_h w_h\, d\boldsymbol{x}=(\mathbb{M}_u Y,Z)_{N_u}, \tag{3.37}$$

$$\forall \boldsymbol{q}_h=\sum_{m=1}^{N_{\boldsymbol{p}}} Q_m\boldsymbol{\kappa}_m,\ \forall \boldsymbol{r}_h=\sum_{m=1}^{N_{\boldsymbol{p}}} R_m\boldsymbol{\kappa}_m,\quad \int_\Omega k^{-1}\boldsymbol{q}_h\cdot\boldsymbol{r}_h\, d\boldsymbol{x}=(\mathbb{M}_{\boldsymbol{p}}Q,R)_{N_{\boldsymbol{p}}}, \tag{3.38}$$

$$\forall v_h=\sum_{j=1}^{N_u} Y_j\phi_j,\ \forall \boldsymbol{q}_h=\sum_{m=1}^{N_{\boldsymbol{p}}} Q_m\boldsymbol{\kappa}_m,\qquad \int_\Omega v_h\,\mathrm{div}\,\boldsymbol{q}_h\, d\boldsymbol{x}=(\mathbb{B}Y,Q)_{N_{\boldsymbol{p}}}. \tag{3.39}$$

Bien sûr, $(\mathbb{B}Y,Q)_{N_{\boldsymbol{p}}}=(\mathbb{B}^TQ,Y)_{N_u}$. Enfin, on a

$$\left\{\begin{array}{l} \forall v_h=\displaystyle\sum_{j=1}^{N_u} Y_j\phi_j,\ \forall w_h=\sum_{j=1}^{N_u} Z_j\phi_j,\ \forall \boldsymbol{q}_h=\sum_{m=1}^{N_{\boldsymbol{p}}} Q_m\boldsymbol{\kappa}_m,\ \forall \boldsymbol{r}_h=\sum_{m=1}^{N_{\boldsymbol{p}}} R_m\boldsymbol{\kappa}_m,\\ a_2((v_h,\boldsymbol{q}_h),(w_h,\boldsymbol{r}_h))\\ \qquad=\left(\mathbb{A}\begin{pmatrix}Y\\Q\end{pmatrix},\begin{pmatrix}Z\\R\end{pmatrix}\right)_{N_u+N_{\boldsymbol{p}}}\\ \qquad=(\mathbb{M}_uY,Z)_{N_u}+(\mathbb{B}^TQ,Z)_{N_u}+(\mathbb{B}Y,R)_{N_{\boldsymbol{p}}}-(\mathbb{M}_{\boldsymbol{p}}Q,R)_{N_{\boldsymbol{p}}}. \end{array}\right. \tag{3.40}$$

Pour le problème à deux inconnues, on sait que la matrice $\mathbb{A}$ n'est ni définie-positive, ni définie-négative. En effet, en reprenant les notations ci-dessus :

$$\left(\mathbb{A}\begin{pmatrix}0\\Q\end{pmatrix},\begin{pmatrix}0\\Q\end{pmatrix}\right)_{N_u+N_{\boldsymbol{p}}}=a_2((0,\boldsymbol{q}_h),(0,\boldsymbol{q}_h))=-\int_\Omega k^{-1}|\boldsymbol{q}_h|^2d\boldsymbol{x}<0 \text{ si } Q\neq 0,$$

$$\left(\mathbb{A}\begin{pmatrix}Y\\0\end{pmatrix},\begin{pmatrix}Y\\0\end{pmatrix}\right)_{N_u+N_{\boldsymbol{p}}}=a_2((v_h,0),(v_h,0))=\int_\Omega q|v_h|^2d\boldsymbol{x}>0 \text{ si } Y\neq 0.$$

Proposition 3.24 *La matrice $\mathbb{A}$ est symétrique par construction. Sous les hypothèses (3.9), elle est en outre inversible.*

Démonstration : Comme la forme a_2 est symétrique, la matrice $\mathbb{A}$ l'est aussi.
Ensuite, on propose une méthode qui permet de résoudre directement le système (3.34) pour un second membre $F=\begin{pmatrix}F_u\\F_{\boldsymbol{p}}\end{pmatrix}\in\mathbb{R}^{N_u+N_{\boldsymbol{p}}}$ quelconque. Précisément, $X=\begin{pmatrix}U\\P\end{pmatrix}\in\mathbb{R}^{N_u+N_{\boldsymbol{p}}}$ est solution de (3.34) si, et seulement si,

$$\mathbb{M}_uU+\mathbb{B}^TP=F_u \text{ et } \mathbb{B}U-\mathbb{M}_{\boldsymbol{p}}P=F_{\boldsymbol{p}}.$$

Or, sous les hypothèses (3.9), $\mathbb{M}_{\boldsymbol{p}}$ est inversible car elle est (symétrique) définie-positive. La résolution précédente est donc équivalente à :

$$\begin{array}{ll} & \mathbb{M}_uU+\mathbb{B}^TP=F_u \text{ et } P=\mathbb{M}_{\boldsymbol{p}}^{-1}(\mathbb{B}U-F_{\boldsymbol{p}})\\ \Longleftrightarrow & \mathbb{M}_uU+\mathbb{B}^T(\mathbb{M}_{\boldsymbol{p}}^{-1}(\mathbb{B}U-F_{\boldsymbol{p}}))=F_u \text{ et } P=\mathbb{M}_{\boldsymbol{p}}^{-1}(\mathbb{B}U-F_{\boldsymbol{p}})\\ \Longleftrightarrow & (\mathbb{M}_u+\mathbb{B}^T\mathbb{M}_{\boldsymbol{p}}^{-1}\mathbb{B})U=F_u+\mathbb{B}^T\mathbb{M}_{\boldsymbol{p}}^{-1}F_{\boldsymbol{p}} \text{ et } P=\mathbb{M}_{\boldsymbol{p}}^{-1}(\mathbb{B}U-F_{\boldsymbol{p}}). \end{array}$$

En éliminant P, on construit le **complément de Schur**, égal ici à $\mathbb{S}_u = \mathbb{M}_u + \mathbb{B}^T\mathbb{M}_{\boldsymbol{p}}^{-1}\mathbb{B}$, ce qui permet de résoudre le problème en U. En effet, sous les hypothèses (3.9), $\mathbb{M}_u$ est (symétrique) définie-positive, et le complément de Schur $\mathbb{S}_u$ est lui aussi (symétrique) définie-positif : il est en particulier inversible. On en conclut qu'on peut déterminer la solution de (3.34) en prenant successivement

$$U = {\mathbb{S}_u}^{-1}(F_u + \mathbb{B}^T\mathbb{M}_{\boldsymbol{p}}^{-1}F_{\boldsymbol{p}}), \text{ puis } P = \mathbb{M}_{\boldsymbol{p}}^{-1}(\mathbb{B}U - F_{\boldsymbol{p}}).$$

La matrice $\mathbb{A}$ est donc inversible. ◇

Remarque 3.25 *Pour résoudre le système (3.34), on peut également raisonner en éliminant U, c'est-à-dire qu'on écrit $U = \mathbb{M}_u^{-1}(F_u - \mathbb{B}^T P)$, puis on construit $\mathbb{S}_{\boldsymbol{p}} = -(\mathbb{M}_{\boldsymbol{p}} + \mathbb{B}\mathbb{M}_u^{-1}\mathbb{B}^T)$, et on a cette fois $P = \mathbb{S}_{\boldsymbol{p}}^{-1}(F_{\boldsymbol{p}} - \mathbb{B}\mathbb{M}_u^{-1}F_u)$, puis $U = \mathbb{M}_u^{-1}(F_u - \mathbb{B}^T P)$.*

Il s'ensuit que la formulation variationnelle discrète (3.33) admet une solution $(u_h, \boldsymbol{p}_h)$ et une seule. Pour comparer cette solution discrète à la solution exacte $(u, \boldsymbol{p})$, on utilise le premier lemme de Strang, voir le théorème C.39. [19] Il faut pour cela que la forme a_2 soit uniformément $\mathcal{V}_h \times \mathcal{V}_h$ stable au sens de la condition (C.44). La difficulté est qu'en général, *on ne peut pas déduire* cette condition discrète de la condition de stabilité exacte (C.7). En effet, dans (C.44) le sup est pris dans $\mathcal{V}_h$, alors qu'il est pris dans $\mathcal{V}$ dans (C.7) : or pour tout h, $\mathcal{V}_h$ est *strictement* inclus dans $\mathcal{V}$. Néanmoins, dans le cas du problème à deux inconnues, on peut reproduire au niveau discret la construction de l'élément qui permet d'obtenir la stabilité exacte, en s'inspirant de la démonstration du théorème 3.2 "en ajoutant des $_h$".

Théorème 3.26 *Sous les hypothèses (3.9), et si les espaces discrets $(M_h)_h$ et $(\boldsymbol{Q}_h)_h$ vérifient la condition de compatibilité*

$$\forall h,\ \forall \boldsymbol{q}_h \in \boldsymbol{Q}_h,\ q^{-1}\operatorname{div}\boldsymbol{q}_h \in M_h, \tag{3.41}$$

alors la forme a_2 est uniformément $\mathcal{V}_h \times \mathcal{V}_h$ stable :

$$\exists \underline{\alpha} > 0,\ \forall h,\ \forall (v_h, \boldsymbol{q}_h) \in \mathcal{V}_h,\ \sup_{(w_h, \boldsymbol{r}_h)\in\mathcal{V}_h\setminus\{0\}} \frac{|a_2((v_h, \boldsymbol{q}_h), (w_h, \boldsymbol{r}_h))|}{\|(w_h, \boldsymbol{r}_h)\|_{\mathcal{V}}} \geq \underline{\alpha}\, \|(v_h, \boldsymbol{q}_h)\|_{\mathcal{V}}. \tag{3.42}$$

Démonstration : Pour tout $(v_h, \boldsymbol{q}_h) \in \mathcal{V}_h$ non-nul, on doit trouver un représentant $(w_h^*, \boldsymbol{r}_h^*) \in \mathcal{V}_h$ permettant de vérifier la condition (3.42). On choisit $\boldsymbol{r}_h^* = -\boldsymbol{q}_h$, et $w_h^* = \frac{1}{2}(v_h + q^{-1}\operatorname{div}\boldsymbol{q}_h)$. D'après la condition de compatibilité (3.41), on a $w_h^* \in M_h$. En outre

$$a_2((v_h, \boldsymbol{q}_h), (\frac{1}{2}(v_h + q^{-1}\operatorname{div}\boldsymbol{q}_h), -\boldsymbol{q}_h)) = \int_\Omega \left(k^{-1}|\boldsymbol{q}_h|^2 + \frac{1}{2}q^{-1}|\operatorname{div}\boldsymbol{q}_h|^2 + \frac{1}{2}\, q\, |v_h|^2 \right) d\boldsymbol{x},$$

et on en déduit que

$$a_2((v_h, \boldsymbol{q}_h), (w_h^*, \boldsymbol{r}_h^*)) \geq \min((k_{max})^{-1}, \frac{1}{2}(q_{max})^{-1}, \frac{1}{2}q_{min})\, \|(v_h, \boldsymbol{q}_h)\|_{\mathcal{V}}^2.$$

Comme dans le cas exact, on obtient la majoration

$$\|(w_h^*, \boldsymbol{r}_h^*)\|_{\mathcal{V}} \leq (1 + \frac{1}{2}(q_{min})^{-2})^{1/2}\, \|(v_h, \boldsymbol{q}_h)\|_{\mathcal{V}},$$

19. La forme a_2 n'étant pas coercive, on ne peut pas utiliser le lemme de Céa.

et ainsi on a une condition uniforme, valable pour tout h et tout $(v_h, \boldsymbol{q}_h) \in \mathcal{V}_h$,

$$\sup_{(w_h,\boldsymbol{r}_h)\in\mathcal{V}\setminus\{0\}} \frac{|a_2((v_h,\boldsymbol{q}_h),(w_h,\boldsymbol{r}_h))|}{\|(w_h,\boldsymbol{r}_h)\|_{\mathcal{V}}} \geq \frac{\min((k_{max})^{-1}, \frac{1}{2}(q_{max})^{-1}, \frac{1}{2}q_{min})}{(1+\frac{1}{2}(q_{min})^{-2})^{1/2}} \|(v_h,\boldsymbol{q}_h)\|_{\mathcal{V}}.$$

C'est la condition d'uniforme $\mathcal{V}_h \times \mathcal{V}_h$ stabilité (3.42) avec

$$\underline{\alpha} = \frac{\min((k_{max})^{-1}, \frac{1}{2}(q_{max})^{-1}, \frac{1}{2}q_{min})}{(1+\frac{1}{2}(q_{min})^{-2})^{1/2}}.$$

$\diamond$

Plutôt que de démontrer la condition d'uniforme $\mathcal{V}_h \times \mathcal{V}_h$ stabilité, on peut passer par la théorie de la T-coercivité discrète uniforme, voir le théorème C.37. La démarche consiste à nouveau à transposer la démonstration du cas exact (voir la fin de §3.2.2) au cadre discret. Pour tout $(v_h, \boldsymbol{q}_h) \in \mathcal{V}_h \setminus \{(0,0)\}$, on cherche en effet $(w_h^*, \boldsymbol{r}_h^*) \in \mathcal{V}_h$ *dépendant linéairement de* $(v_h, \boldsymbol{q}_h)$ et tel que

$$a_2((v_h,\boldsymbol{q}_h),(w_h^*,\boldsymbol{r}_h^*)) \geq \alpha^\star \|(v_h,\boldsymbol{q}_h)\|_{\mathcal{V}}^2,$$

avec $\alpha^\star > 0$ indépendant de h et $(v_h, \boldsymbol{q}_h)$. A partir de là, on choisit $\mathbb{T}_h \in \mathcal{L}(\mathcal{V}_h)$ défini par $\mathbb{T}_h((v_h,\boldsymbol{q}_h)) = (w_h^*, \boldsymbol{r}_h^*)$. De plus, on veut que $|||\mathbb{T}_h||| \leq \beta^\star$ avec $\beta^\star > 0$ indépendant de h (cf. définition C.35). L'idée sous-jacente à la T-coercivité discrète est que, si on peut transposer le cas exact en ajoutant des $_h$, alors la T-coercivité discrète uniforme suit immédiatement puisque la définition est indépendante de h ! Dans notre cas, la transposition de (3.12) donne la définition :

$$\mathbb{T}_h((v_h,\boldsymbol{q}_h)) = (\frac{1}{2}(v_h + q^{-1}\operatorname{div}\boldsymbol{q}_h), -\boldsymbol{q}_h).$$

La difficulté est qu'il faut que $\mathbb{T}_h((v_h,\boldsymbol{q}_h)) \in \mathcal{V}_h$, ce qui impose des contraintes sur le choix des espaces $(M_h)_h$ et $(\boldsymbol{Q}_h)_h$. Or, sous la condition de compatibilité (3.41), la propriété $\mathbb{T}_h((v_h,\boldsymbol{q}_h)) \in \mathcal{V}_h$ suit automatiquement : on a donc établi la T-coercivité discrète uniforme.

Pour le problème à deux inconnues, et sous les hypothèses du théorème 3.26, on laisse finalement h *varier* : on sait qu'il existe des solutions approchées $(u_h, \boldsymbol{p}_h)_h$. D'après le premier lemme de Strang énoncé au théorème C.39, pour que l'erreur tende vers 0 quand h tend vers 0, il suffit d'avoir la propriété d'approximabilité minimale (C.26) dans $\mathcal{V} = L^2(\Omega) \times \mathbf{H}(\operatorname{div},\Omega)$. On réécrit ci-dessous le résultat de convergence pour le problème à deux inconnues.

Théorème 3.27 *Sous les hypothèses (3.9), si la condition de compatibilité (3.41) est vérifiée, et si la propriété d'approximabilité minimale (C.26) est vraie pour* $(\mathcal{V}_h)_h$ *dans* $L^2(\Omega) \times \mathbf{H}(\operatorname{div},\Omega)$*, alors l'erreur tend vers* 0 *quand* h *tend vers* 0 *:*

$$\lim_{h\to 0} \|u-u_h\|_{L^2(\Omega)} + \|\boldsymbol{p}-\boldsymbol{p}_h\|_{\mathbf{H}(\operatorname{div},\Omega)} = 0. \tag{3.43}$$

En pratique, on va maintenant définir les espaces discrets $(\mathcal{V}_h)_h$, sous la forme $\mathcal{V}_h = M_h \times \boldsymbol{Q}_h$, pour résoudre numériquement le problème posé dans un domaine Ω polygonal ($d = 2$) ou polyédrique ($d = 3$). Comme pour la résolution du problème à une inconnue, on

se base sur des maillages $(\mathcal{T}_h)_h$ formés de simplexes : $\overline{\Omega} = \cup_{1\le\ell\le L} T_\ell$, voir (3.21). D'après [4], pour construire un sous-espace M_h de $L^2(\Omega)$, il n'y a pas de condition de continuité à respecter entre deux simplexes partageant une facette, alors qu'il faut imposer la continuité de la composante normale pour construire un sous-espace [20] $\boldsymbol{Q}_h$ de $\mathbf{H}(\mathrm{div},\Omega)$. On choisit l'élément fini de Raviart-Thomas.
On définit tout d'abord, pour $k \in \mathbb{N}$,

$$\mathbf{RT}_k(T) := \{\boldsymbol{q} \in (P(T))^d \ : \ \exists \boldsymbol{a} \in (P_k(T))^d,\ \exists b \in P_k(T), \\ \forall \boldsymbol{x} \in T,\ \boldsymbol{q}(\boldsymbol{x}) = \boldsymbol{a} + b\boldsymbol{x}\}. \tag{3.44}$$

On note que si $\boldsymbol{q} \in \mathbf{RT}_k(T)$, alors on a $(\boldsymbol{q}\cdot\boldsymbol{n})_{|a_e^T} \in P_k(a_e^T)$ pour $1 \le e \le d+1$, où $\boldsymbol{n}$ est la normale unitaire extérieure à T, $(a_e^T)_{e=1,\cdots,d+1}$ sont les facettes de T , et $P_k(a_e^T)$ est l'ensemble des polynômes de degré au plus k définis sur a_e^T.
A partir de là, on définit l'espace discret :

$$\boldsymbol{Q}_{h,k} := \{\boldsymbol{q}_h \in \mathbf{H}(\mathrm{div},\Omega) \ : \ \forall \ell,\ \boldsymbol{q}_{h|T_\ell} \in \mathbf{RT}_k(T_\ell)\}.$$

Pour construire l'espace discret $M_{h,k}$ correspondant (c'est-à-dire de degré d'approximation *compatible*), on choisit :

$$M_{h,k} := \{v_h \in L^2(\Omega) \ : \ \forall \ell,\ v_{h|T_\ell} \in P_k(T_\ell)\}.$$

En effet, pour tout $\boldsymbol{q}_h \in \boldsymbol{Q}_{h,k}$, sa divergence appartient à $L^2(\Omega)$ et, en outre, pour tout ℓ, $\mathrm{div}\,\boldsymbol{q}_{h|T_\ell} \in P_k(T_\ell)$.
On se concentre maintenant sur le cas $k = 0$: $M_h = M_{h,0}$, $\boldsymbol{Q}_h = \boldsymbol{Q}_{h,0}$, et $\mathcal{V}_h = M_{h,0} \times \boldsymbol{Q}_{h,0}$. Un élément v_h de M_h étant constant par simplexe, il est caractérisé par les valeurs $(\int_{T_\ell} v_h\, d\boldsymbol{x})_{\ell=1,L}$. On appelle l'application $v_h \mapsto (\int_{T_\ell} v_h\, d\boldsymbol{x})_{\ell=1,L}$ les **degrés de liberté** de M_h. A partir de là, on peut définir simplement une base de M_h de la forme $(\underline{w}_j)_{1\le j\le L}$:

$$\forall j,\ \underline{w}_j \in M_h,\ \text{et}\ \int_{T_i} \underline{w}_j\, d\boldsymbol{x} = \delta_{ij}\ \ \text{pour } 1 \le i \le L.$$

Par construction, le support de chaque fonction $\underline{w}_j$ est égal au simplexe T_j.
Pour les éléments de $\boldsymbol{Q}_h$, on remarque qu'il y a *a priori* $d+1$ degrés de liberté par simplexe, puisque dans la définition de l'élément $\mathbf{RT}_0(T)$ le couple $(\boldsymbol{a}, b)$ parcourt $\mathbb{R}^{d+1}$. Comme en outre il faut assurer la continuité de la composante normale à la traversée de chaque facette, on choisit "naturellement" les degrés de liberté $(\int_{a_e^T} \boldsymbol{q}_h \cdot \boldsymbol{n}_e^T\, da)_{e=1,d+1}$, où on a noté $\boldsymbol{n}_e^T$ le vecteur unitaire normal à la facette a_e^T, et dirigé vers l'extérieur de T.

Proposition 3.28 *Soit T un simplexe. Tout élément $\boldsymbol{q}$ de $\mathbf{RT}_0(T)$ est déterminé de façon unique par les $d+1$ valeurs $(\int_{a_e^T} \boldsymbol{q}_h \cdot \boldsymbol{n}_e^T\, da)_{e=1,d+1}$.*

On réalise la démonstration lorsque $d = 2$. Le cas $d = 3$ est laissé en exercice.

Démonstration : On écrit $\boldsymbol{q}(\boldsymbol{x}) = \boldsymbol{a} + b\boldsymbol{x}$ pour $\boldsymbol{x} \in T$ et on veut montrer que $(\int_{a_e^T} \boldsymbol{q}_h \cdot \boldsymbol{n}_e^T\, da)_{e=1,3}$ étant données, on peut déterminer a_1, a_2, b. Sans perte de généralité, on se place dans le triangle T de sommets $S_1(0,0)$, $S_2(x_{2,1},0)$ et $S_3(x_{3,1},x_{3,2})$, avec $x_{2,1} \neq 0$ et $x_{3,2} \neq 0$. Les arêtes sont respectivement :

20. Pour définir $\boldsymbol{Q}_h$, on s'appuie sur des fonctions discrètes *à valeurs vectorielles.*

— $a_1^T = [S_1, S_2]$, de normale unitaire extérieure $\boldsymbol{n}_1^T(0,-1)$ et incluse dans la droite d'équation $x_2 = 0$;
— $a_2^T = [S_2, S_3]$, de normale unitaire extérieure $\boldsymbol{n}_2^T(n_{2,1}, n_{2,2})$ et incluse dans la droite d'équation $n_{2,1}(x_1 - x_{2,1}) + n_{2,2}x_2 = 0$, avec $n_{2,1} \neq 0$;
— $a_3^T = [S_3, S_1]$, de normale unitaire extérieure $\boldsymbol{n}_3^T(n_{3,1}, n_{3,2})$ et incluse dans la droite d'équation $n_{3,1}x_1 + n_{3,2}x_2 = 0$, avec $n_{3,1} \neq 0$.

A partir de là, on trouve [21]

$$\begin{array}{ll} \boldsymbol{q}_h \cdot \boldsymbol{n}_1^T(\boldsymbol{x}) = -a_2, & \forall \boldsymbol{x} \in a_1^T ; \\ \boldsymbol{q}_h \cdot \boldsymbol{n}_2^T(\boldsymbol{x}) = a_1 n_{2,1} + a_2 n_{2,2} + b x_{2,1} n_{2,1}, & \forall \boldsymbol{x} \in a_2^T ; \\ \boldsymbol{q}_h \cdot \boldsymbol{n}_3^T(\boldsymbol{x}) = a_1 n_{3,1} + a_2 n_{3,2}, & \forall \boldsymbol{x} \in a_3^T . \end{array}$$

La première équation donne la valeur de a_2, à savoir

$$a_2 = -\frac{1}{|a_1^T|} \int_{a_1^T} \boldsymbol{q}_h \cdot \boldsymbol{n}_1^T \, da.$$

Ensuite, la troisième équation détermine la valeur de a_1, puisque $n_{3,1} \neq 0$. Enfin, la deuxième équation détermine la valeur de b, puisque $n_{2,1} \neq 0$. ◇

A partir de ce résultat local, on en déduit qu'un élément $\boldsymbol{q}_h$ de $\boldsymbol{Q}_h$ est caractérisé par les valeurs $(\int_{a_e} \boldsymbol{q}_h \cdot \boldsymbol{n}_e da)_{e=1,A^+}$, où $(a_e)_{1 \leq e \leq A^+}$ est l'ensemble des facettes du maillage et, pour chaque e, on a noté $\boldsymbol{n}_e$ un vecteur unitaire normal [22] à la facette a_e. On appelle l'application $\boldsymbol{q}_h \mapsto (\int_{a_e} \boldsymbol{q}_h \cdot \boldsymbol{n}_e da)_{e=1,A^+}$ les **degrés de liberté** de $\boldsymbol{Q}_h$. Qui plus est, on peut définir une base de $\boldsymbol{Q}_h$ de la forme $(\underline{\boldsymbol{\omega}}_m)_{1 \leq m \leq A^+}$:

$$\forall m, \ \underline{\boldsymbol{\omega}}_m \in \boldsymbol{Q}_h, \text{ et } \int_{a_l} \underline{\boldsymbol{\omega}}_m \cdot \boldsymbol{n}_e da = \delta_{lm} \ \text{ pour } 1 \leq l \leq A^+.$$

D'après la proposition 3.28, pour chaque facette non-située sur la frontière, le raccord de la composante normale (constante pour l'élément fini $\mathbf{RT}_0$) est assuré si, et seulement si, les degrés de liberté coïncident. En outre, le support de chaque fonction $\underline{\boldsymbol{\omega}}_m$ est égal à la réunion des simplexes auquel la facette a_m appartient : un seul simplexe si a_m est située sur la frontière, ou deux simplexes sinon.

Remarque 3.29 *Comme les degrés de liberté pour M_h et $\boldsymbol{Q}_h$ ne sont plus des valeurs ponctuelles, on parle de degrés de liberté* de type moment.

On se place explicitement dans le cas $d = 2$ pour obtenir une borne supérieure *optimale* du nombre d'éléments non-nuls de $\mathbb{A}_h$. Pour h donné, on a un maillage à N^+ sommets, A^+ arêtes et L triangles. Pour rappel, on décompose A^+ en $A + A_b$, où A_b est le nombre d'arêtes situées sur la frontière. En outre, on sait que $A_b + 2A = 3L$ (voir la proposition 3.13). D'après ce qui précède, la matrice pour le problème à deux inconnues (la matrice $\mathbb{A}_h$ de (3.35)) est de dimension $N^{tot} = N_u + N_{\boldsymbol{p}} = A^+ + L$, puisque $N_u = dim(M_h) = L$ et $N_{\boldsymbol{p}} = dim(\boldsymbol{Q}_h) = A^+$. Comme dans le cas du problème à une inconnue, on peut estimer

21. En particulier, les expressions ci-dessous montrent que les traces normales sont *constantes* par facette, comme prescrit dans la définition de $\mathbf{RT}_0(T)$.

22. Pour chaque facette, il y a exactement deux vecteurs normaux unitaires : on en choisit donc un...

le nombre d'éléments non-nuls de $\mathbb{A}_h$. Encore une fois, on pourra établir des résultats similaires pour $d = 3$, sous réserve que la famille de maillages est régulière. On utilisera pour cela les ingrédients de la remarque 3.16.

Proposition 3.30 *On suppose les propriétés (3.21) vérifiées et que $\Omega \subset \mathbb{R}^2$. Alors le nombre d'éléments non-nuls de la matrice $\mathbb{A}_h \in \mathbb{R}^{N_h^{tot} \times N_h^{tot}}$ est asymptotiquement au plus de l'ordre de $5.8 N_h^{tot}$:*

$$\forall h, \quad \frac{nnz(\mathbb{A}_h)}{N_h^{tot}} \leq 5.8.$$

Démonstration : On rappelle que $\mathbb{A}$ peut être écrite sous la forme d'une matrice par blocs, voir (3.35)). En reprenant les notations utilisées, on va estimer le nombre d'éléments non-nuls par bloc. Pour ce faire, on va procéder comme à la proposition 3.13.

- Bloc $\mathbb{M}_u$: on a $(\mathbb{M}_u)_{i,j} = \int_\Omega q \underline{w}_j \underline{w}_i \, d\boldsymbol{x}$ pour $1 \leq i, j \leq L$. D'après la propriété sur le support des fonctions de base $(\underline{w}_j)_{1 \leq j \leq L}$, on a $(\mathbb{M}_u)_{i,j} \neq 0$ si, et seulement si, $i = j$. On en déduit que $nnz(\mathbb{M}_u) = L$.
- Bloc $\mathbb{B}$: on a $\mathbb{B}_{l,j} = \int_\Omega \underline{w}_j \operatorname{div} \underline{\boldsymbol{\omega}}_l \, d\boldsymbol{x}$ pour $1 \leq l \leq A^+,\ 1 \leq j \leq L$. Pour une ligne l, d'après les propriétés de support de $\underline{\boldsymbol{\omega}}_l$ d'une part, et de $\underline{w}_j$ d'autre part, on vérifie qu'on a (au plus) un seul élément non-nul si l'arête a_l correspondante est située sur la frontière, et deux sinon. On en déduit que $nnz(\mathbb{B}) \leq A_b + 2A = 3L$.
- Bloc $\mathbb{M}_{\boldsymbol{p}}$: on a $(\mathbb{M}_{\boldsymbol{p}})_{l,m} = \int_\Omega k^{-1} \underline{\boldsymbol{\omega}}_m \cdot \underline{\boldsymbol{\omega}}_l \, d\boldsymbol{x}$ pour $1 \leq l, m \leq A^+$. Pour une ligne l, d'après les propriétés de support de $\underline{\boldsymbol{\omega}}_l$ de $\underline{\boldsymbol{\omega}}_m$, deux cas peuvent se présenter. Si l'arête a_l est située sur la frontière, le support de $\underline{\boldsymbol{\omega}}_l$ et réduit à un triangle T_0, et il existe exactement trois fonctions $\underline{\boldsymbol{\omega}}_m$ dont le support contient T_0 : ce sont les trois fonctions de base associées à une des arêtes de T_0. Sinon, le support de $\underline{\boldsymbol{\omega}}_l$ est égal aux deux triangles dont l'intersection est a_l, et il existe exactement cinq fonctions $\underline{\boldsymbol{\omega}}_m$ dont le support contient au moins un de ces deux triangles. On en déduit que $nnz(\mathbb{M}_{\boldsymbol{p}}) \leq 3A_b + 5A = A^+ + 2(A_b + 2A) = A^+ + 6L$.

On en conclut que $nnz(\mathbb{A}_h) = nnz(\mathbb{M}_u) + 2nnz(\mathbb{B}) + nnz(\mathbb{M}_{\boldsymbol{p}})$ est majoré par

$$nnz(\mathbb{A}_h) \leq 13L + A^+ = 5.8(L + A^+) + 2.4(3L - 2A^+) < 5.8(L + A^+),$$

puisque $3L - 2A^+ = -A_b < 0$. ◇

Exercice 3.4 *Lorsque $d = 2$ et à l'aide des relations entre nombre d'arêtes et de triangles, prouver que cette propriété est valable quel que soit l'ordre $k \geq 0$. Si on note $N_{h,k}^{tot}$ la dimension de $\mathcal{V}_{h,k} = M_{h,k} \times \boldsymbol{Q}_{h,k}$, et $\mathbb{A}_{h,k} \in \mathbb{R}^{N_{h,k}^{tot} \times N_{h,k}^{tot}}$ la matrice associée à la forme a_2 dans une base "bien choisie", alors :*

$$\exists C_k' > 0, \quad \forall h, \quad \frac{nnz(\mathbb{A}_{h,k})}{N_{h,k}^{tot}} \leq C_k'.$$

Pour établir la propriété d'approximabilité minimale (C.26) pour le problème à deux inconnues, on choisit $\mathcal{V}_+ = \mathcal{D}(\Omega) \times (C^\infty(\Omega))^2$ comme sous-espace dense de $\mathcal{V} = L^2(\Omega) \times \mathbf{H}(\operatorname{div}, \Omega)$. On peut établir cette propriété pour le problème à deux inconnues, par passage à l'élément fini de référence, à l'aide d'un opérateur d'**interpolation** $\Pi_h : \mathcal{V}_+ \to \mathcal{V}_h$ (voir par exemple [4]).

Nous précisons les similitudes et différences par rapport à au problème à une inconnue concernant le passage à l'élément de référence $\hat{T}$. Celui-ci se fait toujours selon la transformation affine F_ℓ (3.29). Ensuite, on transforme les fonctions à valeurs scalaires u, v comme en (3.31). Mais, pour les fonctions à valeurs vectorielles $\boldsymbol{p}$, $\boldsymbol{q}$, on utilise cette fois la **transformation de Piola** :

$$\boldsymbol{q} = \frac{1}{|det(\mathbb{A}_\ell)|}\mathbb{A}_\ell\hat{\boldsymbol{q}}\circ F_\ell^{-1} \qquad \Longleftrightarrow \qquad \hat{\boldsymbol{q}} = |det(\mathbb{A}_\ell)|\,\mathbb{A}_\ell^{-1}\boldsymbol{q}\circ F_\ell. \tag{3.45}$$

On écrit $\boldsymbol{q}(\boldsymbol{x}) = |det(\mathbb{A}_\ell)|^{-1}\mathbb{A}_\ell\hat{\boldsymbol{q}}(\hat{\boldsymbol{x}})$ pour tout $\boldsymbol{x}\in T_\ell$ (ou tout $\hat{\boldsymbol{x}}\in\hat{T}$). On peut ensuite établir les identités suivantes sur les divergences :

$$\forall\hat{\boldsymbol{x}}\in\hat{T},\ \mathrm{div}\,\boldsymbol{q}(\boldsymbol{x}) = \frac{1}{|det(\mathbb{A}_\ell)|}\hat{\mathrm{div}}\hat{\boldsymbol{q}}(\hat{\boldsymbol{x}})\,; \quad \forall\boldsymbol{x}\in T_\ell,\ \hat{\mathrm{div}}\hat{\boldsymbol{q}}(\hat{\boldsymbol{x}}) = |det(\mathbb{A}_\ell)|\ \mathrm{div}\,\boldsymbol{q}(\boldsymbol{x}). \tag{3.46}$$

Pour des fonctions suffisamment régulières v et $\hat{v}$ en correspondance selon (3.31), respectivement $\boldsymbol{q}$ et $\hat{\boldsymbol{q}}$ en correspondance selon (3.45), on a les formules de changement de variables (voir [4, §2.1.3]) :

$$\int_{T_\ell}\boldsymbol{q}\cdot\nabla v\,d\boldsymbol{x} = \int_{\hat{T}}\hat{\boldsymbol{q}}\cdot\hat{\nabla}\hat{v}\,d\hat{\boldsymbol{x}}, \quad \int_{T_\ell} v\ \mathrm{div}\,\boldsymbol{q}\,d\boldsymbol{x} = \int_{\hat{T}}\hat{v}\,\hat{\mathrm{div}}\hat{\boldsymbol{q}}\,d\hat{\boldsymbol{x}},$$
$$\int_{\partial T_\ell} v\,\boldsymbol{q}\cdot\boldsymbol{n}\,d\Gamma = \int_{\partial\hat{T}}\hat{v}\,\hat{\boldsymbol{q}}\cdot\hat{\boldsymbol{n}}\,d\hat{\Gamma}.$$

Pour établir la propriété d'approximabilité uniforme (C.29) et estimer la vitesse de convergence, on choisit cette fois

$$\tilde{\mathcal{V}} = \{(\tilde{v},\tilde{\boldsymbol{q}})\in\mathcal{V}\ :\ \tilde{v}\in H^1(\Omega),\ \mathrm{div}\,\tilde{\boldsymbol{q}}\in H^{r_{\max}}(\Omega)\},$$

où l'exposant de régularité $r_{\max}\in]0,1]$ a été introduit à la proposition 3.19. On arrive au résultat suivant sur l'ordre de convergence de la définition (3.20) (voir [10, §4.5] pour les détails).

Théorème 3.31 *(**problème à deux inconnues**) On suppose que les coefficients k et q appartiennent à $\mathcal{P}W^{1,\infty}(\Omega)$, qu'ils vérifient les hypothèses (3.9), et enfin que la famille de maillage est régulière et conforme. Si la donnée $f\in H^{r_{\max}}(\Omega)$, l'ordre de convergence α_{cv} pour le problème à deux inconnues est égal à :*

- *$\alpha_{cv} = r_{\max}-\epsilon$ pour tout $\epsilon\in]0,r_{\max}[$ si $r_{\max}\le 1$, ou*
- *$\alpha_{cv} = 1$ si $r_{\max}=1$.*

Qui plus est, la constante C_{cv} ne dépend ni de la solution $(u,\boldsymbol{p})$, ni de la donnée f :

$$\exists C_{cv}>0,\ \forall h,\ \forall f\in H^{r_{\max}}(\Omega),$$
$$\|u-u_h\|_{L^2(\Omega)} + \|\boldsymbol{p}-\boldsymbol{p}_h\|_{\mathbf{H}(\mathrm{div};\Omega)} \le C_{cv}\,h^{\alpha_{cv}}\,\|f\|_{H^{r_{\max}}(\Omega)}.$$

Deuxième partie

Algèbre linéaire numérique

Dans cette partie, on notera indifféremment $\mathbb{K}$ *pour* $\mathbb{K} = \mathbb{R}$ *ou* $\mathbb{C}$. *On s'intéresse à la résolution de systèmes linéaires dans* $\mathbb{K}^n$, *pour* $n \geq 1$. *Pour tout* $n \in \mathbb{N}^*$, *on notera* I_n *la matrice identité d'ordre* n, $\mathbf{0}_n$ *le vecteur nul d'ordre* n *et* $(\mathrm{e}_k)_{k=1,n}$ *la base orthonormale canonique de* $\mathbb{K}^n$, *telle que pour tout* $i \in \{1, \cdots, n\}$, *le vecteur* $\mathrm{e}_i \in \mathbb{K}^n$ *est le* i^{eme} *vecteur de cette base. Pour éviter toute confusion, les vecteurs sont représentés par des lettres minuscules et les matrices par des majuscules. Pour* $\mathbb{A} \in \mathbb{K}^{n \times n}$ *et* $b \in \mathbb{K}^n$ *donnés, on étudie la résolution de systèmes linéaires de la forme :*

$$\textit{Trouver } x \in \mathbb{K}^n \textit{ tel que } \mathbb{A}x = b.$$

On commence par préciser la notion de coût calcul et on rappelle quelques notions algorithmiques au chapitre 4. Des compléments informatiques (programmation, etc.) rédigés par Modave sont disponibles sur la page web [27].

On appelle **méthode directe** *de résolution du système linéaire une méthode qui donne exactement* x *solution après un nombre fini d'opérations élémentaires* $(+, -, *, /, \sqrt{})$. *Ces méthodes font l'objet du chapitre 5.*

On appelle **méthode itérative** *de résolution du système linéaire tout algorithme qui construit à partir d'une estimation initiale* x^0 *une suite de vecteurs* $\{x^k\}_{k \in \mathbb{N}}$ *destinée à converger vers la solution* x *du système. Ces méthodes font l'objet du chapitre 6.*

Dans le chapitre 7, on étudiera plus précisément deux méthodes itératives de ce type : le gradient conjugué et le GMRES.

Enfin, dans le chapitre 8, on s'intéressera à cette méthode itérative, utilisée pour la résolution des problèmes aux valeurs propres.

On trouvera des résultats d'analyse spectrale en dimension finie à l'Annexe A, et quelques définitions concernant les normes a l'Annexe B.

Chapitre 4

Faire des calculs

On donne en préambule les éléments d'algorithmique numérique à connaître absolument avant de se lancer dans la programmation, ou l'utilisation de logiciels de calcul.

4.1 Précision et convergence

Tout d'abord, il faut être conscient, lorsque l'on effectue un calcul *numérique*, que la précision est **finie**, à la différence du calcul *formel*, par exemple. La première question est donc, pourquoi utilise-t-on une méthode numérique, *a priori* moins précise ? La réponse est pragmatique : on ne sait pas résoudre formellement le système linéaire

$$\text{Trouver } x \in \mathbb{K}^N \text{ tel que } \mathbb{A}x = b. \tag{4.1}$$

dès lors que l'ordre $N \geq 1$ de la matrice $\mathbb{A}$ est "trop grand" ; ou de façon encore plus pragmatique, le temps de résolution est de toute façon beaucoup trop important !
Que signifie alors l'association de termes **convergence numérique** ? Avant de répondre à cette question, nous allons détailler quelques problèmes simples, inhérents au calcul numérique, par opposition au calcul formel.
La finitude de la précision vient de la représentation en machine des nombres réels, sous la forme générique

$$\pm a_0, a_1 \cdots a_p \ 10^d, \text{ avec } (a_0, \cdots, a_p) \in \{0, \cdots, 9\}^{p+1}, \ a_0 \neq 0, \ d \in \{-d_{max}, \cdots, d_{max}\},$$

où p et d_{max} dépendent du processeur qui effectue les calculs. Plus précisément, la représentation est du type indiqué ci-dessus, mais en base 2. On dit aussi que $p+1$ est le nombre maximal de chiffres significatifs de la représentation en machine, et que $10^{-d_{max}}$ est la précision machine. Cette représentation génère deux difficultés :

- Tout nombre dont la valeur absolue est plus grande que $10^{d_{max}+1}$ est considéré comme infini, et symétriquement, tout nombre dont la valeur absolue est strictement plus petite que $10^{-d_{max}}$ est considéré comme étant nul ;
- Les opérations sur ces nombres (addition, multiplication, etc. ; extraction de racine, exponentation, etc.) sont effectuées en précision finie. Prenons l'exemple de la multiplication... Si les deux nombres ont respectivement q et q' chiffres significatifs ($q, q' \in \{1, \cdots, p+1\}$), leur produit possède $q+q'-1$ ou $q+q'$ chiffres significatifs. Dès lors que $q+q'-1 > p+1$, une *troncature* est effectuée lors de la mise en mémoire

du résultat même si le calcul était exact, puisque la représentation de tout nombre comporte au plus $p+1$ chiffres significatifs.

C'est la raison pour laquelle les calculs numériques produisent en général des erreurs d'arrondi...

Par voie de conséquence, et pour revenir à notre problème, il devient difficile d'obtenir un résultat du type $\mathbb{A}x - b = 0$. Et même si l'ordinateur affirme que $\mathbb{A}x - b = 0$, ceci signifie uniquement que la différence est plus petite que la précision machine, d'après l'exposé précédent. Par ailleurs, on se contente en général d'une valeur *approchée*, c'est-à-dire à ε près, pour éviter un coût calcul trop élevé (compromis coût calcul-précision). Nous venons d'introduire la notion de *calcul exact à ε près*, qui est très courante chez l'ingénieur. Petit à petit, nous glissons du monde des mathématiques, en passant par celui du calcul scientifique, vers celui de l'art de l'ingénieur. Ces mondes, bien qu'ils ne répondent pas aux mêmes critères, n'en restent pas moins complémentaires, et indissociables.

Revenons aux mathématiques, après cette brêve incursion. Quand on parle de calcul exact à ε près, quel est le sens mathématique sous-jacent ? Typiquement, si on note $\|\cdot\|$ une norme quelconque, pour $\varepsilon > 0$, on cherche x_ε tel que

$$\|\mathbb{A}x_\varepsilon - b\| \leq \varepsilon. \tag{4.2}$$

Il est clair que l'ensemble des x_ε qui satisfont à (4.2) n'est pas réduit à un singleton ! Quoiqu'il en soit, à ε près, l'obtention d'un tel x_ε est suffisante... On parle de *convergence numérique*.

4.2 Comptage des opérations

Donnons deux exemples élémentaires de comptage des opérations dans $\mathbb{K}^N$.

1. Le *produit scalaire (complexe dans $\mathbb{C}^N$)* de deux vecteurs, qui s'écrit

$$(x, y) = \sum_{i=1}^{N} x_i \overline{y_i},$$

est effectué en N multiplications et $(N-1)$ additions. Usuellement, on ne conserve que le terme principal, ce qui signifie que l'on considère que le produit scalaire requiert N additions et N multiplications.

2. La *multiplication matrice vecteur*, qui s'écrit composante par composante,

$$(\mathbb{A}y)_i = \sum_{j=1}^{N} \mathbb{A}_{i,j} y_j, \quad 1 \leq i \leq N,$$

requiert N^2 additions et N^2 multiplications, ce qui laisse à penser qu'un produit matrice-vecteur est équivalent à N produits scalaires... Ceci étant, que se passe-t-il si l'on sait que la matrice $\mathbb{A}$ est creuse [23], c'est-à-dire avec en moyenne $Z_{\neq}$ éléments non nuls par ligne, pour $Z_{\neq}$ constant ou très petit devant N. On ne va stocker que les positions, i. e. les paires d'indices (i, j), et les valeurs $\mathbb{A}_{i,j}$ non nulles ! Lorsque l'on multiplie $\mathbb{A}$ par x, on n'effectue

23. Lorsque la matrice $\mathbb{A}$ possède N^2 éléments non-nuls, ou au moins ZN^2 éléments non-nuls avec $Z \in]0, 1[$ indépendant de N, on dit que la matrice est **pleine**.

que les multiplications pour lesquelles $\mathbb{A}_{i,j} \neq 0$ (et les additions de termes non nuls). En moyenne, on aura donc effectué $Z_{\neq} N$ additions, et autant de multiplications...
Pourquoi un tel exemple ? Lorsque l'on résout un problème par une méthode de différences finies ou d'élément finis, la matrice obtenue comporte très peu d'éléments non nuls par ligne, de l'ordre d'une dizaine. Par exemple, pour la discrétisation d'un Laplacien en 2D par différences finies, on a vu que $Z_{\neq} \leq 5$ à la section 2.3, alors que pour un calcul par éléments finis P_1, on a $Z_{\neq} \leq 7$ (proposition 3.13). Si la dimension de $\mathbb{K}^N$ est de l'ordre de $N = 10^6$ (ce qui est très courant !), on voit que les deux évaluations du coût calcul donnent

$$2N^2 = 2N\,N, \text{ resp. } 2Z_{\neq} N \approx 20\,N,$$

c'est-à-dire l'équivalent de 2.000.000 produits scalaires, contre une vingtaine...

4.3 Temps calcul

Une autre façon d'estimer le coût du calcul est de mesurer le **temps de calcul**, par l'intermédiaire d'une horloge.
A priori, ces deux méthodes semblent tout à fait similaires. De fait, ceci dépend de la machine sur laquelle on effectue le calcul numérique. La première objection concerne les *opérations*. Une addition, une multiplication, une division ont-elles le même coût ? Une réponse possible consiste à compter précisément le nombre de chaque type d'opérations [24]...
Un problème beaucoup plus épineux est que la machine peut (pour simplifier, il existe d'autres modes de fonctionnement), soit travailler *séquentiellement*, soit *en parallèle*. Dans le premier cas, les opérations sont exécutées l'une après l'autre. Dans le second cas, la machine est constituée de plusieurs processeurs, qui peuvent alors exécuter simultanément des opérations, et échanger des données entre eux [25], voir la section 4.5 pour plus de précisions. Dans ce cas, supposons que l'on teste plusieurs fois le même problème, sur une machine disposant de plus en plus de processeurs : le temps horloge diminue, alors que le nombre total d'opérations restera constant ! Les deux estimateurs de coût calcul ne sont donc pas équivalents...
Enfin, il peut également être utile de quantifier le **stockage mémoire** requis pour l'exécution de la méthode. Par exemple, lorsque l'on utilise une méthode itérative, on constate que le stockage est beaucoup plus faible que pour une méthode directe. Ceci ne préjuge cependant pas de la supériorité d'une méthode sur une autre...
Cette discussion est volontairement restée très générale, et elle peut être vue comme une introduction à l'algorithmique numérique. Ce qu'il faut retenir, c'est qu'il convient d'être prudent lorsque l'on évalue la qualité d'une méthode numérique, car celle-ci résulte habituellement de compromis entre les divers critères et contraintes que nous avons évoqués ci-dessus. Pour ce type de problèmes, il est fort utile d'acquérir de l'expérience, notamment en réalisant des comparaisons entre plusieurs méthodes.

Remarque 4.1 *La notion de coût liée à l'énergie dépensée pour réaliser le calcul est apparue plus récemment. Elle répond à deux problématiques, la première étant la prise de*

24. Ceci étant, on raisonne usuellement en opérations flottantes par seconde, ou **FLOPs** = **FL**oating **OP**erations per **s**econd, pour un processeur donné, sans distinguer les opérations entre elles.

25. Ici, on suppose implicitement que l'algorithme de calcul le permet. Le fait qu'un algorithme est effectivement exécutable en parallèle, ou *parallélisable*, doit être étudié avec précision. On donnera quelques exemples dans la suite.

conscience écologique, la seconde étant plus prosaïquement liée aux ressources énergétiques disponibles pour réaliser le calcul en question !

4.4 Construction efficace des matrices différences finies et éléments finis

Avant de résoudre les systèmes linéaires, expliquons brièvement comment les construire en pratique.

En préambule, on note qu'il n'y pas de difficulté particulière pour la construction des matrices et des seconds membres issus de la discrétisation par la méthode des différences finies. Prenons l'exemple de la discrétisation du Laplacien. En dimension 1, les points voisins sont numérotées de façon continue. La matrice $\mathbb{A}_1$ est tridiagonale (2.5). En dimension 2, on a $N = n^2$, et dans la numérotation globale (voir la figure 2.5), entre deux points voisins il y a un décalage de 1 dans la direction x, et un décalage de n (le nombre d'inconnues dans la direction x) dans la direction y. On a ainsi observé que la matrice $\mathbb{A}_2$ était pentadiagonale par points, et tridiagonale par blocs (2.22). Enfin, le calcul de chaque élément non-nul de $\mathbb{A}_1$ ou $\mathbb{A}_2$ est trivial. Pour le Laplacien généralisé, on a la même structure, avec un calcul très simple de chaque élément non-nul de la matrice en fonction des coefficients k et q intervenant dans l'EDP (2.33).
La généralisation en dimension d est immédiate. En particulier on vérifie aisément que, pour le calcul de la matrice $\mathbb{A}_d$, la complexité est linéaire par rapport au nombre de points de discrétisation. Bien sûr, le calcul du second membre f est élémentaire, et de complexité linéaire (évaluations ponctuelles).

Pour la méthode des éléments finis, on doit construire la matrice $\mathbb{A}$ et le second membre F de (3.14). La difficulté est que, même si les matrices sont creuses (voir la proposition 3.13), elles ne pas structurées en général. Prenons l'exemple de la diffusion (problème à une inconnue) posé dans $\Omega \subset \mathbb{R}^d$, $d = 2$ ou 3. Dans le fichier contenant les informations du maillage, on dispose des coordonnées des sommets $(M_i)_{i=1}^N$ ainsi que pour chaque simplexe, de la liste de ses $d+1$ sommets. Pour évaluer les intégrales sur un simplexe, on utilise un schéma d'intégration numérique approprié, qui dépend de l'ordre polynomial de la fonction à intégrer et qui nécessite d'approcher les valeurs de cette fonction en certains points du simplexe. Ainsi, pour un polynôme P de degré au plus k, si on appelle $(\boldsymbol{x}_{k,\ell})_{\ell=1,N_k}$ les points d'intégration et $(\omega_{k,\ell})_{\ell=1,N_k}$ les poids associés, on a [16] :

$$\int_{int(T_\ell)} P(\boldsymbol{x}) d\boldsymbol{x} = \sum_{\ell=1,N_k} \omega_{k,\ell}\, P(\boldsymbol{x}_{k,\ell}).$$

En règle générale, dans un simplexe T_ℓ, la position des points d'intégration est déterminée à partir de celle des sommets. On utilise cette formule pour calculer les intégrales, ce qui demande d'évaluer les fonctions de base w_i et leur gradient $\mathbf{grad}\, w_i$ aux points d'intégration.
Reprenons le calcul des éléments de la matrice $\mathbb{A}$ (3.15) pour l'élément fini de Lagrange

P_1.[26] Une première approche est d'écrire l'algorithme sous la forme suivante cf. (3.24) :

initialisation
  $\mathbb{A} \in \mathbb{R}^N \times \mathbb{R}^N$, $\mathbb{A} = 0$.
itérations : pour $i = 1, \cdots, N$**, faire**
  itérations : pour $j = 1, \cdots, N$**, faire**
    itérations : pour $\ell = 1, \cdots, L$ **et t.q.** M_i, $M_j \in T_\ell$**, faire**

$$\mathbb{A}_{i,j} = \mathbb{A}_{i,j} + \int_{int(T_\ell)} \left(k\, \mathbf{grad}\, w_{j|T_\ell} \cdot \mathbf{grad}\, w_{i|T_\ell} + q\, w_{j|T_\ell} w_{i|T_\ell} \right) d\boldsymbol{x}$$

    fin boucle sur ℓ.
  fin boucle sur j.
fin boucle sur i.

Cette approche de double boucle sur la liste des sommets nécessite de connaître pour tous les couples d'indices (i, j) la liste des simplexes contenant M_i et M_j. Cette liste est vide si M_i et M_j ne sont pas voisins, c'est-à-dire si M_i et M_j n'appartiennent pas au même simplexe. En général, cette information n'est pas fournie, il faut écrire un algorithme spécifique pour l'obtenir. On note que la complexité est supérieure ou égale à $O(N^2)$, à cause de la double boucle sur les indices de sommets : en effet, on effectue au moins une vérification (liste des simplexes vide) par couple d'indices. On sait que cette construction est inefficace, puisque le nombre d'éléments non-nuls de $\mathbb{A}$ est strictement inférieur à $7N$ si $d = 2$ (voir la proposition 3.13).

Pour remédier à toutes ses difficultés, une idée est de parcourir les simplexes, et d'agréger au fur et à mesure les intégrales. Au cours du processus, il faut se souvenir que les contributions sont prises en compte par sommet : sur la i^{eme} ligne de $\mathbb{A}$, on regroupe toutes les contributions faisant intervenir la fonction "chapeau" w_i. Ainsi, l'algorithme optimal de construction de la matrice $\mathbb{A}$ est le suivant[27] :

initialisation
  $\mathbb{A} \in \mathbb{R}^N \times \mathbb{R}^N$, $\mathbb{A} = 0$.
itérations : pour $\ell = 1, \cdots, L$**, faire**
  itérations : pour i **t.q.** $M_i \in T_\ell$**, faire**
    itérations : pour j **t.q.** $M_j \in T_\ell$**, faire**

$$\mathbb{A}_{i,j} = \mathbb{A}_{i,j} + \int_{int(T_\ell)} \left(k\, \mathbf{grad}\, w_{j|T_\ell} \cdot \mathbf{grad}\, w_{i|T_\ell} + q\, w_{j|T_\ell} w_{i|T_\ell} \right) d\boldsymbol{x}$$

    fin boucle sur j.
  fin boucle sur i.
fin boucle sur ℓ.

Cet algorithme est appelé **assemblage** de la matrice éléments finis. Notons que cet algorithme est de complexité $O(L(d+1)^2)$, qu'on utilise uniquement les informations à disposition, et qu'on n'effectue une seule fois le calcul de w_i et $\mathbf{grad}\, w_i$ aux points d'intégration. Comme L dépend linéairement de N (cf. proposition 3.13 si $d = 2$, et remarque 3.16 si

26. Dans le cas d'une approximation par élément fini de Lagrange P_k, $k \geq 2$, on effectue les mêmes calculs, avec N le nombre de degrés de libertés, et $(M_i)_{i=1,N}$ la localisation de ces ddl.

27. Si la matrice $\mathbb{A}$ est symétrique, on peut adapter cet algorithme en réduisant la boucle sur j aux $j \geq i$ et en posant $\mathbb{A}_{j,i} = \mathbb{A}_{i,j}$ à la fin de la boucle sur j.

$d = 3$), on est passé d'une complexité quadratique à une complexité linéaire pour le calcul de la matrice $\mathbb{A}$. On peut bien sûr assembler le second membre F de (3.14) de la même façon ; et il n'est pas difficile de vérifier que l'agorithme d'assemblage du second membre est de complexité $O(L(d+1))$.

4.5 Utilisation du calcul parallèle

Afin de réduire le temps calcul, une solution est d'utiliser le *calcul parallèle*. En algèbre linéaire numérique, nous avons déjà évoqué deux opérations élémentaires au §4.2 : le produit scalaire, et la multiplication matrice-vecteur. Celles-ci constituent la base algorithmique des méthodes itératives. [28] On va examiner comment on peut *paralléliser* ces opérations pour les deux méthodes de discrétisation, à savoir les différences finies et les élements finis.

4.5.1 Un modèle pour l'architecture des machines parallèles

Notre modèle est celui d'une machine disposant de $\mathtt{P}$ nœuds de calcul. Chaque nœud est composé d'un ensemble de processeurs et d'une mémoire "locale", partagée par les processseurs qui le composent.
Les processeurs réalisent les calculs. Les nœuds de calcul sont reliés entre eux par un réseau de communication permettant l'échange de données. Pour simplifier la présentation, on suppose que les nœuds de calcul sont identiques, de même que les liens de communication, et enfin que chaque paire de nœuds dispose d'un lien de communication. Pour plus de détails, voir la page web [27].
Concernant les communications, on suppose que les nœuds ne peuvent envoyer ou recevoir qu'un seul message (i.e. un ensemble de données) à la fois. On note $\mathtt{p} \to \mathtt{q}$ la communication formalisant l'envoi de données du nœud $\mathtt{p}$ au nœud $\mathtt{q}$. Si les deux communications $\mathtt{p}_1 \to \mathtt{q}_1$ et $\mathtt{p}_2 \to \mathtt{q}_2$ doivent se produire simultanément [29], alors : si $\mathtt{p}_1 \neq \mathtt{p}_2$ et $\mathtt{q}_1 \neq \mathtt{q}_2$, les communications sont *parallèles* ; sinon, les communications sont *séquentielles*. L'utilisation de schéma de communications parallèles peut avoir son importance lorsque la part des communications (rapportée au coût total calculs + communications) devient non-négligeable.

4.5.2 Répartition des données

Pour le produit scalaire $\alpha = (x, y)$ avec $x, y \in \mathbb{K}^N$ la répartition des données est triviale : pour chaque indice $1 \leq i \leq N$, les données x_i et y_i sont stockés sur un même nœud.
Pour le produit matrice vecteur $z = \mathbb{A}y$ avec $\mathbb{A} \in \mathbb{K}^{N \times N}$ et $y, z \in \mathbb{K}^N$, il faut stocker le vecteur y, et construire et stocker la matrice $\mathbb{A}$. Concernant le stockage, on retient le modèle suivant. On suppose que les données $(y_i)_{i=1,N}$ et $(\mathbb{A}_{i,j})_{i,j=1,N}$ sont réparties comme suit : pour chaque indice $1 \leq i \leq N$, les données y_i et $(\mathbb{A}_{i,j})_{j=1,N}$ (la i^{eme} ligne de $\mathbb{A}$) et le résultat z_i sont stockés sur un même nœud. Une contrainte est que les données soient réparties "équitablement", ou "le plus équitablement possible", sur chaque nœud, pour équilibrer les calculs.

28. Il y a également la résolution d'un système linéaire par méthode directe que nous étudierons uniquement d'un point de vue séquentiel. Pour l'adaptation au calcul parallèle, nous renvoyons à Meurant [26].
29. Sous réserve que les données nécessaires soient disponibles sur $\mathtt{p}_1$ et $\mathtt{q}_1$.

4.6 Parallélisation du produit scalaire

On suppose que les composantes des deux vecteurs $x, y \in \mathbb{K}^N$ sont réparties équitablement sur les $\mathtt{P}$ nœuds, c'est-à-dire qu'il y a $N/\mathtt{P}$ ou $N/\mathtt{P}+1$ (où / est la division entière) composantes par nœud. Pour effectuer le produit scalaire (x, y), on effectue successivement les opérations suivantes :

- calcul simultané du produit scalaire (partiel) sur chacun des nœuds ;
- envoi des valeurs des produits scalaires partiels sur un nœud donné (communication de type *réduction*) ;
- calcul du produit scalaire (x, y) (la somme des produits scalaires partiels) sur ce nœud ;
- envoi du résultat (x, y) à tous les nœuds (communication de type *diffusion*).

Par définition, les communications de type réduction ou diffusion sont séquentielles.
Comme les processeurs sont identiques et opérent sur le même nombre de données, les calculs sont équirépartis. S'ils débutent simultanément, ils s'achèvent simultanément et le temps calcul est réduit d'un facteur $\mathtt{P}$.
Si on néglige le coût des communications (réduction, diffusion) et de la somme, le temps écoulé est divisé d'un facteur $\mathtt{P}$ par rapport au calcul séquentiel du produit scalaire (x, y) sur un nœud unique. Le parallèlisme est *maximal* sur l'architecture à $\mathtt{P}$ nœuds.
Si au contraire le coût des communications est non-négligeable, on peut choisir de dupliquer certains calculs afin de pouvoir paralléliser les communications. Citons notamment le schéma "papillon" qui permet de passer, pour les réductions et diffusions, de $\mathtt{P}$ communications séquentielles à $\log_2(\mathtt{P}-1)+1$ communications parallèles.

4.7 Parallélisation du produit matrice-vecteur pour les différences finies

Commençons par les différences finies. Comme par essence ces schémas possèdent une structure (tensorielle), on choisit une répartition structurée des données : on découpe le domaine de calcul $\Omega =]0,1[^d$ contenant $N = n^d$ points de discrétisation en sous-domaines $(\Omega_{\mathtt{p}})_{\mathtt{p}=1,\mathtt{P}}$ de forme tensorielle. On fait en sorte que les sous-domaines contiennent le même nombre de points de discrétisation, égal à $N/\mathtt{P}$.[30] Typiquement, si $\mathtt{P} = \mathtt{Q}^d$, on choisit $0 = x^0 < x^1 < x^2 < \cdots < x^{\mathtt{Q}-1} < x^{\mathtt{Q}} = 1$ tel que chaque intervalle $]x^{q-1}, x^q[$ $(1 \leq q \leq \mathtt{Q})$ contient $n/\mathtt{Q}$ point de discrétisation. On pose

$$\Omega_{\mathtt{p}} =]x^{q_1-1}, x^{q_1}[\times]x^{q_2-1}, x^{q_2}[\times\cdots]x^{q_d-1}, x^{q_d}[,$$

avec $1 \leq q_i \leq \mathtt{Q}$ pour $1 \leq i \leq d$.

Ainsi, pour tout $\mathtt{p}$, $\Omega_{\mathtt{p}}$ contient $(n/\mathtt{Q})^d$ points de discrétisation, c'est-à-dire $N/\mathtt{P}$ points comme annoncé. On affecte les données correspondant aux points de discrétisation situés dans $\Omega_{\mathtt{p}}$ au nœud de calcul $\mathtt{p}$.
A l'instar de la numérotation des points de discrétisation (voir la figure 2.5), on a la correspondance : $\mathtt{p} = q_1 + (q_2-1)\mathtt{Q} + \cdots + (q_d-1)\mathtt{Q}^{d-1}$, voir un exemple pour $d = 2$ à la

30. Pour une approche où on examine plus spécifiquement l'obtention théorique d'un "parallélisme maximal", c'est-à-dire où chaque sous-domaine contienne quelques points de discrétisation, on renvoie au §6.13.

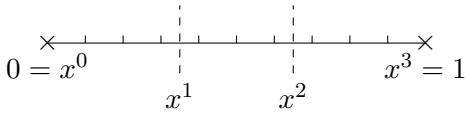


FIGURE 4.1 – Exemple avec $n = 9$ points de discrétisation par direction, et $\mathtt{Q} = 3$.

figure 4.2.

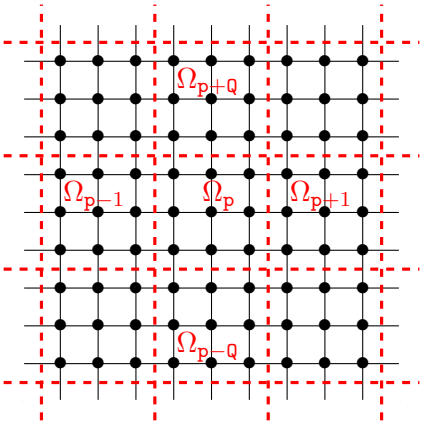


FIGURE 4.2 – Numérotation des sous-domaines pour $d = 2$.

Afin de pouvoir réaliser le produit matrice-vecteur localement (et en parallèle), il faut récupérer les données $(y_i)_{i\in\mathcal{J}_{\mathtt{p}}^{ext}}$, où $\mathcal{J}_{\mathtt{p}}^{ext}$ est l'ensemble des indices des points n'appartenant pas à $\Omega_{\mathtt{p}}$, mais qui interviennent dans au moins un schéma à $(2d+1)$ points d'un point de $\Omega_{\mathtt{p}}$. Pour l'exemple de la figure 4.2, ceci concerne des points de $\Omega_{\mathtt{p}-\mathtt{Q}}$, $\Omega_{\mathtt{p}-1}$, $\Omega_{\mathtt{p}+1}$ et $\Omega_{\mathtt{p}+\mathtt{Q}}$, sauf lorsque le sous-domaine $\Omega_{\mathtt{p}}$ possède une frontière commune avec Ω. L'intérêt est que le schéma de communications résultant est lui aussi structuré. En effet, on doit récupérer les données provenant du :

— sous-domaine voisin bas $[\Omega_{\mathtt{p}-\mathtt{Q}}]$;
— sous-domaine voisin gauche $[\Omega_{\mathtt{p}-1}]$;
— sous-domaine voisin droite $[\Omega_{\mathtt{p}+1}]$;
— sous-domaine voisin haut $[\Omega_{\mathtt{p}+\mathtt{Q}}]$.

(Si ces sous-domaines existent).
Si on prend en compte la possibilité de réaliser des communications parallèles, on note que celles-ci sont très simples à mettre en œuvre ! Par exemple, la récupération de *toutes* les données venant des sous-domaines de gauche se fait en *deux étapes* de communications parallèles, quelle que soit la valeur de $\mathtt{Q}$. Dans la configuration de la figure 4.3, on procède comme suit :

[étape 1] tous les sous-domaines de numéro $\mathtt{p}$ pair reçoivent simultanément les données de leur voisin de gauche ;
[étape 2] tous les sous-domaines de numéro $\mathtt{p}$ impair tel que $\mathtt{p} \neq 1 \bmod [4]$ reçoivent

Ω_{13}	Ω_{14}	Ω_{15}	Ω_{16}
Ω_9	Ω_{10}	Ω_{11}	Ω_{12}
Ω_5	Ω_6	Ω_7	Ω_8
Ω_1	Ω_2	Ω_3	Ω_4

FIGURE 4.3 – Sous-domaines pour $d = 2$ et $\mathtt{Q} = 4$.

simultanément les données de leur voisin de gauche.
Ces deux étapes sont bien parallèles au sens de la définition du §4.5.1. Si on procède de même pour la réception des données des voisins bas, droite et haut, il y a en tout *huit étapes* de communications parallèles. On vérifie sans peine que ce résultat est indépendant de $\mathtt{Q}$. Qui plus est, ces communications parallèles sont structurées. Enfin, en dimension d, on peut vérifier que ce nombre de communications parallèles est égal à $2d$.

A partir de là, les calculs sont réalisés en parallèle sur chaque nœud selon le schéma initial à $2d+1$ points. Comme les sous-domaines contiennent le même nombre de points, les calculs sont bien équirépartis sont les $\mathtt{P}$ nœuds, puisqu'ils disposent de processeurs identiques. S'ils débutent simultanément, ils s'achèvent simultanément et le temps calcul est réduit d'un facteur $\mathtt{P}$.

La conclusion générale est que la structure initiale peut-être utilisée à toutes les phases de la parallélisation :
- la répartition des données est structurée ;
- les communications sont structurées ;
- les calculs procèdent localement selon le schéma initial à $2d + 1$ points.

4.8 Parallélisation du produit matrice-vecteur pour les éléments finis

Pour les éléments finis, la problématique est différente : il n'y a pas de structure "intrinsèque" telle que le schéma à trois ou cinq points pour les différences finies. Par conséquent, on adopte un point de vue un peu différent : l'idée est maintenant de préserver les algorithmes de construction des matrice $\mathbb{A}$ et second membre F de (3.14) découlant de la discrétisation des formulations variationnelles.
Pour cela, on va choisir de partitionner les maillages, ce qui revient à découper le domaine $\Omega \subset \mathbb{R}^d$, $d = 2, 3$. Pour rappel, voir le §3.3, on doit construire un maillage de $\overline{\Omega}$, composé de simplexes fermés. L'idée est de partir d'une partition $\{\Omega_{\mathtt{p}}\}_{\mathtt{p}=1,\cdots,\mathtt{P}}$ de Ω, voir la note de bas de page[17] page 68, avec des interfaces entre sous-domaines polygonales si $d = 2$, respectivement polyédriques si $d = 3$. On maille chacun des sous-domaines $\Omega_{\mathtt{p}}$, de sorte

que les maillages $(\mathcal{T}_h^{\mathtt{p}})_{\mathtt{p}=1,\mathtt{P}}$ soient compatibles aux interfaces entre sous-domaines :

$$\left\{\begin{array}{l} \text{pour tout } \mathtt{p}\in\{1,\cdots,\mathtt{P}\}, \text{ toute facette d'un simplexe de } \mathcal{T}_h^{\mathtt{p}} \text{ située sur } \partial\Omega_{\mathtt{p}} \text{ est} \\ \quad \text{soit facette d'un simplexe de } \mathcal{T}_h^{\mathtt{q}} \text{ pour un } \mathtt{q}\neq\mathtt{p}, \\ \quad \text{soit incluse dans } \partial\Omega. \end{array}\right. \tag{4.3}$$

On appelle $\Sigma_{\mathtt{pq}}$ l'interface entre $\Omega_{\mathtt{p}}$ et $\Omega_{\mathtt{q}}$ pour $\mathtt{p}\neq\mathtt{q}$, telle que $\Sigma_{\mathtt{pq}} = int(\overline{\Omega_{\mathtt{p}}}\cap\overline{\Omega_{\mathtt{q}}})$ si la dimension d'Hausdorff de $\overline{\Omega_{\mathtt{p}}}\cap\overline{\Omega_{\mathtt{q}}}$ est égale à $d-1$: dans ce cas, on dit que les sous-domaines sont voisins ; et $\Sigma_{\mathtt{pq}}=\emptyset$ sinon. Par construction, on a automatiquement $\Sigma_{\mathtt{pq}}=\Sigma_{\mathtt{qp}}$. La réunion des interfaces est notée $\Sigma_S = \cup_{\mathtt{p},\mathtt{q}}\overline{\Sigma_{\mathtt{pq}}}$.

Remarque 4.2 *En pratique, si on dispose d'un logiciel de génération de maillage, celui-ci génère un maillage de $\overline{\Omega}$, partitionné pour définir des sous-domaines $\{\Omega_{\mathtt{p}}\}_{\mathtt{p}=1,\cdots,\mathtt{P}}$ en respectant certains critères (nombre identique de simplexes par sous-domaine), et minimisation de la taille de Σ_S, où la taille est égale au nombre total de facettes se trouvant sur une interface $\Sigma_{\mathtt{pq}}=\Sigma_{\mathtt{qp}}$. On verra une justification de ces critères en fin de chapitre.*

A partir de maintenant et pour fixer les idées, on considère pour le problème (3.1) la discrétisation obtenue pour l'élément fini de Lagrange P_1, et on appelle $(M_i)_{1\leq i\leq N}$ les sommets (internes) du maillage. Pour paralléliser les calculs, l'idée est d'assembler la matrice $\mathbb{A}$ et le second membre F par sous-domaine. Après assemblage parallèle pour chaque $\mathtt{p}$ sur les simplexes appartenant à $\overline{\Omega_{\mathtt{p}}}$, on a construit des matrices $(\mathbb{A}_{\mathtt{p}}^+)_{\mathtt{p}=1,\mathtt{P}}\in\mathbb{R}^{N\times N}$, et des seconds membres $(F_{\mathtt{p}}^+)_{\mathtt{p}=1,\mathtt{P}}\in\mathbb{R}^N$. On se concentre sur les matrices dans la suite, la procédure étant similaire pour les seconds membres. Si on note $N_{\mathtt{p}}^+$ le nombre de sommets situés dans $\overline{\Omega_{\mathtt{p}}}\setminus\partial\Omega$, on a a priori $N_{\mathtt{p}}^+$ lignes non-nulles dans $\mathbb{A}_{\mathtt{p}}^+$. On note également $N_{\mathtt{p}}$ le nombre de sommets situés dans $int(\Omega_{\mathtt{p}})$.
L'assemblage sur le nœud $\mathtt{p}$ consiste en :

initialisation
$\mathbb{A}_{\mathtt{p}}^+ = 0\in\mathbb{R}^N\times\mathbb{R}^N$.
itérations : pour $\ell=1,\cdots,L_{\mathtt{p}}$**, faire**
itérations : pour i **t.q.** $M_i\in T_\ell$**, faire**
itérations : pour j **t.q.** $M_j\in T_\ell$ **faire**

$$(\mathbb{A}_{\mathtt{p}}^+)_{i,j} = (\mathbb{A}_{\mathtt{p}}^+)_{i,j} + \int_{int(T_\ell)} \left(k\,\mathbf{grad}\,w_{j|T_\ell}\cdot\mathbf{grad}\,w_{i|T_\ell} + q\,w_{j|T_\ell}w_{i|T_\ell}\right)d\boldsymbol{x}$$

fin boucle sur j.
fin boucle sur i.
fin boucle sur ℓ.

Si on compare les matrices $(\mathbb{A}_{\mathtt{p}}^+)_{\mathtt{p}=1,\mathtt{P}}$ à la matrice $\mathbb{A}$, on a deux situations.

— Soit le sommet $M_i\notin\Sigma_S$, c'est-à-dire qu'il existe $\mathtt{p}_i$ tel que $M_i\in int(\Omega_{\mathtt{p}_i})$. On dit que M_i est un **sommet interne** (à $\Omega_{\mathtt{p}_i}$). Dans ce cas, la i^{eme} ligne de $\mathbb{A}$ se retrouve intégralement dans la i^{eme} ligne de la matrice $\mathbb{A}_{\mathtt{p}_i}^+$. Enfin, dans toutes les matrices $(\mathbb{A}_{\mathtt{p}}^+)_{\mathtt{p}\neq\mathtt{p}_i}$, la i^{eme} ligne est nulle.

— Soit le sommet $M_i\in\Sigma_S$. On dit alors que M_i est un **sommet d'interface**. Dans ce cas, la i^{eme} ligne de $\mathbb{A}$ se trouve répartie entre les i^{emes} lignes des matrices $(\mathbb{A}_{\mathtt{p}}^+)_{\mathtt{p}\,tq\,M_i\in\partial\Omega_{\mathtt{p}}}$. Pour un sommet d'interface M_i donné : soit il appartient à exactement deux frontières distinctes de sous-domaine, soit il appartient à au moins trois

frontières distinctes. Dans la seconde configuration, on dit que M_i est un **point de croisement**. Cette configuration ne peut se produire que si $\mathtt{P} \geq 3$.

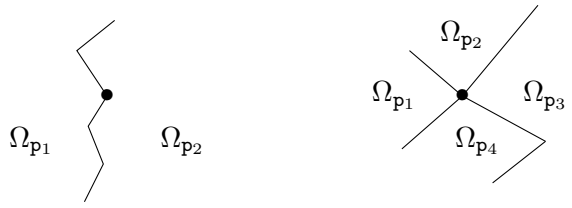


FIGURE 4.4 – Sommets d'interface • localisés sur exactement deux, ou sur au moins trois, frontières distinctes de sous-domaine. A droite • est un point de croisement.

Voir des exemples à la figure 4.4 lorsque $d = 2$. Lorsque $d = 3$ la situation est similaire, mais un peu plus compliquée car les points de croisement peuvent être situés sur des arêtes (ouvertes, à l'exclusion des extrémités), ou des sommets, de l'interface.

On conclut de la discussion qui précède que, pour les indices correspondant aux sommets d'interface, il faut encore assembler des contributions. Il y a deux techniques "naturelles" : la première consiste à assigner l'ensemble des données associées aux sommets d'interface à un nœud dédié ; la seconde consiste à réaliser le produit matrice-vecteur avec les matrices $(\mathbb{A}_{\mathtt{p}}^{+})_{\mathtt{p}=1,\mathtt{P}}$ ainsi construites, et dans ce cas les résultats sont partiels pour les indices correspondant aux sommets d'interface, puis à échanger les données associées aux sommets d'interface entre nœuds afin de finaliser le calcul.

Dans la suite on retient la première technique : on dispose de $\mathtt{P}+1$ nœuds, et on stocke les données correspondant aux sommets d'interface sur le nœud $\mathtt{P}+1$.
Ainsi, après assemblage sur les nœuds $\mathtt{p} = 1, \mathtt{P}$, on va transférer (via une phase de communication séquentielle) les lignes i des matrices $(\mathbb{A}_{\mathtt{p}}^{+})_{\mathtt{p}=1,\mathtt{P}}$ concernées au nœud $\mathtt{P}+1$, afin de construire la matrice $\mathbb{A}_{\mathtt{P}+1} \in \mathbb{R}^{N\times N}$ par assemblage sur ce nœud. En outre, on choisit de remettre à 0 les lignes i des matrices $(\mathbb{A}_{\mathtt{p}}^{+})_{\mathtt{p}=1,\mathtt{P}}$, et le résultat est noté $(\mathbb{A}_{\mathtt{p}})_{\mathtt{p}=1,\mathtt{P}}$. Pour résumer, à l'issue d'un assemblage en deux temps, on dispose :

— dans $(\mathbb{A}_{\mathtt{p}})_{\mathtt{p}=1,\mathtt{P}}$, des informations pour les sommets internes à $(\Omega_{\mathtt{p}})_{\mathtt{p}=1,\mathtt{P}}$;
— dans $\mathbb{A}_{\mathtt{P}+1}$, des informations pour les sommets d'interface.

La construction du second membre est similaire.
On note

$$\left\{\begin{array}{rcl|rcl} \text{Pour } \mathtt{p}=1,\mathtt{P} \ : \ \mathcal{I}_{\mathtt{p}} & = & \{i\in\{1,\cdots,N\}\,|\,M_i\in int(\Omega_{\mathtt{p}})\}, & N_{\mathtt{p}} & = & |\mathcal{I}_{\mathtt{p}}| \text{ et} \\ \mathcal{I}_{\mathtt{p}}^{+} & = & \{i\in\{1,\cdots,N\}\,|\,M_i\in \overline{\Omega_{\mathtt{p}}}\}, & N_{\mathtt{p}}^{+} & = & |\mathcal{I}_{\mathtt{p}}^{+}|\,; \\ \mathcal{I}_{\mathtt{P}+1} & = & \{i\in\{1,\cdots,N\}\,|\,M_i\in \Sigma_S\}, & N_{\mathtt{P}+1} & = & |\mathcal{I}_{\mathtt{P}+1}|. \end{array}\right. \tag{4.4}$$

Pour $1 \leq \mathtt{p} \leq \mathtt{P}$ fixé, on suppose l'assemblage de $\mathbb{A}_{\mathtt{p}}^{+}$ réalisé. D'un point de vue matriciel, si on ne retient que les $N_{\mathtt{p}}^{+}$ lignes et $N_{\mathtt{p}}^{+}$ colonnes non-nulles de $\mathbb{A}_{\mathtt{p}}^{+}$ (ie. d'indices appartenant à $\mathcal{I}_{\mathtt{p}}^{+}$), on peut la décomposer par blocs en

$$\begin{pmatrix} \overset{\circ}{\mathbb{A}}_{\mathtt{p},\mathtt{p}} & \mathbb{A}_{\mathtt{p},\Sigma} \\ (\mathbb{A}_{\mathtt{p},\Sigma})^T & \mathbb{A}_{\Sigma,\Sigma}^{\mathtt{p}} \end{pmatrix} = \begin{pmatrix} \overset{\circ}{\mathbb{A}}_{\mathtt{p},\mathtt{p}} & \mathbb{A}_{\mathtt{p},\Sigma} \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ (\mathbb{A}_{\mathtt{p},\Sigma})^T & \mathbb{A}_{\Sigma,\Sigma}^{\mathtt{p}} \end{pmatrix},$$

avec

— $\overset{\circ}{\mathbb{A}}_{\mathtt{p},\mathtt{p}} \in \mathbb{R}^{N_\mathtt{p} \times N_\mathtt{p}}$ correspondant aux interactions entre sommets internes à $\Omega_\mathtt{p}$;

— $\mathbb{A}_{\mathtt{p},\Sigma} \in \mathbb{R}^{N_\mathtt{p} \times (N_\mathtt{p}^+ - N_\mathtt{p})}$ correspondant aux interactions entre sommets internes à $\Omega_\mathtt{p}$ et sommets d'interface situés sur $\partial\Omega_\mathtt{p}$;

— $(\mathbb{A}_{\mathtt{p},\Sigma})^T \in \mathbb{R}^{N_\mathtt{p} \times (N_\mathtt{p}^+ - N_\mathtt{p})}$ correspondant aux interactions entre sommets d'interface situés sur $\partial\Omega_\mathtt{p}$ et sommets internes à $\Omega_\mathtt{p}$;

— $\mathbb{A}^\mathtt{p}_{\Sigma,\Sigma} \in \mathbb{R}^{(N_\mathtt{p}^+ - N_\mathtt{p}) \times (N_\mathtt{p}^+ - N_\mathtt{p})}$ correspondant aux interactions entre sommets d'interface calculées pour les simplexes inclus dans $\overline{\Omega_\mathtt{p}}$.

Un sommet interne contribuant à $\mathbb{A}_{\mathtt{p},\Sigma}$ (et $(\mathbb{A}_{\mathtt{p},\Sigma})^T$) est dit **sommet voisin de l'interface**.

La matrice $\mathbb{A}_\mathtt{p}$ correspond aux blocs $\overset{\circ}{\mathbb{A}}_{\mathtt{p},\mathtt{p}}$ et $\mathbb{A}_{\mathtt{p},\Sigma}$. La contribution mettant en jeu des sommets de $\partial\Omega_\mathtt{p} \cap \Sigma_S$, c'est-à-dire d'une part des sommets d'interface situés sur $\partial\Omega_\mathtt{p}$ et d'autre part des sommets voisins de l'interface situés dans $\Omega_\mathtt{p}$, correspond aux blocs $(\mathbb{A}_{\mathtt{p},\Sigma})^T$ et $\mathbb{A}^\mathtt{p}_{\Sigma,\Sigma}$.

Si on écrit maintenant la matrice $\mathbb{A}$ (globale) par blocs, avec par ordre d'indice ceux de $\mathcal{I}_1$ (1er bloc, sommets internes à Ω_1), de $\mathcal{I}_2$ (2ème bloc, sommets internes à Ω_2), ... jusqu'à ceux de $\mathcal{I}_\mathtt{P}$ (Pème bloc, sommets internes à $\Omega_\mathtt{P}$), et enfin ceux de $\mathcal{I}_{\mathtt{P}+1}$ ((P + 1)ème bloc, sommets d'interface), on a

$$\mathbb{A} = \begin{pmatrix} \overset{\circ}{\mathbb{A}}_{1,1} & 0 & & & \mathbb{A}_{1,\Sigma} \\ 0 & \overset{\circ}{\mathbb{A}}_{2,2} & \ddots & & \mathbb{A}_{2,\Sigma} \\ & \ddots & \ddots & 0 & \vdots \\ & & 0 & \overset{\circ}{\mathbb{A}}_{\mathtt{P},\mathtt{P}} & \mathbb{A}_{\mathtt{P},\Sigma} \\ (\mathbb{A}_{1,\Sigma})^T & (\mathbb{A}_{2,\Sigma})^T & \cdots & (\mathbb{A}_{\mathtt{P},\Sigma})^T & \mathbb{A}_{\Sigma,\Sigma} \end{pmatrix}, \tag{4.5}$$

où on commet un (petit) abus de notation pour les matrices $\mathbb{A}_{\mathtt{P},\Sigma}$, contenant les mêmes informations que précédemment mais appartenant maintenant à $\mathbb{R}^{N_\mathtt{P} \times N_{\mathtt{P}+1}}$, et où la matrice d'interface $\mathbb{A}_{\Sigma,\Sigma} \in \mathbb{R}^{N_{\mathtt{P}+1} \times N_{\mathtt{P}+1}}$ est l'assemblage de toutes les matrices $(\mathbb{A}^\mathtt{p}_{\Sigma,\Sigma})_{\mathtt{p}=1,\mathtt{P}}$. Il est intéressant de noter que les blocs non-diagonaux d'indices $(\mathtt{p},\mathtt{q})$ pour $1 \leq \mathtt{p} \neq \mathtt{q} \leq \mathtt{P}$ sont automatiquement nuls. Cette propriété découle du fait que, si M_i est un sommet interne de $\Omega_\mathtt{p}$ et M_j est un sommet interne de $\Omega_\mathtt{q}$, alors ils ne peuvent pas appartenir à un même simplexe du maillage. D'après la propriété de support des fonctions chapeau, $\mathbb{A}_{i,j} = 0$.

Pour résumer, dans la configuration à $\mathtt{P}+1$ nœuds, on a, pour $1 \leq \mathtt{p} \leq \mathtt{P}$, les informations concernant les sommets internes à $\Omega_\mathtt{p}$ sur le nœud $\mathtt{p}$, et les informations concernant les sommets d'interface sur le nœud $\mathtt{P}+1$. Pour le produit matrice-vecteur $z = \mathbb{A}y$:

— sur le nœud $\mathtt{p}$ ($1 \leq \mathtt{p} \leq \mathtt{P}$) on dispose de $\overset{\circ}{\mathbb{A}}_{\mathtt{p},\mathtt{p}}$ et $\mathbb{A}_{\mathtt{p},\Sigma}$, et du bloc vecteur $y_\mathtt{p}$ contenant $(y_i)_{i \in \mathcal{I}_\mathtt{p}}$; et on veut calculer le bloc vecteur $z_\mathtt{p}$;

— sur le nœud $\mathtt{P}+1$ on dispose de $\left((\mathbb{A}_{\mathtt{p},\Sigma})^T\right)_{\mathtt{p}=1,\mathtt{P}}$ et $\mathbb{A}_{\Sigma,\Sigma}$ et du bloc vecteur $y_{\mathtt{P}+1}$ contenant $(y_i)_{i \in \mathcal{I}_{\mathtt{P}+1}}$; et on veut calculer le bloc vecteur $z_{\mathtt{P}+1}$.

Pour $1 \leq \mathtt{p} \leq \mathtt{P}$, on doit réaliser sur le nœud $\mathtt{p}$ les multiplications par blocs

$$z_\mathtt{p} = \overset{\circ}{\mathbb{A}}_{\mathtt{p},\mathtt{p}} y_\mathtt{p} + \mathbb{A}_{\mathtt{p},\Sigma} y_{\mathtt{P}+1}, \tag{4.6}$$

et on doit réaliser sur le nœud $\mathtt{P}+1$ les multiplications par blocs

$$z_{\mathtt{P}+1} = \mathbb{A}_{\Sigma,\Sigma} y_{\mathtt{P}+1} + \sum_{\mathtt{p}=1,\mathtt{P}} (\mathbb{A}_{\mathtt{p},\Sigma})^T y_{\mathtt{p}}. \tag{4.7}$$

Afin de pouvoir réaliser les calculs en parallèles (localement sur chaque nœud), il faut disposer de $(y_{\mathtt{q}})_{\mathtt{q}\neq\mathtt{p}}$. Et, si c'est le cas, alors on peut effectuer les produits matrice-vecteur simultanément. Comme on l'a déjà vu, il y a une dissymétrie entre les $\mathtt{P}$ premiers nœuds, et le nœud $\mathtt{P}+1$. Ceci se retrouve dans les communications, puisque tous les sous-domaines doivent communiquer des données à l'interface (et vice-versa), mais que les sous-domaines ne communiquent pas entre eux. Ainsi on effectue les communications parallèles, pour $1 \leq \mathtt{p} \leq \mathtt{P}$:

- $(\mathtt{P}+1 \to \mathtt{p})$ le nœud $\mathtt{p}$ reçoit les données $y_{\mathtt{P}+1}$ du nœud $\mathtt{P}+1$ (en pratique, on se restreint aux indices i tels que $M_i \in \partial\Omega_{\mathtt{p}} \cap \Sigma_S$) ;
- $(\mathtt{p} \to \mathtt{P}+1)$ le nœud $\mathtt{P}+1$ reçoit les données $y_{\mathtt{p}}$ du nœud $\mathtt{p}$ (en pratique, on se restreint aux indices correspondants aux voisins de l'interface situés dans $\Omega_{\mathtt{p}}$).

Une fois ces $\mathtt{P}$ communications parallèles effectuées, on peut effectuer les produits matrice-vecteur (4.6) et (4.7) simultanément.

Pour conclure, notons que les algorithmes que nous avons proposés dans le §4.8 sont couramment utilisés pour paralléliser les produits matrice-vecteur intervenant dans les méthodes de décomposition de domaine, voir la partie III.

Exercice 4.1 *Détailler la procédure (sous forme algorithmique) permettant de réaliser la parallélisation du produit matrice-vecteur selon la seconde technique de la page 91, c'est-à-dire celle basée sur les calculs matrice-vecteur avec les matrices* $(\mathbb{A}^+_{\mathtt{p}})_{\mathtt{p}=1,\mathtt{P}}$*, puis l'échange de données pour finaliser le calcul du second membre.*

On a vu que, pour les différences finies, les calculs sont automatiquement équirépartis grâce à leur structure intrinsèque. Qu'en est-il pour les éléments finis ?
Soit $\mathtt{p}$ donné. Pour tous les indices $i \in \mathcal{I}_{\mathtt{p}}$, on effectue le calcul $\sum_{j=1,N} \mathbb{A}_{i,j} y_j$, où, pour la matrice, les informations sont regroupées dans les blocs $\overset{\circ}{\mathbb{A}}_{\mathtt{p},\mathtt{p}}$ et $\mathbb{A}_{\mathtt{p},\Sigma}$. On va estimer le nombre total d'opérations réalisées en fonction des caractéristiques du maillage de $\overline{\Omega_{\mathtt{p}}}$ (on examine le cas $d = 2$ ci-dessous).
Pour chaque arête a interne à $\Omega_{\mathtt{p}}$, c'est-à-dire telle que $\overline{a} \subset \overline{\Omega_{\mathtt{p}}}$ et $\overline{a} \not\subset \partial\Omega_{\mathtt{p}}$, on effectue 4 opérations (2 multiplications et 2 additions), sauf si une extrémité se trouve sur $\partial\Omega_{\mathtt{p}}$, auquel cas on effectue 2 opérations (1 multiplication et 1 addition). Si on revient aux blocs, le premier cas correspond à des calculs dans le bloc $\overset{\circ}{\mathbb{A}}_{\mathtt{p},\mathtt{p}}$, et le second à des calculs dans le bloc $\mathbb{A}_{\mathtt{p},\Sigma}$.
Dans $\overline{\Omega_{\mathtt{p}}}$, on a $L_{\mathtt{p}}$ triangles et $A^+_{\mathtt{p}}$ arêtes avec $A^+_{\mathtt{p}} = A_{\mathtt{p}} + A_{\mathtt{p},b}$, où $A_{\mathtt{p},b}$ est le nombre d'arêtes incluses dans $\partial\Omega_{\mathtt{p}}$. On a également $N^+_{\mathtt{p}} = N_{\mathtt{p}} + N_{\mathtt{p},b}$ sommets, avec $N_{\mathtt{p},b} = A_{\mathtt{p},b}$ sommets situés sur $\partial\Omega_{\mathtt{p}}$. Si on néglige les contributions concernant la frontière (asymptotiquement, si on note h le pas du maillage, on a $\lim_{h\to 0} A_{\mathtt{p},b}/A_{\mathtt{p}} = 0$), on effectue donc $4A_{\mathtt{p}}$ opérations. Si on reprend le décompte des arêtes de la démonstration de la proposition 3.13 qui donne la relation $3L = 2A_{\mathtt{p}} + A_{\mathtt{p},b}$, on en conclut qu'on va asymptotiquement effectuer $6L_{\mathtt{p}}$ opérations.

Ainsi, pour avoir des calculs asymptotiquement équirépartis sur les P premiers nœuds, il faut donc construire des maillages comportant le même nombre de triangles par sous-domaine, c'est-à-dire que $L_{\mathtt{p}} = L/\mathtt{P}$ pour $1 \leq \mathtt{p} \leq \mathtt{P}$. Ceci justifie le premier critère à respecter par les logiciels de génération de maillages (voir la remarque 4.2). Quant au second critère de cette remarque, il exprime simplement le fait qu'on veut minimiser le coût des calculs réalisés sur le nœud $\mathtt{P}+1$ qui gère l'interface. Si toutefois ce dernier coût reste trop important comparé à celui des autres nœuds, une solution possible consiste à réserver 2 nœuds, ou plus, pour gérer l'interface.

Exercice 4.2 *Evaluer la répartition des calculs lorsque la parallélisation du produit matrice-vecteur selon la seconde technique de la page 91.*

Chapitre 5

Les méthodes directes

5.1 Introduction

En préambule, notons que le calcul de la solution d'un système linéaire d'ordre 20 par les formules de Cramer est impossible à réaliser sur un ordinateur, car il requiert environ 10^{21} opérations élémentaires $(+, -, *, /)$ [31]. En conséquence, il est nécessaire de déterminer d'autres méthodes, qui soient utilisables en pratique, par exemple pour résoudre les systèmes linéaires obtenus après discrétisation.

On rappelle pour commencer que si la matrice d'un système linéaire est diagonale ou de forme triangulaire, ceci apporte une simplification importante dans le calcul explicite de la solution de ce système. Comment utiliser cette particularité pour traiter le cas général ? Une première approche conduit à la méthode dite d'élimination, une seconde à la méthode de factorisation. Ces méthodes sont décrites dans ce chapitre. Les algorithmes classiques qui en découlent sont appelés **méthodes directes** ; les méthodes de Gauss, Crout et Cholesky font partie de cet ensemble d'algorithmes, leur étude fait l'objet de ce chapitre dans les paragraphes 5.9, 5.10, 5.11 respectivement. Le §5.14 constitue une introduction à l'utilisation pratique de ces méthodes pour la résolution de systèmes linéaires à matrice à faible largeur de bande, tels que ceux obtenus après discrétisation. Enfin, dans le §5.15, on indique le coût calcul des méthodes utilisées.

31. En effet, pour calculer un déterminant d'ordre n il faut faire la somme de $n!$ produits de n facteurs, soit au total $n \times n!$ opérations :

$$\det(\mathbb{A}) = \sum_{\sigma \in \mathcal{S}(n)} \varepsilon(\sigma) \mathbb{A}_{1,\sigma(1)}\, \mathbb{A}_{2,\sigma(2)} \cdots \mathbb{A}_{n,\sigma(n)}.$$

Le coût calcul des n composantes du vecteur x, solution du système linéaire $\mathbb{A}x = b$ par les formules de Cramer, est donc de $n(n+1) \times n!$ opérations. Ainsi pour la résolution d'un système linéaire d'ordre 10 (respectivement 20), environ 4.10^8 opérations (resp. 10^{21} opérations) sont nécessaires pour calculer la solution par les formules de Cramer.

5.2 Systèmes linéaires simples à résoudre

5.2.1 Système linéaire à matrice diagonale

Définition 5.1 *On appelle matrice* **diagonale** *une matrice* $\mathbb{D} \in \mathbb{K}^{n\times n}$ *($\mathbb{D}$ comme Diagonale), telle que* $\mathbb{D}_{i,j} = 0$ *pour tout couple* (i,j) *tel que* $i \neq j$.

Proposition 5.2 *Le déterminant d'une matrice diagonale* $\mathbb{D} \in \mathbb{K}^{n\times n}$ *est égal au produit des éléments diagonaux :*

$$\det(\mathbb{D}) = \prod_{i=1}^{n} \mathbb{D}_{i,i}.$$

Proposition 5.3 *Soit* $\mathbb{D} \in \mathbb{K}^{n\times n}$ *une matrice diagonale inversible, la solution du système linéaire* $\mathbb{D}y = b$ *dans* $\mathbb{K}^n$ *est obtenue par les formules :*

pour $i = 1, \ldots, n$ **faire**
  $y_i = b_i/\mathbb{D}_{i,i}$.
fin

5.2.2 Système linéaire à matrice triangulaire

Définition 5.4 *On appelle matrice* **triangulaire inférieure** *une matrice* $\mathbb{L} \in \mathbb{K}^{n\times n}$ *($\mathbb{L}$ comme Lower), telle que* $\mathbb{L}_{i,j} = 0$ *pour tout couple* (i,j) *tel que* $i < j$.
On appelle matrice **triangulaire supérieure** *une matrice* $\mathbb{U} \in \mathbb{K}^{n\times n}$ *($\mathbb{U}$ comme Upper), telle que* $\mathbb{U}_{i,j} = 0$ *pour tout couple* (i,j) *tel que* $j < i$.

$$\mathbb{L} = \begin{bmatrix} \mathrm{x} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \mathrm{x} & \mathrm{x} & 0 & 0 & 0 & 0 & 0 & 0 \\ \mathrm{x} & \mathrm{x} & \mathrm{x} & 0 & 0 & 0 & 0 & 0 \\ \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & 0 & 0 & 0 & 0 \\ \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & 0 & 0 & 0 \\ \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & 0 & 0 & 0 \\ \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & 0 & 0 \\ \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \end{bmatrix} \quad \text{et} \quad \mathbb{U} = \begin{bmatrix} \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & 0 & 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathrm{x} \end{bmatrix}$$

Proposition 5.5 *Le déterminant d'une matrice triangulaire* $\mathbb{T} \in \mathbb{K}^{n\times n}$ *(supérieure ou inférieure) est égal au produit des éléments diagonaux :*

$$\det(\mathbb{T}) = \prod_{i=1}^{n} \mathbb{T}_{i,i}.$$

Remarque 5.6 *La Proposition 5.5 montre que le déterminant d'une matrice triangulaire est très facile à calculer. Cette propriété peut être exploitée pour le calcul du déterminant d'une matrice carrée quelconque* $\mathbb{A} \in \mathbb{K}^{n\times n}$. *Si on suppose que l'on peut écrire la matrice* $\mathbb{A}$ *sous la forme d'un produit* $\mathbb{A} = \mathbb{L}\mathbb{U}$ *dans lequel* $\mathbb{L}$ *est un matrice triangulaire inférieure et* $\mathbb{U}$ *une matrice triangulaire supérieure (voir la suite du chapitre), on peut alors écrire*

$$\det(\mathbb{A}) = \det(\mathbb{L})\det(\mathbb{U}) = \prod_{i=1}^{n} \mathbb{L}_{i,i} \prod_{i=1}^{n} \mathbb{U}_{i,i}.$$

On montre facilement par récurrence les résultats ci-dessous.

Proposition 5.7 *Soit $\mathbb{L} \in \mathbb{K}^{n\times n}$ une matrice triangulaire inférieure inversible, la solution du système linéaire $\mathbb{L}y = b$ dans $\mathbb{K}^n$ est obtenue par les formules :*

$$\left\|\begin{array}{l} \textbf{pour } i = 1, \ldots, n \textbf{ faire} \\ \quad y_i = [b_i - \sum_{j<i} \mathbb{L}_{i,j} y_j]/\mathbb{L}_{i,i}. \\ \textbf{fin} \end{array}\right.$$

Soit $\mathbb{U} \in \mathbb{K}^{n\times n}$ une matrice triangulaire supérieure inversible, la solution du système linéaire $\mathbb{U}x = y$ dans $\mathbb{K}^n$ est obtenue par les formules :

$$\left\|\begin{array}{l} \textbf{pour } i = n, \ldots, 1 \textbf{ faire} \\ \quad x_i = [y_i - \sum_{j>i} \mathbb{U}_{i,j} x_j]/\mathbb{U}_{i,i}. \\ \textbf{fin} \end{array}\right.$$

Pour une matrice triangulaire supérieure, ces relations montrent que l'on peut calculer le vecteur y de proche en proche, en commençant par la dernière composante y_n ; on dit alors que l'on résout le système linéaire **en remontant**. Pour le système linéaire à matrice triangulaire inférieure, on calcule également le vecteur x de proche en proche, en commençant par la première composante x_1 ; on dit que l'on résout le système linéaire **en descendant**.

Remarque 5.8 *Noter que l'hypothèse d'inversibilité entraîne que les éléments diagonaux de deux matrices sont tous différents de zéro, puisque*

$$\det(\mathbb{L}) = \prod_{i=1}^{n} \mathbb{L}_{i,i} \quad \textit{et} \quad \det(\mathbb{U}) = \prod_{i=1}^{n} \mathbb{U}_{i,i}.$$

5.2.3 Conclusion

Ces propriétés sur les matrices diagonales ou triangulaires nous conduisent naturellement à construire d'autres approches de la résolution des systèmes linéaires que celle basée sur les formules de Cramer : on essaiera de se ramener au cas de matrices diagonales ou triangulaires.

5.3 Partition des matrices et vecteurs en blocs

5.3.1 Définition des blocs

Les notions précédentes concernent une approche des matrices élément par élément ; il est souvent utile d'effectuer une partition de la matrice $\mathbb{A} \in \mathbb{K}^{n\times n}$ en $P \times P$ blocs, pour écrire formellement

$$\begin{bmatrix} [\mathbb{A}]_{1,1} & [\mathbb{A}]_{1,2} & \cdots & [\mathbb{A}]_{1,P} \\ [\mathbb{A}]_{2,1} & [\mathbb{A}]_{2,2} & \ddots & [\mathbb{A}]_{2,P} \\ \vdots & \ddots & \ddots & \vdots \\ [\mathbb{A}]_{P,1} & [\mathbb{A}]_{P,2} & \cdots & [\mathbb{A}]_{P,P} \end{bmatrix}.$$

Dans cette écriture les $[\mathbb{A}]_{p,q} \in \mathbb{K}^{n_p \times n_q}$ sont P^2 matrices appelées **blocs** avec $(n_q)_{q=1,P}$ des nombres entiers non-nuls tels que $\sum_{q=1,P} n_q = n$. Seuls les P blocs diagonaux sont *nécessairement carrés*. Pour effectuer une multiplication matrice-vecteur avec $x \in \mathbb{K}^n$, ce dernier est alors partitionné selon

$$\begin{bmatrix} [x]_1 \\ [x]_2 \\ \vdots \\ [x]_P \end{bmatrix}.$$

où $[x]_p \in \mathbb{K}^{n_p}$ sont P vecteurs.
Les définitions précédentes se généralisent naturellement suivant :

Définition 5.9 *On appelle matrice* **triangulaire inférieure par blocs** *une matrice* $\mathbb{L} \in \mathbb{K}^{n\times n}$ *telle que* $[\mathbb{L}]_{p,q} = 0$ *pour tout couple* (p,q) *tel que* $p < q$ *;*

On appelle matrice **triangulaire supérieure par blocs** *une matrice* $\mathbb{U} \in \mathbb{K}^{n\times n}$ *telle que* $[\mathbb{U}]_{p,q} = 0$ *pour tout couple* (p,q) *tel que* $p > q$*.*

$$\mathbb{L} = \begin{bmatrix} [\mathbb{L}]_{1,1} & & & & \\ [\mathbb{L}]_{2,1} & [\mathbb{L}]_{2,2} & & & \\ \ddots & \ddots & \ddots & & \\ \ddots & \ddots & \ddots & \ddots & \\ [\mathbb{L}]_{P,1} & \ddots & \ddots & [\mathbb{L}]_{P,P-1} & [\mathbb{L}]_{P,P} \end{bmatrix} \quad et \quad \mathbb{U} = \begin{bmatrix} [\mathbb{U}]_{1,1} & [\mathbb{U}]_{1,2} & \ddots & \ddots & [\mathbb{U}]_{1,P} \\ & [\mathbb{U}]_{2,2} & \ddots & \ddots & \ddots \\ & & \ddots & \ddots & \ddots \\ & & & \ddots & [\mathbb{U}]_{P-1,P} \\ & & & & [\mathbb{U}]_{P,P} \end{bmatrix}.$$

Enfin on appelle matrice **diagonale par blocs** *une matrice* $\mathbb{D} \in \mathbb{K}^{n\times n}$ *telle que* $[\mathbb{D}]_{p,q} = 0$ *pour tout couple* (p,q) *tel que* $p \neq q$*.*

$$\mathbb{D} = \begin{bmatrix} [\mathbb{D}]_{1,1} & & & & \\ & [\mathbb{D}]_{2,2} & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & [\mathbb{D}]_{P,P} \end{bmatrix}.$$

La plupart des résultats démontrés pour les matrices se généralisent au cas des matrices définies par blocs, en prenant la précaution dans l'écriture des formules de noter que les termes qui interviennent sont des matrices ; ceci nécessite en particulier une adaptation spécifique aux règles de calcul algébrique. D'une part, si le produit de deux blocs est bien un bloc, cette opération n'est pas commutative. D'autre part, on doit remplacer la division par les éléments diagonaux par la multiplication à gauche par l'inverse du bloc diagonal. Voir le paragraphe suivant pour des illustrations.

5.3.2 Parallélisation du produit matrice-vecteur par blocs

Pour paralléliser le produit matrice-vecteur $y = \mathbb{A}x$ lorsque la matrice $\mathbb{A}$ et les vecteurs x, y sont découpés en blocs, on procède comme suit. On suppose qu'on dispose d'une machine avec $\mathtt{P}$ nœuds de calcul (voir le §4.5). On affecte au nœud $\mathtt{p}$ les blocs $([\mathbb{A}]_{\mathtt{p},\mathtt{q}})_{\mathtt{q}=1,\mathtt{P}}$ ainsi que le bloc $[x]_{\mathtt{p}}$. Sur ce nœud, on doit effectuer le produit matrice-vecteur

$$[y]_{\mathtt{p}} = \sum_{\mathtt{q}=1,\mathtt{P}} [\mathbb{A}]_{\mathtt{p},\mathtt{q}}[x]_{\mathtt{q}} = [\mathbb{A}]_{\mathtt{p},\mathtt{p}}[x]_{\mathtt{p}} + \sum_{\mathtt{q}\neq\mathtt{p}} [\mathbb{A}]_{\mathtt{p},\mathtt{q}}[x]_{\mathtt{q}}. \tag{5.1}$$

Pour pouvoir réaliser ce produit localement sur le nœud $\mathtt{p}$, il faut disposer de $([x]_{\mathtt{q}})_{\mathtt{q}\neq\mathtt{p}}$. Si c'est le cas pour $\mathtt{p} = 1, \mathtt{P}$, alors on peut effectuer les produits matrice-vecteur simultanément. Ainsi :

- Il y a d'abord une première phase de communications où chaque nœud $\mathtt{q}$ envoie aux autres nœuds les données $[x]_{\mathtt{q}}$. On peut réaliser certaines communications en parallèle, pour arriver au total à $\mathtt{P} - 1$ communications parallèles en tout : $\mathtt{q} \to \mathtt{q} + 1 \bmod [\mathtt{P}]$; $\mathtt{q} \to \mathtt{q} + 2 \bmod [\mathtt{P}]$; $\cdots$; $\mathtt{q} \to \mathtt{q} + \mathtt{P} - 1 \bmod [\mathtt{P}]$.
- A partir de là, le calcul (5.1) est réalisé en parallèle.
- Enfin, il y a une deuxième phase de communications où chaque nœud $\mathtt{q}$ envoie aux autres nœuds les résultats $[y]_{\mathtt{q}}$ ($\mathtt{P} - 1$ communications parallèles comme précédemment).

Prenons l'exemple d'une matrice pleine. Afin d'optimiser l'algorithme, il est pertinent de construire des blocs qui possèdent tous la même taille : d'une part, le nombre de données à transmettre est constant (phases de communications) ; d'autre part, le nombre d'opérations pour réaliser le calcul (5.1) est également constant.

5.4 Résultats sur les matrices triangulaires

On rappelle les résultats suivants :

Proposition 5.10 *Le produit de deux matrices triangulaires supérieures (resp. inférieures) $\mathbb{T}', \mathbb{T}'' \in \mathbb{K}^{n\times n}$ est une matrice triangulaire supérieure (resp. inférieure) de $\mathbb{K}^{n\times n}$.*

Par contre, le produit d'une matrice triangulaire inférieure par une matrice triangulaire supérieure est une *matrice à structure quelconque*. Il en est de même du produit d'une matrice triangulaire supérieure par une matrice triangulaire inférieure.

Proposition 5.11 *La matrice inverse d'une matrice triangulaire inférieure (resp. supérieure) inversible de $\mathbb{K}^{n\times n}$ est une matrice triangulaire inférieure (resp. supérieure).*

Proposition 5.12 *Le déterminant d'une matrice triangulaire par blocs de $\mathbb{K}^{n\times n}$ est égal au produit des déterminants des blocs diagonaux.*

En particulier, une matrice triangulaire par blocs est inversible si, et seulement si, tous ses blocs diagonaux sont inversibles.

Proposition 5.13 *Soit $\mathbb{L} \in \mathbb{K}^{n\times n}$ une matrice triangulaire inférieure par blocs inversible. La solution du système linéaire $\mathbb{L}y = b$ est obtenue par les formules :*

pour $p = 1, \ldots, P$ **faire**
$\quad [y]_p = [\mathbb{L}]_{p,p}^{-1}\Big([b]_p - \sum_{q<p} [\mathbb{L}]_{p,q}[y]_q\Big).$
fin

Soit $\mathbb{U} \in \mathbb{K}^{n\times n}$ une matrice triangulaire supérieure par blocs inversible. La solution du système linéaire $\mathbb{U}x = y$ est obtenue par les formules :

pour $p = P, \ldots, 1$ **faire**
$\quad [x]_p = [\mathbb{U}]_{p,p}^{-1}\Big([y]_p - \sum_{q>p} [\mathbb{U}]_{p,q}[x]_q\Big).$
fin

5.5 La méthode d'élimination

On considère le système linéaire $\mathbb{A}x = b$ dans $\mathbb{K}^n$ (dont la matrice $\mathbb{A}$ est inversible)

$$\begin{pmatrix} \mathbb{A}_{1,1} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \mathbb{A}_{1,n} \\ \mathbb{A}_{2,1} & \mathbb{A}_{2,2} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \mathbb{A}_{i,1} & \cdot & \cdot & \mathbb{A}_{i,i} & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \mathbb{A}_{n,1} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \mathbb{A}_{n,n} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_i \\ \cdot \\ \cdot \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ b_i \\ \cdot \\ \cdot \\ b_n \end{pmatrix}$$

dans lequel on *suppose* $\mathbb{A}_{1,1} \neq 0$; à l'aide de la première ligne de ce système, on exprime la composante x_1 en fonction des autres :

$$x_1 = [b_1 - \sum_{2\leq j\leq n} \mathbb{A}_{1,j}x_j]/\mathbb{A}_{1,1}$$

en reportant cette identité dans la ligne i du système linéaire ($2 \leq i \leq n$), on obtient

$$\mathbb{A}_{i,1}x_1 + \sum_{j=2}^{n} \mathbb{A}_{i,j}x_j = b_i, \iff \sum_{j=2}^{n} \left(\mathbb{A}_{i,j} - \mathbb{A}_{i,1}\mathbb{A}_{1,j}/\mathbb{A}_{1,1}\right) x_j = b_i - \mathbb{A}_{i,1}/\mathbb{A}_{1,1}b_1.$$

On introduit alors les notations $\mathbb{A}^{(0)} = \mathbb{A}$ et $b^{(0)} = b$, puis on définit la matrice $\mathbb{A}^{(1)}$ et le vecteur $b^{(1)}$ suivant :

- pour la première ligne :
  $\mathbb{A}_{1,j}^{(1)} = \mathbb{A}_{1,j}^{(0)}$, $1 \leq j \leq n$, et $b_1^{(1)} = b_1^{(0)}$;
- pour les $n-1$ autres lignes :
  $\mathbb{A}_{i,j}^{(1)} = \mathbb{A}_{i,j}^{(0)} - \mathbb{A}_{i,1}^{(0)} \times \mathbb{A}_{1,j}^{(0)}/\mathbb{A}_{1,1}^{(0)}$, et $b_i^{(1)} = b_i^{(0)} - \mathbb{A}_{i,1}^{(0)} \times b_1^{(0)}/\mathbb{A}_{1,1}^{(0)}$, $2 \leq i \leq n$, $1 \leq j \leq n$.

On obtient un système linéaire *équivalent* au précédent, au sens où les deux systèmes admettent la même solution. Noter que la première colonne de la matrice $\mathbb{A}^{(1)}$ est **nulle** par construction à l'exception du élément $\mathbb{A}^{(1)}_{1,1}$.

Si $\mathbb{A}^{(1)}_{2,2} \neq 0$, on peut réitérer le procédé en éliminant cette fois l'inconnue x_2 des $n-2$ lignes $i = 3, 4, \ldots, n$, et ainsi de suite... On génère de cette façon une suite de matrices et de seconds membres par l'algorithme :

1) **initialisation :**
$\mathbb{A}^{(0)} = \mathbb{A} \in \mathbb{K}^{n\times n}$.
$b^{(0)} = b \in \mathbb{K}^n$.
2) **itérations : pour** $k = 1, 2, \ldots, n-1$ **faire**
(1) élimination de l'inconnue x_k et mise à jour de la matrice
$\mathbb{A}^{(k)}_{i,j} = \mathbb{A}^{(k-1)}_{i,j} \qquad 1 \leq i \leq k, \quad 1 \leq j \leq n$
$\mathbb{A}^{(k)}_{i,j} = \mathbb{A}^{(k-1)}_{i,j} - \mathbb{A}^{(k-1)}_{i,k} \times \mathbb{A}^{(k-1)}_{k,j} / \mathbb{A}^{(k-1)}_{k,k} \qquad k+1 \leq i \leq n, \quad 1 \leq j \leq n$
(2) modification du second membre
$b^{(k)}_i = b^{(k-1)}_i \qquad 1 \leq i \leq k$
$b^{(k)}_i = b^{(k-1)}_i - \mathbb{A}^{(k-1)}_{i,k} \times b^{(k-1)}_k / \mathbb{A}^{(k-1)}_{k,k} \qquad k+1 \leq i \leq n$
fin

Noter que les éléments $\mathbb{A}^{(k)}_{i,j}$ pour $k+1 \leq i \leq n$ et $1 \leq j \leq k$ définis par ces formules, sont nuls par construction. En effet, pour de tels couples (i, j), si $j < k$: on sait que d'une part $\mathbb{A}^{(k-1)}_{i,j} = 0$, et d'autre part que $\mathbb{A}^{(k-1)}_{k,j} = 0$ d'après l'itération précédente ; et, si $j = k$: alors la formule donne $\mathbb{A}^{(k)}_{i,k} = 0$. En d'autres termes, on peut écrire, pour $k = 1, \cdots, n-1$:

$$\mathbb{A}^{(k)} = \begin{pmatrix} [\mathbb{U}]_{1,1} & [\mathbb{U}]_{1,2} \\ 0 & \mathcal{S}^{(k)} \end{pmatrix},$$

avec $[\mathbb{U}]_{1,1} \in \mathbb{K}^{k\times k}$ une matrice *triangulaire supérieure*, $[\mathbb{U}]_{1,2} \in \mathbb{K}^{k\times(n-k)}$, et $\mathcal{S}^{(k)} \in \mathbb{K}^{(n-k)\times(n-k)}$. On appelle $\mathcal{S}^{(k)}$ le **complément de Schur**.

Après $n-1$ itérations de cet algorithme (en supposant que les différents éléments $\mathbb{A}^{(k)}_{k,k}$ sont non nuls pour chaque k) la matrice $\mathbb{A}^{(n-1)}$ obtenue est une matrice triangulaire supérieure et le système linéaire

$$\mathbb{A}^{(n-1)} x = b^{(n-1)}$$

peut être résolu à l'aide de la Proposition 5.7. De plus, il a la même solution que le système initial $\mathbb{A}x = b$, puisque tous les systèmes linéaires $\mathbb{A}^{(k)}x = b^{(k)}$ sont équivalents entre eux, pour $k = 1, \cdots, n-1$.

On introduit ensuite la matrice *triangulaire inférieure* $\mathbb{L}^{(k)}$ de rang n, identique à la

matrice I_n, à l'exception de la colonne k :

$$\mathbb{L}^{(k)} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathrm{x} & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathrm{x} & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & \mathrm{x} & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \mathrm{x} & 0 & 0 & 0 & 1 \end{pmatrix},$$

avec $\mathbb{L}_{i,k}^{(k)} = -\mathbb{A}_{i,k}^{(k-1)}/\mathbb{A}_{k,k}^{(k-1)}$ pour $i \geq k+1$. Par construction, on a la relation :

$$\det(\mathbb{L}^{(k)}) = 1.$$

De plus, on vérifie facilement que pour tout $k < n$, $\mathbb{A}^{(k)} = \mathbb{L}^{(k)}\mathbb{A}^{(k-1)}$ et $b^{(k)} = \mathbb{L}^{(k)}b^{(k-1)}$.

Finalement, en posant $\mathbb{U} = \mathbb{A}^{(n-1)}$ et $\tilde{\mathbb{L}} = \mathbb{L}^{(n-1)} \cdots \mathbb{L}^{(1)}$, on peut écrire

$$\mathbb{U} = \tilde{\mathbb{L}}\mathbb{A} \quad \text{et} \quad \mathbb{U}x = \tilde{\mathbb{L}}b, \tag{5.2}$$

où $\mathbb{U}$ et $\tilde{\mathbb{L}}$ sont des matrices inversibles, resp. triangulaire supérieure pour $\mathbb{U}$ et triangulaire inférieure pour $\tilde{\mathbb{L}}$. Le calcul de la solution x par les formules de **remontée** est alors immédiat.
En outre, on peut montrer sans difficulté le résultat ci-dessous.

Proposition 5.14 *Pour tout k l'inverse $\mathbb{L}^{-(k)}$ de la matrice $\mathbb{L}^{(k)}$ est une matrice* triangulaire inférieure *de rang n, identique à I_n, à l'exception de la colonne k, avec, pour $i > k$, $\mathbb{L}_{i,k}^{-(k)} = \mathbb{A}_{i,k}^{(k-1)}/\mathbb{A}_{k,k}^{(k-1)}$. En d'autres termes, on a la relation*

$$\mathbb{L}^{-(k)} = 2I_n - \mathbb{L}^{(k)}.$$

La seule question qui se pose alors est de savoir si on peut toujours calculer cette matrice $\mathbb{A}^{(n-1)}$ par les formules précédentes :

il faut pour cela que $\mathbb{A}_{k,k}^{(k-1)} \neq 0$ pour $k = 1, 2, \ldots, n-1$.

Si en cours de calcul, on rencontre un élément diagonal $\mathbb{A}_{k,k}^{(k-1)}$ nul, on peut procéder de la façon suivante : on recherche dans la colonne k de la matrice $\mathbb{A}^{(k-1)}$ un élément $\mathbb{A}_{i_k,k}^{(k-1)}$ non nul pour $i_k > k$, et *s'il en existe un*, on échange alors les lignes i_k et k de la matrice. Cette modification revient à multiplier à gauche la matrice courante $\mathbb{A}^{(k-1)}$ par

une matrice de permutation $\mathbb{P}(i_k,k)$ qui amène le élément $\mathbb{A}_{i_k,k}^{(k-1)}$ sur la diagonale [32]

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mathbb{A}_{1,1}^{(k-1)} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & \mathbb{A}_{2,2}^{(k-1)} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & \mathbb{A}_{3,3}^{(k-1)} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & \mathbf{0} & \mathrm{x} & \mathrm{x} & \mathbb{A}_{k,j}^{(k-1)} & \mathrm{x} \\ 0 & 0 & 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & \mathbb{A}_{i_k,k}^{(k-1)} & \mathrm{x} & \mathrm{x} & \mathbb{A}_{i_k,j}^{(k-1)} & \mathrm{x} \\ 0 & 0 & 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \end{pmatrix}$$

soit

$$\mathbb{P}(i_k,k)\mathbb{A}^{(k-1)} = \begin{pmatrix} \mathbb{A}_{1,1}^{(k-1)} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & \mathbb{A}_{2,2}^{(k-1)} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & \mathbb{A}_{3,3}^{(k-1)} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & \mathbb{A}_{i_k,k}^{(k-1)} & \mathrm{x} & \mathrm{x} & \mathbb{A}_{i_k,j}^{(k-1)} & \mathrm{x} \\ 0 & 0 & 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & \mathbf{0} & \mathrm{x} & \mathrm{x} & \mathbb{A}_{k,j}^{(k-1)} & \mathrm{x} \\ 0 & 0 & 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \end{pmatrix}$$

Dans la factorisation en cours, cette multiplication n'affecte pas les lignes d'indice strictement inférieur à k déjà calculées, mais seulement les lignes i et k de la matrice $\mathbb{A}^{(k)}$; noter que les composantes $b_{i_k}^{(k-1)}$ et $b_k^{(k-1)}$ doivent aussi être échangées pour que le nouveau système linéaire soit équivalent au précédent.

Que se passe-t-il si à l'étape k (k fixé) *tous les éléments* $\mathbb{A}_{i,k}^{(k-1)}$ de la colonne k sont nuls ? Cela veut dire que l'on a obtenu une matrice de la forme

$$\begin{pmatrix} \mathbb{A}_{1,1}^{(k-1)} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & \mathbb{A}_{2,2}^{(k-1)} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & \mathbb{A}_{3,3}^{(k-1)} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & \mathbf{0} & \mathrm{x} & \mathrm{x} & \mathbb{A}_{k,j}^{(k-1)} & \mathrm{x} \\ 0 & 0 & 0 & \mathbf{0} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & \mathbf{0} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & \mathbf{0} & \mathrm{x} & \mathrm{x} & \mathbb{A}_{i,j}^{(k-1)} & \mathrm{x} \\ 0 & 0 & 0 & \mathbf{0} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \end{pmatrix}$$

32. Notons que pour toute matrice de permutation $\mathbb{P}(i_k,k)$, on a la relation $\mathbb{P}(i_k,k)\mathbb{P}(i_k,k) = I_n$. De plus, pour $k \geq 2$, on peut toujours écrire $\mathbb{P}(i_k,k)$ sous la forme par blocs

$$\mathbb{P}(i_k,k) = \begin{pmatrix} I_{k-1} & 0 \\ 0 & \mathbb{P} \end{pmatrix},$$

puisque $i_k > k$.

et la matrice $\mathbb{A}^{(k-1)}$ est donc au plus de rang $n-1$, ce qui est contraire à l'hypothèse $\mathbb{A}$ inversible car la relation

$$\mathbb{A}^{(k-1)} = \mathbb{L}^{(k-1)} \cdots \mathbb{L}^{(1)} \mathbb{A}$$

entraîne bien sûr

$$\det(\mathbb{A}^{(k-1)}) = \det(\mathbb{A}) \neq 0.$$

On en conclut que l'on peut *toujours* réaliser une permutation de lignes pour la matrice $\mathbb{A}^{(k-1)}$ de manière à pouvoir poursuivre l'algorithme sur la matrice permutée $\mathbb{P}(i_k, k)\mathbb{A}^{(k-1)}$.

5.6 La méthode de factorisation

Dans le paragraphe précédent on a obtenu pour une matrice $\mathbb{A}$ donnée, les matrices triangulaires inversibles $\mathbb{U}$ et $\tilde{\mathbb{L}}$. En l'absence de permutations de lignes, on aboutit à (5.2). A l'aide de la Proposition 5.14, on définit

$$\mathbb{L} = \tilde{\mathbb{L}}^{-1} = \mathbb{L}^{-(1)} \cdots \mathbb{L}^{-(n-1)}.$$

La matrice $\mathbb{L}$ est triangulaire inférieure à diagonale unité, qui vérifie la relation

$$\mathbb{A} = \mathbb{L}\,\mathbb{U}.$$

Cette relation est appelée **factorisation de Gauss** de la matrice $\mathbb{A}$. A partir de cette factorisation la solution du système linéaire $\mathbb{A}x = b$ est obtenue en deux étapes :

- la **descente**, qui consiste à calculer le vecteur y solution de $\mathbb{L}y = b$;
- la **remontée**, dans laquelle on calcule le vecteur x solution de $\mathbb{U}x = y$.

Néanmoins, au cours des calculs, on peut avoir à effectuer une certain nombre de permutations de lignes pour amener des éléments non nuls sur la diagonale. Si on doit effectuer une permutation à l'itération k, on factorise $\mathbb{P}(i_k, k)\mathbb{A}^{(k-1)}$ sous la forme

$$\mathbb{A}^{(k)} = \mathbb{L}^{(k)}\mathbb{P}(i_k, k)\mathbb{A}^{(k-1)}.$$

Au terme des itérations, on a alors factorisé la matrice $\mathbb{A}$ de la façon suivante [33]

$$\mathbb{U} = \mathbb{L}^{(n-1)}\mathbb{P}(i_{n-1}, n-1) \cdots \mathbb{L}^{(k)}\mathbb{P}(i_k, k) \cdots \mathbb{L}^{(1)}\mathbb{P}(i_1, 1)\mathbb{A}.$$

On a donc un produit "mêlé" de matrices du type $\mathbb{L}^{(k)}$ et de matrices de permutation $\mathbb{P}(i_p, p)$. Fort heureusement, on peut prouver le résultat suivant

Proposition 5.15 *Soient k et p tels que $k < p$. Alors il existe une matrice $\mathbb{L}^{(k,p)}$ possédant la même structure que $\mathbb{L}^{(k)}$ et telle que*

$$\mathbb{P}(i_p, p)\mathbb{L}^{(k)} = \mathbb{L}^{(k,p)}\mathbb{P}(i_p, p). \tag{5.3}$$

Démonstration : Il suffit de considérer la matrice $\mathbb{L}^{(k,p)}$ définie par

$$\mathbb{L}^{(k,p)}_{i_p,k} = \mathbb{L}^{(k,p)}_{p,k},\ \mathbb{L}^{(k,p)}_{p,k} = \mathbb{L}^{(k,p)}_{i_p,k},\ \mathbb{L}^{(k,p)}_{i,j} = \mathbb{L}^{(k,p)}_{i,j} \text{ sinon.}$$

◇

33. Avec la convention $\mathbb{P}(i_k, k) = I_n$ lorsqu'on n'effectue pas de permutation à l'étape k...

A partir de ce résultat, on écrit tout simplement

$$\begin{aligned}
\mathbb{U} &= \mathbb{L}^{(n-1)}\mathbb{P}(i_{n-1},n-1)\mathbb{L}^{(n-2)}\mathbb{P}(i_{n-2},n-2)\cdots\mathbb{A}\\
&= \mathbb{L}^{(n-1)}\bar{\mathbb{L}}^{(n-2)}\mathbb{P}(i_{n-1},n-1)\mathbb{P}(i_{n-2},n-2)\mathbb{L}^{(n-3)}\mathbb{P}(i_{n-3},n-3)\cdots\mathbb{A}\\
&= \mathbb{L}^{(n-1)}\bar{\mathbb{L}}^{(n-2)}\bar{\mathbb{L}}^{(n-3)}\mathbb{P}(i_{n-1},n-1)\mathbb{P}(i_{n-2},n-2)\mathbb{P}(i_{n-3},n-3)\cdots\mathbb{A}\\
&\vdots\\
&= \mathbb{L}^{(n-1)}\bar{\mathbb{L}}^{(n-2)}\cdots\bar{\mathbb{L}}^{(1)}\mathbb{P}(i_{n-1},n-1)\cdots\mathbb{P}(i_1,1)\mathbb{A}.
\end{aligned}$$

Ci-dessus, on a bien sûr, pour $1 \leq k \leq n-2$,

$$\bar{\mathbb{L}}^{(k)} = \prod_{\ell=1}^{n-(1+k)} \mathbb{P}(i_{n-\ell},n-\ell)\,\mathbb{L}^{(k)} \prod_{\ell=k+1}^{n-1} \mathbb{P}(i_\ell,\ell),$$

ce qui définit des matrices triangulaires inférieures possédant la même structure que $\mathbb{L}^{(k)}$, cf. la Proposition 5.15.

On aboutit pour finir à $\mathbb{P}\,\mathbb{A} = \mathbb{L}\,\mathbb{U}$, où $\mathbb{L}$ est triangulaire inférieure, $\mathbb{U}$ triangulaire supérieure, et $\mathbb{P}$ est un produit de matrices de permutation, $\mathbb{P} = \mathbb{P}(i_{n-1},n-1)\ \cdots\ \mathbb{P}(i_1,1)$. Notons que $\mathbb{P}^{-1} = \mathbb{P}(i_1,1)\ \cdots\ \mathbb{P}(i_{n-1},n-1) = \mathbb{P}^T$. On a ainsi démontré le résultat suivant

Théorème 5.16 *Soit une matrice $\mathbb{A}$ inversible de $\mathbb{K}^{n\times n}$, alors il existe une matrice de permutation $\mathbb{P}$ telle que*

$$\mathbb{P}\ \mathbb{A} = \mathbb{L}\ \mathbb{U},$$

avec $\mathbb{L}$ matrice triangulaire inférieure à diagonale unité, $\mathbb{U}$ matrice triangulaire supérieure.

5.7 Stabilité numérique et stratégies de pivotage

On voit bien que cette écriture n'est pas unique puisqu'à chaque échange, on peut avoir le choix entre plusieurs lignes pour effectuer la permutation. On peut alors introduire un critère supplémentaire pour déterminer la ligne à permuter, par exemple un critère de stabilité numérique [34]. Supposons que l'élément courant $\mathbb{A}_{k,k}^{(k-1)}$ soit petit, de l'ordre de ε, alors les formules de calcul

$$\mathbb{A}_{i,j}^{(k)} = \mathbb{A}_{i,j}^{(k-1)} - \frac{1}{\varepsilon}\ \mathbb{A}_{i,k}^{(k-1)}\ \mathbb{A}_{k,j}^{(k-1)}$$

montrent que dans le complément de Schur $\mathcal{S}^{(k)}$ résultant, le second terme est dominant, c'est-à-dire que l'on a

$$\mathcal{S}^{(k)} \approx -\frac{1}{\varepsilon} u\cdot v^T$$

les vecteurs $u, v \in \mathbb{K}^{n-k}$ représentant respectivement la colonne k et la ligne k de la matrice $\mathbb{A}^{(k-1)}$, pour les indices strictement supérieurs à k. Comme $u \neq 0$ et $v \neq 0$, la matrice $u\cdot v^T \in \mathbb{K}^{(n-k)\times(n-k)}$ est de rang 1. Dans ce cas, la matrice $\mathcal{S}^{(k)} \in \mathbb{K}^{(n-k)\times(n-k)}$

34. En se sens, on se rapproche des considérations algorithmiques développées au chapitre 4.

est *quasi-singulière* dès que $n-1 > k$! Autrement dit, le choix d'un petit élément diagonal peut conduire à une *instabilité numérique* de la factorisation.

Le choix de cet élément, appelé **pivot** doit donc être effectué avec la plus grande attention, et cela introduit naturellement deux variantes de la factorisation de Gauss :

- la factorisation avec **pivot partiel** revient à rechercher à chaque étape de l'algorithme le plus grand élément en valeur absolue parmi les $\mathbb{A}_{i,k}^{(k-1)}$ pour $i \geq k$. La permutation de lignes associée à ce choix correspond à une multiplication à gauche de la matrice $\mathbb{A}$ par une matrice de permutation $\mathbb{P}$.
- la factorisation avec **pivot total** revient à rechercher à chaque étape de l'algorithme le plus grand élément en valeur absolue parmi tous les $\mathbb{A}_{i,j}^{(k-1)}$ pour $i \geq k$ et $j \geq k$. Si on choisit un élément $\mathbb{A}_{i_k,j_k}^{(k-1)}$ en dehors de la colonne k, il faut ajouter à la permutation de lignes $\mathbb{P}(i_k, k)$ une permutation de colonnes qui correspond à une multiplication à droite de la matrice $\mathbb{A}$ par une matrice de permutation $\mathbb{Q}(j_k, k)$ (avec $j_k > k$). En fin de factorisation, on a obtenu

$$\mathbb{P}\,\mathbb{A}\,\mathbb{Q} = \mathbb{L}\,\mathbb{U}.$$

- Si enfin on effectue une recherche avec **pivot total** sur les *éléments diagonaux* $\mathbb{A}_{i,i}^{(k-1)}$ uniquement (pour $i \geq k$), on a alors une permutation de ligne et une permutation de colonne identiques, i. e. $\mathbb{Q}(i_k, k) = \mathbb{P}(i_k, k)$. En fin de factorisation, on arrive à $\mathbb{Q} = \mathbb{P}^T = \mathbb{P}^*$, soit

$$\mathbb{P}\,\mathbb{A}\,\mathbb{P}^* = \mathbb{P}\,\mathbb{A}\,\mathbb{P}^T = \mathbb{L}\,\mathbb{U}.$$

  Cette propriété sera utilisée pour factoriser des matrices hermitiennes ou symétriques.

Remarque 5.17 *En reprenant les notations précédentes, on voit qu'une permutation de lignes $\mathbb{P}(i_k, k)$ ou de colonnes $\mathbb{Q}(j_k, k)$ de la matrice $\mathcal{S}^{(k)}$ à l'étape k, ne modifie pas les blocs déjà calculés $[\mathbb{L}]_{1,1}$ et $[\mathbb{U}]_{1,1}$:*

$$\mathbb{P}(i_k,k)\,\mathbb{A}\,\mathbb{Q}(j_k,k) = \begin{pmatrix} I_{k-1} & 0 \\ 0 & \mathbb{P} \end{pmatrix} \begin{pmatrix} [\mathbb{L}]_{1,1} & 0 \\ [\mathbb{L}]_{2,1} & I_{n-k+1} \end{pmatrix} \begin{pmatrix} [\mathbb{U}]_{1,1} & [\mathbb{U}]_{1,2} \\ 0 & \mathcal{S}^{(k)} \end{pmatrix} \begin{pmatrix} I_{k-1} & 0 \\ 0 & \mathbb{Q} \end{pmatrix}$$

soit encore

$$\mathbb{P}(i_k,k)\,\mathbb{A}\,\mathbb{Q}(j_k,k) = \begin{pmatrix} [\mathbb{L}]_{1,1} & 0 \\ \mathbb{P}[\mathbb{L}]_{2,1} & \mathbb{P} \end{pmatrix} \begin{pmatrix} [\mathbb{U}]_{1,1} & [\mathbb{U}]_{1,2}\mathbb{Q} \\ 0 & \mathcal{S}^{(k)}\mathbb{Q} \end{pmatrix}$$

On énonce pour finir un résultat d'unicité.

Proposition 5.18 *Soit $\mathbb{A} \in \mathbb{K}^{n\times n}$ une matrice inversible, pour des matrices de permutation $\mathbb{P}$ et $\mathbb{Q}$ données, la factorisation $\mathbb{P}\,\mathbb{A}\,\mathbb{Q} = \mathbb{L}\,\mathbb{U}$ est unique.*

Démonstration : Cela est évident par construction des matrices $\mathbb{L}$ et $\mathbb{U}$, leurs éléments étant déterminés de manière unique par les formules de l'algorithme. Mais on peut aussi démontrer ce résultat par une méthode qui sera utile par la suite. Supposons qu'il existe des matrices triangulaires inférieures $\mathbb{L}$ et $\mathbb{L}'$, et triangulaires supérieures $\mathbb{U}$ et $\mathbb{U}'$ qui vérifient $\mathbb{P}\,\mathbb{A}\,\mathbb{Q} = \mathbb{L}\,\mathbb{U} = \mathbb{L}'\,\mathbb{U}'$. Par voie de conséquence,

$$(\mathbb{L}')^{-1}\mathbb{L} = \mathbb{U}'\mathbb{U}^{-1} (= \mathbb{M}).$$

Or, $(\mathbb{L}')^{-1}\mathbb{L}$ (resp. $\mathbb{U}'\mathbb{U}^{-1}$) est une matrice triangulaire inférieure (resp. triangulaire supérieure). Ainsi $\mathbb{M}$ est nécessairement une matrice diagonale. De plus, $\mathbb{L}$ et $\mathbb{L}'$ étant à diagonale unité, il en est de même pour $(\mathbb{L}')^{-1}\mathbb{L}$. Finalement, $\mathbb{M} = I_n$. $\diamond$

5.8 Les méthodes directes

Dans la suite on appellera **méthode directe** de résolution d'un système linéaire $\mathbb{A}x = b$ dans $\mathbb{K}^n$ tout algorithme qui calcule la solution x en un nombre d'opérations déterminé *a priori*[35].

Les méthodes d'élimination et de factorisation sont à ce titre des méthodes directes, car il est évident que le nombre d'opérations nécessaires au calcul des matrices $\mathbb{L}$ et $\mathbb{U}$ est fini – ce nombre d'opérations sera calculé précisément – Par ailleurs le coût d'une descente et d'une remontée est égal à $2n^2$ opérations (voir le §5.15).

Noter que ce nombre d'opérations dépend des propriétés de la matrice $\mathbb{A}$, car on peut tirer parti d'une éventuelle symétrie par exemple. On distingue ainsi plusieurs types de factorisation :

- La méthode de Gauss $\mathbb{A} = \mathbb{L}\mathbb{U}$, avec $\mathbb{L}$ matrice triangulaire inférieure à diagonale unité, et $\mathbb{U}$ matrice triangulaire supérieure. $\mathbb{A}$ doit être inversible.
- La méthode de Crout $\mathbb{A} = \mathbb{L}\mathbb{D}\mathbb{L}^*$, avec $\mathbb{L}$ matrice triangulaire inférieure à diagonale unité, et $\mathbb{D}$ matrice diagonale. $\mathbb{A}$ doit être hermitienne ($\mathbb{K} = \mathbb{C}$), resp. symétrique ($\mathbb{K} = \mathbb{R}$), et inversible.
- La méthode de Cholesky $\mathbb{A} = \mathbb{L}\mathbb{L}^*$, avec $\mathbb{L}$ matrice triangulaire inférieure. $\mathbb{A}$ doit être hermitienne ($\mathbb{K} = \mathbb{C}$), resp. symétrique ($\mathbb{K} = \mathbb{R}$), et définie-positive.

Dans ce qui suit, on reprend l'étude de la factorisation de la matrice $\mathbb{A}$ dans une formulation plus générale, en supposant uniquement que cette matrice est inversible.

5.9 Algorithme de factorisation de Gauss

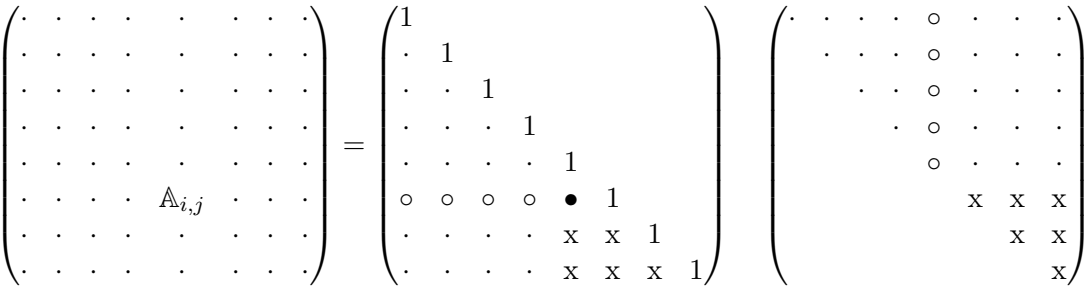
Maintenant que l'existence des matrices $\mathbb{L}$ et $\mathbb{U}$ est établie, on vérifie que l'on peut calculer leurs éléments directement par identification, à partir des relations

$$\forall\, i,j \quad 1 \le i,j \le n \qquad \mathbb{A}_{i,j} = \sum_{\ell=1}^{n} \mathbb{L}_{i,\ell}\, \mathbb{U}_{\ell,j} = \sum_{\ell=1}^{\ell=\min(i,j)} \mathbb{L}_{i,\ell}\, \mathbb{U}_{\ell,j}$$

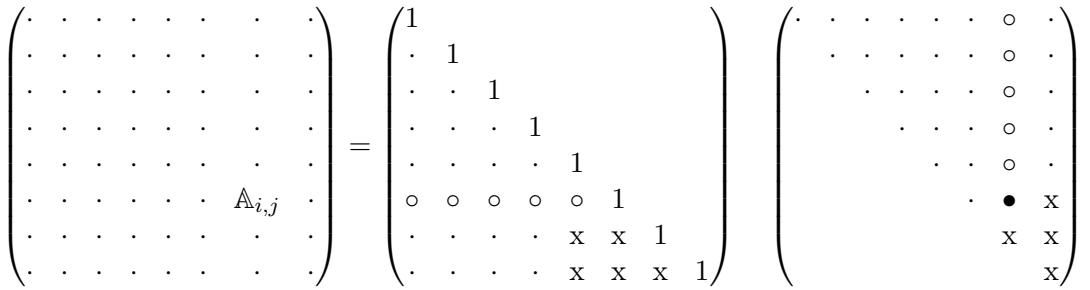
puisque $\mathbb{L}_{i,\ell} = 0$ si $\ell > i$, resp. $\mathbb{U}_{\ell,j} = 0$ si $\ell > j$.
On procède en calculant pour un indice k donné, tous les éléments $\mathbb{L}_{.,k}$ de la colonne k de la matrice $\mathbb{L}$, puis tous les éléments $\mathbb{U}_{k,.}$ de la ligne k de la matrice $\mathbb{U}$. Ce processus peut être représenté par les schémas suivants, dans lesquels les éléments $\cdot$ sont supposés connus, les éléments x sont inconnus et l'élément $\bullet$ est en cours de calcul à l'aide des éléments

35. Il s'agit ici de faire la distinction avec les méthodes itératives, étudiées au chapitre 6, pour lesquelles le nombre d'opérations dépend du nombre d'itérations de la méthode, nombre qu'il est impossible de connaître à l'avance car il est lié au choix de la solution initiale $x^0 \in \mathbb{K}^n$ relativement au second membre b.

connus $\circ$.



Calcul d'un élément de $\mathbb{L}$.



Calcul d'un élément de $\mathbb{U}$.

En résumé, pour une matrice $\mathbb{A}$ donnée, les éléments des matrices $\mathbb{L}$ et $\mathbb{U}$ sont calculés (à une permutation de lignes et de colonnes près) par les formules

pour $k = 1, \ldots, n-1$ **faire**
  $\mathbb{L}_{k,k} = 1$
  $\mathbb{U}_{k,k} = \mathbb{A}_{k,k} - \sum_{j<k} \mathbb{L}_{k,j}\mathbb{U}_{j,k}$
  pour $i = k+1, \ldots, n$ **faire**
    $\mathbb{L}_{i,k} = [\mathbb{A}_{i,k} - \sum_{j<k} \mathbb{L}_{i,j}\mathbb{U}_{j,k}] \;/\; \mathbb{U}_{k,k}$
  fin
  pour $i = k+1, \ldots, n$ **faire**
    $\mathbb{U}_{k,i} = \mathbb{A}_{k,i} - \sum_{j<k} \mathbb{L}_{k,j}\mathbb{U}_{j,i}$
  fin
fin

Les formules précédentes calculent les éléments de la matrice $\mathbb{L}$ colonne par colonne, et ceux de la matrice $\mathbb{U}$ ligne par ligne. L'algorithme suivant définit les mêmes matrices, bien que les éléments de la matrice $\mathbb{L}'$ soient calculés ligne par ligne et ceux de la matrice $\mathbb{U}'$

colonne par colonne.

pour $k = 1, \ldots, n-1$ **faire**
  pour $i = 1, \ldots, k-1$ **faire**
    $\mathbb{L}'_{k,i} = [\mathbb{A}_{k,i} - \sum_{j<i} \mathbb{L}'_{k,j}\mathbb{U}'_{j,i}]$ / $\mathbb{U}'_{i,i}$
  fin
  pour $i = 1, \ldots, k-1$ **faire**
    $\mathbb{U}'_{i,k} = \mathbb{A}_{i,k} - \sum_{j<i} \mathbb{L}'_{i,j}\mathbb{U}'_{j,k}$
  fin
  $\mathbb{L}'_{k,k} = 1$
  $\mathbb{U}'_{k,k} = \mathbb{A}_{k,k} - \sum_{j<k} \mathbb{L}'_{k,j}\mathbb{U}'_{j,k}$
fin

5.10 Factorisation de Gauss-Jordan. Factorisation de Crout

On écrit les algorithmes ainsi que les démonstrations uniquement dans le cas $\mathbb{K} = \mathbb{C}$. *Les mêmes idées s'appliquent dans* $\mathbb{K} = \mathbb{R}$ *si on remplace* * *par* ^{T}, *resp. hermitienne par symétrique (et si omet la conjugaison).*
Dans la factorisation précédente $\mathbb{A} = \mathbb{L}\mathbb{U}$, on a imposé le choix d'une matrice $\mathbb{L}$ triangulaire inférieure à diagonale unité. Si on note $\mathbb{D}$ la matrice diagonale formée à partir des éléments diagonaux de $\mathbb{U}$: $\mathbb{D}_{i,i} = \mathbb{U}_{i,i}$, alors

$$\mathbb{A} = \mathbb{L}\,\mathbb{U} = \mathbb{L}\,\mathbb{D}\,\tilde{\mathbb{U}}.$$

Cette factorisation est appelée *factorisation de Gauss-Jordan*, dans laquelle la matrice $\tilde{\mathbb{U}}$ est triangulaire supérieure à diagonale unité. Les éléments des matrices $\mathbb{D}$ et $\tilde{\mathbb{U}}$ sont déterminés à partir des éléments de $\mathbb{U}$ par les relations

pour $k = 1, \ldots, n$ **faire**
  $\mathbb{D}_{k,k} = \mathbb{U}_{k,k}$
  pour $i = 1, \ldots, k$ **faire**
    $\tilde{\mathbb{U}}_{k,i} = \mathbb{U}_{i,k}$ / $\mathbb{U}_{k,k} = \mathbb{U}_{i,k}$ / $\mathbb{D}_{k,k}$.
  fin
fin

Cette écriture s'avère utile dans le cas où la matrice $\mathbb{A}$ est hermitienne. En effet, lorsqu'on emploie une stratégie de *pivot total sur les éléments diagonaux* ($\mathbb{Q} = \mathbb{P}^T = \mathbb{P}^*$), on a alors

Proposition 5.19 *Soit* $\mathbb{A} \in \mathbb{C}^{n\times n}$ *une matrice inversible et hermitienne, à une matrice de permutation* $\mathbb{P}$ *près, la factorisation* $\mathbb{P}\,\mathbb{A}\,\mathbb{P}^* = \mathbb{L}\,\mathbb{D}\,\mathbb{L}^*$ *est unique.*

Démonstration : On a par construction $(\mathbb{P}\,\mathbb{A}\,\mathbb{P}^*)^* = \mathbb{P}\,\mathbb{A}^*\,\mathbb{P}^* = \mathbb{P}\,\mathbb{A}\,\mathbb{P}^*$, d'où

$$\mathbb{L}\,\mathbb{D}\,\tilde{\mathbb{U}} = \tilde{\mathbb{U}}^*\,\mathbb{D}\,\mathbb{L}^*,$$

et en utilisant la Proposition 5.18 sur l'unicité de la factorisation de Gauss, on trouve bien

$$\mathbb{L} = \tilde{\mathbb{U}}^*.$$

$\diamond$

On obtient ainsi la **factorisation de Crout** pour une matrice hermitienne :

$$\mathbb{A} = \mathbb{L}\,\mathbb{D}\,\mathbb{L}^*.$$

Par identification, les éléments de $\mathbb{L}$ et $\mathbb{D}$ sont calculés suivant

pour $k = 1, \ldots, n$ **faire**
  $\mathbb{L}_{k,k} = 1 \quad$ et $\quad \mathbb{D}_{k,k} = \mathbb{A}_{k,k} - \sum_{j<k} \mathbb{D}_{j,j}\,|\mathbb{L}_{k,j}|^2$
  pour $i = k+1, \ldots, n$ **faire**
    $\mathbb{L}_{i,k} = [\mathbb{A}_{i,k} - \sum_{j<k} \mathbb{L}_{i,j}\mathbb{D}_{j,j}\overline{\mathbb{L}_{k,j}}]\ /\ \mathbb{D}_{k,k}.$
  fin
fin

La résolution du système linéaire $\mathbb{A}x = b$ s'effectue alors en trois étapes : le calcul du vecteur z solution de $\mathbb{L}z = b$, puis du vecteur y par $\mathbb{D}y = z$ et enfin du vecteur x par $\mathbb{L}^*x = y$. Le coût calcul de ces trois résolutions est identique à celui de l'étape correspondante de la méthode de Gauss puisque la matrice $\mathbb{L}$ est à diagonale unité.

On s'est limité au cas $\mathbb{Q} = \mathbb{P}^T = \mathbb{P}^*$. Ceci signifie que dans la stratégie du pivot total la recherche du meilleur pivot est limité aux éléments diagonaux de la matrice en cours de calcul, pour conserver la symétrie. Cette contrainte peut être levée dans le cadre de la factorisation par blocs, qui sera étudiée plus loin.

Pour finir, comme il n'est plus besoin de calculer la matrice triangulaire supérieure $\mathbb{U}$ $(= \mathbb{L}^*)$, on déduit de la Proposition 5.38 le résultat sur le **coût calcul**.

5.11 Factorisation de Cholesky

On écrit les algorithmes ainsi que les démonstrations uniquement dans le cas $\mathbb{K} = \mathbb{C}$. Les mêmes idées s'appliquent dans $\mathbb{K} = \mathbb{R}$ si on remplace * par ^{T}, resp. hermitienne par symétrique (et si omet la conjugaison).

Supposons maintenant que $\mathbb{A} \in \mathbb{C}^{n\times n}$ soit une matrice hermitienne[36] définie-positive, il résulte de la Proposition 5.19 que l'on peut écrire $\mathbb{A} = \mathbb{P}^*\,\mathbb{L}\,\mathbb{D}\,\mathbb{L}^*\,\mathbb{P}$, puisque $\mathbb{P}^{-1} = \mathbb{P}^* = \mathbb{P}^T$. De plus, par définition, on a la propriété

$$\forall x \in \mathbb{C}^n,\ x \neq 0 \qquad 0 < (\mathbb{A}x, x) = (\mathbb{P}^*\mathbb{L}\mathbb{D}\mathbb{L}^*\mathbb{P}x, x) = (\mathbb{D}\,\mathbb{L}^*\mathbb{P}x, \mathbb{L}^*\mathbb{P}x).$$

Pour tout $y \neq 0$, il existe $x \neq 0$ tel que $y = \mathbb{L}^*\mathbb{P}x$, et on a par conséquent $(\mathbb{D}y, y) > 0$. Pour tout $1 \leq i \leq n$, si on choisit pour y le i^{eme} vecteur de base e_i, on en déduit que $\mathbb{D}_{i,i} \in \mathbb{R}$,

36. Lorsque $\mathbb{A} \in \mathbb{C}^{n\times n}$ est hermitienne, on a pour tout $x \in \mathbb{C}^n$: $(\mathbb{A}x, x) = (x, \mathbb{A}^*x) \overset{\mathbb{A}^*=\mathbb{A}}{=} (x, \mathbb{A}x) = \overline{(\mathbb{A}x, x)}$, ainsi $(\mathbb{A}x, x) \in \mathbb{R}$ et on peut considérer son signe, et par voie de conséquence le caractère **définipositif** d'une matrice hermitienne $\mathbb{A}$...

et de plus $\mathbb{D}_{i,i} > 0$; la matrice diagonale $\mathbb{D}$ est donc une matrice réelle définie-positive, ce qui permet de définir la matrice "racine carrée" diagonale $\mathbb{D}^{1/2}$ par $(\mathbb{D}^{1/2})_{i,i} = \sqrt{\mathbb{D}_{i,i}}$. En posant alors $\mathcal{L} = \mathbb{L}\mathbb{D}^{1/2}$, on obtient finalement

$$\mathbb{P}\ \mathbb{A}\ \mathbb{P}^* = \mathbb{L}\mathcal{L}^*.$$

Proposition 5.20 *Soit $\mathbb{A} \in \mathbb{C}^{n\times n}$ une matrice hermitienne définie-positive, il existe une matrice $\mathcal{L}$ triangulaire inférieure, telle que la factorisation $\mathbb{P}\mathbb{A}\mathbb{P}^* = \mathcal{L}\mathcal{L}^*$ est unique à une matrice de permutation $\mathbb{P}$ près.*

En utilisant les formules générales, les éléments de $\mathcal{L}$ sont calculés *colonne par colonne* par les relations

pour $k = 1, \ldots, n$ **faire**
$$\mathcal{L}_{k,k} = [\mathbb{A}_{k,k} - \sum_{j<k} |\mathcal{L}_{k,j}|^2]^{1/2}$$
pour $i = k+1, \ldots, n$ **faire**
$$\mathcal{L}_{i,k} = [\mathbb{A}_{i,k} - \sum_{j<k} \mathcal{L}_{k,j}\overline{\mathcal{L}_{i,j}}] \ / \ \mathcal{L}_{k,k}.$$
fin
fin

On peut aussi calculer la matrice $\mathcal{L}$ *ligne par ligne* suivant

pour $k = 1, \ldots, n$ **faire**
pour $i = 1, \ldots, k-1$ **faire**
$$\mathcal{L}_{k,i} = [\mathbb{A}_{k,i} - \sum_{j<i} \mathcal{L}_{k,j}\overline{\mathcal{L}_{i,j}}] \ / \ \mathcal{L}_{i,i}$$
fin
$$\mathcal{L}_{k,k} = [\mathbb{A}_{k,k} - \sum_{j<k} |\mathcal{L}_{k,j}|^2]^{1/2}.$$
fin

Une conséquence importante de la Proposition 5.20, dans le cas où la matrice $\mathbb{A}$ est hermitienne et *définie-positive*, est que l'on n'a pas besoin de vérifier que les termes

$$\mathbb{A}_{k,k} - \sum_{j<k} |\mathcal{L}_{k,j}|^2$$

sont tous strictement positifs, pour tout k. L'existence des éléments diagonaux $\mathcal{L}_{k,k}$ étant assurée par cette Proposition, il suffit de vérifier par identification que l'on peut les calculer de cette manière. Noter que cette propriété n'est pas satisfaite si $\mathbb{A}$ est supposée seulement hermitienne : dans la factorisation de Crout certains éléments diagonaux $\mathbb{D}_{i,i}$ peuvent en effet être négatifs.

Une autre conséquence remarquable de la Proposition 5.18 peut être obtenue en faisant intervenir à nouveau le complément de Schur $\mathcal{S}$:

$$\mathbb{A} = \begin{pmatrix} [\mathbb{A}]_{1,1} & [\mathbb{A}]_{1,2} \\ [\mathbb{A}]_{2,1} & [\mathbb{A}]_{2,2} \end{pmatrix}$$

avec $[\mathbb{A}]_{1,1} = [\mathbb{A}]_{1,1}^* \in \mathbb{C}^{n_1 \times n_1}$, $[\mathbb{A}]_{2,1} \in \mathbb{C}^{n_2 \times n_1}$, $[\mathbb{A}]_{1,2} = [\mathbb{A}]_{2,1}^* \in \mathbb{C}^{n_1 \times n_2}$, $[\mathbb{A}]_{2,2} = [\mathbb{A}]_{2,2}^* \in \mathbb{C}^{n_2 \times n_2}$, $n = n_1 + n_2$. Par construction, $[\mathbb{A}]_{1,1}$ et $[\mathbb{A}]_{2,2}$ sont définies-positives. Par exemple :

$$\forall \hat{x} \in \mathbb{C}^{n_1}, \hat{x} \neq 0 \quad 0 < \left(\mathbb{A} \begin{pmatrix} \hat{x} \\ 0 \end{pmatrix}, \begin{pmatrix} \hat{x} \\ 0 \end{pmatrix} \right) = \left(\begin{pmatrix} [\mathbb{A}]_{1,1} & [\mathbb{A}]_{2,1}^* \\ [\mathbb{A}]_{2,1} & [\mathbb{A}]_{2,2} \end{pmatrix} \begin{pmatrix} \hat{x} \\ 0 \end{pmatrix}, \begin{pmatrix} \hat{x} \\ 0 \end{pmatrix} \right) = ([\mathbb{A}]_{1,1}\hat{x}, \hat{x}).$$

On suppose que l'on connait une factorisation de Cholesky du premier bloc : $[\mathbb{A}]_{1,1} = [\mathcal{L}]_{1,1}[\mathcal{L}]_{1,1}^*$, avec $[\mathcal{L}]_{1,1} \in \mathbb{C}^{n_1 \times n_1}$ matrice triangulaire inférieure, alors (cf. Remarque 5.17)

$$\mathbb{A} = \begin{pmatrix} [\mathbb{A}]_{1,1} & [\mathbb{A}]_{1,2} \\ [\mathbb{A}]_{2,1} & [\mathbb{A}]_{2,2} \end{pmatrix} = \begin{pmatrix} [\mathcal{L}]_{1,1} & 0 \\ [\mathcal{L}]_{2,1} & I_{n_2} \end{pmatrix} \times \begin{pmatrix} I_{n_1} & 0 \\ 0 & \mathcal{S} \end{pmatrix} \times \begin{pmatrix} [\mathcal{L}]_{1,1}^* & [\mathcal{L}]_{2,1}^* \\ 0 & I_{n_2} \end{pmatrix}. \tag{5.4}$$

Par identification, on trouve facilement que $[\mathcal{L}]_{2,1} = [\mathbb{A}]_{2,1}([\mathcal{L}]_{1,1}^*)^{-1}$, et $\mathcal{S} = [\mathbb{A}]_{2,2} - [\mathcal{L}]_{2,1}[\mathcal{L}]_{2,1}^* = [\mathbb{A}]_{2,2} - [\mathbb{A}]_{2,1}[\mathbb{A}]_{1,1}^{-1}[\mathbb{A}]_{2,1}^*$.

Proposition 5.21 *Soit $\mathbb{A} \in \mathbb{C}^{n \times n}$ une matrice hermitienne définie-positive, pour tout couple (n_1, n_2) avec $n_1 + n_2 = n$, le complément de Schur $\mathcal{S} \in \mathbb{C}^{n_2 \times n_2}$ est une matrice hermitienne définie-positive.*

Démonstration : Par construction la matrice $\mathcal{S} = [\mathbb{A}]_{2,2} - [\mathcal{L}]_{2,1}[\mathcal{L}]_{2,1}^*$ est hermitienne. En outre, on peut reformuler (5.4) sous la forme équivalente

$$\begin{pmatrix} I_{n_1} & 0 \\ 0 & \mathcal{S} \end{pmatrix} = \underline{\mathcal{L}}^{-1} \times \mathbb{A} \times (\underline{\mathcal{L}}^{-1})^*, \text{ avec } \underline{\mathcal{L}} = \begin{pmatrix} [\mathcal{L}]_{1,1} & 0 \\ [\mathcal{L}]_{2,1} & I_{n_2} \end{pmatrix}.$$

Ainsi, pour tout $\tilde{x} \in \mathbb{C}^{n_2}, \tilde{x} \neq 0$, on a $x' = (\underline{\mathcal{L}}^{-1})^* \begin{pmatrix} 0 \\ \tilde{x} \end{pmatrix} \neq 0$ et

$$\begin{aligned} (\mathcal{S}\tilde{x}, \tilde{x}) &= \left(\begin{pmatrix} I_{n_1} & 0 \\ 0 & \mathcal{S} \end{pmatrix} \begin{pmatrix} 0 \\ \tilde{x} \end{pmatrix}, \begin{pmatrix} 0 \\ \tilde{x} \end{pmatrix} \right) \\ &= \left(\mathbb{A} \times (\underline{\mathcal{L}}^{-1})^* \begin{pmatrix} 0 \\ \tilde{x} \end{pmatrix}, (\underline{\mathcal{L}}^{-1})^* \begin{pmatrix} 0 \\ \tilde{x} \end{pmatrix} \right) = (\mathbb{A}x', x') > 0. \end{aligned}$$

◇

En conséquence, on en déduit qu'il est toujours possible de réaliser la factorisation (de Cholesky) d'une matrice hermitienne définie-positive *sans permutation*, ce qui est *faux* pour la factorisation (de Gauss) d'une matrice inversible quelconque, et qui reste *faux* en général pour la factorisation (de Crout) d'une matrice hermitienne inversible quelconque.

Corollaire 5.22 *Soit $\mathbb{A} \in \mathbb{C}^{n \times n}$ une matrice hermitienne définie-positive. Il existe une matrice $\mathcal{L}$ triangulaire inférieure unique, telle que $\mathbb{A} = \mathcal{L}\mathcal{L}^*$.*

Démonstration : Pour réaliser la factorisation sans permutation, on raisonne par récurrence sur k, qui varie de 1 à $n-1$, et l'on montre que $\mathbb{A}_{k,k}^{(k-1)} = (\mathbb{A}^{(k-1)}e_k, e_k) > 0$.
Avec la convention $\mathcal{S}^{(0)} = \mathbb{A}$, ceci revient à vérifier que $(\mathcal{S}^{(k-1)}e_k, e_k) > 0$.

Pour cela, il suffit de montrer que $\mathcal{S}^{(k-1)}$ est hermitienne définie-positive, par récurrence sur k. Or, $\mathcal{S}^{(0)} = \mathbb{A}$ est hermitienne définie-positive et, d'après la Proposition précédente, l'assertion "$\mathcal{S}^{(k-1)}$ hermitienne définie-positive entraîne $\mathcal{S}^{(k)}$ hermitienne définie-positive" est vraie. ◇

5.12 Factorisation par blocs

On écrit les algorithmes ainsi que les démonstrations uniquement dans le cas $\mathbb{K} = \mathbb{C}$. Les mêmes idées s'appliquent dans $\mathbb{K} = \mathbb{R}$ si on remplace * par ^{T}, resp. hermitienne par symétrique (et si omet la conjugaison).
Dans ce qui précède les matrices triangulaires qui définissent les différentes factorisations ont été calculées élément par élément, on peut à ce titre les qualifier de **factorisations ponctuelles** ; cependant si on effectue une une partition des matrices $\mathbb{A}$, $\mathbb{L}$ et $\mathbb{U}$ en $P \times P$ blocs, on obtient formellement

$$\begin{bmatrix} [\mathbb{A}]_{1,1} & [\mathbb{A}]_{1,2} & \dots & [\mathbb{A}]_{1,P} \\ [\mathbb{A}]_{2,1} & [\mathbb{A}]_{2,2} & \ddots & [\mathbb{A}]_{2,P} \\ \vdots & \ddots & \ddots & \vdots \\ [\mathbb{A}]_{P,1} & [\mathbb{A}]_{P,2} & \dots & [\mathbb{A}]_{P,P} \end{bmatrix} = \begin{bmatrix} [\mathbb{L}]_{1,1} & 0 & \dots & 0 \\ [\mathbb{L}]_{2,1} & [\mathbb{L}]_{2,2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ [\mathbb{L}]_{P,1} & [\mathbb{L}]_{P,2} & \dots & [\mathbb{L}]_{P,P} \end{bmatrix} \times \begin{bmatrix} [\mathbb{U}]_{1,1} & [\mathbb{U}]_{1,2} & \dots & [\mathbb{U}]_{1,P} \\ 0 & [\mathbb{U}]_{2,2} & \ddots & [\mathbb{U}]_{2,P} \\ \dots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & [\mathbb{U}]_{P,P} \end{bmatrix}$$

Dans cette écriture on rappelle que seuls les blocs diagonaux sont nécessairement carrés. Par identification on obtient la **factorisation par blocs** de Gauss

pour $k = 1, \dots, P$ **faire** (*factorisation ponctuelle* du bloc $[\mathbb{A}]_{k,k}$)
  $[\mathbb{L}]_{k,k}\ [\mathbb{U}]_{k,k} = [\mathbb{A}]_{k,k} - \sum_{q<k} [\mathbb{L}]_{k,q}[\mathbb{U}]_{q,k}$
  pour $p = k+1, \dots, P$ **faire**
    $[\mathbb{L}]_{p,k}[\mathbb{U}]_{k,k} = [\mathbb{A}]_{p,k} - \sum_{q<k} [\mathbb{L}]_{p,q}[\mathbb{U}]_{q,k}$
  fin
  pour $p = k+1, \dots, P$ **faire**
    $[\mathbb{L}]_{k,k}[\mathbb{U}]_{k,p} = [\mathbb{A}]_{k,p} - \sum_{q<k} [\mathbb{L}]_{k,q}[\mathbb{U}]_{q,p}$
  fin
fin

dans cette écriture les produits $[\mathbb{L}]_{p,q}[\mathbb{U}]_{q,k}$ sont des produits de matrices. Les blocs $[\mathbb{L}]_{p,k}$, resp. $[\mathbb{U}]_{k,p}$, sont obtenus par résolution de systèmes linéaires carrés à matrice triangulaire supérieure, et dont les seconds membres sont des matrices $[\mathbb{B}]$ et $[\mathbb{B}']$. Pour cela, on écrit pour $k = 1, \cdots, P$:

$$[\mathbb{L}]_{p,k}[\mathbb{U}]_{k,k} = [\mathbb{B}] \iff [\mathbb{U}]^*_{k,k}[\mathbb{L}]^*_{p,k} = [\mathbb{B}]^* \ ; \quad [\mathbb{L}]_{k,k}[\mathbb{U}]_{k,p} = [\mathbb{B}'].$$

Comme pour la factorisation de Gauss ponctuelle, il peut être nécessaire d'effectuer des permutations de lignes ou de colonnes afin de placer des blocs réguliers sur la diagonale.

Mentionnons également la factorisation de Crout par blocs

pour $k = 1, \ldots, P$ **faire** (*factorisation ponctuelle* du bloc $[\mathbb{A}]_{k,k}$)
  $[\mathbb{L}]_{k,k}[\mathbb{D}]_{k,k}[\mathbb{L}]^*_{k,k} = [\mathbb{A}]_{k,k} - \sum_{q<k} [\mathbb{L}]_{k,q}[\mathbb{L}]^*_{k,q}$
  pour $p = k+1, \ldots, P$ **faire**
    $[\mathbb{D}]_{k,k}[\mathbb{L}]^*_{p,k} = [\mathbb{A}]_{p,k} - \sum_{q<k} [\mathbb{L}]_{p,q}[\mathbb{L}]^*_{k,q}$
  fin
fin

et enfin la factorisation de Cholesky par blocs

pour $k = 1, \ldots, P$ **faire** (*factorisation ponctuelle* du bloc $[\mathbb{A}]_{k,k}$)
  $[\mathcal{L}]_{k,k}[\mathcal{L}]^*_{k,k} = [\mathbb{A}]_{k,k} - \sum_{q<k} [\mathcal{L}]_{k,q}[\mathcal{L}]^*_{k,q}$
  pour $p = k+1, \ldots, p$ **faire**
    $[\mathcal{L}]_{k,k}[\mathcal{L}]^*_{p,k} = [\mathbb{A}]_{p,k} - \sum_{q<k} [\mathcal{L}]_{p,q}[\mathcal{L}]^*_{k,q}$
  fin
fin

Cette formulation est récursive puisque l'on peut y remplacer à nouveau les factorisations ponctuelles par des factorisations par blocs... Cette approche est intéressante pour un calcul concret sur ordinateur dans le cas de matrices de très grande taille qui ne tiennent pas en mémoire (on les découpe alors en blocs suffisamment petits) ou le plus souvent pour des matrices qui ne sont effectivement connues que sous la forme d'une partition.

Une autre application de cette formulation est la recherche d'un pivot extra-diagonal pour des matrices hermitiennes dont la factorisation pose des problèmes numériques, comme les matrices non définies. En effet dans le cas d'une stratégie de pivot total par points, il faut, pour *préserver la "symétrie"*, limiter la recherche de l'élément de plus grande valeur absolue aux seuls termes diagonaux. Cette procédure peut s'avérer inefficace si les éléments diagonaux sont trop petits. On applique alors la technique des **pivots jumeaux** : supposons qu'à l'étape k de la factorisation, le pivot idéal $\mathcal{S}_{k,k'} = \overline{\mathcal{S}_{k',k}}$ soit en position extra-diagonale

$$\mathcal{S} = \begin{pmatrix} \mathcal{S}_{k,k} & \cdot & \cdot & \overline{\mathcal{S}_{k,k'}} & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \mathcal{S}_{k',k} & \cdot & \cdot & \mathcal{S}_{k',k'} & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix}$$

alors à l'aide d'une permutation symétrique des lignes et des colonnes, on commence par écrire

$$\mathbb{P}\mathcal{S}\mathbb{P}^* = \begin{pmatrix} \mathcal{S}_{k,k} & \overline{\mathcal{S}_{k,k'}} & \cdot & \cdot & \cdot \\ \mathcal{S}_{k',k} & \mathcal{S}_{k',k'} & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix}$$

puis on effectue une factorisation par blocs "symétrique" de la matrice $\mathcal{S}$ avec un premier **bloc diagonal** 2×2, et les autres blocs diagonaux de la partition d'ordre 1. On assure ainsi la stabilité numérique de la factorisation, tout en conservant la "symétrie".

5.13 Profil et conservation du profil

Les matrices qui proviennent de l'approximation de la solution d'une équation aux dérivées partielles par la méthode des différences finies ou la méthode des éléments finis possèdent en règle générale deux propriétés fondamentales [37].
D'une part leur caractère creux : une matrice d'ordre n est **creuse** lorsque le nombre moyen d'éléments non nuls par ligne est petit devant n. Ceci permet de réduire le coût de la multiplication matrice-vecteur (voir le §6.1 pour plus de précisions).
D'autre part, les éléments non-nuls sont rassemblés autour de la diagonale ; en d'autres termes, elles possèdent beaucoup de 0 au début des lignes et colonnes [38]. L'exploitation de cette seconde propriété apporte une économie considérable lors de la factorisation, tant sur le plan du temps calcul (par la réduction du nombre d'opérations réalisées), que de la place mémoire, comme nous allons le voir.

Pour introduire la notion de profil, considérons la matrice d'ordre 8 :

$$\mathbb{A} = \begin{bmatrix} \bullet & \bullet & 0 & \bullet & 0 & 0 & 0 & \bullet \\ \bullet & \bullet & \bullet & 0 & 0 & \bullet & 0 & \bullet \\ 0 & \bullet & \bullet & \bullet & 0 & \bullet & 0 & \bullet \\ \bullet & 0 & \bullet & \bullet & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \bullet & \bullet & 0 & 0 \\ 0 & \bullet & \bullet & 0 & \bullet & \bullet & 0 & \bullet \\ 0 & 0 & 0 & 0 & 0 & 0 & \bullet & 0 \\ \bullet & \bullet & \bullet & 0 & 0 & \bullet & 0 & \bullet \end{bmatrix}$$

Les éléments non nuls de cette matrice sont représentés par le symbole $\bullet$ et les autres par 0. On remarque tout de suite que la matrice $\mathbb{A}$ ci-dessus est à *structure symétrique*, puisque les $\bullet$ sont placés symétriquement par rapport à la diagonale. En d'autres termes,

$$\mathbb{A}_{i,j} \neq 0 \Longleftrightarrow \mathbb{A}_{j,i} \neq 0. \tag{5.5}$$

On voudrait dans la suite que cette propriété (5.5) soit satisfaite (notamment pour simplifier les notations !). A cette fin, on *"symétrisera"* la structure de la matrice, en ajoutant un *couple* de symbole $\bullet$ non-diagonaux, dès lors que *l'un au moins* des $\mathbb{A}_{i,j}$ et $\mathbb{A}_{j,i}$ correspondants est non nul.

Pour chaque ligne k de la matrice $\mathbb{A}$, on peut définir $il(k)$ le plus petit indice de colonne l tel que $\mathbb{A}_{k,l} \neq 0$; on définit de même pour chaque colonne k, $ic(k)$ le plus petit indice de ligne l, tel que $\mathbb{A}_{l,k} \neq 0$. On note que pour pouvoir définir $il(k)$ et $ic(k)$ pour tout k, une

37. Les approximations par des méthodes intégrales sont un cas limite : les matrices sont "presque creuses", puisque le nombre moyen d'éléments non nuls par ligne est de l'ordre de $\log(n)$.

38. Il est important de noter que ces deux propriétés ne sont pas comparables : une matrice peut-être creuse sans qu'il y ait beaucoup de 0 en début des lignes et colonnes.

condition *suffisante*[39] est l'inversibilité de $\mathbb{A}$, ce qui tombe bien, puisque on s'intéresse à la factorisation en vue de la résolution de systèmes linéaires !
Comme (5.5) est satisfaite (éventuellement par "symétrisation" de la structure), on a automatiquement $il(k) = ic(k)$ pour tout k, ce qui permet de s'affranchir des indices $ic(k)$ dans la suite.

On introduit la propriété suivante :

$$il(k) \leq k, \ \forall k. \tag{5.6}$$

On a le résultat ci-dessous.

Proposition 5.23 *Si les factorisations de Gauss et Crout se font sans permutation, alors la propriété (5.6) est vraie.*

Démonstration : Supposons qu'il existe k tel que $il(k) > k$. Dans ce cas, on a par définition $\mathbb{A}_{k,1} = \mathbb{A}_{k,2} = \cdots = \mathbb{A}_{k,k} = 0$. Ecrivons que $\mathbb{A} = \mathbb{L}\mathbb{U}$ avec $\mathbb{L}$ triangulaire inférieure et $\mathbb{U}$ triangulaire supérieure, sur la $k^{ème}$ ligne

$$\begin{aligned}
0 &= \mathbb{A}_{k,1} = \mathbb{L}_{k,1}\mathbb{U}_{1,1}, \\
0 &= \mathbb{A}_{k,2} = \mathbb{L}_{k,1}\mathbb{U}_{1,2} + \mathbb{L}_{k,2}\mathbb{U}_{2,2}, \\
&\vdots \\
0 &= \mathbb{A}_{k,k-1} = \mathbb{L}_{k,1}\mathbb{U}_{1,k-1} + \mathbb{L}_{k,2}\mathbb{U}_{2,k-1} + \cdots + \mathbb{L}_{k,k-1}\mathbb{U}_{k-1,k-1} \\
0 &= \mathbb{A}_{k,k} = \mathbb{L}_{k,1}\mathbb{U}_{1,k} + \mathbb{L}_{k,2}\mathbb{U}_{2,k} + \cdots + \mathbb{L}_{k,k-1}\mathbb{U}_{k-1,k} + \mathbb{L}_{k,k}\mathbb{U}_{k,k}.
\end{aligned}$$

On déduit de la première équation que $\mathbb{L}_{k,1} = 0$, puisque $\mathbb{U}$ est inversible. De la deuxième équation, on déduit alors que $\mathbb{L}_{k,2} = 0$, et ainsi de suite jusqu'à la $(k-1)^{ème}$ équation, qui fournit le résultat $\mathbb{L}_{k,k-1} = 0$. En passant enfin à la $k^{ème}$ équation, on arrive à $\mathbb{L}_{k,k}\mathbb{U}_{k,k} = 0$, ce qui contredit l'hypothèse $\mathbb{L}$ et $\mathbb{U}$ inversibles. ◇

En d'autres termes, la propriété (5.6) est *nécessaire* pour pouvoir factoriser une matrice sans permutation, mais elle n'est pas *suffisante*...

Exercice 5.1 *Soit $\mathbb{A} \in \mathbb{R}^{3\times 3}$ la matrice symétrique et inversible*

$$\mathbb{A} = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 1 & 3 \\ 2 & 3 & 3 \end{bmatrix}.$$

Vérifier que $\mathbb{A}$ ne peut être factorisée sans permutation, bien que la propriété (5.6) soit satisfaite.

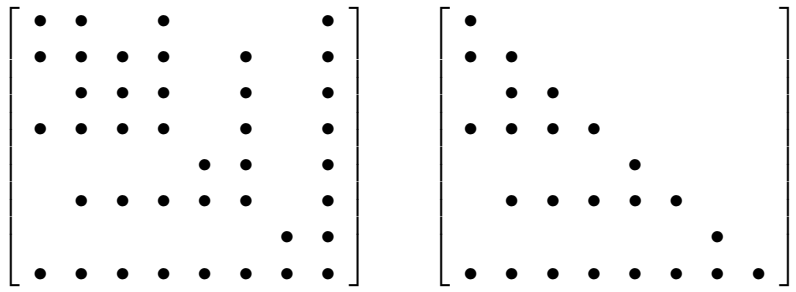
Sous réserve que les propriétés (5.5) et (5.6) soient satisfaites pour $\mathbb{A} \in \mathbb{K}^{n\times n}$, on introduit les ensembles

$$Pl(\mathbb{A}) = \{(k,l), 1 \leq k \leq n, il(k) \leq l \leq k\} \text{ et } Pc(\mathbb{A}) = \{(l,k), 1 \leq k \leq n, il(k) \leq l \leq k\}.$$

39. En effet, lorsque $\mathbb{A}$ est inversible, ses lignes et ses colonnes sont *toutes* non nulles.

Définition 5.24 *on appelle* **profil** *de la matrice* $\mathbb{A}$
– *l'ensemble* $Pr(\mathbb{A}) = Pl(\mathbb{A})$ *si* $\mathbb{A}$ *est hermitienne ou symétrique*
– *l'ensemble* $Pr(\mathbb{A}) = Pl(\mathbb{A}) \cup Pc(\mathbb{A})$ *sinon.*

Reprenons l'exemple ci-dessus et décrivons l'ensemble $Pr(\mathbb{A})$ correspondant.



$Pr(\mathbb{A})$: à gauche si $\mathbb{A}$ vérifie uniquement (5.5), à droite si $\mathbb{A}$ est de plus hermitienne ou symétrique.

Remarque 5.25 *On a défini le* **profil ponctuel**, *il est évident que l'on peut associer à toute partition de la matrice* $\mathbb{A}$ *un* **profil par blocs**.

Par définition, $(i,j) \in Pr(\mathbb{A})$ n'entraîne pas que $\mathbb{A}_{i,j} \neq 0$. En d'autres termes il peut exister des éléments de $\mathbb{A}$ nuls à l'intérieur du profil ! Il faut donc distinguer l'ensemble $Pr(\mathbb{A})$ de l'ensemble

$$\mathcal{S}q(\mathbb{A}) = \{(k,l), 1 \leq l \leq k \leq n, \mathbb{A}_{k,l} \neq 0\}$$

qui est appelé le **squelette** de la matrice $\mathbb{A}$. Lorsque la matrice $\mathbb{A}$ est creuse, alors le cardinal de son squelette est petit devant n^2.

Proposition 5.26 *Les factorisations de Gauss et Crout,* lorsqu'elles se font sans permutation*, ainsi que la factorisation de Cholesky, conservent le profil.*
En d'autres termes, $(i,j) \notin Pr(\mathbb{A})$ *entraîne* $\mathbb{L}_{i,j} = \mathbb{U}_{i,j} = 0$.

Démonstration : On raisonne pour la factorisation de Gauss $\mathbb{L}\,\mathbb{U} = \mathbb{A}$. On vérifie tout d'abord que $\mathbb{L}$ conserve le profil en reprenant la preuve de la Proposition 5.23, i. e. en écrivant que, pour la ligne k, $\mathbb{A}_{i,k} = 0$ tant que $i < il(k)$. Ensuite, pour $\mathbb{U}$, on écrit cette fois des égalités pour la colonne k de $\mathbb{A}$, sachant que $\mathbb{A}_{k,j} = 0$ tant que $j < il(k)$.
Cette propriété est commune aux trois factorisations, la structure des calculs étant identique. ◇

Remarque 5.27 *Ce résultat est valable pour le profil par points et pour le profil par blocs.*

Pour finir, on introduit la **largeur de bande** de la ligne k par la relation $lb(k) = k - il(k) + 1$. On dit qu'une matrice $\mathbb{A} \in \mathbb{K}^{n\times n}$ est **à faible largeur de bande** lorsque sa largeur de bande moyenne $l = [\sum_{k=1}^{n} l(k)]/n$ est petite devant n. Si de plus on a $lb(k) \eqsim l$ pour tout k, on peut vérifier que le coût calcul de la factorisation est de l'ordre de $O(n\,l^2)$.

5.14 Factorisation QR

5.14.1 Introduction

Nous allons montrer un résultat très utile en pratique, concernant cette fois la factorisation d'une matrice quelconque :

Proposition 5.28 *Soit une matrice $\mathbb{A} \in \mathbb{K}^{n\times m}$ ($n \geq m$), alors il existe une matrice carrée $\mathbb{Q} \in \mathbb{K}^{n\times n}$ et une matrice triangulaire supérieure $\mathbb{R} \in \mathbb{K}^{n\times m}$ telles que*

$$\mathbb{A} = \mathbb{Q}\,\mathbb{R}, \text{ avec } \mathbb{Q}^*\mathbb{Q} = I_n.$$

La preuve est constructive. Nous allons exhiber la factorisation par la méthode de Householder, où $\mathbb{Q}$ est obtenue par produits successifs de matrices orthogonales élémentaires. Puis nous expliquerons plus brièvement la factorisation par la méthode de Givens où $\mathbb{Q}$ est obtenue par produits successifs de matrices de rotations planes.

5.14.2 Factorisation de Householder

Définition 5.29 *Soit $v \in \mathbb{K}^n\backslash\{0\}$. On appelle matrice de Householder la matrice de la forme :*

$$\mathbb{H}(v) = I_n - 2\frac{v\,v^*}{v^*\,v} = I_n - 2\frac{v\,v^*}{\|v\|^2}.$$

Par convention, la matrice identité est considérée comme étant une matrice de Householder.

Proposition 5.30 *Les matrices de Householder sont Hermitiennes et unitaires, de déterminant égal à -1.*

Démonstration : Soit $v \in \mathbb{K}^n\backslash\{0\}$ et $v^\perp$ un vecteur orthogonal à v. La matrice $\mathbb{H}(v)$ est *hermitienne*. En effet, on a :

$$(\mathbb{H}(v))^* = I_n - 2\frac{(v\,v^*)^*}{\|v\|^2} = I_n - 2\frac{v\,v^*}{\|v\|^2} = \mathbb{H}(v).$$

Elle est également *unitaire*, car on a :

$$\begin{aligned}
\mathbb{H}(v)\,(\mathbb{H}(v))^* &= \left(I_n - 2\frac{v\,v^*}{\|v\|^2}\right)\left(I_n - 2\frac{v\,v^*}{\|v\|^2}\right) \\
&= I_n - 4\,\frac{v\,v^*}{\|v\|^2} + 4\frac{(v\,v^*)\,(v\,v^*)}{\|v\|^4} \\
&= I_n - 4\,\frac{v\,v^*}{\|v\|^2} + 4\frac{v\,(v\,v^*)\,v^*}{\|v\|^4} \\
&= I_n.
\end{aligned}$$

On remarque que $\mathbb{H}(v)v^\perp = v^\perp$. Ainsi, 1 est valeur propre de $\mathbb{H}(v)$ de multiplicité $n-1$. De plus, $\mathbb{H}(v)v = -v$. Ainsi, -1 est valeur propre de $\mathbb{H}(v)$ de multiplicité 1. Le déterminant d'un matrice étant égal au produit de ses valeurs propres, on obtient bien $\det(\mathbb{H}(v)) = -1$. $\diamond$

L'interprétation géométrique est la suivante : pour $u \in \mathbb{K}^n$, $\mathbb{H}(v)u$ est le vecteur symétrique à u par rapport à l'hyperplan orthogonal au vecteur v.

Nous réécrivons ci-dessous le théorème [13, thm 4.5-1, p. 152] ainsi que sa preuve :

Théorème 5.31 *Soit $v \in \mathbb{K}^n$ tel que $\sum_{i=2}^{n} |v_i| > 0$. Alors on peut construire deux matrices de Householder $\mathbb{H}(w_+)$, $\mathbb{H}(w_-)$ telles que le vecteur $\mathbb{H}(w_\pm)v$ soit colinéaire à e_1, le premier vecteur de la base canonique de $\mathbb{K}^n$. Plus précisément, soit α tel que $v_1 = (v_1, e_1) = |v_1| \exp(i\alpha)$, alors pour $w_\pm := v \pm \|v\| \exp(i\alpha)$, on a :*

$$\mathbb{H}(w_\pm)v = \mp \|v\| \exp(i\alpha)\, \mathrm{e}_1.$$

Démonstration : Par calcul, on a :

$$\mathbb{H}\left(v \pm \|v\| \exp(i\alpha)\, \mathrm{e}_1\right) v = v - 2\frac{(v \pm \|v\| \exp(i\alpha)\, \mathrm{e}_1)(v^* \pm \|v\| \exp(-i\alpha)\, \mathrm{e}_1^*)v}{(v^* \pm \|v\| \exp(-i\alpha)\, \mathrm{e}_1^*)(v \pm \|v\| \exp(i\alpha)\, \mathrm{e}_1)},$$

avec :

$$\begin{aligned}(v \pm \|v\| \exp(i\alpha)\, \mathrm{e}_1)(v^* \pm \|v\| \exp(-i\alpha)\, \mathrm{e}_1^*)v &= (v \pm \|v\| \exp(i\alpha)\, \mathrm{e}_1)(\|v\|^2 \pm \|v\|\, |v_1|), \\ &= \|v\|(\|v\| \pm |v_1|)(v \pm \|v\| \exp(i\alpha)\, \mathrm{e}_1).\end{aligned}$$

et :

$$\begin{aligned}(v^* \pm \|v\| \exp(-i\alpha)\, \mathrm{e}_1^*)(v \pm \|v\| \exp(i\alpha)\, \mathrm{e}_1) &= \|v\|^2 \pm \|v\|\, |v_1| \pm \|v\|\, |v_1| + \|v\|^2, \\ &= 2\, \|v\|(\|v\| \pm |v_1|)\,.\end{aligned}$$

◇

La remarque suivante est importante en pratique pour l'implémentation de la méthode :

Remarque 5.32 *Pour éviter que le dénominateur $2\,\|v\|(\|v\| \pm |v_1|)$ associé à $w_\pm$ soit trop petit, on construira $\mathbb{H}(w_+)$ plutôt que $\mathbb{H}(w_-)$. Pour $\mathbb{K} = \mathbb{R}$, $\exp(i\alpha) = \pm 1$, c'est le signe de v_1 ; on construit alors $\mathbb{H}(w_+)$ avec :* $\left\{ \begin{array}{lcl} w_+ &=& v + \|v\|\, \mathrm{e}_1 \quad \textit{si } v_1 \geq 0, \\ w_+ &=& v - \|v\|\, \mathrm{e}_1 \quad \textit{si } v_1 < 0. \end{array} \right.$

Nous allons maintenant procéder à la factorisation de $\mathbb{A}$. On pose $k_{max} = n - 1$ si $m = n$ et $k_{max} = m$ si $m < n$.

Initialisation : on pose $\mathbb{A}^{(0)} = \mathbb{A}$.

Pour $k = 1$ **:**

— Soit a_1 la première colonne de la matrice $\mathbb{A}^{(0)}$: $a_1 := \mathbb{A}^{(0)}_{:,1} \in \mathbb{K}^n$.
Soit $\alpha_1 \in \mathbb{R}$ tel que $(a_1)_1 = |(a_1)_1| \exp(i\alpha_1)$.
Si $\sum_{i=2}^{n} |\mathbb{A}^{(0)}_{i,1}| = 0$, on pose $\mathbb{H}_1 := I_n$ et $\mathbb{A}^{(1)} = \mathbb{A}^{(0)}$.
Si $\sum_{i=2}^{n} |\mathbb{A}^{(0)}_{i,1}| > 0$, on calcule $\mathbb{H}_1 = \mathbb{H}\left(a_1 + \|a_1\| \exp(i\alpha_1)\, \mathrm{e}_1\right)$.
On calcule $\mathbb{A}^{(1)} := \mathbb{H}_1 \mathbb{A}^{(0)}$, qui est telle que $\mathbb{A}^{(1)}_{i>1,1} = 0$.

Pour $2 \leq k \leq k_{max}$ **:** pour construire $\mathbb{A}^{(k)}$, on suppose qu'à l'étape $k-1$, on a calculé la matrice $\mathbb{A}^{(k-1)} = \mathbb{H}_{k-1} \cdots \mathbb{H}_1 \mathbb{A}^{(0)}$ telle que pour $j \in \{1, \cdots, k-1\}$, $\mathbb{A}^{(k-1)}_{i>j,j} = 0$.

— Soit $a_k := \mathbb{A}^{(k-1)}_{k:n,k} \in \mathbb{K}^{n-k+1}$, soit $\alpha_k \in \mathbb{R}$ tel que $(a_k)_1 = (a_k, \mathrm{e}_1)_{n-k+1} = |(a_k)_1| \exp(i\alpha_k)$, où e_1 est le premier vecteur de la base orthonormale canonique de $\mathbb{K}^{n-k+1}$.
Si $\sum_{i=k+1}^{n} |\mathbb{A}^{(k-1)}_{i,k}| = 0$, on pose $\mathbb{H}_k := I_n$ et $\mathbb{A}^{(k)} = \mathbb{A}^{(k-1)}$.
Si $\sum_{i=k+1}^{n} |\mathbb{A}^{(k-1)}_{i,k}| > 0$, on calcule $\mathbb{H}_k = \begin{pmatrix} I_{k-1} & 0 \\ 0 & \widetilde{\mathbb{H}}_k \end{pmatrix}$, avec :
$\widetilde{\mathbb{H}}_k = \mathbb{H}\left(a_k + \|a_k\| \exp(i\alpha_k)\, \mathrm{e}_1\right) \in \mathbb{K}^{(n-k+1)\times(n-k+1)}$.
On calcule $\mathbb{A}^{(k)} := \mathbb{H}_k\, \mathbb{A}^{(k-1)} = \prod_{i=1}^{k} \mathbb{H}_i\, \mathbb{A}^{(0)}$, qui est telle que pour $j \in \{1, \cdots, k\}$ $\mathbb{A}^{(k)}_{i>j,j} = 0$.

D'après la proposition 5.30, $\mathbb{H}_k$ est hermitienne pour $1 \leq k \leq k_{max}$. On a alors $\mathbb{A} = \mathbb{Q}\mathbb{R}$, où $\mathbb{Q} = \mathbb{H}_1^* \cdots \mathbb{H}_{k_{max}}^* = \mathbb{H}_1 \cdots \mathbb{H}_{k_{max}}$ est une matrice unitaire et $\mathbb{R} = \mathbb{A}^{(k_{max})}$ est une matrice triangulaire supérieure.
L'algorithme de la factorisation de Householder s'écrit donc :

Initialisation : $\mathbb{R} = \mathbb{A}$, $\mathbb{Q} = I_n$
$k_{max} = m$ si $m < n$; $k_{max} = n - 1$ si $m = n$
Itérations : pour $k = 1, \ldots, k_{max}$**, faire**
$a_k = \mathbb{R}_{k:n,k} \in \mathbb{K}^{n-k+1}$,
si $\sum_{i=k+1}^{n} |(a_k)_i| = 0, \quad \mathbb{H}_k = I_n$
sinon,
calculer $\|a_k\|$ et α_k tel que $(a_k)_1 = (a_k, \mathrm{e}_1)_{n-k+1} = |(a_k)_1| \exp(i\alpha_k), \quad \mathrm{e}_1 \in \mathbb{K}^{n-k+1}$

calculer
$$\left\{ \begin{array}{rcl} v_k & = & a_k + \|a_k\| \exp(i\alpha_k)\, \mathrm{e}_1 \\ \\ \delta & := & \dfrac{v_k^* v_k}{2} = \|a_k\|(\|a_k\| + |(a_k)_1|). \end{array} \right. \tag{5.7}$$

calculer $\mathbb{H}_k = \begin{pmatrix} I_{k-1} & 0 \\ 0 & \mathbb{H}(v_k) \end{pmatrix}$ et $\mathbb{Q} = \mathbb{Q}\mathbb{H}_k^*$

Itérations : pour $l = k, \cdots, m$**, faire**
$b = (\mathbb{R})_{k:n,l}, \quad (\mathbb{R})_{k:n,l} = b - \dfrac{v_k^* b}{\delta} v_k$
fin pour l
fin si
fin pour k

Dans l'algorithme lorsque $k = 1$, on a $\mathbb{H}_1 = \mathbb{H}(v_1)$. Par abus de notation, ceci signifie que dans l'expression de $\mathbb{H}_1$ ci-dessus, la matrice I_0 ainsi que la colonne et la ligne composées de 0 disparaissent.
Les matrices $\mathbb{A}$, $\mathbb{Q}$, $\mathbb{R}$ sont liées par la relation $\mathbb{A} = \mathbb{Q}\mathbb{R}$, dans laquelle $\mathbb{Q} \in \mathbb{K}^{n\times n}$ est une matrice unitaire et $\mathbb{R} \in \mathbb{K}^{n\times m}$ est une matrice triangulaire supérieure.

Remarque 5.33 *On peut effectuer au préalable des permutations de lignes ou colonnes pour commencer l'algorithme avec une matrice $\mathbb{A}_{perm}$ telle que, pour $1 \leq j \leq p$ on a $(\mathbb{A}_{perm})_{i,j} = 0$ si $i > j$, et pour p maximal.*

5.14.3 Factorisation de Givens

Définition 5.34 *Soient $1 \leq i < j \leq n$. On appelle matrice de Givens associée aux paramètres (c, s) et aux indices (i, j) la matrice de la forme :*

$$\mathbb{G} = \begin{pmatrix} I_{i-1} & \cdots & \cdots & \cdots & 0 \\ 0 & \overline{c} & 0 & \overline{s} & \vdots \\ \vdots & 0 & I_{j-i-1} & 0 & \vdots \\ \vdots & -s & 0 & c & \vdots \\ 0 & \cdots & \cdots & \cdots & I_{n-j} \end{pmatrix} \in \mathbb{K}^{n,n}, \tag{5.8}$$

où c et s apparaissent aux intersections des lignes et colonnes d'indices i et j ; et sont tels que : $|c|^2 + |s|^2 = 1$.
Dans l'expression (5.8) lorsque $i = 1$, ou $j = i + 1$, ou $j = n$, la matrice I_0 ainsi que la colonne et la ligne composées de 0 en appui disparaissent.
Pour $i < j$ fixés, les éléments non nuls de $\mathbb{G}$ sont tels que : $\left\{ \begin{array}{rcl} \mathbb{G}_{k,k} & = & 1 \text{ pour } k \neq i, j \\ \mathbb{G}_{i,i} & = & \overline{\mathbb{G}}_{j,j} = \overline{c} \\ \mathbb{G}_{i,j} & = & -\overline{\mathbb{G}}_{j,i} = \overline{s} \end{array} \right.$.

Pour $\mathbb{K} = \mathbb{R}$, le produit $\mathbb{G}v$ représente la rotation d'angle $\arctan(s/c)$ du vecteur v dans le plan $(\mathrm{e}_i, \mathrm{e}_j)$.

Soit a_1 la première colonne de $\mathbb{A}$, telle que les deux derniers éléments non nuls soient d'indices i et j, $i < j$: $a_1 = ((a_1)_1, \cdots, (a_1)_i, 0, \cdots, (a_1)_j, 0, \cdots)^T$.
On construit la matrice de Givens $\mathbb{G}_{1,i}$ telle que : $\left\{ \begin{array}{rcl} r & = & (|(a_1)_i|^2 + |(a_1)_j|^2)^{1/2} \\ c & = & (a_1)_i / r \\ s & = & (a_1)_j / r \end{array} \right.$.
Le vecteur $\mathbb{G}_{1,i}a_1$ est tel que $\left\{ \begin{array}{rcll} (\mathbb{G}_{1,i}a_1)_k & = & (a_1)_k & \text{pour } k < i \\ (\mathbb{G}_{1,i}a_1)_i & = & r & \\ (\mathbb{G}_{1,i}a_1)_k & = & 0 & \text{pour } k > i \end{array} \right.$.

Pour obtenir une factorisation QR de la matrice $\mathbb{A}$, on répète cette procédure sur le vecteur $\mathbb{G}_{1,i}a_1$ jusqu'à obtenir un vecteur colinéaire à e_1. Soit $\prod_i \mathbb{G}_{1,i}$ le produit des matrices utilisées, on a $\left(\left(\prod_i \mathbb{G}_{1,i} \right) \mathbb{A} \right)_{i>1,1} = 0$. On applique de nouveau ces étapes au second vecteur

de la matrice $(\prod_i \mathbb{G}_{1,i})\,\mathbb{A}$. L'algorithme final est le suivant :

Initialisation : $\mathbb{R} = \mathbb{A}$, $\mathbb{Q}^* = I_n$
$k_{max} = m$ si $m < n$; $k_{max} = n-1$ si $m = n$
Itérations : pour $k = 1, \ldots, k_{max}$**, faire**
  $a_k = \mathbb{R}_{k:n,k} \in \mathbb{K}^{n-k+1}, \quad j = n-k+1$
  Tant que $j > k$, **trouver** i, $i < j$ **tel que**
  les deux derniers éléments non nuls soient d'indices i **et** j **:**
    $a_k = ((a_k)_1, \cdots, (a_k)_i, 0, \cdots, (a_k)_j, 0, \cdots)^T$,
    calculer :

$$\begin{cases} r &= (|(a_k)_i|^2 + |(a_k)_j|^2)^{1/2} \\ c &= (a_k)_i / r \\ s &= (a_k)_j / r \end{cases} \qquad \widetilde{\mathbb{G}} = \begin{pmatrix} I_{i-1} & 0 & \cdots & \cdots & 0 \\ 0 & \overline{c} & 0 & \overline{s} & \vdots \\ \vdots & 0 & I_{j-i-1} & 0 & \vdots \\ \vdots & -s & 0 & c & \vdots \\ 0 & \cdots & \cdots & \cdots & I_{n-k+1-j} \end{pmatrix} \tag{5.9}$$

    calculer $\mathbb{G} = \begin{pmatrix} I_{k-1} & 0 \\ 0 & \widetilde{\mathbb{G}} \end{pmatrix}$, $\mathbb{Q} = \mathbb{Q}\,\mathbb{G}^*$ et $\mathbb{R}_{i:n,:} = \mathbb{G}\,\mathbb{R}_{i:n,:}$
    poser $j = i$
  fin tant que
fin pour k

On a bien construit $\mathbb{Q}$ et $\mathbb{R}$ telles que $\mathbb{A} = \mathbb{Q}\mathbb{R}$, avec $\mathbb{Q} \in \mathbb{K}^{n\times n}$ unitaire et $\mathbb{R} \in \mathbb{K}^{n\times m}$ triangulaire supérieure. La remarque 5.33 s'applique également à cette méthode.

5.14.4 Factorisation de Gram-Schmidt

On suppose que $\mathbb{A} \in \mathbb{K}^{n\times n}$ est inversible. On note $(a_i)_{1\le i\le n}$ les vecteurs colonnes de $\mathbb{A}$. On peut construire une base orthonormale de $vect(a_1, \cdots, a_n)$ en utilisant l'algorithme d'orthonormalisation de Gram-Schmidt :

initialisation
  $\widetilde{q}_1 = a_1 \in \mathbb{K}^n, \quad \widetilde{q}_1 \neq 0$
  $q_1 = \widetilde{q}_1 / \|\widetilde{q}_1\|, \quad \mathbb{R} = 0$
itérations : pour $k = 2, \ldots, n$**, faire**
  $\widetilde{q}_k = a_k$
  pour $l = 1, \ldots, k-1$**, faire**
    $\mathbb{R}_{l,k} = (a_k, q_l)$
    $\widetilde{q}_k = \widetilde{q}_k - \mathbb{R}_{l,k}\, q_l,$
  fin pour l
  $\mathbb{R}_{k,k} = \|\widetilde{q}_k\|, \quad q_k = \widetilde{q}_k / \mathbb{R}_{k,k}.$
fin pour k

(5.10)

Remarque 5.35 *On a bien :* $(a_k, q_l) = \sum_{i=1}^{n} a_{i,k}\, \overline{q}_{i,l} = (\mathbb{Q}^*\mathbb{A})_{l,k} = \mathbb{R}_{l,k}.$

On peut améliorer la stabilité numérique des calculs avec *l'algorithme de Gram-Schmidt modifié* :

$$
\begin{array}{l}
\textbf{initialisation} \\
\quad \widetilde{q}_1 = a_1 \in \mathbb{K}^n, \quad \widetilde{q}_1 \neq 0 \\
\quad q_1 = \widetilde{q}_1 / \|\widetilde{q}_1\|, \quad \mathbb{R} = 0 \\
\textbf{itérations : pour } k = 2, \ldots, n\textbf{, faire} \\
\quad \widetilde{q}_k = a_k \\
\quad \textbf{pour } l = 1, \ldots, k-1\textbf{, faire} \\
\qquad \mathbb{R}_{l,k} = (\widetilde{q}_k, q_l) \\
\qquad \widetilde{q}_k = \widetilde{q}_k - \mathbb{R}_{l,k}\, q_l, \\
\quad \textbf{fin pour } l \\
\quad \mathbb{R}_{k,k} = \|\widetilde{q}_k\|, \quad q_k = \widetilde{q}_k / \mathbb{R}_{k,k}. \\
\textbf{fin pour } k
\end{array}
\tag{5.11}
$$

On obtient un algorithme équivalent à l'algorithme 5.10 du point de vue algébrique, et qui dans certains cas donne de meilleurs résultats numériques car les directions calculées respectent mieux les relations d'orthogonalité.

Dans les deux cas, on a :

— Pour $k \leq n$: $vect(q_1, \cdots, q_k) = vect(a_1, \cdots, a_k)$.

— Par construction $(q_i)_{i=1,n}$ est une base orthonormale : $\forall i, j \in \{1, \cdots, n\}$, $(q_i, q_j) = \delta_{i,j}$.

On a bien construit $\mathbb{Q}$ et $\mathbb{R}$ telles que $\mathbb{A} = \mathbb{Q}\mathbb{R}$, avec $\mathbb{Q} \in \mathbb{K}^{n\times n}$ unitaire et $\mathbb{R} \in \mathbb{K}^{n\times n}$ triangulaire supérieure.

5.15 Coûts calculs

Dans ce paragraphe, on indique le coût calcul des différents algorithmes pour une exécution séquentielle. Pour une méthode directe de résolution, il suffit de compter le nombre d'opérations de l'algorithme correspondant.

Proposition 5.36 *Soit $\mathbb{D} \in \mathbb{K}^{n\times n}$ une matrice diagonale inversible, et $b \in \mathbb{K}^n$ un vecteur. Le coût calcul de la résolution directe du système linéaire $\mathbb{D}y = b$ est de n opérations.*

Proposition 5.37 *Soit $\mathbb{T} \in \mathbb{K}^n \times \mathbb{K}^n$ une matrice triangulaire inversible, et $b \in \mathbb{K}^n$ un vecteur. Le coût calcul de la résolution directe du système linéaire $\mathbb{T}y = b$ est de n^2 opérations.*

Démonstration : On peut utiliser les formules de la Proposition 5.7, pour un système linéaire à matrice triangulaire, $\mathbb{T}y = b$. Pour un système linéaire à matrice triangulaire inférieure $\mathbb{L}y = b$, le calcul de la composante y_i requiert :

- pour $i = 1$: 1 division ;
- pour $i > 1$: 1 soustraction, $(i-1)$ multiplications, $(i-2)$ additions, 1 division.

Soit un total de $(2i-1)$ opérations élémentaires $(+, -, *, /)$. Le coût calcul des n composantes du vecteur y, est donc égal à

$$\sum_{i=1}^{n}(2i-1) = 2\frac{n(n+1)}{2} - n = n^2 \text{ opérations,}$$

ce qui reste "raisonnable" car il peut y avoir $(n^2+n)/2$ éléments non-nuls dans la matrice $\mathbb{L}$. On obtient le même coût pour un système linéaire à matrice triangulaire supérieure $\mathbb{U}x = y$. $\diamond$

Proposition 5.38 *Soit $\mathbb{A} \in \mathbb{K}^{n\times n}$ une matrice inversible, et $b \in \mathbb{K}^n$ un vecteur. Le nombre d'opérations élémentaires $(+,-,*,/)$ nécessaires au calcul de la solution du système linéaire $\mathbb{A}x = b$ par la méthode de Gauss est de l'ordre de $2n^3/3$.*

Démonstration : Traitons d'abord le coût de la factorisation de Gauss $\mathbb{A} = \mathbb{L}\mathbb{U}$ à l'aide du premier algorithme du §5.9. Calculons les n colonnes de $\mathbb{L}$. Soit donc k variant de 1 à n. Pour déterminer la colonne k de $\mathbb{L}$, il faut calculer un élément diagonal et $n-k$ éléments non-diagonaux. Or, $\mathbb{L}_{k,k} = 1$, et chaque élément $(\mathbb{L}_{i,k})_{i>k}$ requiert $2k-1$ opérations élémentaires $(+,-,*,/)$. Le nombre total d'opérations est donc égal à

$$\sum_{k=1}^{n}(2k-1)(n-k) = 2n\sum_{k=1}^{n}k - 2\sum_{k=1}^{n}k^2 + O(n^2) = n^3 - \frac{2}{3}n^3 + O(n^2) = \frac{1}{3}n^3 + O(n^2),$$

soit un nombre total d'opérations d'environ $n^3/3$ pour déterminer $\mathbb{L}$. Bien sûr, une étude de l'algorithme de calcul des lignes de $\mathbb{U}$ fournit un nombre total d'opérations lui aussi égal à

$$\frac{1}{3}n^3 + O(n^2).$$

Une fois que les matrices $\mathbb{L}$ et $\mathbb{U}$ sont calculées, la résolution du systéme linéaire peut s'effectuer par une descente-remontée, dont le coût est égal à $2n^2$ opérations. Le coût total de la résolution du système linéaire $\mathbb{A}x = b$ est donc de l'ordre de $2n^3/3$ opérations élémentaires $(+,-,*,/)$. $\diamond$

La partie la plus coûteuse est la factorisation de la matrice, puisque l'algorithme de descente-remontée nécessite $2n^2$ opérations ; par conséquent lorsqu'on a $O(1)$ systèmes linéaires à résoudre avec la même matrice (mais des seconds membres différents) ; le coût total reste de l'ordre de $2n^3/3$. On déduit de la Proposition 5.38 les résultats coût calcul des algorithmes de Crout et Cholesky.

Proposition 5.39 *Soit $\mathbb{A} \in \mathbb{K}^{n\times n}$ une matrice hermitienne inversible, et $b \in \mathbb{K}^n$ un vecteur. Le coût calcul de la résolution du système linéaire $\mathbb{A}y = b$ à l'aide de la méthode de Crout est de l'ordre de $n^3/3$ opérations.*

Proposition 5.40 *Soit $\mathbb{A} \in \mathbb{K}^{n\times n}$ une matrice hermitienne définie-positive. Le coût calcul de la résolution du système linéaire $\mathbb{A}y = b$ à l'aide de la méthode de Cholesky est de l'ordre de $n^3/3$.*

Dans certains cas, on peut réduire très significativement réduire le coût calcul de la résolution du système linéaire $\mathbb{A}y = b$, en prenant en compte le profil de la matrice $\mathbb{A}$. Cette observation est basée sur le résultat de conservation du profil, voir la proposition 5.26. Un exemple représentatif est donné par les matrices issues de la discrétisation du Laplacien, ou plus généralement de la diffusion, dans $]0,1[^d$ (d est la dimension spatiale) par différences finies, voir le chapitre 2. Notons $\mathbb{A}_d$ la matrice obtenue pour une discrétisation de pas $h = 1/(n+1)$ dans chaque direction, avec $n \geq 1$. Par construction, on a donc

$\mathbb{A}_d \in \mathbb{R}^{n_d \times n_d}$, où $n_d = n^d$.
Ci-dessous, on traite les cas $d = 1, 2$ en détail, avec les notations du §5.13 pour le profil.

Pour $\mathbb{A}_1 \in \mathbb{R}^{n_1 \times n_1}$ avec $n_1 = n$, on a

$$il(1) = 1, \text{ et } \forall i \geq 2,\ il(i) = i - 1.$$

Regardons l'algorithme de la factorisation de Cholesky colonne par colonne du §5.11. D'après la proposition 5.26, la matrice $\mathbb{L}_1$ possède le même profil que $\mathbb{A}_1$. Ainsi, on voit que les sommes sur j ne comprennent qu'un terme pour $k \geq 2$, à savoir celui pour $j = k - 1$, car sinon $(\mathbb{L}_1)_{k,j} = 0$ d'après la conservation du profil. De même, les boucles sur i ne comprennent qu'un indice, $i = k+1$, car sinon $(\mathbb{L}_1)_{i,k} = 0$. On en conclut que pour chaque valeur de k on effectue un nombre d'opérations indépendant de n_1. Et, comme k varie de 1 à n_1, le coût de la factorisation de Cholesky est finalement de l'ordre de $O(n_1)$.
Si on reprend maintenant l'algorithme de descente de la proposition 5.7, on constate que son coût est en $O(n_1)$. Et de même pour la remontée. Ainsi, pour $\mathbb{A}_1$, le coût calcul de résolution est comparable à celui d'un système linéaire avec une matrice diagonale, $O(n_1)$!

Pour $\mathbb{A}_2 \in \mathbb{R}^{n_2 \times n_2}$ avec $n_2 = n^2$, on a

$$il(1) = 1,\ \forall i \in \{2, \cdots, n\},\ il(i) = i - 1, \text{ et } \forall i \geq n + 1,\ il(i) = i - n.$$

Reprenons l'analyse ci-dessus en l'adaptant à ce profil. Dans l'algorithme de la factorisation de Cholesky colonne par colonne de $\mathbb{A}_2$, on voit que les sommes sur j ne comprennent que n termes au plus, avec $\max(0, k - n) \leq j < k$, car sinon $(\mathbb{L}_2)_{k,j} = 0$ d'après la conservation du profil. De même, les boucles sur i ne comprennent que n indices au plus, avec $k + 1 \leq i \leq \max(k + n, n_2)$, car sinon $(\mathbb{L}_2)_{i,k} = 0$. On en conclut cette fois que pour chaque valeur de k on effectue un nombre d'opérations de l'ordre de $n \times n$. Et, comme k varie de 1 à n_2, le coût de la factorisation de Cholesky est finalement de l'ordre de $O(n_2 \times n \times n) = O(n_2^2)$.
Si on reprend l'algorithme de descente de la proposition 5.7, on constate que son coût est en $O(n_2 \times n) = O(n_2^{3/2})$. Et de même pour la remontée. Le coût calcul de résolution n'est plus linéaire, mais il reste beaucoup plus favorable que l'estimation $O(n_2^3)$ de la proposition 5.40 !

Exercice 5.2 *En dimension $d \geq 3$, montrer que le coût de la factorisation de Cholesky de $\mathbb{A}_d \in \mathbb{R}^{n_d \times n_d}$ est de l'ordre de $O(n_d^{3-2/d})$, et que le coût de la résolution du système linéaire par descente-remontée est de l'ordre de $O(n_d^{2-1/d})$.*

5.16 Utilisation du calcul parallèle

On reprend le formalisme des §4.5 et §4.8, à savoir d'une part qu'on dispose d'une machine avec $\mathtt{P} + 1$ nœuds de calcul et d'autre part, qu'on va s'intéresser spécifiquement à la parallélisation de la méthode de descente-remontée (factorisation de Gauss) pour une matrice obtenue après discrétisation par éléments finis de Lagrange P_1. [40] Si on appelle

40. Pour une étude beaucoup plus systématique de la résolution directe d'un système linéaire en parallèle, nous renvoyons à [26], ouvrage dans lequel de très nombreuses solutions sont proposées.

$(M_i)_{1\leq i\leq N}$ les sommets internes du maillage, on obtient une matrice $\mathbb{A}\in\mathbb{R}^{N\times N}$ et un vecteur $b\in\mathbb{R}^N$ après discrétisation, et on doit résoudre $\mathbb{A}x=b$. En suivant §4.8, après découpage de l'ensemble des indices $\{1,2,\cdots,N\}$ en $\mathtt{P}+1$ sous-ensembles disjoints, on aboutit à une matrice du type

$$\mathbb{A}=\begin{pmatrix}[\mathbb{A}]_{1,1} & 0 & \cdots & 0 & [\mathbb{A}]_{1,\mathtt{P}+1}\\ 0 & [\mathbb{A}]_{2,2} & \ddots & \vdots & [\mathbb{A}]_{2,\mathtt{P}+1}\\ \vdots & \ddots & \ddots & 0 & \vdots\\ 0 & \cdots & 0 & [\mathbb{A}]_{\mathtt{P},\mathtt{P}} & [\mathbb{A}]_{\mathtt{P},\mathtt{P}+1}\\ [\mathbb{A}]_{\mathtt{P}+1,1} & [\mathbb{A}]_{\mathtt{P}+1,2} & \cdots & [\mathbb{A}]_{\mathtt{P}+1,\mathtt{P}} & [\mathbb{A}]_{\mathtt{P}+1,\mathtt{P}+1}\end{pmatrix}. \tag{5.12}$$

On rappelle que les blocs non-diagonaux d'indices $(\mathtt{p},\mathtt{q})$ pour $1\leq\mathtt{p}\neq\mathtt{q}\leq\mathtt{P}$ sont automatiquement nuls : on dit que le dernier sous-ensemble des indices (de numéro $\mathtt{P}+1$) réalise un **séparateur** de la matrice $\mathbb{A}$. Par ailleurs, les matrices diagonales $[\mathbb{A}]_{\mathtt{p},\mathtt{p}}$ sont inversibles pour tout $1\leq\mathtt{p}\leq\mathtt{P}+1$. Par rapport à la matrice (4.5), on a considéré ici un cas un peu plus "général" : on ne suppose plus nécessairement que la matrice est symétrique. Grâce au séparateur, si on reprend la factorisation de Gauss par blocs, on vérifie directement par un calcul similaire à celui de la démonstration de la proposition 5.23 (ou bien en invoquant la conservation du profil) que la factorisation $\mathbb{LU}$ de $\mathbb{A}$ produit les matrices

$$\mathbb{L}=\begin{pmatrix} I_{N_1} & 0 & \cdots & 0 & 0\\ 0 & I_{N_2} & \ddots & \vdots & 0\\ \vdots & \ddots & \ddots & 0 & \vdots\\ 0 & \cdots & 0 & I_{N_{\mathtt{P}}} & 0\\ [\mathbb{A}]_{\mathtt{P}+1,1}[\mathbb{A}]_{1,1}^{-1} & [\mathbb{A}]_{\mathtt{P}+1,2}[\mathbb{A}]_{2,2}^{-1} & \cdots & [\mathbb{A}]_{\mathtt{P}+1,\mathtt{P}}[\mathbb{A}]_{\mathtt{P},\mathtt{P}}^{-1} & I_{N_{\mathtt{P}+1}}\end{pmatrix} \tag{5.13}$$

et

$$\mathbb{U}=\begin{pmatrix}[\mathbb{A}]_{1,1} & 0 & \cdots & 0 & [\mathbb{A}]_{1,\mathtt{P}+1}\\ 0 & [\mathbb{A}]_{2,2} & \ddots & \vdots & [\mathbb{A}]_{2,\mathtt{P}+1}\\ \vdots & \ddots & \ddots & 0 & \vdots\\ 0 & \cdots & 0 & [\mathbb{A}]_{\mathtt{P},\mathtt{P}} & [\mathbb{A}]_{\mathtt{P},\mathtt{P}+1}\\ 0 & 0 & \cdots & 0 & \mathcal{S}\end{pmatrix}. \tag{5.14}$$

Le complément de Schur $\mathcal{S}$ est égal à

$$\mathcal{S}=[\mathbb{A}]_{\mathtt{P}+1,\mathtt{P}+1}-\sum_{\mathtt{p}=1,\mathtt{P}}[\mathbb{A}]_{\mathtt{P}+1,\mathtt{p}}[\mathbb{A}]_{\mathtt{p},\mathtt{p}}^{-1}[\mathbb{A}]_{\mathtt{p},\mathtt{P}+1}. \tag{5.15}$$

A partir de là, on peut résoudre le système linéaire $\mathbb{A}x=b$ à l'aide d'une descente-remontée par blocs (cf. proposition 5.7) : $\mathbb{L}y=b$, puis $\mathbb{U}x=y$. Du fait de l'existence du séparateur, l'algorithme se simplifie en

1. résoudre $\mathbb{L}y=b$
(i) **itérations : pour** $\mathtt{p}=1,\cdots,\mathtt{P}$ $[y]_{\mathtt{p}}=[b]_{\mathtt{p}}$, **fin boucle sur** $\mathtt{p}$
(ii) $[y]_{\mathtt{P}+1}=[b]_{\mathtt{P}+1}-\sum_{\mathtt{p}=1,\mathtt{P}}[\mathbb{A}]_{\mathtt{P}+1,\mathtt{p}}[\mathbb{A}]_{\mathtt{p},\mathtt{p}}^{-1}[b]_{\mathtt{p}}$
2. résoudre $\mathbb{U}x=y$
(iii) $\mathcal{S}[x]_{\mathtt{P}+1}=[y]_{\mathtt{P}+1}$
(iv) **itérations : pour** $\mathtt{p}=1,\cdots,\mathtt{P}$ $[\mathbb{A}]_{\mathtt{p},\mathtt{p}}[x]_{\mathtt{p}}=[y]_{\mathtt{p}}-[\mathbb{A}]_{\mathtt{p},\mathtt{P}+1}[x]_{\mathtt{P}+1}$, **fin boucle sur** $\mathtt{p}$

Pour rappel, pour $1 \leq \mathtt{p} \leq \mathtt{P}+1$ on affecte au nœud $\mathtt{p}$ les blocs $([\mathbb{A}]_{\mathtt{p},\mathtt{q}})_{\mathtt{q}=1,\mathtt{P}+1}$ ainsi que le bloc $[b]_{\mathtt{p}}$. On raisonne maintenant en termes de parallélisation sur les $\mathtt{P}$ premiers nœuds de la machine.

Il est clair que l'étape (i) est complètement parallèle !
Concernant l'étape (ii), on peut la découper en trois phases :

- résolution de systèmes linéaires simultanée pour $1 \leq \mathtt{p} \leq \mathtt{P}$: trouver $[z]_{\mathtt{p}}$ tel que $[\mathbb{A}]_{\mathtt{p},\mathtt{p}}[z]_{\mathtt{p}} = [b]_{\mathtt{p}}$;
- une phase de communications ($\mathtt{p} \to \mathtt{P}+1$) séquentielle où chaque nœud $\mathtt{p}$ ($1 \leq \mathtt{p} \leq \mathtt{P}$) envoie les données $[z]_{\mathtt{p}}$ au nœud $\mathtt{P}+1$;
- calcul sur le nœud $\mathtt{P}+1$: $[y]_{\mathtt{P}+1} = [b]_{\mathtt{P}+1} - \sum_{\mathtt{p}=1,\mathtt{P}} [\mathbb{A}]_{\mathtt{P}+1,\mathtt{p}}[z]_{\mathtt{p}}$.

L'étape (iv) est découpée en deux phases.

- une phase de communications ($\mathtt{P}+1 \to \mathtt{p}$) séquentielle où chaque nœud $\mathtt{p}$ ($1 \leq \mathtt{p} \leq \mathtt{P}$) reçoit les données $[x]_{\mathtt{P}+1}$ du nœud $\mathtt{P}+1$;
- résolution de systèmes linéaires simultanée pour $1 \leq \mathtt{p} \leq \mathtt{P}$: trouver $[x]_{\mathtt{p}}$ tel que $[\mathbb{A}]_{\mathtt{p},\mathtt{p}}[x]_{\mathtt{p}} = [y]_{\mathtt{p}} - [\mathbb{A}]_{\mathtt{p},\mathtt{P}+1}[x]_{\mathtt{P}+1}$.

L'étape (iii) est la plus complexe à analyser. Cette complexité provient du calcul du complément de Schur $\mathcal{S}$. Si on examine la formule (5.15) on note qu'on doit connaître les produits matriciels $[\mathbb{A}]^{-1}_{\mathtt{p},\mathtt{p}}[\mathbb{A}]_{\mathtt{p},\mathtt{P}+1}$ pour $1 \leq \mathtt{p} \leq \mathtt{P}$ pour pouvoir calculer $\mathcal{S}$. Or, si on revient à la modélisation et au vocabulaire du §4.8, ceci demande de résoudre des systèmes linéaires du type $[\mathbb{A}]_{\mathtt{p},\mathtt{p}}[z]^k_{\mathtt{p}} = [b]^k_{\mathtt{p}}$, où $([b]^k_{\mathtt{p}})_{k=1,n_{\mathtt{p}}}$ sont les colonnes de $[\mathbb{A}]_{\mathtt{p},\mathtt{P}+1}$ correspondant aux $n_{\mathtt{p}}$ sommets d'interface situés sur la frontière $\partial\Omega_{\mathtt{p}}$. Si on concatène les résultats $([z]^k_{\mathtt{p}})_{k=1,n_{\mathtt{p}}}$ dans la matrice $[\mathbb{Z}]_{\mathtt{p},\mathtt{P}+1}$, le complément de Schur est égal à

$$\mathcal{S} = [\mathbb{A}]_{\mathtt{P}+1,\mathtt{P}+1} - \sum_{\mathtt{p}=1,\mathtt{P}} [\mathbb{A}]_{\mathtt{P}+1,\mathtt{p}}[\mathbb{Z}]_{\mathtt{p},\mathtt{P}+1}.$$

Soit $n_{\max} = \max_{1\leq \mathtt{p}\leq \mathtt{P}} n_{\mathtt{p}}$. On note que, si on connaît $([b]^k_{\mathtt{p}})_{1\leq \mathtt{p}\leq \mathtt{P},k=1,n_{\mathtt{p}}}$ alors, pour chaque valeur de $k \leq n_{\max}$, on peut réaliser les calculs en parallèle, comme aux étapes (ii) et (iv). On commence donc par une phase de communications ($\mathtt{P}+1 \to \mathtt{p}$) séquentielle où chaque nœud $\mathtt{p}$ ($1 \leq \mathtt{p} \leq \mathtt{P}$) reçoit les données $[\mathbb{A}]_{\mathtt{P}+1,\mathtt{p}}$ du nœud $\mathtt{P}+1$. A partir de là, on peut calculer $[\mathbb{A}]_{\mathtt{P}+1,\mathtt{p}}[\mathbb{Z}]_{\mathtt{p},\mathtt{P}+1}$ en $n_{\mathtt{p}}$ calculs sur le nœud $\mathtt{p}$. Ces calculs peuvent être réalisés en parallèle !
Toutefois, il y a trois difficultés majeures. Tout d'abord, pour un nœud $\mathtt{p}$, on peut avoir $n_{\mathtt{p}} < n_{\max}$, auquel cas le parallélisme se détériore pour les étapes $k > \mathtt{p}$ puisqu'il n'y a plus de calculs à réaliser sur le nœud $\mathtt{p}$, et la détérioration est d'autant plus conséquente que le nombre de nœuds concernés augmente. Par ailleurs le nombre total d'étapes de calculs (en parallèle), égal à $n_{\max}$, peut être très grand, c'est-à-dire qu'au final le coût de l'étape (iii) domine très largement le coût total des étapes (i), (ii) et (iv). Enfin, il reste à résoudre le système linéaire avec la matrice $\mathcal{S}$...

Pour résoudre ces difficultés, une première approche est de demander au logiciel de génération de maillage (voir la remarque 4.2) de respecter un critère additionel, à savoir l'obtention d'une partition du maillage avec des nombres $n_{\mathtt{p}}$ égaux, ou très proches afin de maximiser le parallélisme. Mais ceci ne règle pas la difficulté liée à la valeur de $n_{\max}$, sauf si

$n_{\max}$ est petit, ou si les systèmes linéaires de matrice $[\mathbb{A}]_{\mathtt{p},\mathtt{p}}$ sont peu coûteux à résoudre ? ! Et il faut encore calculer $[x]_{\mathtt{P}+1}$ solution de $\mathcal{S}[x]_{\mathtt{P}+1} = [y]_{\mathtt{P}+1}$.

C'est pourquoi on privilégie une seconde approche, à savoir la détermination d'une "bonne approximation" de $\mathcal{S}^{-1}$. Ceci est justifié par le fait que le système linéaire se réécrit de manière équivalente $[x]_{\mathtt{P}+1} = \mathcal{S}^{-1}[y]_{\mathtt{P}+1}$, et dans ce cas on utilisera une méthode itérative pour résoudre le système. Cette branche de l'algèbre linéaire numérique est très vaste, on renvoie par exemple à [32, 17] pour la conception de ces approximations. Voir aussi la partie III. Les méthodes itératives seront quant à elles étudiées aux chapitres 6 et 7.

Chapitre 6

Les méthodes itératives

6.1 Critère d'arrêt, convergence et coût calcul

A la notion de calcul à ε près (décrite en (4.2)) correspond, par dualité, celle de la précision requise, ce qui permet de déterminer un **critère d'arrêt** pour une méthode de résolution itérative. En effet, pour $\varepsilon > 0$ et x^0 donnés, on va effectuer des itérations,

$$\left\| \begin{array}{l} \textbf{initialisation} \\ \quad x^0 \in \mathbb{K}^n \\ \textbf{itérations : pour } k = 0, 1, \ldots, \textbf{ faire} \\ \quad x^{k+1} \leftarrow x^k \\ \textbf{tant que } \|\mathbb{A}x^k - b\| > \varepsilon \, \|\mathbb{A}x^0 - b\| \end{array} \right. \tag{6.1}$$

En d'autres termes, les itérations stoppent à la plus petite valeur de k telle que $\|Ax^k - b\| \leq \varepsilon \, \|Ax^0 - b\|$. On ajoute la norme $\|\mathbb{A}x^0 - b\|$ à droite pour prendre en compte la précision initiale ($k = 0$) du calcul.
A partir de là, la voie est libre pour évaluer le **coût calcul** d'une méthode itérative.
La première quantité est le **nombre d'itérations** nécessaire à la validation du critère d'arrêt. Intuitivement, on aura tendance à privilégier une méthode nécessitant peu d'itérations. C'est effectivement un critère, mais ça n'est pas le seul. Baser une analyse de la qualité d'une méthode itérative sur le nombre d'itérations uniquement est *incorrect*. Un second critère, complémentaire du premier, est le **coût d'une itération**. Typiquement, il s'agit du nombre d'opérations nécessaires à la réalisation d'une itération, c'est-à-dire au calcul de x^{k+1}, connaissant x^k. A partir de ces deux critères, on obtient une idée du **coût calcul** en multipliant le nombre d'itérations par le coût d'une itération.

6.2 Décomposition régulière

Les méthodes itératives sont associées à la notion de décomposition régulière d'une matrice, qui nécessite quelques rappels sur l'analyse numérique matricielle. Nous nous sommes inspirés de [13] pour la démonstration de certains résultats.

Définition 6.1 *on appelle* **décomposition régulière** *d'une matrice* $\mathbb{A} \in \mathbb{K}^{n \times n}$ *la donnée de deux matrices* $\mathbb{M}, \mathbb{N} \in \mathbb{K}^{n \times n}$ *telles que*

(i) $\mathbb{A} = \mathbb{M} - \mathbb{N}$ *;*
(ii) $\mathbb{M}$ *est inversible.*

On associe à toute décomposition régulière la méthode itérative

$$\left\| \begin{array}{l} \textbf{initialisation} \\ \quad x^0 \in \mathbb{K}^n \\ \textbf{itérations : pour } k = 0, 1, \ldots, \textbf{faire} \\ \quad \mathbb{M}x^{k+1} = \mathbb{N}x^k + b \\ \textbf{fin} \end{array} \right. \tag{6.2}$$

On voit que si cette méthode converge vers un vecteur x, celui–ci vérifie néssairement la relation

$$\mathbb{M}x = \mathbb{N}x + b \quad \text{ soit encore } \quad \mathbb{A}x = b.$$

Si on appelle $r^k = b - \mathbb{A}x^k$ le **vecteur résidu**, cette méthode peut aussi s'écrire sous la forme

$$\left\| \begin{array}{l} \textbf{initialisation} \\ \quad x^0 \in \mathbb{K}^n \\ \textbf{itérations : pour } k = 0, 1, \ldots, \textbf{faire} \\ \quad x^{k+1} = x^k + \mathbb{M}^{-1} r^k \\ \textbf{fin} \end{array} \right. \tag{6.3}$$

On note si que $\mathbb{M} = \mathbb{A}$ la méthode itérative converge en une seule itération quel que soit le vecteur initial : $x^1 = x^0 + \mathbb{A}^{-1}(b - \mathbb{A}x^0) = \mathbb{A}^{-1}b$, mais il s'agit là d'une méthode directe qui nécessite la résolution du système linéaire $\mathbb{A}z = r$. Dans la pratique on recherche donc des matrices $\mathbb{M}$ pour lesquelles cette résolution n'est pas trop coûteuse. L'objet de ce chapitre est l'étude générale de ce type de méthode et de leur convergence vers x solution du système linéaire suivant le choix de la matrice $\mathbb{M}$.

On introduit le **vecteur erreur** $e^k = x - x^k$, alors

$$\left.\begin{array}{c} \mathbb{M}x = \mathbb{N}x + b \\ \mathbb{M}x^{k+1} = \mathbb{N}x^k + b \end{array}\right\} \Longrightarrow e^{k+1} = \mathbb{M}^{-1}\mathbb{N}e^k.$$

Afin de déterminer un critère de convergence, on utilise la notion de rayon spectral.

Proposition 6.2 *La méthode itérative converge si et seulement si*

$$\rho(\mathbb{M}^{-1}\mathbb{N}) < 1.$$

Démonstration : Par construction le vecteur e^k vérifie $e^{k+1} = \mathbb{M}^{-1}\mathbb{N}e^k = \left[\mathbb{M}^{-1}\mathbb{N}\right]^{k+1} e^0$. Une condition nécessaire et suffisante pour que e^{k+1} tende vers 0 quel que soit le vecteur initial x^0 est que $\rho(\mathbb{M}^{-1}\mathbb{N}) < 1$, d'après le Théorème B.18. ⋄

Ainsi, pour définir une méthode itérative convergente il faut donc choisir la matrice $\mathbb{M}$ de façon que

(i) $\rho(\mathbb{M}^{-1}\mathbb{N}) < 1$,

(ii) la résolution du système linéaire $\mathbb{M}x^{k+1} = \mathbb{N}x^k + b$ ne soit pas trop coûteuse car il faudra l'effectuer à chaque itération !

Nous énonçons un résultat général garantissant la convergence.

Théorème 6.3 *[Householder–John] Soient $\mathbb{A} \in \mathbb{C}^{n\times n}$ une matrice hermitienne définie-positive et une décomposition régulière $\mathbb{A} = \mathbb{M} - \mathbb{N}$. Si la matrice hermitienne $\mathbb{M}^* + \mathbb{N}$ est définie-positive alors*

$$\rho(\mathbb{M}^{-1}\mathbb{N}) < 1.$$

Démonstration : Puisque $\mathbb{A}$ est hermitienne, par construction $\mathbb{M}^* + \mathbb{N} = \mathbb{M}^* + \mathbb{M} - \mathbb{A}$ est aussi hermitienne. On va utiliser la Proposition B.12, qui permet d'affirmer que $\rho(\mathbb{M}^{-1}\mathbb{N}) \leq \|\mathbb{M}^{-1}\mathbb{N}\|$ pour toute norme matricielle $\|\cdot\|$. On choisit la norme vectorielle induite par la norme $\|v\|_{\mathbb{A}} = \sqrt{(\mathbb{A}v, v)}$ qui est bien une norme vectorielle puisque $\mathbb{A}$ est hermitienne définie-positive. On étudie alors la norme matricielle

$$\|\mathbb{M}^{-1}\mathbb{N}\|_{\mathbb{A}} = \max_{v\neq 0} \frac{\|\mathbb{M}^{-1}\mathbb{N}v\|_{\mathbb{A}}}{\|v\|_{\mathbb{A}}} = \max_{v\in S_{\mathbb{A}}} \|v - \mathbb{M}^{-1}\mathbb{A}v\|_{\mathbb{A}},$$

avec $S_{\mathbb{A}} := \{v \in \mathbb{C}^n \ : \ \|v\|_{\mathbb{A}} = 1\}$. Soit maintenant $v \in S_{\mathbb{A}}$ et $w = \mathbb{M}^{-1}\mathbb{A}v$; alors

$$\begin{array}{ll}
\|\mathbb{M}^{-1}\mathbb{N}v\|_{\mathbb{A}}^2 & = \|v - \mathbb{M}^{-1}\mathbb{A}v\|_{\mathbb{A}}^2 = \|v - w\|_{\mathbb{A}}^2 \\
& = (\mathbb{A}(v-w), v-w) \\
& = (\mathbb{A}v, v) - (\mathbb{A}w, v) - (\mathbb{A}v, w) + (\mathbb{A}w, w) \\
(\mathbb{A} \text{ hermitienne}) & = 1 - (w, \mathbb{A}v) - (\mathbb{M}w, w) + (\mathbb{A}w, w) \\
& = 1 - (w, \mathbb{M}w) - ((\mathbb{A} - \mathbb{M})w, w) \\
& = 1 - (\mathbb{M}^*w, w) - (\mathbb{N}w, w) \\
& = 1 - ((\mathbb{M}^* + \mathbb{N})w, w) \\
& \leq 1 - \lambda_{min}(\mathbb{M}^* + \mathbb{N})\|w\|_2^2.
\end{array}$$

On rappelle que par définition $\|\cdot\|_2^2 = (\cdot,\cdot)$. Soit la fonction continue de $S_{\mathbb{A}}$ dans $\mathbb{R}$, $v \mapsto \|\mathbb{M}^{-1}\mathbb{A}v\|_2$. Comme $S_{\mathbb{A}}$ est un compact de $\mathbb{C}^n$, la fonction atteint son minimum : soit $v_{min} \in S_{\mathbb{A}}$ un point de minimum et C_{min} la valeur minimum. Comme $v_{min} \neq 0$, et comme $\mathbb{M}^{-1}\mathbb{A}$ est inversible, on en déduit que $C_{min} > 0$. Par ailleurs $\lambda_{min}(\mathbb{M}^* + \mathbb{N}) > 0$ car la matrice $\mathbb{M}^* + \mathbb{N}$ est (hermitienne) définie-positive par hypothèse. A l'aide de la Proposition B.12, on conclut que

$$\rho(\mathbb{M}^{-1}\mathbb{N}) \leq \|\mathbb{M}^{-1}\mathbb{N}\|_{\mathbb{A}} \leq 1 - (C_{min})^2\lambda_{min}(\mathbb{M}^* + \mathbb{N}) < 1.$$

◇

Il est raisonnable de rechercher la meilleure convergence possible : pour cela il faut savoir comparer les vitesses de convergence de deux méthodes itératives associées aux décompositions régulières $\mathbb{A} = \mathbb{M}_1 - \mathbb{N}_1 = \mathbb{M}_2 - \mathbb{N}_2$. Un bon indicateur est le rayon spectral : si $\rho(\mathbb{M}_1^{-1}\mathbb{N}_1) < \rho(\mathbb{M}_2^{-1}\mathbb{N}_2) < 1$, alors on s'attend à ce que la première méthode converge plus vite que la seconde (voir §6.4).

Définition 6.4 *La vitesse de convergence d'une méthode itérative est la quantité*

$$R(\mathbb{M}^{-1}\mathbb{N}) = -\log \rho(\mathbb{M}^{-1}\mathbb{N}).$$

On a : $R(\mathbb{M}^{-1}\mathbb{N}) > 0 \iff \rho(\mathbb{M}^{-1}\mathbb{N}) < 1$. En outre, la vitesse est d'autant plus grande que le rayon spectral de la matrice est petit.

6.3 Itérations par points – Itérations par blocs

On présente maintenant quelques décompositions régulières classiques, en écrivant la matrice $\mathbb{A}$ sous la forme $\mathbb{A} = \mathbb{D} - \mathbb{E} - \mathbb{F}$ où $\mathbb{D}, \mathbb{E}, \mathbb{F} \in \mathbb{K}^{n\times n}$ sont les matrices définies par

$$\begin{array}{l} \mathbb{D}_{i,j} = \mathbb{A}_{i,i} \text{ si } i = j \text{ et } \mathbb{D}_{i,j} = 0 \text{ si } i \neq j \\ \mathbb{E}_{i,j} = -\mathbb{A}_{i,j} \text{ si } i > j \text{ et } \mathbb{E}_{i,j} = 0 \text{ si } i \leq j \\ \mathbb{F}_{i,j} = -\mathbb{A}_{i,j} \text{ si } i < j \text{ et } \mathbb{F}_{i,j} = 0 \text{ si } i \geq j. \end{array} \tag{6.4}$$

$\mathbb{D}$ est une matrice diagonale,
$\mathbb{E}$ est une matrice triangulaire inférieure stricte,
$\mathbb{F}$ une matrice triangulaire supérieure stricte.

Cette écriture est dite **par points** puisque les indices i et j varient de 1 à n, elle se généralise à l'écriture **par blocs** en utilisant un découpage (ou partition) par blocs de la matrice $\mathbb{A}$ en P^2 blocs (cf. §5.3) : on écrit $\mathbb{A} = \mathbb{D}_B - \mathbb{E}_B - \mathbb{F}_B$ où on a ajouté des indices $_B$ pour rappeler qu'il s'agit de blocs.

$$\begin{array}{ll} [\mathbb{D}_B]_{p,q} & = [\mathbb{A}]_{p,q} \text{ si } p = q \text{ et } [\mathbb{D}_B]_{p,q} = [0] \text{ si } p \neq q \\ [\mathbb{E}_B]_{p,q} & = -[\mathbb{A}]_{p,q} \text{ si } p > q \text{ et } [\mathbb{E}_B]_{p,q} = [0] \text{ si } p \leq q \\ [\mathbb{F}_B]_{p,q} & = -[\mathbb{A}]_{p,q} \text{ si } p < q \text{ et } [\mathbb{F}_B]_{p,q} = [0] \text{ si } p \geq q, \end{array} \tag{6.5}$$

où cette fois les indices p et q varient de 1 à P nombre de blocs de la partition. Dans ce formalisme les blocs extra-diagonaux peuvent être rectangulaires, mais les *blocs diagonaux sont nécessairement carrés.*

Remarque 6.5 *Une méthode par points peut être interprétée comme une méthode par blocs dans lequel chaque bloc est réduit à un seul élément de la matrice (choisir $P = n$).*

6.4 Critère de convergence

Au paragraphe 4.1, on a introduit la notion de convergence d'un algorithme à la précision ε après k itérations (avec $\varepsilon > 0$), par la majoration :

$$\|r^k\| \leq \varepsilon \|r^0\|, \tag{6.6}$$

est atteinte pour une norme vectorielle $\|\cdot\|$ à déterminer. D'après la relation $e^k = (\mathbb{M}^{-1}\mathbb{N})^k e^0$ et la Proposition B.13, on déduit qu'il existe une norme $\|\cdot\|$ qui en pratique ne dépend que de la matrice $\mathbb{A}$ (et pas du second membre b) telle que

$$\|e^k\| \leq (\rho(\mathbb{M}^{-1}\mathbb{N}))^k \|e^0\|, \text{ soit } \|\mathbb{A}^{-1}r^k\| \leq (\rho(\mathbb{M}^{-1}\mathbb{N}))^k \|\mathbb{A}^{-1}r^0\|,$$

puisque $r^k = b - \mathbb{A}x^k = -\mathbb{A}e^k$. Si on introduit le **conditionnement**, ou **nombre de conditionnement** de $\mathbb{A}$, dans la norme $\|\cdot\|$, défini par :

$$\kappa(\mathbb{A}) = \|\mathbb{A}^{-1}\| \, \|\mathbb{A}\|, \tag{6.7}$$

on en déduit

$$\|r^k\| = \|\mathbb{A}\mathbb{A}^{-1}r^k\| \leq \|\mathbb{A}\|(\rho(\mathbb{M}^{-1}\mathbb{N}))^k \|\mathbb{A}^{-1}r^0\| \leq \kappa(\mathbb{A})(\rho(\mathbb{M}^{-1}\mathbb{N}))^k \|r^0\|. \tag{6.8}$$

En comparant (6.6) et (6.8), on *estime* le nombre d'itérations K nécessaires pour vérifier le critère de convergence par la formule

$$\varepsilon = \kappa(\mathbb{A})(\rho(\mathbb{M}^{-1}\mathbb{N}))^K, \text{ soit } K = \frac{\log(\kappa(\mathbb{A})/\varepsilon)}{R(\mathbb{M}^{-1}\mathbb{N})}. \tag{6.9}$$

On peut aller plus loin et prouver qu'asymptotiquement la norme $\|e^k\|$ se comporte "au pire" comme $(\rho(\mathbb{M}^{-1}\mathbb{N}))^k$.

Proposition 6.6 *Soit $\mathbb{B} \in \mathbb{K}^{n\times n}$, $c \in \mathbb{K}^n$ et $u \in \mathbb{K}^n$ tels que*

$$u = \mathbb{B}u + c.$$

On considère la méthode itérative

$$\left\| \begin{array}{l} \textbf{initialisation} \\ \quad u^0 \in \mathbb{K}^n \\ \textbf{itérations : pour } k = 0, 1, \ldots, \textbf{faire} \\ \quad u^{k+1} = \mathbb{B}u^k + c \\ \textbf{fin} \end{array} \right. \tag{6.10}$$

Alors $\lim_{k\to+\infty}\{\sup_{\|u^0-u\|=1} \|u^k - u\|^{1/k}\} = \rho(\mathbb{B})$.

Remarque 6.7 *Si on choisit $\mathbb{B} = \mathbb{M}^{-1}\mathbb{N}$ et $c = \mathbb{M}^{-1}b$, on en déduit*

$$\lim_{k\to+\infty} \{ \sup_{\|e^0\|=1} \|e^k\|^{1/k}\} = \rho(\mathbb{M}^{-1}\mathbb{N}).$$

Démonstration : Pour tout u^0 et $k \geq 0$, on a par récurrence $u^k - u = \mathbb{B}^k(u^0 - u)$. D'où

$$\sup_{\|u^0-u\|=1} \|u^k - u\|^{1/k} = \sup_{\|u^0-u\|=1} \|\mathbb{B}^k(u^0 - u)\|^{1/k} = \{ \sup_{\|u^0-u\|=1} \|\mathbb{B}^k(u^0 - u)\|\}^{1/k} = \|\mathbb{B}^k\|^{1/k},$$

avec $\|\cdot\|$ la norme matricielle induite. Le résultat est alors une conséquence du Corollaire B.20, car $\lim_{k\to+\infty} \|\mathbb{B}^k\|^{1/k} = \rho(\mathbb{B})$. ◇

6.5 Méthode de Jacobi par points

C'est la méthode itérative la plus ancienne. Elle correspond au choix le "plus simple" de $\mathbb{M}$, à savoir $\mathbb{M} = \mathbb{D}$ et $\mathbb{N} = \mathbb{E} + \mathbb{F}$. L'algorithme itératif de Jacobi s'écrit

$$\left\| \begin{array}{l} \textbf{initialisation} \\ \quad x^0 \in \mathbb{K}^n \\ \textbf{itérations : pour } k = 0, 1, \ldots, \textbf{faire} \\ \quad \mathbb{D}x^{k+1} = (\mathbb{E} + \mathbb{F})x^k + b \\ \textbf{fin} \end{array} \right. \tag{6.11}$$

En effet, à chaque itération on doit résoudre un système linéaire avec une *matrice diagonale*. Au sein d'une itération, l'algorithme s'écrit composante par composante

$$\left\| \begin{array}{l} \textbf{pour } i = 1, 2, \ldots, n \textbf{ faire} \\ \quad \mathbb{A}_{i,i}x_i^{k+1} = b_i - \displaystyle\sum_{j\neq i} \mathbb{A}_{i,j}x_j^k \\ \textbf{fin} \end{array} \right. \tag{6.12}$$

La matrice d'itération associée est notée

$$\mathbb{J} = \mathbb{D}^{-1}(\mathbb{E} + \mathbb{F}).$$

6.6 Méthode de Gauss-Seidel par points

Dans la formule précédente, on peut prendre en compte les nouvelles valeurs des composantes de x^{k+1} au fur et à mesure de leur calcul, en commençant par la première ($i = 1$), puis la deuxième ($i = 2$), etc. On obtient alors l'algorithme

$$\left\| \begin{array}{l} \textbf{pour } i = 1, 2, \ldots, n \textbf{ faire} \\ \quad \mathbb{A}_{i,i} x_i^{k+1} = b_i - \sum_{j<i} \mathbb{A}_{i,j} x_j^{k+1} - \sum_{j>i} \mathbb{A}_{i,j} x_j^k \\ \textbf{fin} \end{array} \right. \tag{6.13}$$

Cette méthode correspond au choix $\mathbb{M} = \mathbb{D} - \mathbb{E}$ et $\mathbb{N} = \mathbb{F}$. L'algorithme itératif de Gauss-Seidel s'écrit

$$\left\| \begin{array}{l} \textbf{initialisation} \\ \quad x^0 \in \mathbb{K}^n \\ \textbf{itérations : pour } k = 0, 1, \ldots, \textbf{faire} \\ \quad (\mathbb{D} - \mathbb{E}) x^{k+1} = \mathbb{F} x^k + b \\ \textbf{fin} \end{array} \right. \tag{6.14}$$

La matrice d'itération associée est notée

$$\mathbb{G} = (\mathbb{D} - \mathbb{E})^{-1}\mathbb{F}.$$

Dans ce cas, l'*ordre de numérotation* des inconnues à une influence sur l'algorithme, contrairement à l'algorithme de la méthode de Jacobi.

6.7 Méthode de relaxation par points

On introduit un paramètre réel $\omega \neq 0$ et on écrit que chaque composante x_i^{k+1} est une combinaison de la valeur connue x_i^k et de celle fournie par la méthode de Gauss-Seidel $\tilde{x}_i^{k+1}$:

$$\begin{aligned} x_i^{k+1} &= (1-\omega) x_i^k + \omega \tilde{x}_i^{k+1} \\ \Longrightarrow \mathbb{A}_{i,i} x_i^{k+1} &= (1-\omega)\mathbb{A}_{i,i} x_i^k + \omega \left(b_i - \sum_{j<i} \mathbb{A}_{i,j} x_j^{k+1} - \sum_{j>i} \mathbb{A}_{i,j} x_j^k \right). \end{aligned}$$

Ce qui revient à prendre

$$\mathbb{M} = \frac{1}{\omega}(\mathbb{D} - \omega\mathbb{E}) \quad \text{et} \quad \mathbb{N} = \frac{1}{\omega}\left(\omega\mathbb{F} + (1-\omega)\mathbb{D}\right).$$

L'algorithme itératif s'écrit alors

$$\left\| \begin{array}{l} \textbf{initialisation} \\ \quad x^0 \in \mathbb{K}^n \\ \textbf{itérations : pour } k = 0, 1, \ldots, \textbf{faire} \\ \quad (\mathbb{D} - \omega\mathbb{E}) x^{k+1} = \left(\omega\mathbb{F} + (1-\omega)\mathbb{D}\right) x^k + \omega b \\ \textbf{fin} \end{array} \right. \tag{6.15}$$

La matrice d'itération est notée

$$\mathbb{L}_\omega = (\mathbb{D} - \omega\mathbb{E})^{-1}\left((1-\omega)\mathbb{D} + \omega\mathbb{F}\right).$$

On remarque que pour $\omega = 1$ on retrouve bien la méthode de Gauss-Seidel, i. e. $\mathbb{L}_1 = \mathbb{G}$.

Pour cette matrice, on peut écrire les relations

$$\det\left(\mathbb{L}_\omega\right) = \det\left((\mathbb{D})^{-1}\right)(1-\omega)^n(\mathbb{D}) = (1-\omega)^n,$$

car les matrices $\mathbb{E}$ et $\mathbb{F}$ sont triangulaires à diagonale nulle, ainsi que

$$|\det\left(\mathbb{L}_\omega\right)| = \prod_i |\lambda_i(\mathbb{L}_\omega)| \leq \rho(\mathbb{L}_\omega)^n.$$

On obtient finalement l'encadrement

$$|1-\omega| \leq |\det\left(\mathbb{L}_\omega\right)|^{1/n} \leq \rho(\mathbb{L}_\omega).$$

Ainsi la méthode diverge pour $\omega \notin]0,2[$. La méthode est dite

- de sous–relaxation quand $0 < \omega < 1$;
- de Gauss-Seidel quand $\omega = 1$;
- de sur–relaxation quand $1 < \omega < 2$.

En général on prend $\omega \in]1,2[$ et la méthode s'appelle en anglais Successive Over Relaxation (S.O.R.). On peut démontrer le théorème :

Théorème 6.8 *[Ostrowski–Reich] Soit $\mathbb{A} \in \mathbb{C}^{n\times n}$ une matrice hermitienne définie-positive, la méthode de relaxation converge si et seulement si $\omega \in]0,2[$.*

Démonstration : La condition est nécessaire puisque pour $\omega \notin]0,2[$ la méthode diverge. Pour démontrer que la condition est suffisante, on note que la matrice $\mathbb{M}^* + \mathbb{N} = \dfrac{2-\omega}{\omega}\,\mathbb{D}$ est hermitienne définie-positive quand $\omega \in]0,2[$, puisque $\mathbb{A}_{i,i} > 0$ pour tout $i = 1, \cdots, n$. Le résultat s'en déduit par application du Théorème 6.3. ◇

Remarque 6.9 *Ce résultat est encore valable si la matrice $\mathbb{A}$ n'est plus hermitienne, mais si $\mathbb{A} + \mathbb{A}^*$ reste définie-positive [30].*

6.8 Méthodes par blocs

Nous avons proposés plusieurs méthodes itératives basées sur une écriture par points de la matrice $\mathbb{A}$. Nous reprenons maintenant ces méthodes en supposant maintenant que $\mathbb{A}$ est découpée en P^2 blocs, cf. (6.5).

— Méthode de Jacobi par blocs :

$$\left\|\begin{array}{l} \textbf{initialisation} \\ \quad x^0 \in \mathbb{K}^n \\ \textbf{itérations : pour } k = 0, 1, \ldots, \textbf{faire} \\ \quad \mathbb{D}_B x^{k+1} = (\mathbb{E}_B + \mathbb{F}_B)x^k + b \\ \textbf{fin} \end{array}\right. \tag{6.16}$$

A chaque itération on doit résoudre un système linéaire avec une *matrice diagonale par blocs*. La matrice d'itération associée est notée

$$\mathbb{J}_B = (\mathbb{D}_B)^{-1}(\mathbb{E}_B + \mathbb{F}_B).$$

— Méthode de Gauss-Seidel par blocs :

$$\left\|\begin{array}{l}\textbf{initialisation}\\ \quad x^0 \in \mathbb{K}^n\\ \textbf{itérations : pour } k = 0, 1, \ldots, \textbf{faire}\\ \quad (\mathbb{D}_B - \mathbb{E}_B)x^{k+1} = \mathbb{F}_B x^k + b\\ \textbf{fin}\end{array}\right. \tag{6.17}$$

A chaque itération on doit résoudre un système linéaire avec une *matrice triangulaire inférieure par blocs*. La matrice d'itération associée est notée

$$\mathbb{G}_B = (\mathbb{D}_B - \mathbb{E}_B)^{-1}\mathbb{F}_B.$$

— Méthode de relaxation par blocs ($\omega \neq 0$) :

$$\left\|\begin{array}{l}\textbf{initialisation}\\ \quad x^0 \in \mathbb{K}^n\\ \textbf{itérations : pour } k = 0, 1, \ldots, \textbf{faire}\\ \quad (\mathbb{D}_B - \omega\mathbb{E}_B)x^{k+1} = (\omega\mathbb{F}_B + (1-\omega)\mathbb{D}_B)\, x^k + \omega b\\ \textbf{fin}\end{array}\right. \tag{6.18}$$

A chaque itération on doit résoudre un système linéaire avec une *matrice triangulaire inférieure par blocs*. La matrice d'itération est notée

$$\mathbb{L}_{\omega,B} = (\mathbb{D}_B - \omega\mathbb{E}_B)^{-1}\left((1-\omega)\mathbb{D}_B + \omega\mathbb{F}_B\right).$$

Pour $\omega = 1$ on retrouve bien la méthode de Gauss-Seidel par blocs, i. e. $\mathbb{L}_{1,B} = \mathbb{G}_B$.

On peut généraliser le Théorème 6.8 au cas où la matrice $\mathbb{A}$ est écrite par blocs. En effet, sous les hypothèses de ce théorème on remarque que $\mathbb{D}_B$ est toujours hermitienne définie-positive, avec en outre

$$Spe(\mathbb{D}_B) = \cup_{p=1,P} Spe([\mathbb{A}]_{p,p}).$$

On peut alors reprendre le raisonnement du §6.7 pour obtenir le résultat ci-dessous.

Théorème 6.10 *[Ostrowski–Reich] Soit $\mathbb{A} \in \mathbb{C}^{n\times n}$ une matrice hermitienne définie-positive, la méthode de relaxation par blocs converge si et seulement si $\omega \in]0, 2[$.*

Comme il est presque toujours possible de définir une méthode *par points* là où on a défini une méthode *par blocs* – il suffit qu'il n'y ait pas d'élément diagonal nul – on a tendance à penser intuitivement que la méthode *par blocs* convergera alors plus vite.

6.9 Matrices tridiagonales par blocs

On considère dans ce paragraphe les matrices dont la structure est de la forme

$$\mathbb{A} = \begin{bmatrix} [\mathbb{D}]_1 & -[\mathbb{F}]_1 & & & \\ -[\mathbb{E}]_1 & [\mathbb{D}]_2 & -[\mathbb{F}]_2 & & \\ & \ddots & \ddots & \ddots & \\ & & -[\mathbb{E}]_{P-2} & [\mathbb{D}]_{P-1} & -[\mathbb{F}]_{P-1} \\ & & & -[\mathbb{E}]_{P-1} & [\mathbb{D}]_P \end{bmatrix}$$

Les matrices tridiagonales par blocs sont très courantes dans le cadre de l'approximation de solution d'équations différentielles par les méthodes des différences finies ou des éléments finis (sur des maillages structurés).

Proposition 6.11 *Soit $\mathbb{A} \in \mathbb{K}^{n\times n}$ une matrice tridiagonale par blocs, alors*

$$\rho(\mathbb{G}_B) = \rho(\mathbb{J}_B)^2.$$

Démonstration : Les valeurs propres de la matrice de Jacobi (resp. de Gauss-Seidel) sont les racines du polynôme caractéristique $p_{\mathbb{J}_B}(\lambda)$ (resp. $p_{\mathbb{G}_B}(\lambda)$) :

$$p_{\mathbb{J}_B}(\lambda) = \det\left((\mathbb{D}_B)^{-1}(\mathbb{E}_B + \mathbb{F}_B) - \lambda\mathbb{I}_n\right) = \det\left(\lambda\mathbb{D}_B - \mathbb{E}_B - \mathbb{F}_B\right)/\det\left(-\mathbb{D}_B\right);$$

$$\begin{aligned} p_{\mathbb{G}_B}(\lambda) &= \det\left((\mathbb{D}_B - \mathbb{E}_B)^{-1}\mathbb{F}_B - \lambda\mathbb{I}_n\right) = \det\left(\lambda\mathbb{D}_B - \lambda\mathbb{E}_B - \mathbb{F}_B\right)/\det\left(\mathbb{E}_B - \mathbb{D}_B\right), \\ &= \det\left(\lambda\mathbb{D}_B - \lambda\mathbb{E}_B - \mathbb{F}_B\right)/\det\left(-\mathbb{D}_B\right). \end{aligned}$$

Introduisons la matrice diagonale $\mathbb{Q}(\mu)$, définie par les blocs tels que pour tout $p \in \{1,\ldots,P\}$, $[\mathbb{Q}]_{p,p} = \mu^p\,\mathbb{I}_{n_p}$. On montre facilement que :

$$\lambda^2\mathbb{D}_B - \lambda^2\mathbb{E}_B - \mathbb{F}_B = \mathbb{Q}(\lambda)(\lambda^2\mathbb{D}_B - \lambda\mathbb{E}_B - \lambda\mathbb{F}_B)(\mathbb{Q}(\lambda))^{-1},$$

ce qui permet d'écrire :

$$\begin{aligned} p_{\mathbb{G}_B}(\lambda^2) &= \det\left(\lambda^2\mathbb{D}_B - \lambda^2\mathbb{E}_B - \mathbb{F}_B\right)/\det\left(-\mathbb{D}_B\right), \\ &= \det\left(\lambda^2\mathbb{D}_B - \lambda\mathbb{E}_B - \lambda\mathbb{F}_B\right)/\det\left(-\mathbb{D}_B\right), \\ &= \lambda^n\det\left(\lambda\mathbb{D}_B - \mathbb{E}_B - \mathbb{F}_B\right)/\det\left(-\mathbb{D}_B\right) = \lambda^n p_{\mathbb{J}_B}(\lambda). \end{aligned}$$

Ainsi lorsque $\lambda \neq 0$ est racine de $p_{\mathbb{J}_B}$, $\lambda^2 \neq 0$ est racine de $p_{\mathbb{G}_B}$, et réciproquement. Noter que si $\lambda = 0$ est valeur propre de $\mathbb{G}_B$, cela n'influe pas sur le résultat car on étudie $\rho(\mathbb{G}_B) = \max_i |\lambda_i(\mathbb{G}_B)|$.

◇

Ce résultat montre que la méthode de Gauss–Seidel converge (ou diverge !) deux fois plus vite que la méthode de Jacobi pour les matrices tridiagonales par blocs.

On admet le résultat ci-dessous [13].

Théorème 6.12 *Soit $\mathbb{A} \in \mathbb{K}^{n\times n}$ une matrice tridiagonale par blocs, telle que les valeurs propres de $\mathbb{J}_B$ soient toutes réelles, alors les méthodes par blocs de Jacobi, de Gauss-Seidel et de relaxation avec $\omega \in]0,2[$, divergent ou convergent simultanément.*

Etudions maintenant la valeur "optimale" du paramètre de relaxation ω, c'est-à-dire celle minimisant $\rho(\mathbb{L}_{\omega,B})$. Comme pour la Proposition 6.11 on écrit

$$p_{\mathbb{L}_{\omega,B}}(\lambda) = \det\left((\frac{1}{\omega}\mathbb{D}_B - \mathbb{E}_B)^{-1}(\frac{1-\omega}{\omega}\mathbb{D}_B + \mathbb{F}_B) - \lambda\mathbb{I}_n\right)$$

d'où, en utilisant la relation $\det(\mathbb{E}_B - \frac{1}{\omega}\mathbb{D}_B) = \omega^{-n}\det(-\mathbb{D}_B)$,

$$\begin{aligned}
p_{\mathbb{L}_{\omega,B}}(\lambda) &= \det\left(\frac{\lambda+\omega-1}{\omega}\mathbb{D}_B - \lambda\mathbb{E}_B - \mathbb{F}_B\right)\omega^n/\det(-\mathbb{D}_B)\,; \\
\Rightarrow\ p_{\mathbb{L}_{\omega,B}}(\lambda^2) &= \omega^n\det\left(\frac{\lambda^2+\omega-1}{\omega}\mathbb{D}_B - \lambda^2\mathbb{E}_B - \mathbb{F}_B\right)/\det(-\mathbb{D}_B) \\
&= \omega^n\det\left(\frac{\lambda^2+\omega-1}{\omega}\mathbb{D}_B - \lambda\mathbb{E}_B - \lambda\mathbb{F}_B\right)/\det(-\mathbb{D}_B) \\
&= \lambda^n\,\omega^n\det\left(\frac{\lambda^2+\omega-1}{\omega}\mathbb{D}_B - \mathbb{E}_B - \mathbb{F}_B\right)/\det(-\mathbb{D}_B) \\
&= \lambda^n\,\omega^n\,p_{\mathbb{J}_B}\left(\frac{\lambda^2+\omega-1}{\lambda\omega}\right).
\end{aligned}$$

NB. *dans ce qui suit* $\zeta^{1/2}$ *représente une racine carrée complexe de* ζ.

Si λ est valeur propre non nulle de $\mathbb{L}_{\omega,B}$ alors $\nu = \dfrac{\lambda+\omega-1}{\lambda^{1/2}\omega}$ est valeur propre de $\mathbb{J}_B$. Réciproquement si ν est valeur propre de $\mathbb{J}_B$, alors les racines $\lambda_\pm$ de l'équation

$$\lambda\nu^2\omega^2 = (\lambda+\omega-1)^2$$

sont valeurs propres de $\mathbb{L}_{\omega,B}$. Cette équation se met encore sous la forme

$$\lambda^2 + \lambda\left(2(\omega-1) - \nu^2\omega^2\right) + (\omega-1)^2 = 0$$

et on en déduit

$$\lambda_\pm = \frac{1}{2}(\nu^2\omega^2 - 2\omega + 2) \pm \frac{\nu\omega}{2}(\nu^2\omega^2 - 4\omega + 4)^{1/2}.$$

On suppose dans la suite que les valeurs propres ν de $\mathbb{J}_B$ sont réelles. On a : $\rho(\mathbb{L}_{\omega,B}) = \max_{\nu\in Spe(\mathbb{J}_B)} M(\nu,\omega)$, avec

$$M(\nu,\omega) : \left\{\begin{array}{rcl} \mathbb{R}^+\times]0,2[& \to & [0,1] \\ (\nu,\omega) & \mapsto & \max(|\lambda_+(\nu,\omega)|, |\lambda_-(\nu,\omega)|) \end{array}\right. .$$

Pour connaître la valeur de $\rho(\mathbb{L}_{\omega,B})$, il faut étudier les variations de $M(\nu,\omega)$ en fonction de ω. La courbe représentative des variations de $M(\nu,\omega)$ est représentée sur la figure 6.9 pour différentes valeurs de ν.

Théorème 6.13 *Soit* $\mathbb{A} \in \mathbb{C}^{n\times n}$ *une matrice hermitienne définie-positive et tridiagonale par blocs, alors les méthodes par blocs de Jacobi, de Gauss-Seidel et de relaxation pour* $\omega \in]0,2[$ *convergent. De plus, il existe une valeur optimale du paramètre* ω

$$\omega_{opt} = \frac{2}{1+\sqrt{1-\rho(\mathbb{J}_B)^2}}$$

telle que

$$\rho(\mathbb{L}_{\omega_{opt},B}) = \min_{\omega\in]0,2[}\rho(\mathbb{L}_{\omega,B}) < \rho(\mathbb{G}_B) = \rho(\mathbb{J}_B)^2 < \rho(\mathbb{J}_B) < 1.$$

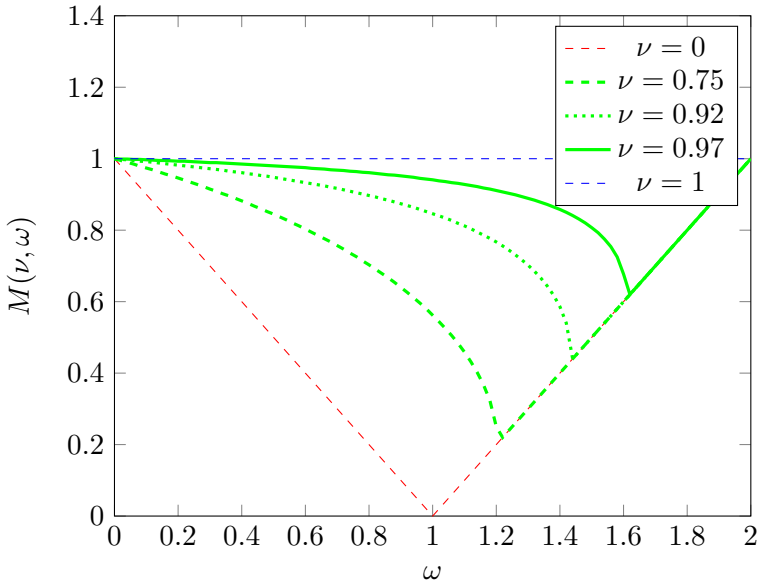


FIGURE 6.1 – La fonction $M(\nu, \omega)$ pour différentes valeurs de ν.

Démonstration : Pour appliquer le Théorème 6.12 il suffit de vérifier que les valeurs propres ν de $\mathbb{J}_B$ sont réelles : il existe $v \neq 0$ tel que

$$\begin{array}{rrl} \mathbb{J}_B v = (\mathbb{D}_B)^{-1}(\mathbb{E}_B + \mathbb{F}_B)v & = \nu v \Longrightarrow (\mathbb{E}_B + \mathbb{F}_B)v = \nu \mathbb{D}_B v \\ \text{soit encore} \qquad\qquad \mathbb{A}v & = (1-\nu)\mathbb{D}_B v \Longrightarrow (\mathbb{A}v, v) = (1-\nu)(\mathbb{D}_B v, v). \end{array}$$

Si $\mathbb{A}$ est hermitienne définie-positive alors nécessairement $\mathbb{D}_B$ l'est aussi, ainsi $(\mathbb{A}v, v)$ et $(\mathbb{D}_B v, v)$ sont des réels strictement positifs, et ν valeur propre de $\mathbb{J}_B$ est réelle et plus petite que 1. Alors d'après le Théorème 6.12 les méthodes par blocs de Jacobi, de Gauss-Seidel et de relaxation pour $0 < \omega < 2$ convergent. ◇

Remarque 6.14 *si on ne connait pas exactement l'expression de la valeur optimale ω_{opt} l'étude des variations de $\omega \mapsto \rho(\mathbb{L}_{\omega,B})$ montre qu'il vaut mieux l'approcher par valeurs supérieures puisque la dérivée $\rho'(\mathbb{L}_{\omega,B})$ vaut 1 quand $\omega \to \omega_{opt+}$ mais tend vers $-\infty$ quand $\omega \to \omega_{opt-}$!*

6.10 Méthodes de Richardson

Pour résoudre un système linéaire avec une matrice $\mathbb{A} \in \mathbb{R}^{n\times n}$ (et un second membre $b \in \mathbb{R}^n$, le cas échéant en considérant séparément ses parties réelle et imaginaire), une méthode itérative très utilisée en optimisation (cf. [12]) est la méthode de gradient à pas constant, qui calcule le nouvel itéré sous la forme

$$x^{k+1} = x^k + \alpha(b - \mathbb{A}x^k), \quad \alpha \neq 0.$$

Le résidu $r^k = b - \mathbb{A}x^k$ est le gradient d'une fonctionnelle que l'on cherche à minimiser et le paramètre α est le pas de descente.... Lorsque $\mathbb{A}$ est une matrice symétrique définie-positive,

on peut en effet résoudre le problème $\mathbb{A}x = b$ en minimisant la fonctionnelle

$$v \mapsto \frac{1}{2}(\mathbb{A}v, v) - (b, v)$$

sur $\mathbb{R}^n$. La méthode de Richardson correspond à la décomposition

$$\mathbb{M} = \frac{1}{\alpha}\mathbb{I}_n \quad \text{et} \quad \mathbb{N} = \frac{1}{\alpha}\mathbb{I}_n - \mathbb{A},$$

régulière pour tout $\alpha \neq 0$. En effet, l'itération est

$$x^{k+1} = \mathbb{M}^{-1}\mathbb{N}x^k + \mathbb{M}^{-1}b = x^k - \alpha\mathbb{A}x^k + \alpha b.$$

On l'appelle **méthode de Richardson du premier ordre à pas constant** (le paramètre α est fixé) et on l'écrit :

$$\left\| \begin{array}{l} \textbf{initialisation} \\ \quad x^0 \in \mathbb{R}^n \\ \textbf{itérations : pour } k = 0, 1, \ldots, \textbf{faire} \\ \quad x^{k+1} = x^k + \alpha(b - \mathbb{A}x^k) \\ \textbf{fin} \end{array} \right. \tag{6.19}$$

Proposition 6.15 *Soit $\mathbb{A} \in \mathbb{R}^{n\times n}$ une matrice symétrique définie-positive, la méthode de Richardson à pas constant converge si et seulement si*

$$0 < \alpha < \frac{2}{\rho(\mathbb{A})}.$$

Le pas optimal est $\alpha_{opt} = \frac{2}{\lambda_1+\lambda_n}$.

Démonstration : On note $0 < \lambda_n \leq \lambda_{n-1} \leq \ldots \leq \lambda_1$ les valeurs propres de $\mathbb{A}$, les valeurs propres de

$$R_\alpha = \mathbb{M}^{-1}\mathbb{N} = \mathbb{I}_n - \alpha\mathbb{A}$$

sont les $1 - \alpha\lambda_j$. Par définition, $\rho(R_\alpha) = \max_j |1 - \alpha\lambda_j|$ et on a

$$|1 - \alpha\lambda_j| < 1 \;\Leftrightarrow\; -1 < 1 - \alpha\lambda_j < 1 \;\Leftrightarrow\; 0 < \alpha\lambda_j < 2 \;\Leftrightarrow\; 0 < \alpha < \frac{2}{\lambda_j}.$$

La convergence a donc lieu si, et seulement si,

$$0 < \alpha < \min_j \frac{2}{\lambda_j} = \frac{2}{\lambda_1} = \frac{2}{\rho(\mathbb{A})}.$$

Par ailleurs, on minimise $\alpha \mapsto \rho(R_\alpha)$ lorsque $|1 - \alpha\lambda_n| = |1 - \alpha\lambda_1|$, c'est-à-dire lorsque $1 - \alpha\lambda_n = \alpha\lambda_1 - 1$, ce qui conduit à la valeur annoncée de α_{opt}. ⋄

En général on dispose de peu d'informations sur le spectre des matrices que l'on traite, et il est difficile de donner au paramètre α une valeur qui assure la convergence. Une variante de la méthode de Richardson fournit une solution à ce problème : on modifie le paramètre α à chaque itération, en lui donnant la valeur qui minimise la norme du

résidu $r^{k+1} = b - \mathbb{A}x^{k+1}$, pour une norme bien choisie (voir plus bas). La **méthode de Richardson à pas variable** s'écrit

$$\left\| \begin{array}{l} \textbf{initialisation} \\ \quad x^0 \in \mathbb{R}^n \\ \textbf{itérations : pour } k = 0, 1, \ldots, \textbf{faire} \\ \quad x^{k+1} = x^k + \alpha^k (b - \mathbb{A}x^k) \\ \textbf{fin} \end{array} \right. \tag{6.20}$$

Proposition 6.16 *Soit $\mathbb{A} \in \mathbb{R}^{n\times n}$ une matrice symétrique définie-positive, si*

$$\alpha^k = \alpha^k_{opt} \text{ avec } \alpha^k_{opt} = \frac{\|r^k\|_2^2}{\|r^k\|_{\mathbb{A}}^2}$$

alors la méthode de Richardson à pas variable converge.

Démonstration : Par construction $r^k = b - \mathbb{A}x^k$ et donc :

$$r^{k+1} = b - \mathbb{A}x^{k+1} = r^k - \alpha^k \mathbb{A}r^k.$$

Pour toute norme vectorielle $\|\cdot\|_{\mathbb{C}}$ avec $\mathbb{C}$ matrice symétrique définie-positive, on peut écrire

$$\begin{aligned} \|r^{k+1}\|_{\mathbb{C}}^2 &= \big(\mathbb{C}r^{k+1}, r^{k+1}\big) = \big(\mathbb{C}(r^k - \alpha^k \mathbb{A}r^k), r^k - \alpha^k \mathbb{A}r^k\big) \\ &= \|r^k\|_{\mathbb{C}}^2 - 2\alpha^k(\mathbb{C}\mathbb{A}r^k, r^k) + (\alpha^k)^2 \|\mathbb{A}r^k\|_{\mathbb{C}}^2. \end{aligned}$$

Cette expression atteint son minimum

$$\|r^{k+1}\|_{\mathbb{C}}^2 = \|r^k\|_{\mathbb{C}}^2 - \frac{|(\mathbb{C}\mathbb{A}r^k, r^k)|^2}{\|\mathbb{A}r^k\|_{\mathbb{C}}^2}$$

quand

$$\alpha^k_{opt} = \frac{(\mathbb{C}\mathbb{A}r^k, r^k)}{\|\mathbb{A}r^k\|_{\mathbb{C}}^2}.$$

Lorsque la matrice $\mathbb{A}$ est symétrique définie-positive un choix simple pour le calcul de α^k est de prendre $\mathbb{C} = \mathbb{A}^{-1}$, alors $\|\mathbb{A}r^k\|_{\mathbb{A}^{-1}}^2 = (\mathbb{A}^{-1}\mathbb{A}r^k, \mathbb{A}r^k) = \|r^k\|_{\mathbb{A}}^2$, d'où :

$$\alpha^k_{opt} = \frac{\|r^k\|_2^2}{\|r^k\|_{\mathbb{A}}^2} \quad \text{et} \quad \|r^{k+1}\|_{\mathbb{A}^{-1}}^2 = \|r^k\|_{\mathbb{A}^{-1}}^2 - \frac{\|r^k\|_2^4}{\|r^k\|_{\mathbb{A}}^2}.$$

Considérons tout d'abord le cas où $\exists \mu > 0$ tel que $\mathbb{A} = \mu\, \mathbb{I}_n$, alors

$$\|r^k\|_{\mathbb{A}^{-1}}^2 - \frac{\|r^k\|_2^4}{\|r^k\|_{\mathbb{A}}^2} = (\mu^{-2} - \mu^{-2})\|r^k\|^2 = 0.$$

En d'autres termes la méthode converge en une itération.
Supposons maintenant que $\mathbb{A} \notin \{\mu\,\mathbb{I}_n,\ \mu > 0\}$. On obtient la majoration [41]

$$\begin{aligned}\|r^{k+1}\|^2_{\mathbb{A}^{-1}} &= \|r^k\|^2_{\mathbb{A}^{-1}} \left[1 - \frac{\|r^k\|_2^4}{\|r^k\|^2_{\mathbb{A}}\ \|r^k\|^2_{\mathbb{A}^{-1}}}\right] \\ &\leq \|r^k\|^2_{\mathbb{A}^{-1}} \left[1 - 1/\kappa_2(\mathbb{A})\right] \leq \cdots \leq \|r^0\|^2_{\mathbb{A}^{-1}} \left[1 - 1/\kappa_2(\mathbb{A})\right]^{k+1}\end{aligned}$$

Puisque $\mathbb{A}$ est symétrique définie-positive (et que $\mathbb{A} \notin \{\mu\,\mathbb{I}_n,\ \mu > 0\}$), alors d'après la Proposition B.14 on a $\kappa_2(\mathbb{A}) = \lambda_1(\mathbb{A})/\lambda_n(\mathbb{A}) > 1$ et donc $0 < 1 - (\kappa_2(\mathbb{A}))^{-1} < 1$, ce qui entraîne que $\|r^{k+1}\|_{\mathbb{A}^{-1}}$ tend vers 0 quand k tend vers $+\infty$, et la méthode converge. ⋄

6.11 Matrices à diagonale dominante

Il existe une catégorie de matrices importante dans l'histoire de l'étude des méthodes itératives : les matrices à diagonale dominante. Pour ces matrices, on peut établir la convergence des méthodes itératives par points.

Définition 6.17 *une matrice $\mathbb{A} \in \mathbb{K}^{n\times n}$ est dite à* **diagonale dominante** *si et seulement si*

$$\forall i, \in \{1, \cdots, n\} \qquad \sum_{j\neq i} |\mathbb{A}_{i,j}| \leq |\mathbb{A}_{i,i}|.$$

Une matrice $\mathbb{A} \in \mathbb{K}^{n\times n}$ est dite à **diagonale strictement dominante** *si et seulement si*

$$\forall i \in \{1, \cdots, n\} \qquad \sum_{j\neq i} |\mathbb{A}_{i,j}| < |\mathbb{A}_{i,i}|.$$

Proposition 6.18 *Soit $\mathbb{A} \in \mathbb{K}^{n\times n}$ une une matrice à diagonale strictement dominante, alors $\mathbb{A}$ est inversible.*

Démonstration : Soit $v \in \mathbb{K}^n$ tel que $\mathbb{A}v = 0$. Supposons que $v \neq 0$: on note $1 \leq i \leq n$ l'indice tel que $|v_i| = \max_j |v_j| \neq 0$. On écrit que $(\mathbb{A}v)_i = 0$, soit $\mathbb{A}_{i,i}v_i = -\sum_{j\neq i} \mathbb{A}_{i,j}v_j$. On en déduit que

$$|\mathbb{A}_{i,i}| \leq \sum_{j\neq i} \frac{|v_j|}{|v_i|}\,|\mathbb{A}_{i,j}| \leq \sum_{j\neq i} |\mathbb{A}_{i,j}|,$$

ce qui contredit le fait que $\mathbb{A}$ est à diagonale strictement dominante. ⋄

Proposition 6.19 *Soit $\mathbb{A} \in \mathbb{K}^{n\times n}$ une matrice à diagonale strictement dominante, alors les méthodes de Jacobi et Gauss-Seidel par points convergent.*

41. Pour $v \neq 0$, on a

$$\frac{\|v\|^2_{\mathbb{A}}}{\|v\|_2^2} \times \frac{\|v\|^2_{\mathbb{A}^{-1}}}{\|v\|_2^2} = \frac{(\mathbb{A}v, v)}{\|v\|_2^2} \times \frac{(\mathbb{A}^{-1}v, v)}{\|v\|_2^2} \leq \frac{\|\mathbb{A}v\|_2}{\|v\|_2} \times \frac{\|\mathbb{A}^{-1}v\|_2}{\|v\|_2} \leq \kappa_2(\mathbb{A})$$

avec $\kappa_2(\mathbb{A}) = \|\mathbb{A}\|_2\ \|\mathbb{A}^{-1}\|_2$ le nombre de conditionnement de $\mathbb{A}$ dans la norme $\|\cdot\|_2$.

Démonstration : Par définition de la matrice de Jacobi, $\mathbb{J} = \mathbb{D}^{-1}(\mathbb{E} + \mathbb{F})$ et d'après la Proposition B.8 :

$$\|\mathbb{J}\|_\infty = \max_i \sum_{j \neq i} \frac{|\mathbb{A}_{i,j}|}{|\mathbb{A}_{i,i}|}.$$

Donc si $\mathbb{A}$ est une matrice à diagonale strictement dominante on a $\|\mathbb{J}\|_\infty < 1$ et d'après la Proposition B.12 :

$$\rho(\mathbb{J}) \leq \|\mathbb{J}\|_\infty < 1.$$

Soit maintenant λ une valeur propre de la matrice de Gauss-Seidel $\mathbb{G} = (\mathbb{D} - \mathbb{E})^{-1}\mathbb{F}$:

$$(\mathbb{D} - \mathbb{E})^{-1}\mathbb{F}v = \lambda v \quad \Longleftrightarrow \quad \mathbb{F}v = \lambda(\mathbb{D} - \mathbb{E})v,$$

et soit i la composante telle que $|v_i| = \max_j |v_j| \neq 0$, alors on écrit

$$\begin{aligned} \lambda \mathbb{A}_{i,i} v_i &= \lambda \sum_{j<i} \mathbb{A}_{i,j} v_j - \sum_{j>i} \mathbb{A}_{i,j} v_j \\ \Longrightarrow |\lambda||\mathbb{A}_{i,i}| &\leq |\lambda| \sum_{j<i} |\mathbb{A}_{i,j}| + \sum_{j>i} |\mathbb{A}_{i,j}|. \\ \Longrightarrow |\lambda| \left(|\mathbb{A}_{i,i}| - \sum_{j<i} |\mathbb{A}_{i,j}| \right) &\leq \sum_{j>i} |\mathbb{A}_{i,j}|. \end{aligned}$$

Par ailleurs on sait que

$$\begin{aligned} |\mathbb{A}_{i,i}| - \sum_{j<i} |\mathbb{A}_{i,j}| &> \sum_{j>i} |\mathbb{A}_{i,j}| \geq 0, \\ \Longrightarrow |\lambda| \leq & \frac{\sum_{j>i} |\mathbb{A}_{i,j}|}{|\mathbb{A}_{i,i}| - \sum_{j<i} |\mathbb{A}_{i,j}|} < 1; \end{aligned}$$

ainsi $\rho(\mathbb{G}) < 1$ et la méthode converge. ◇

6.12 Méthode de relaxation symétrique (S.S.O.R.)

L'ordre des inconnues a-t-il une influence sur la convergence de la méthode ? Pour la méthode de Jacobi, la réponse est non puisque les mises à jour des composantes de la solution approchée x^{k+1} sont indépendantes les unes des autres. Par contre, pour les méthodes de Gauss-Seidel et de relaxation, cette question est pertinente puisque la numérotation des inconnues joue un rôle dans leur définition : chaque composante x_i^{k+1} de x^{k+1} est définie à partir des composantes d'indice inférieur x_j^{k+1} pour $j < i$ (sur ce sujet voir Adams et Jordan [1]). Pour éviter les problèmes liés à la numérotation des inconnues quand on n'a pas d'information utile à exploiter, il est donc préférable de *symétriser* les itérations de S.O.R. en inversant l'ordre des calculs à chaque itération : on effectue une demi-itération dans l'ordre croissant des inconnues et la demi-itération suivante dans l'ordre décroissant.

On obtient ainsi la méthode de **sur-relaxation symétrique**, Symmetric Successive Over Relaxation (S.S.O.R. en abrégé), qui s'écrit pour la méthode par blocs (on peut également écrire la méthode par points)

$$\left\|\begin{array}{l} \textbf{initialisation} \\ \quad x^0 \in \mathbb{K}^n \\ \textbf{itérations : pour } k = 0, 1, \ldots, \textbf{faire} \\ \quad (\mathbb{D}_B - \omega\mathbb{E}_B)x^{k+1/2} = (\omega\mathbb{F}_B + (1-\omega)\mathbb{D}_B)\, x^k + \omega b \\ \quad (\mathbb{D}_B - \omega\mathbb{F}_B)x^{k+1} = (\omega\mathbb{E}_B + (1-\omega)\mathbb{D}_B)\, x^{k+1/2} + \omega b \\ \textbf{fin} \end{array}\right. \tag{6.21}$$

On note

$$S_{\omega,B} = (\frac{1}{\omega}\mathbb{D}_B - \mathbb{F}_B)^{-1}\left((\frac{1-\omega}{\omega})\mathbb{D}_B + \mathbb{E}_B\right)(\frac{1}{\omega}\mathbb{D}_B - \mathbb{E}_B)^{-1}\left((\frac{1-\omega}{\omega})\mathbb{D}_B + \mathbb{F}_B\right)$$

la matrice d'itération associée. L'étude directe des valeurs propres de cette matrice est assez compliquée, mais on peut néanmoins vérifier le résultat suivant (voir [37, §15]).

Proposition 6.20 *Soit $\mathbb{A} \in \mathbb{C}^{n\times n}$ une matrice hermitienne définie-positive, la méthode S.S.O.R. converge si et seulement si $\omega \in]0,2[$. Qui plus est, il existe une unique valeur ω_{opt}^S minimisant $\rho(S_{\omega,B})$.*

Remarque 6.21 *En général, on ne sait pas en donner une expression analytique de ω_{opt}^S. Pour certains systèmes linéaires, la valeur*

$$\omega'_{opt} = \frac{2}{1 + \sqrt{2\left(1 - \rho(\mathbb{J}_B)\right)^2}}$$

est une valeur pour laquelle S.S.O.R. converge environ deux fois plus vite que S.O.R. avec ω_{opt}, mais il faut se souvenir qu'une itération de S.S.O.R. coûte environ deux fois plus cher qu'une itération de S.O.R.... Le véritable intérêt de cette méthode est de diminuer l'influence de la numérotation des inconnues sur la convergence.

6.13 Utilisation du calcul parallèle

On se concentre sur la parallélisation des méthodes itératives par points. Précisément, on considère des systèmes linéaires avec des matrices issues de la discrétisation par différences finies pour $d = 1$ et $d = 2$, et on étudie le cas des méthodes itératives par points de Jacobi et Gauss-Seidel. On examine en particulier jusqu'où on peut pousser le parallélisme dans les calculs, c'est-à-dire en supposant qu'on dispose d'un nombre "illimité" de nœuds $\mathtt{P}$, la seule contrainte étant que $\mathtt{P} \leq N$, où $N = n^d$ est le nombre d'inconnues, avec n points de discrétisation par direction dans le domaine de calcul $]0,1[^d$ (on reprend le formalisme du §4.7). Dans la suite, on discrétise le Laplacien (ou le Laplacien généralisé).
Pour finir, on considère brièvement le cas des méthodes par blocs, ainsi que la discrétisation par éléments finis.

6.13.1 Parallélisation de Jacobi par points

On étudie ici comment paralléliser (6.11)-(6.12), pour $\mathbb{A} = \mathbb{A}_1$ ou $\mathbb{A}_2$, avec $\mathbb{A}_d$ la matrice issue des différences finies pour $d = 1, 2$. On suppose que l'indice k d'itération externe (6.11) est donné.

Cas $d = 1$

D'après l'expression (2.5) de $\mathbb{A}_1$, la boucle interne (6.12) s'écrit, avec la convention $x_0^k = x_{N+1}^k = 0$,

$$\left\| \begin{array}{l} \textbf{pour } i = 1, 2, \ldots, N \textbf{ faire} \\ \quad x_i^{k+1} = \frac{1}{2}(x_{i-1}^k + x_{i+1}^k) + \frac{h^2}{2} b_i \\ \textbf{fin} \end{array} \right. \tag{6.22}$$

Examinons le "parallélisme maximal" extrême, c'est-à-dire avec $\mathtt{P} = N$: pour $1 \leq i \leq N$, on affecte au nœud i les données d'indice i. A l'itération externe k, le nœud i possède les données x_i^k et $\frac{1}{2}h^2 b_i$ en mémoire et, pour calculer x_i^{k+1}, il doit récupérer les données x_{i-1}^k de son voisin gauche, et x_{i+1}^k de son voisin droite. (Les nœuds sont considérés comme étant placés sur une ligne horizontale.) Cette phase de communications peut être réalisée en deux étapes parallèles, à l'instar de ce qui est proposé au §4.7. A partir de là, les calculs sont équirépartis, si on réalise pour $i = 1$ les opérations superflues avec $x_0^k = 0$ et, pour $i = N$, celles avec $x_{N+1}^k = 0$. Le "parallélisme maximal" avec $\mathtt{P} = N$ est donc accessible ! De plus, la matrice $\mathbb{A}_1$ étant tridiagonale, la méthode de Jacobi converge (voir le théorème 6.13).

Cas $d = 2$

Afin de préserver la structure du schéma à cinq points, on conserve la numérotation à deux indices (i, j), où $0 \leq i, j \leq n + 1$. D'après l'expression (2.22) de $\mathbb{A}_2$, la boucle interne (6.12) s'écrit, avec la convention $x_{0,j}^k = x_{n+1,j}^k = x_{i,0}^k = x_{i,n+1}^k = 0$,

$$\left\| \begin{array}{l} \textbf{pour } i, j = 1, 2, \ldots, n \textbf{ faire} \\ \quad x_{i,j}^{k+1} = \frac{1}{4}(x_{i,j-1}^k + x_{i-1,j}^k + x_{i+1,j}^k + x_{i,j+1}^k) + \frac{h^2}{4} b_{i,j} \\ \textbf{fin} \end{array} \right. \tag{6.23}$$

Examinons à nouveau le "parallélisme maximal" extrême, c'est-à-dire avec $\mathtt{P} = N$: pour $1 \leq i, j \leq n$, on affecte au nœud $\mathtt{p}_{i,j} = i + (j-1)n$ les données d'indice (i, j). A l'itération externe k, le nœud $\mathtt{p}_{i,j}$ possède donc les données $x_{i,j}^k$ et $\frac{1}{2}h^2 b_{i,j}$ en mémoire et, pour calculer $x_{i,j}^{k+1}$, il doit récupérer les données :

- $x_{i,j-1}^k$ de son voisin bas,
- $x_{i-1,j}^k$ de son voisin gauche,
- $x_{i+1,j}^k$ de son voisin droite,
- $x_{i,j+1}^k$ de son voisin haut.

(Les nœuds sont maintenant considérés comme étant placés sur une grille bidimensionnelle, avec n nœuds dans chaque direction.) Cette phase de communications peut être réalisée en quatre étapes parallèles, encore une fois à l'instar de ce qui est proposé au §4.7. A partir de là, les calculs sont équirépartis, avec une convention similaire au cas $d = 1$. Le "parallélisme maximal" avec $\mathtt{P} = N$ reste donc accessible ! L'inconvénient est que, la $\mathbb{A}_1$ n'étant plus tridiagonale par blocs, on ne sait pas si la méthode de Jacobi converge.

Résultats complémentaires

Tout d'abord, on peut vérifier qu'on aboutit des résultats similaires de "parallélisme maximal" avec $\mathtt{P} = N$ pour le Laplacien généralisé (ou diffusion), puisque la structure des matrices $\mathbb{A}'_1$ et $\mathbb{A}'_2$ est la même que celle des matrices $\mathbb{A}_1$ et $\mathbb{A}_2$. Et, si les coefficients sont positifs et bornés inférieurement par une constante strictement positive, alors les matrices $\mathbb{A}'_1$ et $\mathbb{A}'_2$ sont à diagonale strictement dominante. Dans ce cas, la proposition 6.19 assure la convergence de la méthode de Jacobi.

Exercice 6.1 *Montrer que pour la méthode de Jacobi par points appliquée au différences finies pour $d \geq 3$, on peut établir le "parallélisme maximal" avec $\mathtt{P} = N$.*

Exercice 6.2 *Dans le cadre du "parallélisme maximal" avec $\mathtt{P} = N$, proposer une évaluation rapide du vecteur résidu $r^k = b - \mathbb{A}x^k$ pour la méthode de Jacobi par points appliquée au différences finies.*

6.13.2 Parallélisation de Gauss-Seidel par points

On reprend la démarche, à savoir l'étude du "parallélisme maximal" de la méthode de Gauss-Seidel (6.13)-(6.14), pour $\mathbb{A} = \mathbb{A}_1$ ou $\mathbb{A}_2$.

Cas $d = 1$

La boucle interne (6.13) s'écrit

$$\left\| \begin{array}{l} \textbf{pour } i = 1, 2, \ldots, N \textbf{ faire} \\ \quad x_i^{k+1} = \frac{1}{2}(x_{i-1}^{k+1} + x_{i+1}^{k}) + \frac{h^2}{2} b_i \\ \textbf{fin} \end{array} \right.$$

La difficulté majeure est que cet algorithme est séquentiel ! En effet, pour calculer x_i^{k+1}, il faut disposer de x_{i-1}^{k+1}, et donc de x_{i-2}^{k+1}, et ainsi de suite jusqu'à x_1^{k+1}. Comment remédier à cette difficulté ? Il faut ré-examiner le schéma à trois points (voir la figure 2.2) : on note que si i est pair, ses deux voisins sont impairs, et vice versa... Une idée est alors de séparer les points de discrétisation en deux sous-ensembles :

— les points d'indice impair, dits *rouges*,
— les points d'indice pair, dits *noirs*.

A partir de là, on choisit de *renuméroter* les N_R points rouges (de gauche à droite), puis les N_N points noirs (de gauche à droite). Dans la suite, on suppose que $n = 2n'$ pour alléger

FIGURE 6.2 – Pour $d = 1$ et $n = 8$, les points rouges et noirs / La renumérotation.

les notations, et ainsi $N_R = N_N = n'$. Selon cette numérotation, la matrice s'écrit

$$\mathbb{A}_1^{RN} = \frac{1}{h^2}\begin{pmatrix} 2I_{n'} & -\mathbb{B}_1' \\ -(\mathbb{B}_1')^T & 2I_{n'} \end{pmatrix}, \quad \text{avec } \mathbb{B}_1' = \begin{pmatrix} 1 & & & & \\ 1 & 1 & & & \\ & \ddots & \ddots & & \\ & & 1 & 1 & \\ & & & 1 & 1 \end{pmatrix} \in \mathbb{R}^{n'\times n'}. \tag{6.24}$$

D'après cette expression (6.24), la matrice intervenant dans la décomposition régulière pour Gauss-Seidel par points est par définition égale à

$$\mathbb{M}_{GS}^{RN} = \frac{1}{h^2}\begin{pmatrix} 2I_{n'} & 0 \\ -(\mathbb{B}_1')^T & 2I_{n'} \end{pmatrix}.$$

Notons $\mathbb{P}$ la matrice de permutation associée à la renumérotation : on a $\mathbb{A}_1^{RN} = \mathbb{P}^{-1}\mathbb{A}_1\mathbb{P}$. D'après cette expression et puisque $\mathbb{A}_1$ est symétrique définie-positive, $\mathbb{A}_1^{RN}$ l'est aussi.

Remarque 6.22 *On peut vérifier que $\mathbb{M}_{GS}^{RN}$ n'est pas égale à $\mathbb{P}^{-1}\mathbb{M}_{GS}\mathbb{P}$, où $\mathbb{M}_{GS}$ est la matrice intervenant dans la décomposition régulière de $\mathbb{A}_1$ pour Gauss-Seidel. En effet, dans $\mathbb{M}_{GS}$ on a conservé tous les éléments de $\mathbb{A}_1$ correspondant à la contribution du voisin gauche, alors que dans $\mathbb{M}_{GS}^{RN}$ on n'a conservé cette contribution que pour les voisins gauches des nœuds noirs. Par contre, on note que $\mathbb{M}_{Jac}^{RN} = 2h^{-2}I_n$ est inchangée. La renumérotation rouge-noire* ne modifie pas *la méthode de Jacobi.*

Ecrite par blocs, la renumérotation donne pour les vecteurs

$$x := \begin{pmatrix} x_R \\ x_N \end{pmatrix} \text{ avec } x_R, x_N \in \mathbb{R}^{n'},$$

c'est-à-dire que $x_{R,i} = x_i$ pour $1 \le i \le n'$, et $x_{N,i} = x_{i+n'}$ pour $1 \le i \le n'$. On en conclut que la boucle interne de Gauss-Seidel avec renumérotation rouge-noir s'écrit

$$\begin{pmatrix} 2I_{n'} & 0 \\ -(\mathbb{B}_1')^T & 2I_{n'} \end{pmatrix}\begin{pmatrix} x_R^{k+1} \\ x_N^{k+1} \end{pmatrix} = \begin{pmatrix} 0 & \mathbb{B}_1' \\ 0 & 0 \end{pmatrix}\begin{pmatrix} x_R^{k} \\ x_N^{k} \end{pmatrix} + h^2\begin{pmatrix} b_R \\ b_N \end{pmatrix}.$$

On a donc la boucle interne

$$\begin{cases} x_R^{k+1} = \dfrac{1}{2}\mathbb{B}_1' x_N^k + \dfrac{1}{2}h^2 b_R \\ x_N^{k+1} = \dfrac{1}{2}(\mathbb{B}_1')^T x_R^{k+1} + \dfrac{1}{2}h^2 b_N. \end{cases}$$

Soit, avec renumérotation rouge-noir par points, les deux boucles (de longueur $N/2$)

pour $i = 1, 2, \ldots, n'$ **faire**
  $x_{R,i}^{k+1} = \frac{1}{2}(x_{N,i-1}^k + x_{N,i}^k) + \frac{h^2}{2}b_{R,i}$
fin

et

pour $i = 1, 2, \ldots, n'$ **faire**
  $x_{N,i}^{k+1} = \frac{1}{2}(x_{R,i}^{k+1} + x_{R,i+1}^{k+1}) + \frac{h^2}{2}b_{N,i}$
fin

A partir de là, on vérifie que l'on a un "parallélisme maximal" avec $\mathtt{P} = N/2$. En effet, supposons que le nœud i possède les données $x^k_{R,i}$, $\frac{1}{2}h^2 b_{R,i}$, $x^k_{N,i}$ et $\frac{1}{2}h^2 b_{N,i}$ en mémoire. Pour calculer $x^{k+1}_{R,i}$, il doit récupérer les données :

— $x^k_{N,i-1}$ de son voisin gauche,
— $x^k_{N,i}$ de son voisin droite.

Pour calculer $x^{k+1}_{N,i}$, il doit récupérer les données :

— $x^{k+1}_{R,i}$ de son voisin gauche,
— $x^{k+1}_{R,i+1}$ de son voisin droite.

Il y a donc quatre étapes de communications parallèles (deux pour les voisins gauches ; deux pour les voisins droits), avec des calculs équirépartis à l'instar de l'algorithme de la méthode de Jacobi par points. Enfin, la matrice $\mathbb{A}_1^{RN}$ étant symétrique définie-positive, la méthode de Gauss-Seidel converge d'après le théorème 6.10.

Cas $d = 2$

Si on conserve la numérotation "naturelle" (ci-après à deux indices (i, j)), on est à nouveau confronté à un algorithme séquentiel puisque, pour calculer $x^{k+1}_{i,j}$, on a besoin de $x^{k+1}_{i-1,j}$ et $x^{k+1}_{i,j-1}$, etc. jusqu'à $x^{k+1}_{1,1}$. Pour remédier à cette difficulté, on va reprendre la même idée que pour $d = 1$, voir la figure 6.3. On suppose toujours que $n = 2n'$, avec

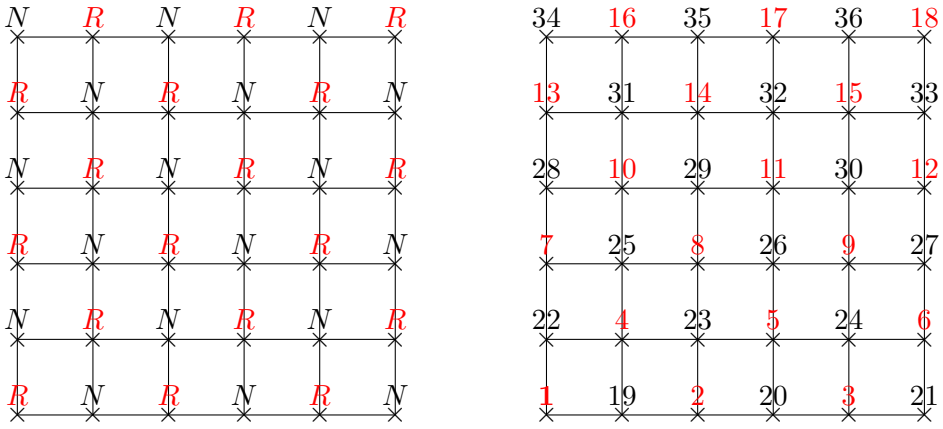


FIGURE 6.3 – Pour $d = 2$ et $n = 6$, les points rouges et noir / La renumérotation.

$N = n^2 = 4(n')^2$ points de discrétisation. On pose $N' = N^2/2$. Après coloriage avec $N_R = N'$ points rouges, et $N_B = N'$ points noirs, on constate qu'un point rouge à pour voisins quatre points noirs, et qu'un point noir à pour voisins quatre points rouges. Ceci va à nouveau permettre une parallèlisation de la méthode de Gauss-Seidel utilisée sur la matrice après renumérotation rouge-noir. Selon cette numérotation, la matrice s'écrit en effet

$$\mathbb{A}_2^{RN} = \frac{1}{h^2}\begin{pmatrix} 4I_{N'} & -\mathbb{B}'_2 \\ -(\mathbb{B}'_2)^T & 4I_{N'} \end{pmatrix}. \tag{6.25}$$

La matrice $\mathbb{A}_2^{RN}$ est toujours symétrique définie-positive. La structure de $\mathbb{B}'_2$ est un peu plus complexe que celle de $\mathbb{B}'_1$, car le numéro des voisins gauche et droite va dépendre de la parité de l'indice j (voir la figure 6.3) : lorsque j est impair on commence la ligne par un point rouge, et lorsque j est pair on commence la ligne par un point noir. Elle est omise ici. D'après (6.25), dans la décomposition régulière de $\mathbb{A}_2^{RN}$ pour Gauss-Seidel par points, on a

$$\mathbb{M}_{GS}^{RN} = \frac{1}{h^2}\begin{pmatrix} 4I_{N'} & 0 \\ -(\mathbb{B}'_2)^T & 4I_{N'} \end{pmatrix}.$$

Le point crucial est que la boucle interne qui en résulte est très similaire à celle du cas $d = 1$. Ecrite par blocs, elle s'écrit :

$$\begin{cases} x_R^{k+1} = \dfrac{1}{4}\mathbb{B}'_2 x_N^k + \dfrac{1}{4}h^2 b_R \\ x_N^{k+1} = \dfrac{1}{4}(\mathbb{B}'_2)^T x_R^{k+1} + \dfrac{1}{4}h^2 b_N. \end{cases}$$

Par points, avec renumérotation rouge-noir par points, on a nouveau deux boucles de longueur N', dont on omet l'écriture pour les raisons évoquées au-dessus. On peut vérifier que l'on a un "parallélisme maximal" avec $\mathtt{P} = N/2$, avec huit étapes de communications parallèles (deux pour les voisins bas, resp. voisins gauches, resp. voisins droits, resp. voisins hauts), avec des calculs équirépartis. Enfin, la matrice $\mathbb{A}_2^{RN}$ étant symétrique définie-positive, on a convergence de la méthode de Gauss-Seidel converge (théorème 6.10).

6.13.3 Parallélisation – Autres configurations

En préambule, on observe que la structure des matrices est inchangée entre les méthodes de Gauss-Seidel et de relaxation. Ainsi, on peut procéder à l'identique de Gauss-Seidel pour paralléliser la méthode de relaxation par points pour une matrice issue de la discrétisation par différences finies. Et la convergence reste garantie si les conditions du théorème 6.10 sont remplies.

Pour les méthodes par blocs, on utilise §4.7 pour en déduire des algorithmes parallèles. Pour la méthode de Jacobi, il s'agit d'une simple adaptation des algorithmes proposés au §6.13.1. Pour la méthode de Gauss-Seidel (et les méthodes de relaxation), pour obtenir une parallélisation optimale, on procède à une renumérotation rouge-noir des blocs. Dans ce dernier cas, la convergence est garantie par le théorème 6.10.

Pour une matrice issue de la discrétisation par éléments finis :

— Pour la méthode de Jacobi : même procédure qu'avant mais l'équirépartition des calculs n'est pas automatique, puisque le nombre de sommets voisins (dans le maillage) n'est pas constant en général... Si la famille de maillages est régulière, on peut toutefois garantir une borne supérieure uniforme du nombre de voisins, valable pour tous les maillages, et tous les sommets les composant.

— Pour la méthode Gauss-Seidel (ou une méthode de relaxation) : en l'absence de structure dans le maillage, la renumérotation n'est pas aussi simple que pour les différences finies. Il faut utiliser un algorithme de coloriage tel que deux sommets du maillage voisins sont nécessairement de deux couleurs différentes. Ceci s'apparente au problème du coloriage d'un graphe, et requiert en général plus de quatre

couleurs. Ainsi, si les communications restent parallèles pour passer des sommets d'une couleur A aux sommets d'une couleur B, il y a en tout 2 fois le nombre de couleurs étapes de communications. En outre, une fois le coloriage réalisé, rien ne garantit que le nombre de sommets d'une couleur donnée est le même pour toutes les couleurs. L'équirépartition des données par nœud n'est donc pas automatiquement équilibrée. Enfin, pour les calculs, il y a le même difficulté d'équirépartition que pour l'algorithme de Jacobi.

Comme on le voit ci-dessus, l'obtention théorique du "parallélisme maximal" est beaucoup plus ardue en l'absence de structure, ce qui est en général le cas pour une discrétisation par éléments finis. L'utilisation des algorithmes par blocs tels que proposés au §4.8 semble plus indiquée dans ce cadre. La convergence de la méthode de Gauss-Seidel (ou de relaxation) reste garantie par le théorème 6.10.

Chapitre 7

Les méthodes de Krylov

7.1 Les espaces de Krylov

Définition 7.1 *Pour une matrice $\mathbb{B} \in \mathbb{K}^{n\times n}$ et un vecteur $y \in \mathbb{K}^n$ donnés, on appelle espace de Krylov associé à $\mathbb{B}$ et y l'espace $\mathcal{K}_m(\mathbb{B}, y)$, $m \in \mathbb{N}^*$ tel que :*

$$\mathcal{K}_m(\mathbb{B}, y) := vect\left(y, \mathbb{B}y, \cdots, \mathbb{B}^{m-1}y\right).$$

Les espaces de Krylov forment un famille croissante de sous-espaces vectoriels de $\mathbb{K}^n$, nécessairement stationnaire, c'est-à-dire constante à partir d'un certain indice m (car on est en dimension finie).

Lemme 7.2 *Par construction, on a :*

1. $\forall p \leq m$, $\mathcal{K}_p(\mathbb{B}, y) \subset \mathcal{K}_m(\mathbb{B}, y)$,
2. $dim(\mathcal{K}_m) \leq min(n, m)$,
3. $\mathbb{B}[\mathcal{K}_m(\mathbb{B}, y)] \subset \mathcal{K}_{m+1}(\mathbb{B}, y)$.

Lemme 7.3 *Si $\mathbb{B}^m y \in \mathcal{K}_m(\mathbb{B}, y)$, alors $\forall p \geq 0$, $\mathbb{B}^{m+p}y \in \mathcal{K}_m(\mathbb{B}, y)$ et $\mathcal{K}_{m+p}(\mathbb{B}, y) = \mathcal{K}_m(\mathbb{B}, y)$.*

Démonstration : Montrons le lemme 7.3 par récurrence. Supposons que $\mathbb{B}^m y \in \mathcal{K}_m(\mathbb{B}, y)$. Montrons que $\mathbb{B}^{m+1}y \in \mathcal{K}_m(\mathbb{B}, y)$. Il existe des coefficients $(\alpha_p)_{p=0,\cdots,m-1} \in \mathbb{K}^m$ tels que : $\mathbb{B}^m y = \sum_{p=0}^{m-1} \alpha_p \mathbb{B}^k y$, d'où :

$$\begin{aligned} \mathbb{B}^{m+1}y &= \mathbb{B}(\mathbb{B}^m y) = \mathbb{B}\left(\sum_{p=0}^{m-1} \alpha_p \mathbb{B}^p y\right) \\ &= \sum_{p=0}^{m-1} \alpha_p \mathbb{B}^{p+1} y = \sum_{p=1}^{m-1} \alpha_{p-1} \mathbb{B}^p y + \alpha_{m-1}\mathbb{B}^m y \end{aligned}$$

Les deux termes de la dernière ligne appartiennent à $\mathcal{K}_m(\mathbb{B}, y)$, d'où : $\mathbb{B}^{m+1}y \in \mathcal{K}_m(\mathbb{B}, y)$ et d'après le lemme 7.2 [item 1], $\mathcal{K}_{m+1}(\mathbb{B}, y) = \mathcal{K}_m(\mathbb{B}, y)$.

On raisonne par récurrence sur p : on suppose que pour $p \geq 1$ donné, $\mathbb{B}^{m+p}y \in \mathcal{K}_m(\mathbb{B}, y)$ et $\mathcal{K}_{m+p}(\mathbb{B}, y) = \mathcal{K}_m(\mathbb{B}, y)$. Montrons qu'alors $\mathbb{B}^{m+p+1}y \in \mathcal{K}_m(\mathbb{B}, y)$ et $\mathcal{K}_{m+p+1}(\mathbb{B}, y) = \mathcal{K}_m(\mathbb{B}, y)$. Par hypothèse, il existe des coefficients $(\beta_q)_{q=0,\cdots,m-1} \in \mathbb{K}^m$ tels que : $\mathbb{B}^{m+p}y = \sum_{q=0}^{m-1} \beta_q \mathbb{B}^q y$, d'où :

$$\begin{aligned} \mathbb{B}^{m+p+1}y &= \mathbb{B}(\mathbb{B}^{m+p}y) = \mathbb{B}\left(\sum_{q=0}^{m-1} \beta_q \mathbb{B}^q y\right) \\ &= \sum_{q=0}^{m-1} \beta_q \mathbb{B}^{q+1} y = \sum_{q=1}^{m-1} \beta_{q-1} \mathbb{B}^q y + \beta_{m-1}\mathbb{B}^m y \end{aligned}$$

Les deux termes de la dernière ligne appartiennent à $\mathcal{K}_m(\mathbb{B}, y)$, d'où : $\mathbb{B}^{m+p+1}y \in \mathcal{K}_m(\mathbb{B}, y)$ et d'après le lemme 7.2 [item 1], on a $\mathcal{K}_{m+p+1}(\mathbb{B}, y) = \mathcal{K}_m(\mathbb{B}, y)$. ◇

Lemme 7.4 *Soit m_{max} la dimension maximale des espaces de Krylov $(\mathcal{K}_m(\mathbb{B}, y))_m$. Alors la suite des espaces de Krylov est strictement croissante de 1 à m_{max}, puis elle est constante à partir de m_{max}, c'est-à-dire que :*

1. *on a :*
$$dim(\mathcal{K}_m(\mathbb{B}, y)) = \begin{cases} m & si\ m \leq m_{max}, \\ m_{max} & si\ m > m_{max}. \end{cases}$$
2. *De plus :*
$$\mathcal{K}_1(\mathbb{B}, y) \subsetneq \mathcal{K}_2(\mathbb{B}, y) \subsetneq \cdots \subsetneq \mathcal{K}_m(\mathbb{B}, y) \subsetneq \mathcal{K}_{m+1}(\mathbb{B}, y) \subsetneq \cdots \subsetneq \mathcal{K}_{m_{max}}(\mathbb{B}, y),$$
$$\mathcal{K}_{m_{max}}(\mathbb{B}, y) = \mathcal{K}_{m_{max}+1}(\mathbb{B}, y) = \cdots = \mathcal{K}_n(\mathbb{B}, y).$$
3. *Enfin, pour $1 < m < m_{max}$ et $y' \in \mathcal{K}_m(\mathbb{B}, y)$, linéairement indépendant des vecteurs de $\mathcal{K}_{m-1}(\mathbb{B}, y)$, alors $\mathbb{B}y' \in \mathcal{K}_{m+1}(\mathbb{B}, y)$, et $\mathbb{B}y'$ est linéairement indépendant des vecteurs de $\mathcal{K}_m(\mathbb{B}, y)$.*

Cette dernière propriété est utile en pratique pour construire une base de $\mathcal{K}_m(\mathbb{B}, y)$ (cf. algorithme d'Arnoldi).

Démonstration : Le nombre m_{max} existe car $(\mathcal{K}_m(\mathbb{B}, y))_m$ est une suite croissante de sous-espaces vectoriels de $\mathbb{K}^n$ (lemme 7.2 [item 1]).

Soit m le plus petit entier pour lequel $\mathbb{B}^m y$ est dépendant des vecteurs précédents (i.e. $\mathbb{B}^m y \in \mathcal{K}_m(\mathbb{B}, y)$) : les vecteurs $(y, \mathbb{B}y, \cdots, \mathbb{B}^{m-1}y)$ sont linéairement indépendants et $\mathcal{K}_p(\mathbb{B}, y)$ est de dimension p pour tout $p \leq m$. En particulier, $\mathcal{K}_m(\mathbb{B}, y)$ est de dimension m. D'après le lemme 7.3, pour tout $p > 0$, $\mathcal{K}_{m+p}(\mathbb{B}, y) = \mathcal{K}_m(\mathbb{B}, y)$. On a donc, pour tout $p > 0$, $\mathcal{K}_1(\mathbb{B}, y) \subsetneq \cdots \subsetneq \mathcal{K}_m(\mathbb{B}, y) = \mathcal{K}_{m+p}(\mathbb{B}, y)$. Par définition de m_{max}, on a nécessairement : $m = m_{max}$, ceci démontre les items 1 et 2.

Soit $m \in \{1, \cdots, m_{max} - 1\}$, et $y' \in \mathcal{K}_m(\mathbb{B}, y)$ tel que : $y' = \sum_{p=0}^{m-1} \alpha_p \mathbb{B}^p y$, avec $\alpha_{m-1} \neq 0$.

D'après l'item 2, $y' \notin \mathcal{K}_{m-1}(\mathbb{B}, y)$. On a : $\mathbb{B}y' = \sum_{p=1}^{m} \alpha_{p-1} \mathbb{B}^p y$. D'après l'item 2, on a : $\sum_{p=1}^{m-1} \alpha_{p-1} \mathbb{B}^p y \in \mathcal{K}_m(\mathbb{B}, y)$ et $\alpha_{m-1}\mathbb{B}^m y \in \mathcal{K}_{m+1}(\mathbb{B}, y)$. ◇

Considérons la résolution de $\mathbb{A}x = b$ ($\mathbb{A} \in \mathbb{K}^{n\times n}$ inversible, $b \in \mathbb{K}^n$) avec une méthode itérative, en utilisant une décomposition régulière de $\mathbb{A}$ ($\mathbb{A} = \mathbb{M} - \mathbb{N}$ voir la définition 6.1). Rappelons l'algorithme associé :

$$\left\|\begin{array}{l} \textbf{initialisation} \\ \quad x^0 \in \mathbb{K}^n,\ r^0 = b - \mathbb{A}x^0 \\ \textbf{itérations : pour } k = 0, 1, \cdots, \textbf{faire} \\ \quad x^{k+1} = x^k + \mathbb{M}^{-1}r^k \\ \quad r^{k+1} = b - \mathbb{A}x^{k+1} \\ \quad \textbf{tant que } \|r^{k+1}\| \neq 0 \text{ ou } \|r^{k+1}\| \geq \varepsilon\|r^0\| \\ \textbf{fin} \end{array}\right. \tag{7.1}$$

Faisons trois remarques importantes en pratique :

- La première itération correspond à $k = 0$!
- On a supposé implicitement dans l'algorithme (7.1) que $r^0 \neq 0$ (sinon, il n'est pas nécessaire d'itérer les k !).
- Les deux critères d'arrêt indiqués correspondent respectivement : à une résolution exacte ($\|r^{k+1}\| = 0$) ; à une résolution approchée ($\|r^{k+1}\| \geq \varepsilon\|r^0\|$, pour $\varepsilon > 0$ donné).

C'est un algorithme de recherche de point fixe de la fonction $f(x) = x - \mathbb{M}^{-1}(b - \mathbb{A}x)$. Si l'algorithme est convergent, il converge vers la solution du système linéaire préconditionné $\mathbb{M}^{-1}\mathbb{A}x = \mathbb{M}^{-1}b$.

Nous allons montrer pourquoi l'utilisation des espaces de Krylov est pertinente.

Posons $z^k := \mathbb{M}^{-1}r^k$, de sorte que $x^{k+1} = x^k + z^k$, et $z^k = x^{k+1} - x^k$.

Par récurrence sur k, on a : $x^{k+1} = x^0 + \sum_{i=0}^{k} z^i$. Remarquons d'autre part que :

$$\left\{\begin{array}{rcl} \mathbb{M}x^{k+1} & = & \mathbb{N}x^k + b \\ \mathbb{M}x^k & = & \mathbb{N}x^{k-1} + b \end{array}\right. \text{, d'où : } \mathbb{M}(x^{k+1} - x^k) = \mathbb{N}(x^k - x^{k-1}).$$

On en déduit que $\mathbb{M}z^k = \mathbb{N}z^{k-1}$, soit par récurrence : $z^k = (\mathbb{M}^{-1}\mathbb{N})^k z^0$. Finalement :

$$x^{k+1} = x^0 + \sum_{i=0}^{k} (\mathbb{M}^{-1}\mathbb{N})^i z^0$$

On voit apparaître les puissances successives de $\mathbb{M}^{-1}\mathbb{N}$: la solution x^{k+1} à l'itération $k+1$ de la méthode itérative (7.1) se décompose sur l'espace affine $x^0 + \mathcal{K}_{k+1}(\mathbb{M}^{-1}\mathbb{N}, z^0)$, avec $z^0 = \mathbb{M}^{-1}r^0$.

Lemme 7.5 *La solution de l'algorithme* (7.1) *à l'itération* $k+1$ *est telle que :*

$$x^{k+1} \in x^0 + \mathcal{K}_{k+1}(\mathbb{M}^{-1}\mathbb{A}, \mathbb{M}^{-1}r^0).$$

Démonstration : On raisonne sur les itérations. On a vu que $x^{k+1} \in x^0 + \mathcal{K}_{k+1}(\mathbb{M}^{-1}\mathbb{N}, z^0)$. Or $\mathbb{A} = \mathbb{M} - \mathbb{N}$, et donc $\mathbb{M}^{-1}\mathbb{N} = I_n - \mathbb{M}^{-1}\mathbb{A}$.

On en déduit que : $\mathcal{K}_{k+1}(\mathbb{M}^{-1}\mathbb{N}, z^0) = \mathcal{K}_{k+1}(I_n - \mathbb{M}^{-1}\mathbb{A}, z^0) = \mathcal{K}_{k+1}(\mathbb{M}^{-1}\mathbb{A}, z^0)$. En effet,

d'après la formule de Newton :

$$\begin{aligned}(\mathbb{M}^{-1}\mathbb{N})^k z^0 &= (I_n - \mathbb{M}^{-1}\mathbb{A})^k z^0, \\ &= \sum_{l=0}^{k} (-1)^l \frac{l!}{k!(k-l)!} (\mathbb{M}^{-1}\mathbb{A})^l z^0.\end{aligned}$$

D'où pour tout $k \in \mathbb{N}$, $(\mathbb{M}^{-1}\mathbb{N})^k z^0 \in \mathcal{K}_{k+1}(\mathbb{M}^{-1}\mathbb{A}, z^0)$, et en échangeant les rôles de $\mathbb{N}$ et $\mathbb{A}$, on a réciproquement que $\forall k$, $(\mathbb{M}^{-1}\mathbb{A})^k z^0 \in \mathcal{K}_{k+1}(\mathbb{M}^{-1}\mathbb{N}, z^0)$.

$\diamond$

Les lemmes 7.3 et 7.4 permettent d'établir le théorème ci-dessous, cf. [25] :

Théorème 7.6 *La solution exacte x du système linéaire $\mathbb{A}x = b$ initialisé par x^0 appartient à l'espace affine $x^0 + \mathcal{K}_{m_{max}}(\mathbb{A}, r^0)$, où $r^0 := b - \mathbb{A}x^0$ est le résidu initial.*

Remarque 7.7 *Si on applique le théorème au système linéaire préconditionné $\mathbb{M}^{-1}\mathbb{A}x = \mathbb{M}^{-1}b$ initialisé par x^0, on aboutit à $x \in x^0 + \mathcal{K}_{m'_{max}}(\mathbb{M}^{-1}\mathbb{A}, \mathbb{M}^{-1}r^0)$, avec $m'_{max} \neq m_{max}$. En effet, si le système est "bien préconditionné", on espère que $m'_{max} < m_{max}$: moins d'itérations seront nécessaires pour converger.*

Démonstration : D'après les lemmes 7.3 et 7.4, les vecteurs $\left(r^0, \mathbb{A}r^0, \cdots, \mathbb{A}^{m_{max}-1}r^0\right)$ sont linéairement indépendants, et $\mathbb{A}^{m_{max}}r^0 \in \mathcal{K}_{m_{max}}(\mathbb{A}, r^0)$.
On peut décomposer $\mathbb{A}^{m_{max}}r^0$ sur cette base : $\mathbb{A}^{m_{max}}r^0 = \sum_{l=0}^{m_{max}-1} \alpha_l \mathbb{A}^l r^0$, avec $\alpha_0 \neq 0$. En effet, si $\alpha_0 = 0$, on a alors que $\mathbb{A}^{m_{max}-1}r^0 \in \mathcal{K}_{m_{max}-1}(\mathbb{A}, r^0)$, ce qui contredit que les vecteurs $\left(r^0, \mathbb{A}r^0, \cdots, \mathbb{A}^{m_{max}-1}r^0\right)$ soient linéairement indépendants. On peut donc écrire (on rappelle que $r^0 = b - \mathbb{A}x^0 = \mathbb{A}x - \mathbb{A}x^0$) :

$$\begin{aligned}\frac{1}{\alpha_0}\mathbb{A}^{m_{max}}r^0 &= \sum_{l=1}^{m_{max}-1} \frac{\alpha_l}{\alpha_0}\mathbb{A}^l r^0 + r^0 \\ &= \sum_{l=1}^{m_{max}-1} \frac{\alpha_l}{\alpha_0}\mathbb{A}^l r^0 + \mathbb{A}x - \mathbb{A}x^0\end{aligned} \quad \Leftrightarrow x = x^0 + \frac{1}{\alpha_0}\mathbb{A}^{m_{max}-1}r^0 - \sum_{l=1}^{m_{max}-1} \frac{\alpha_l}{\alpha_0}\mathbb{A}^{l-1}r^0,$$

ce qui se met aussi sous la forme :

$$x = x^0 + \sum_{l=0}^{m_{max}-1} \beta_l \mathbb{A}^l r^0, \text{ avec } \beta_l = \begin{cases} -\dfrac{\alpha_{l+1}}{\alpha_0}, & l = 0, \cdots, k_{max-2}, \\ \dfrac{1}{\alpha_0}, & l = m_{max} - 1. \end{cases}$$

On a donc bien $x \in x^0 + \mathcal{K}_{m_{max}}(\mathbb{A}, r^0)$.

$\diamond$

La suite d'égalité ci-dessus montre également que si $x \in x^0 + \mathcal{K}_k(\mathbb{A}, r^0)$, alors $\mathbb{A}^k r^0 \in \mathcal{K}_k(\mathbb{A}, r^0)$. Autrement dit, les vecteurs $\left(r^0, \mathbb{A}r^0, \cdots, \mathbb{A}^k r^0\right)$ sont linéairement dépendants et d'après le lemme 7.4, $\mathcal{K}_k(\mathbb{A}, r^0) = \mathcal{K}_{m_{max}}(\mathbb{A}, r^0)$. En conclusion, il semble pertinent, à

l'itération k, de chercher la solution itérative du problème $\mathbb{A}x = b$ comme une combinaison linéaire optimale d'éléments de $\mathcal{K}_k(\mathbb{A}, r^0)$. Différentes méthodes, classées sous le nom générique de *méthodes de Krylov* existent. La méthode du gradient conjugué, étudiée au §7.2 est une méthode de Krylov adaptée aux matrices *symétriques définies-positives*. Pour être en mesure de résoudre efficacement des systèmes linéaires plus généraux, on peut utiliser la méthode du résidu minimal généralisée, connue communément sous l'acronyme GMRES pour *Generalized Minimal Residual Method*, et proposée par Y. Saad [33]. Nous étudierons cette méthode au §7.3.

7.2 Méthode du gradient conjugué

7.2.1 Problème de minimisation

Soit à résoudre le problème linéaire suivant :

Problème 7.1 *Trouver $x \in \mathbb{R}^n$ solution de $\mathbb{A}\,x = b$, avec $\mathbb{A} \in \mathbb{R}^{n\times n}$ une matrice* symétrique définie-positive*, et $b \in \mathbb{R}^n$.*

Introduisons la fonctionnelle continue suivante : $\left\{ \begin{array}{rcl} J : \mathbb{R}^n & \to & \mathbb{R}, \\ v & \mapsto & \frac{1}{2}(\mathbb{A}v, v) - (b, v). \end{array} \right.$

Nous allons prouver le théorème 7.8 ci-dessous, qui fait le lien entre la résolution du problème 7.1 et la minimisation de la fonctionnelle J.

Théorème 7.8 *Supposons* $\mathbb{A}$ symétrique définie-positive*, alors :*

1. $\lim\limits_{\|v\|\to+\infty} J(v) = +\infty$ *et J est strictement convexe sur $\mathbb{R}^n$.*
2. *J admet un minimum* unique *en v_{min} tel que* $\mathbf{grad}\, J(v_{min}) = \mathbf{0}_n$.
3. *$v_{min} = x$ (la solution au problème 7.1).*

Démonstration :

1. Considérons $(\mathrm{v}_i)_{i\in\{1,\cdots,n\}}$ une base orthonormale de vecteurs propres de $\mathbb{A}$, avec $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$ les valeurs propres associées ordonnées (voir la proposition A.20). Comme $\mathbb{A}$ est définie positive, $\lambda_i > 0$: $\mathbb{A}\mathrm{v}_1 = \lambda_1\,\mathrm{v}_1$, de sorte que $0 < (\mathbb{A}\mathrm{v}_1, \mathrm{v}_1) = \lambda_1\|\mathrm{v}_1\|^2 = \lambda_1$.
   Si on écrit $v = \sum_{i=1}^{n} v_i\,\mathrm{v}_i$, on a alors : $(\mathbb{A}v, v) = \sum_i \lambda_i v_i^2 \ge \lambda_1\|v\|^2$, d'où :
   $J(v) \ge \frac{1}{2}\lambda_1\|v\|^2 - (b, v)$, et donc $\lim\limits_{\|v\|\to\infty} J(v) = +\infty$.
   Montrons que J est strictement convexe.
   Considérons $\theta \in [0, 1]$ et v, $w \in \mathbb{R}^n$ tels que $v \ne w$. On a d'une part :

$$J(\theta v+(1-\theta)w) = \frac{1}{2}\theta^2(\mathbb{A}v, v)+\theta(1-\theta)(\mathbb{A}v, w)+\frac{1}{2}(1-\theta)^2(\mathbb{A}w, w)-\theta(b, v)-(1-\theta)(b, w),$$

   et d'autre part :

$$\theta J(v) + (1-\theta)J(w) = \frac{1}{2}\theta(\mathbb{A}v, v) + \frac{1}{2}(1-\theta)(\mathbb{A}w, w) - \theta(b, v) - (1-\theta)(b, w),$$

de sorte que la différence entre ces deux derniers calculs donne :

$$J(\theta v + (1-\theta)w) - \theta J(v) - (1-\theta)J(w)$$

$$= -\frac{1}{2}\theta(1-\theta)(\mathbb{A}v, v) - \frac{1}{2}\theta(1-\theta)(\mathbb{A}w, w) + \theta(1-\theta)(\mathbb{A}v, w)$$

$$= -\frac{1}{2}\theta(1-\theta)(\mathbb{A}(v-w), (v-w)) < 0.$$

2. Montrons d'abord l'*existence* d'un minimum. Par définition de l'infinum, il existe $(v_k)_{k\in\mathbb{N}}$ une suite minimisante : $\lim_{k\to+\infty} J(v_k) = \inf_{v\in\mathbb{R}^n} J(v)$. Et, comme $\lim_{\|v\|\to\infty} J(v) = +\infty$, $(v_k)_{k\in\mathbb{N}}$ est bornée : $(v_k)_{k\in\mathbb{N}} \in K$ où $K \subset \mathbb{R}^n$ est un compact. D'après le théorème de Bolzano-Weierstrass, il existe une sous-suite $(v_{k'})_{k'\in\mathbb{N}}$ convergeant vers v_{min} dans K.
   Puisque J est continue, on a que $J(v_{min}) = \lim_{k'\to+\infty} J(v_{k'}) = \inf_{v\in\mathbb{R}^n} J(v)$.
   Le résultat d'*unicité* résulte de la convexité *stricte* de J, et peut être prouvé par contradiction : soient v_1 et v_2 tels que $v_1 \neq v_2$ et $J(v_1) = J(v_2) = \min_{v\in\mathbb{R}^n} J(v)$.
   Soit $\theta \in]0,1[$. On a : $J(\theta v_1 + (1-\theta)v_2) < \theta J(v_1) + (1-\theta)J(v_2) := \min_{v\in\mathbb{R}^n} J(v)$, ce qui est absurde.
   Montrons que $\mathbf{grad}\, J(v_{min}) = 0$. Soit $h \in \mathbb{R}^n$, $h \neq \mathbf{0}_n$.
   On a par définition : $(\mathbf{grad}\, J(v_{min}), h) = \lim_{\varepsilon\to 0} \frac{1}{\varepsilon}\left(J(v_{min} + \varepsilon h) - J(v_{min})\right) \geq 0$, et également $(\mathbf{grad}\, J(v_{min}), -h) \geq 0$, d'où le résultat.

3. Calculons $(\mathbf{grad}\, J(v), h) = \lim_{\varepsilon\to 0} \frac{1}{\varepsilon}\left(J(v+\varepsilon h) - J(v)\right)$, pour $h \in \mathbb{R}^n$:

$$\begin{aligned}(\mathbf{grad}\, J(v), h) &= \lim_{\varepsilon\to 0} \frac{1}{\varepsilon}\left(\frac{1}{2}(\mathbb{A}v, v) + \varepsilon(\mathbb{A}v, h) + \varepsilon^2(\mathbb{A}h, h) - (b, v) - \varepsilon(b, h)\right. \\ &\qquad \left. -\frac{1}{2}(\mathbb{A}v, v) + (b, v)\right) \\ &= (\mathbb{A}v - b, h).\end{aligned}$$

   On en déduit que $\mathbb{A}v_{min} - b = \mathbf{0}_n$, c'est-à-dire $v_{min} = x$.

◇

7.2.2 Caractérisation du minimum

On a donc établi que $x = \underset{y\in\mathbb{R}^n}{\operatorname{argmin}} J(y)$. Considérons $(p^l)_{l\in\{0,\cdots,n-1\}}$ une base de $\mathbb{R}^n$, orthogonale par rapport au produit scalaire $(\mathbb{A}\cdot, \cdot)$, ce que l'on note $\perp_{\mathbb{A}}$. Soit $x^0 \in \mathbb{R}^n$ et x la solution du problème 7.1. On écrit : $x - x^0 := \sum_{l=0}^{n-1} \alpha_l p^l$, et on considère la fonction F :

$$\left\{\begin{array}{rcl} F : \mathbb{R}^n & \to & \mathbb{R} \\ (\alpha_l^*)_l & \mapsto & J\left(x^0 + \sum_{l=0}^{n-1} \alpha_l^* p^l\right) \end{array}\right. \text{, de sorte que :}$$

$$\begin{aligned} F((\alpha_l^*)_l) &= \frac{1}{2}(\mathbb{A}x^0, x^0) + \frac{1}{2}\sum_{k,l} \alpha_k^*\, \alpha_l^*(\mathbb{A}p^k, p^l) \\ &\quad + \sum_{l=0}^{n-1} \alpha_l^*(\mathbb{A}x^0, p^l) - (b, x^0) - \sum_{l=0}^{n-1} \alpha_l^*(b, p^l) \\ &= J(x^0) + \frac{1}{2}\sum_{k,l} \alpha_k^*\, \alpha_l^*(\mathbb{A}p^k, p^l) + \sum_{l=0}^{n-1} \alpha_l^*(\mathbb{A}x^0 - b, p^l). \end{aligned}$$

En utilisant le fait que $(p^l)_l$ est $\perp_{\mathbb{A}}$, on obtient : $\sum_{k,l} \alpha_k^*\, \alpha_l^*(\mathbb{A}p^k, p^l) = \sum_l (\alpha_l^*)^2(\mathbb{A}p^l, p^l)$,

et finalement : $F((\alpha_l^*)_l) = J(x^0) + \frac{1}{2}\sum_l (\alpha_l^*)^2(\mathbb{A}p^l, p^l) + \sum_l \alpha_l^*(\mathbb{A}x^0 - b, p^l)$.

La fonction F atteint son minimum pour $(\alpha_l)_l$ tel que pour tout l, $\dfrac{\partial F}{\partial \alpha_l^*}((\alpha_l)_l) = 0$, c'est-à-dire pour $(\alpha_l)_l$ tel que pour tout l, $\alpha_l(\mathbb{A}p^l, p^l) + (\mathbb{A}x^0 - b, p^l) = 0$. Ainsi, $x = v_{min}$ se met sous la forme :

$$x = x^0 + \sum_{l=0}^{n-1} \alpha_l\, p^l \text{ avec } \alpha_l := \frac{(b - \mathbb{A}x^0, p^l)}{(\mathbb{A}p^l, p^l)} \text{ pour } l = 0, \cdots, n-1. \tag{7.2}$$

A partir d'une base $(p^l)_l$ et avec un vecteur initial x^0, on peut construire une solution au problème 7.1. Pour $k = 1, \cdots, n-1$, on pose :

$$\left\{\begin{array}{rcll} x^k & := & x^0 + \sum_{l=0}^{k-1} \alpha_l\, p^l & \text{la solution à l'itération } k, \\ r^k & := & b - \mathbb{A}x^k & \text{le résidu à l'ordre } k. \end{array}\right.$$

Remarquons que pour $k = 0, \cdots, n-2$:

$$\left\{\begin{array}{rcl} x^{k+1} &=& x^0 + \sum_{l=0}^{k} \alpha_l\, p^l = x^k + \alpha_k\, p^k, \\ r^{k+1} &=& b - \mathbb{A}x^{k+1} = b - \mathbb{A}x^k - \alpha_k \mathbb{A}p^k = r^k - \alpha_k \mathbb{A}p^k. \end{array}\right. \tag{7.3}$$

En outre, pour $k = 0, \cdots, n-2$:

$$\begin{aligned} (r^{k+1}, p^{k+1}) = (b - \mathbb{A}x^{k+1}, p^{k+1}) &= (b - \mathbb{A}(x^0 + \textstyle\sum_{l \le k} \alpha_l\, p^l), p^{k+1}), \\ &= (b - \mathbb{A}x^0, p^{k+1}) - \textstyle\sum_{l \le k} \alpha_l(\mathbb{A}p^l, p^{k+1}), \\ &= (b - \mathbb{A}x^0, p^{k+1}) = (r^0, p^{k+1}) \text{ car } (p^k)_k \perp_{\mathbb{A}}. \end{aligned}$$

On en déduit que $\forall k = 0, \cdots, n-1$, $(r^0, p^k) = (r^k, p^k)$.
On propose alors l'algorithme suivant pour calculer x :

$$
\begin{array}{l}
\textbf{initialisation} \\
\quad x^0 \in \mathbb{R}^n \\
\quad r^0 = b - \mathbb{A}x^0 \\
\quad p^0 = r^0 \\
\textbf{itérations : pour } k = 0, \cdots, n-2, \textbf{ faire} \\
\quad \alpha_k = \dfrac{(r^0, p^k)}{(\mathbb{A}p^k, p^k)},\ x^{k+1} = x^k + \alpha_k p^k \\
\quad r^{k+1} = r^k - \alpha_k \mathbb{A}p^k \\
\quad \text{construire } p^{k+1} \text{ tel que } (p^l)_{l=0,\cdots,k+1} \text{ est } \perp_{\mathbb{A}} \\
\textbf{fin}
\end{array}
\tag{7.4}
$$

Pour le choix de p^{k+1}, on propose dans la suite une récurrence courte, à deux termes, faisant intervenir uniquement des quantités déterminées à l'itération courante et l'itération précédente. Choisissons donc p^{k+1} tel que $p^{k+1} = r^{k+1} + \beta_k p^k$ vérifiant $(\mathbb{A}p^{k+1}, p^k) = 0$.
Par définition : $(\mathbb{A}p^{k+1}, p^k) = (\mathbb{A}r^{k+1}, p^k) + \beta_k(\mathbb{A}p^k, p^k) = 0$, d'où :

$$\beta_k = -\frac{(\mathbb{A}r^{k+1}, p^k)}{(\mathbb{A}p^k, p^k)} \text{ pour } k = 0, \cdots, n-2.$$

Remarque 7.9 *La condition $(p^k)_k$ est $\perp_{\mathbb{A}}$ implique $(\mathbb{A}p^{k+1}, p^k) = 0$ et on verra ci-dessous que la réciproque est vraie pour $\mathbb{A}$ symétrique définie-positive avec le choix $x^{k+1} = r^{k+1} + \beta^k p^k$! On renvoie à la propriété (i) de la proposition 7.10. Cette récurrence courte (à deux termes) ne fonctionne pas en général, c'est-à-dire avec $\mathbb{A}$ quelconque !*

On note que l'algorithme (7.4) converge en au plus $n-1$ itérations d'après la formule (7.2). Mais il peut converger en moins d'itérations : en effet, si $r^{k+1} = 0$ pour $k < n-2$, alors on a $b = \mathbb{A}x^{k+1}$ et x^{k+1}est la solution cherchée. Dans ce cas, on interrompt l'algorithme.

7.2.3 Algorithme du gradient conjugué

On parle de *conjugaison* car les directions $(p^k)_k$ sont conjuguées par rapport à $\mathbb{A}$: $(\mathbb{A}p^k, p^l) = 0$ si $k \neq l$. L'algorithme du gradient conjugué, lorsque la matrice $\mathbb{A}$ est symétrique définie-positive, s'écrit alors :

$$
\begin{array}{l}
\textbf{initialisation} \\
\quad x^0 \in \mathbb{R}^n \\
\quad r^0 = b - \mathbb{A}x^0 \\
\quad p^0 = r^0 \\
\textbf{itérations : pour } k = 0, 1, \cdots, \textbf{ faire} \\
\quad \alpha_k = \dfrac{(r^k, p^k)}{(\mathbb{A}p^k, p^k)},\ x^{k+1} = x^k + \alpha_k p^k \\
\quad r^{k+1} = r^k - \alpha_k \mathbb{A}p^k \\
\quad \beta_k = -\dfrac{(\mathbb{A}r^{k+1}, p^k)}{(\mathbb{A}p^k, p^k)},\ p^{k+1} = r^{k+1} + \beta_k p^k \\
\quad \textbf{tant que } \|r^{k+1}\| \neq 0. \\
\textbf{fin}
\end{array}
\tag{7.5}
$$

On vérifie par récurrence sur $k = 0, \cdots$ que r^k défini par (7.5) est bien le résidu, c'est-à-dire que $r^k = b - \mathbb{A}x^k$. C'est vrai pour $k = 0$ (initialisation). Supposons que c'est vrai pour k. Alors, si on exécute l'algorithme, on a : $r^{k+1} = r^k - \alpha_k \mathbb{A}p^k = r^k + \mathbb{A}(x^k - x^{k+1}) = (r^k + \mathbb{A}x^k) - \mathbb{A}x^{k+1} = b - \mathbb{A}x^{k+1}$.

Proposition 7.10 *Supposons que la matrice $\mathbb{A}$ soit symétrique définie-positive. Alors pour $k = 0, \cdots$:*

$$(\emptyset)\ (r^k, p^k) = \|r^k\|^2.$$

Si la convergence n'est pas atteinte à l'itération k, alors :

$$(i) \quad (\mathbb{A}p^{k+1}, p^l) = 0,\ l \leq k.$$

$$(ii) \quad (r^{k+1}, r^l) = 0,\ l \leq k.$$

$$(iii) \quad (r^{k+1}, p^l) = 0,\ l \leq k.$$

Avant d'effectuer la démonstration, remarquons qu'une conséquence de $(\emptyset)$ est que

$$(r^k, p^k) = 0 \Rightarrow \|r^k\| = 0,$$

c'est-à-dire que si $(r^k, p^k) = 0$, la convergence est atteinte à l'itération k. Pour la démonstration, nous allons utiliser la contraposée : *si la convergence n'est pas atteinte à l'itération k*, on a alors : $(r^k, p^k) \neq 0$, de sorte que $p^k \neq \mathbf{0}_n$ et $(\mathbb{A}p^k, p^k) \neq 0$ *puisque $\mathbb{A}$ est définie-positive*. En particulier, on peut définir α_k et β_k.

Démonstration : Montrons la proposition 7.10 par récurrence sur k.
A l'itération $k = 0$, on dispose de x^0, $r^0 = b - \mathbb{A}x^0$, $p^0 = r^0$, d'où :

$(\emptyset)_{k=0}$ $\quad (r^0, p^0) = \|r^0\|^2$

Si $x^0 \neq x$ (pas de convergence) : α_0 et β_0 existent.

$(i)_{k=0}$ $\quad p^1 = r^1 + \beta_0 p^0 := r^1 - \dfrac{(\mathbb{A}r^1, p^0)}{(\mathbb{A}p^0, p^0)} p^0$:

$$(\mathbb{A}p^1, p^0) = (\mathbb{A}r^1, p^0) - \frac{(\mathbb{A}r^1, p^0)}{(\mathbb{A}p^0, p^0)}(\mathbb{A}p^0, p^0) = 0.$$

$(ii)_{k=0}$ $\quad (r^1, r^0) = (b - \mathbb{A}x^1, r^0) = (b - \mathbb{A}(x^0 + \alpha_0 p^0), r^0)$ cf. déf. x^1,

$$= (r^0 - \alpha_0 \mathbb{A}p^0, r^0) = (r^0 - \alpha_0 \mathbb{A}p^0, p^0) \text{ car } p^0 = r^0,$$

$$= (r^0, p^0) - \alpha_0 (\mathbb{A}p^0, p^0) = 0 \text{ cf. déf. } \alpha_0.$$

$(iii)_{k=0}$ $\quad (r^1, p^0) = (r^1, r^0) = 0$ cf. $(ii)_{k=0}$.

Supposons que $(\emptyset) - (i) - (ii) - (iii)$ soient vraies jusqu'à l'itération k, et que *la convergence ne soit pas atteinte*. Qu'en est-il à l'itération $k+1$?

$$(\emptyset)_{k+1}\ (r^{k+1}, p^{k+1}) = (r^{k+1}, r^{k+1}) + \beta_k (r^{k+1}, p^k) \text{ cf. déf. } p^{k+1},$$

$$= \|r^{k+1}\|^2 \text{ cf. } (iii)_k.$$

Comme *la convergence n'est pas atteinte à l'itération* k, α_{k+1} et β_{k+1} existent.

$$
\begin{array}{lllll}
(iii)_{k+1} & l \leq k+1 : & (r^{k+2}, p^l) & = & (b - \mathbb{A}x^{k+2}, p^l) \text{ cf. déf. } r^{k+2}, \\
 & & & = & (b - \mathbb{A}x^{k+1} - \alpha_{k+1}\mathbb{A}p^{k+1}, p^l) \text{ cf. déf. } x^{k+2}, \\
 & & \Rightarrow (r^{k+2}, p^l) & = & (r^{k+1}, p^l) - \alpha_{k+1}\,(\mathbb{A}p^{k+1}, p^l) \text{ cf. déf. } r^{k+2}, \\
 & \text{si } l \leq k & (r^{k+2}, p^l) & = & 0 \text{ cf. } (iii)_k \text{ et } (i)_k, \\
 & \text{si } l = k+1 & (r^{k+2}, p^l) & = & 0 \text{ cf. déf. } \alpha_{k+1}. \\
(ii)_{k+1} & 0 < l \leq k+1 : & (r^{k+2}, r^l) & = & (r^{k+2}, p^l - \beta_{l-1}p^{l-1}) \text{ cf. déf. } p^{l>0}, \\
 & & & = & (r^{k+2}, p^l) - \beta_{l-1}(r^{k+2}, p^{l-1}) = 0 \text{ cf. } (iii)_{k+1}, \\
 & \text{si } l = 0 & (r^{k+2}, r^0) & = & (r^{k+2}, p^0) = 0 \text{ cf. } (iii)_{k+1}. \\
(i)_{k+1} & l \leq k+1 & (\mathbb{A}p^{k+2}, p^l) & = & (\mathbb{A}r^{k+2}, p^l) + \beta_{k+1}(\mathbb{A}p^{k+1}, p^l) \text{ cf. déf. } p^{k+2}, \\
 & \text{si } l = k+1 & (\mathbb{A}p^{k+2}, p^l) & = & 0 \text{ cf. déf. } \beta_{k+1}, \\
 & \text{si } l \leq k & (\mathbb{A}p^{k+2}, p^l) & = & (\mathbb{A}r^{k+2}, p^l) \text{ cf. } (i)_k, \\
 & & & = & (r^{k+2}, \mathbb{A}p^l) \text{ car } \mathbb{A} \text{ est symétrique}, \\
 & & & = & \dfrac{1}{\alpha_l}(r^{k+2}, r^l - r^{l+1}),\ \alpha_l \neq 0 \text{ cf. } (\emptyset)_{l \leq k}, \\
 & & & = & 0 \text{ cf. } (ii)_{k+1}.
\end{array}
$$

Ceci achève la démonstration. ◇

En outre, on peut vérifier des propriétés additionnelles *tant qu'on n'a pas convergé* :

$$(iv) \qquad \alpha_k = \frac{(r^k, p^k)}{(\mathbb{A}p^k, p^k)} = \frac{\|r^k\|^2}{(\mathbb{A}p^k, p^k)} \text{ cf. } (\emptyset).$$

$$(v) \qquad \beta_k = -\frac{(\mathbb{A}r^{k+1}, p^k)}{(\mathbb{A}p^k, p^k)} = -\frac{(r^{k+1}, \mathbb{A}p^k)}{(\mathbb{A}p^k, p^k)} \text{ car } \mathbb{A} \text{ est symétrique,}$$

$$= -\frac{1}{\alpha_k}\frac{(r^{k+1}, r^k - r^{k+1})}{(\mathbb{A}p^k, p^k)} \text{ voir plus haut,}$$

$$= \frac{1}{\alpha_k}\frac{\|r^{k+1}\|^2}{(\mathbb{A}p^k, p^k)} \text{ cf. } (ii),$$

$$= \frac{\|r^{k+1}\|^2}{\|r^k\|^2}.$$

$$(vi) \qquad dim(\mathtt{E}_k) = k, \text{ où on a défini } \mathtt{E}_k = vect(r^0, \cdots, r^{k-1}), \text{ et d'après } (ii).$$

$$(vii) \qquad \mathtt{E}_k = vect(p^0, \cdots, p^{k-1}) = vect(r^0, \mathbb{A}r^0, \cdots, \mathbb{A}^{k-1}r^0) \text{ par récurrence.}$$

$$(viii) \qquad x_k = \underset{y \in x_0 + \mathtt{E}_k}{\operatorname{argmin}} J(y) = \underset{y \in x_0 + \mathcal{K}_k(\mathbb{A}, r^0)}{\operatorname{argmin}} J(y), \text{ d'après la section 7.2.1.}$$

Montrons la propriété (vii), en se souvenant qu'on calcule r^{k-1} à l'itération $k-2$.

Proposition 7.11 *Si on n'a pas convergé à l'itération $k-2$, alors $dim(\mathtt{E}_k) = k$, avec la convention que l'initialisation correspond à l'itération -1. De plus, $\mathtt{E}_k = vect(p^0, \cdots, p^{k-1}) = vect(r^0, \mathbb{A}r^0, \cdots, \mathbb{A}^{k-1}r^0)$.*

Démonstration : La preuve se fait par récurrence sur la dimension de $\mathtt{E}_k$.
Pour $k = 1$, on a $\mathtt{E}_1 = vect(r^0) = vect(p^0)$, par définition de p^0 et $dim(\mathtt{E}_1) = 1$ si $x^0 \neq x$.
Pour $k = 2$, on a : $p^1 = r^1 + \beta_0 p^0 = r^1 + \beta_0 r^0$.
Si on n'a pas convergé à l'itération 0 $(= k-2)$, $r^1 \neq \mathbf{0}_n$, $\beta_0 := \dfrac{\|r^1\|^2}{\|r^0\|^2} \neq 0$ et d'après (iii) (voir la proposition 7.10) r^1 est orthogonal à r^0, d'où $\mathtt{E}_2 = vect(r^0, r^1) = vect(p^0, p^1)$ et $dim(\mathtt{E}_2) = 2$.
De plus, on a : $r^1 = r^0 - \alpha_0 \mathbb{A}p^0 = r^0 - \alpha_0 \mathbb{A}r^0$, où $\alpha_0 := \dfrac{\|r^0\|^2}{(\mathbb{A}p^0, p^0)} \neq 0$ (on rappelle que α_0 est bien défini car $(\mathbb{A}p^0, p^0) > 0$, la matrice $\mathbb{A}$ étant *définie-positive*.
Comme $(r^0, r^1) = 0$, le vecteur $\mathbb{A}r^0$ est linéairement indépendant de r^0 (et donc des vecteurs de $\mathtt{E}_1$), d'où $\mathbb{A}r^0 \in \mathtt{E}_2$ et $\mathtt{E}_2 = vect(r^0, \mathbb{A}r^0)$.
Soit $k > 2$. On suppose que l'espace vectoriel $\mathtt{E}_k = vect(r^0, \cdots, r^{k-1})$ est tel $dim(\mathtt{E}_k) = k$, et qu'on a l'identité $\mathtt{E}_k = vect(p^0, \cdots, p^{k-1}) = vect(r^0, \mathbb{A}r^0, \cdots, \mathbb{A}^{k-1}r^0)$.
Si on n'a pas convergé à l'itération $k-1$, alors $r^k \neq \mathbf{0}_n$, et d'après l'égalité (iii) de la proposition 7.10, les vecteurs r^l, $l = 0, \cdots, k$ étant orthogonaux entre eux, $dim(\mathtt{E}_{k+1}) = k+1$.
Comme $p^{k-1} \in \mathtt{E}_k$, et que le vecteur $r^k \in \mathtt{E}_{k+1}$ est tel que $r^k \notin \mathtt{E}_k$, alors le vecteur

$p^k := r^k + \beta_{k-1} p^{k-1} \in \mathtt{E}_{k+1}$ est tel que $p^k \notin \mathtt{E}_k$. D'où : $\mathtt{E}_{k+1} = vect(p^0, \cdots, p^k)$.
On a : $r^k = r^{k-1} - \alpha_{k-1} \mathbb{A} p^{k-1}$ (avec $\alpha_{k-1} \neq 0$ car on n'a pas convergé).
Décomposons p^{k-1} sur la base $(r^0, \mathbb{A}r^0, \cdots, \mathbb{A}^{k-2}r^0)$: $p^{k-1} = \sum_{l=0}^{k-2} \mu_l \mathbb{A}^l r^0$.
Comme $p^{k-1} \in \mathtt{E}_k = \mathcal{K}_k(\mathbb{A}, r^0)$ et $p^{k-1} \notin \mathtt{E}_{k-1} = \mathcal{K}_{k-1}(\mathbb{A}, r^0)$, alors $\mu_{k-2} \neq 0$ (cf. lemme 7.4). On a :

$$\begin{aligned} r^k &= r^{k-1} - \alpha_{k-1} \sum_{l=0}^{k-2} \mu_l \mathbb{A}^{l+1} r^0, \\ &= r^{k-1} - \alpha_{k-1} \sum_{l=1}^{k-2} \mu_{l-1} \mathbb{A}^l r^0 - \alpha_{k-1}\, \mu_{k-2} \mathbb{A}^{k-1} r^0. \end{aligned}$$

Les deux premiers termes appartiennent à $\mathtt{E}_{k-1}$, et donc à $\mathcal{K}_{k-1}(\mathbb{A}, r^0)$. Comme $\alpha_{k-1}\, \mu_{k-2} \neq 0$, on a : $r^k \in \mathcal{K}_k(\mathbb{A}, r^0)$. Une analyse dimensionnelle élémentaire permet de conclure que $\mathbb{A}^{k-1} r^0$ est linéairement indépendant des vecteurs de $\mathcal{K}_{k-1}(\mathbb{A}, r^0)$, et que $\mathtt{E}_k = \mathcal{K}_k(\mathbb{A}, r^0)$. $\diamond$

D'après ce qui précède, on a le résultat ci-dessous.

Corollaire 7.12 *L'algorithme du gradient conjugué converge en $n-1$ itérations au plus.*

Remarque 7.13 *Si on résout le Laplacien en $1D$ par la méthode des différences finies, avec $x^0 = \mathbf{0}_n$, on converge en général en* exactement $n-1$ *itérations.*

En pratique, on se donne un critère d'arrêt sur la norme du résidu, qui interrompt l'algorithme avant l'itération $n-2$. En utilisant les expressions (iv) et (v), l'algorithme du gradient conjugué (GC) 7.5 s'écrit alors (avec $\varepsilon > 0$ donné) :

$$\begin{array}{|ll} \textbf{initialisation} & \\ \quad x^0 \in \mathbb{R}^n & \text{l'approximation de la solution } x \\ \quad r^0 = b - \mathbb{A}x^0 & \text{le résidu} \\ \quad p^0 = r^0 & \text{la direction de descente} \\ \textbf{itérations : pour } k = 0, 1, \cdots, \textbf{faire} & \\ \quad \alpha_k = \dfrac{\|r^k\|^2}{(\mathbb{A}p^k, p^k)} & \\ \quad x^{k+1} = x^k + \alpha_k p^k & \text{l'approximation de la solution } x \\ \quad r^{k+1} = r^k - \alpha_k \mathbb{A} p^k & \text{le résidu} \\ \quad \beta_k = \dfrac{\|r^{k+1}\|^2}{\|r^k\|^2} & \\ \quad p^{k+1} = r^{k+1} + \beta_k p^k & \text{la direction de descente} \\ \quad \textbf{tant que } \|r^{k+1}\| \geq \varepsilon \|r^0\| & \text{on n'a pas convergé !} \\ \textbf{fin} & \end{array} \tag{7.6}$$

Remarque 7.14 *Il est judicieux de sauvegarder le calcul de $q^k := \mathbb{A}p^k$ afin d'éviter de le réaliser deux fois par itération !*

7.2.4 Préconditionnement

On introduit : $e(x^k) = (\mathbb{A}(x^k - x), x^k - x)^{1/2}$, qui mesure l'erreur entre la solution calculée à l'itération k et la solution exacte. Ce critère porte à la fois sur le résidu et sur l'erreur de l'itération k.
Le théorème 7.15 ci-dessous permet de quantifier $e(x^k)$ (voir [26]) :

Théorème 7.15 *Soit $\kappa(\mathbb{A})$ le nombre de conditionnement de $\mathbb{A}$, voir* (6.7)*. On a :*

$$e(x^k) \leq 2 \left(\frac{\sqrt{\kappa(\mathbb{A})} - 1}{\sqrt{\kappa(\mathbb{A})} + 1} \right)^k e(x^0).$$

Remarquons que plus $\kappa(\mathbb{A})$ est proche de 1, plus la convergence du gradient conjugué est rapide. On rappelle de plus que $\mathbb{A}$ étant supposée symétrique définie-positive, alors $\kappa(\mathbb{A}) = \dfrac{\lambda_{max}(\mathbb{A})}{\lambda_{min}(\mathbb{A})}$, où $\lambda_{max}(\mathbb{A})$ est la plus grande valeur propre de $\mathbb{A}$ et $\lambda_{min}(\mathbb{A})$ est la plus petite valeur propre de $\mathbb{A}$ (avec la notation de la démonstration du théorème 7.8, on a : $\kappa(\mathbb{A}) = \dfrac{\lambda_1}{\lambda_n}$).

Il est possible d'améliorer le conditionnement de la matrice à l'aide d'un *préconditionnement*. Le principe est le suivant : on considère une matrice $\mathbb{M} \in \mathbb{R}^{n \times n}$ inversible, telle que $\kappa(\mathbb{M}^{-1}\mathbb{A}) << \kappa(\mathbb{A})$, et on multiplie le système linéaire à résoudre par $\mathbb{M}^{-1}$ (on peut alors introduire $e_{PREC}(x^k) = (\mathbb{M}^{-1}\mathbb{A}(x^k - x), x^k - x)^{1/2}$ où $\mathbb{M}^{-1}\mathbb{A}(x^k - x)$ est le résidu du système linéaire préconditionné). Le problème 7.1 se réécrit : trouver $x \in \mathbb{R}^n$ tel que $\mathbb{M}^{-1}\mathbb{A}x = \mathbb{M}^{-1}b$. Comme $\mathbb{M}$ est symétrique définie-positive, d'après la proposition A.20, il existe deux matrices $\mathbb{Q} \in \mathbb{R}^{n \times n}$ orthogonale et $\mathbb{D} \in \mathbb{R}^{n \times n}$ diagonale à coefficients $d_i = \mathbb{D}_{i,i}$, $i = 1, \cdots, n$ strictement positifs, telle que : $\mathbb{M} = \mathbb{Q}\mathbb{D}\mathbb{Q}^T$. On peut définir la matrice égale à $\mathbb{M}^{1/2} = \mathbb{Q}\mathbb{D}^{1/2}\mathbb{Q}^T$, avec $\mathbb{D}^{1/2} = diag(\sqrt{d_i})$; on a donc : $\mathbb{M}^{1/2}\mathbb{M}^{1/2} = \mathbb{Q}\mathbb{D}^{1/2}\mathbb{Q}^T\mathbb{Q}\mathbb{D}^{1/2}\mathbb{Q}^T = \mathbb{M}$ (car $\mathbb{Q}^T\mathbb{Q} = I_n$). Dans ce cas, résoudre le problème 7.1 est équivalent à résoudre :

Problème 7.2 *Trouver $x \in \mathbb{R}^n$ tel que $\mathbb{M}^{-1/2}\mathbb{A}\,\mathbb{M}^{-1/2}y = \mathbb{M}^{-1/2}b$, avec $y = \mathbb{M}^{1/2}x$.*

La matrice $\widetilde{\mathbb{A}} = \mathbb{M}^{-1/2}\mathbb{A}\,\mathbb{M}^{-1/2}$ est symétrique définie-positive : on peut donc utiliser l'algorithme du gradient conjugué. De plus, elle est *semblable* à $\mathbb{M}^{-1}\mathbb{A}$: $\widetilde{\mathbb{A}}$ et $\mathbb{M}^{-1}\mathbb{A}$ ont donc les mêmes valeurs propres et le même nombre de conditionnement.

Tous calculs faits, et en revenant à l'inconnue x, l'algorithme du gradient conjugué

préconditionné (GCP) s'écrit (avec $\varepsilon > 0$ donné) :

$$
\begin{array}{l}
\textbf{initialisation} \\
\quad x^0 \in \mathbb{R}^n \qquad\qquad \text{l'approximation de la solution } x \\
\quad r^0 = b - \mathbb{A}x^0 \qquad\qquad \text{le résidu} \\
\quad \mathbb{M}z^0 = r^0 \\
\quad p^0 = z^0 \qquad\qquad \text{la direction de descente} \\
\textbf{itérations : pour } k = 0, 1, \cdots, \textbf{faire} \\
\quad \alpha_k = \dfrac{(r^k, z^k)}{(\mathbb{A}p^k, p^k)} \\
\quad x^{k+1} = x^k + \alpha_k p^k \qquad\qquad \text{l'approximation de la solution } x \\
\quad r^{k+1} = r^k - \alpha_k \mathbb{A}p^k \qquad\qquad \text{le résidu} \\
\quad \mathbb{M}z^{k+1} = r^{k+1} \\
\quad \beta_k = \dfrac{(r^{k+1}, z^{k+1})}{(r^k, z^k)} \\
\quad p^{k+1} = z^{k+1} + \beta_k p^k \qquad\qquad \text{la direction de descente} \\
\quad \textbf{tant que } \|r^{k+1}\| \geq \varepsilon \|r^0\| \qquad\qquad \text{on n'a pas convergé !} \\
\textbf{fin}
\end{array}
\tag{7.7}
$$

Remarquons que si $\mathbb{M} = I_n$, les algorithmes du GC (7.6) et du GCP (7.7) sont bien identiques. à chaque itération du GCP, on doit résoudre un système linéaire de matrice $\mathbb{M}$ (ligne $\mathbb{M}z^{k+1} = r^{k+1}$), il est judicieux de choisir $\mathbb{M}$ de sorte que ce calcul ait un coût réduit.

En résumé, les propriétés imposées (tout au moins souhaitées !) pour que le préconditionnement $\mathbb{M}$ soit efficace sont :

1. $\mathbb{M}$ symétrique, définie-positive,
2. $\mathbb{M}z = r$ est "facile" à résoudre (via un algorithme parallèle par exemple),
3. $\kappa(\mathbb{M}^{-1}\mathbb{A}) << \kappa(\mathbb{A})$.

La parallélisation de l'algorithme du gradient conjugué (préconditionné ou non) se fait de la même façon que celle des algorithmes itératifs vus précédemment : on parallélise les produits matrice-vecteurs.

7.2.5 Conclusion

D'après les propriétés $(vii) - (viii)$, on a le théorème suivant :

Théorème 7.16 *On considère* $\mathbb{A} \in \mathbb{R}^{n\times n}$ symétrique définie-positive*, et J la fonctionnelle associée au problème 7.1. Alors x^k, la solution à l'itération $k-1$ du gradient conjugué est telle que :*

$$x^k = \underset{y \in x^0 + \mathcal{K}_k(\mathbb{A}, r^0)}{\operatorname{argmin}} J(y).$$

7.3 Le GMRES

D'après ce qui précède, il semble intéressant de chercher la solution itérative du problème $\mathbb{A}x = b$ comme une combinaison linéaire optimale d'éléments de $(\mathcal{K}_k(\mathbb{A}, r^0))_k$. Et, pour les matrices symétriques définies-positives, on a vu qu'on peut en outre réaliser cette recherche à l'aide d'une récurrence courte (à deux termes). Lorsque la matrice n'est pas symétrique définie-positive, différentes méthodes itératives existent, basées sur la recherche (optimale) d'éléments de $(\mathcal{K}_k(\mathbb{A}, r^0))_k$. Elles sont classées sous le nom générique de *méthodes de Krylov*. Selon cette terminologie, la méthode du gradient conjugué est donc une méthode de Krylov adaptée aux matrices symétriques définies-positives. Nous allons maintenant étudier la méthode du résidu minimal généralisé, connue communément sous l'acronyme GMRES pour "Generalized Minimal Residual Method", et proposée par Y. Saad [33].

7.3.1 Problème de minimisation

Soit à résoudre le problème linéaire suivant :

Problème 7.3 *Trouver* $x \in \mathbb{K}^n \,|\, \mathbb{A}\,x = b$, *avec* $\mathbb{A} \in \mathbb{K}^{n\times n}$ *une matrice* inversible, *et* $b \in \mathbb{K}^n$.

On a vu au théorème 7.6 que la solution x du problème 7.3 était telle que $x \in x^0 + \mathcal{K}_{m_{max}}(\mathbb{A}, r^0)$. Lorsque la matrice $\mathbb{A}$ est symétrique-définie positive, d'après le théorème 7.16, à l'itération $k-1$, la solution approchée x^k est égale à $\underset{y\in x^0+\mathcal{K}_k(\mathbb{A},r^0)}{\operatorname{argmin}} J(y)$.
Dans le cas général, on cherche à minimiser la fonctionnelle

$$K : \left\{ \begin{array}{rcl} x^0 + \mathcal{K}_k(\mathbb{A}, r^0) & \to & \mathbb{R} \\ z & \mapsto & \|b - \mathbb{A}z\| \end{array} \right. .$$

Si on écrit $z = x_0 + y$, $y \in \mathcal{K}_k(\mathbb{A}, r^0)$, on a $\|b - \mathbb{A}(x_0 + y)\| = \|r^0 - \mathbb{A}y\|$.
Soit $\varepsilon > 0$ donné. Considérons l'algorithme suivant :

$$\left\| \begin{array}{l} \textbf{initialisation} \\ \quad x^0 \in \mathbb{K}^n \\ \quad r^0 = b - \mathbb{A}x^0 \\ \textbf{itérations : pour } k = 0, 1, \cdots, \textbf{ faire} \\ \quad y^{k+1} = \underset{y\in\mathcal{K}_{k+1}(\mathbb{A},r^0)}{\operatorname{argmin}} \|r^0 - \mathbb{A}y\| \\ \quad x^{k+1} = x^0 + y^{k+1} \\ \quad r^{k+1} = b - \mathbb{A}x^{k+1} \\ \quad \textbf{tant que } \|r^{k+1}\| \geq \varepsilon\|r^0\| \\ \textbf{fin} \end{array} \right. \tag{7.8}$$

Ainsi, pour $k \in \mathbb{N}$, le vecteur $\mathbb{A}y^{k+1}$ est la projection orthogonale de r^0 sur l'espace $\mathbb{A}\mathcal{K}_{k+1}(\mathbb{A}, r^0)$, et y^{k+1} est la projection orthogonale de $\mathbb{A}^{-1}\,r^0$ sur l'espace $\mathcal{K}_{k+1}(\mathbb{A}, r^0)$ pour la norme $\|\cdot\|_{\mathbb{A}}$. Par construction, on a $r^{k+1} = r^0 + \mathbb{A}(x^0 - x^{k+1}) = r^0 - \mathbb{A}y^{k+1}$, avec $y^{k+1} \in \mathcal{K}_{k+1}(\mathbb{A}, r^0)$. On a donc $\mathbb{A}y^{k+1} \in \mathcal{K}_{k+2}(\mathbb{A}, r^0)$ et $r^{k+1} \in \mathcal{K}_{k+2}(\mathbb{A}, r^0)$.

Proposition 7.17 *Tant que $r^{k+1} \neq \mathbf{0}_n$ (on n'a pas convergé), alors $k+1 < m_{max}$.*
Si $r^{k+1} = \mathbf{0}_n$ (on a convergé), alors $m_{max} = k+1$.

Démonstration : Si $r^{k+1} = \mathbf{0}_n$, alors on a $x^{k+1} = x$. Or, d'après le théorème 7.6, $x - x^0 \in \mathcal{K}_{m_{max}}(\mathbb{A}, r^0)$, et d'après l'algorithme (7.8), $x - x^0 = x^{k+1} - x^0 \in \mathcal{K}_{k+1}(\mathbb{A}, r^0)$. D'où : $k+1 = m_{max}$. ◇

Soit $k \in \mathbb{N}$, tel qu'on n'ait pas convergé, et considérons $k+1$ vecteurs de $\mathbb{K}^n$ notés $(v^0, \ldots, v^k)$ et formant une base de $\mathcal{K}_{k+1}(\mathbb{A}, r^0)$.

Décomposons $y^{k+1} \in \mathcal{K}_{k+1}(\mathbb{A}, r^0)$ sur cette base : $y^{k+1} = \sum_{l=0}^{k} \mathrm{y}_l^{k+1} v^l$.

Soit $\mathbb{V}_{k+1} \in \mathbb{K}^{n\times(k+1)}$ la matrice telle que

$$\forall i \in \{1, \cdots, n\}, \forall j \in \{0, \cdots, k-1\}, \quad (\mathbb{V}_{k+1})_{i,j+1} = (v^j)_i,$$

c'est-à-dire dont la $(j+1)^{eme}$ colonne est le vecteur v^j. Soit $\mathrm{y}^{k+1} = (\mathrm{y}_0^{k+1}, \cdots, \mathrm{y}_k^{k+1})^T \in \mathbb{K}^{k+1}$. Remarquons que par construction : $y^{k+1} = \mathbb{V}_{k+1}\mathrm{y}^{k+1}$. Comment déterminer y^{k+1} ?

Proposition 7.18 *On considère une itération k telle que $k+1 \leq m_{max}$.*
La fonctionnelle $f_{k+1} : \left\{ \begin{array}{rcl} \mathbb{K}^{k+1} & \to & \mathbb{R}^+ \\ \mathrm{y} & \mapsto & \|r^0 - \mathbb{A}\mathbb{V}_{k+1}\mathrm{y}\|^2 \end{array} \right.$ est strictement convexe et réalise son minimum pour $\mathrm{y} = \mathrm{y}^{k+1} \in \mathbb{K}^{k+1}$ tel que :

$$\mathbb{Z}_{k+1}\mathrm{y}^{k+1} = \mathrm{q}^{k+1} \text{ avec } \left\{ \begin{array}{rcl} \mathbb{Z}_{k+1} & := & \mathbb{V}_{k+1}^*\mathbb{A}^*\mathbb{A}\mathbb{V}_{k+1}, \\ \mathrm{q}^{k+1} & = & \mathbb{V}_{k+1}^*\mathbb{A}^* r^0. \end{array} \right. \tag{7.9}$$

La matrice $\mathbb{Z}_{k+1} \in \mathbb{C}^{(k+1)\times(k+1)}$ (resp. $\mathbb{R}^{(k+1)\times(k+1)}$) est une matrice hermitienne (resp. symétrique) définie-positive.

Démonstration :

1. Montrons que f_{k+1} est strictement convexe. On procède comme dans la preuve du théorème 7.8. Soit $\theta \in]0,1[$. Soient $\mathrm{x}, \mathrm{y} \in \mathbb{K}^{k+1}$.
   Posons $x = r^0 - \mathbb{A}\mathbb{V}_{k+1}\mathrm{x} \in \mathbb{K}^n$ et $y = r^0 - \mathbb{A}\mathbb{V}_{k+1}\mathrm{y} \in \mathbb{K}^n$. On a :

$$\begin{array}{rcl} f_{k+1}(\theta\mathrm{x} + (1-\theta)\mathrm{y}) & = & \|r^0 - \theta\mathbb{A}\mathbb{V}_{k+1}\mathrm{x} - (1-\theta)\mathbb{A}\mathbb{V}_{k+1}\mathrm{y}\|^2 \\ & = & \|\theta(r^0 - \mathbb{A}\mathbb{V}_{k+1}\mathrm{x}) + (1-\theta)(r^0 - \mathbb{A}\mathbb{V}_{k+1}\mathrm{y})\|^2 \\ & = & \|\theta x + (1-\theta)y\|^2 \\ \text{et : } \theta f_{k+1}(\mathrm{x}) + (1-\theta)f(\mathrm{y}) & = & \theta\|x\|^2 + (1-\theta)\|y\|^2. \end{array}$$

   D'où la stricte convexité , puisque $x \mapsto \|x\|^2$ est strictement convexe.
   Il y a égalité si et seulement si $x = y$, c'est-à-dire $\mathbb{A}\mathbb{V}_{k+1}\mathrm{x} = \mathbb{A}\mathbb{V}_{k+1}\mathrm{y}$, ce qui équivaut à $\mathbb{V}_{k+1}\mathrm{x} = \mathbb{V}_{k+1}\mathrm{y}$ car $\mathbb{A}$ est inversible. Comme $k+1 \leq m_{max}$, $\mathcal{K}_{k+1}(\mathbb{A}, r^0)$ est de dimension $k+1$, et $vect(v^0, \cdots, v^k)$ est une famille libre de $\mathbb{K}^n$. On en déduit que la matrice associée $\mathbb{V}_{k+1}$ est de rang $k+1$, et que son noyau est le vecteur nul. Ainsi, $\mathbb{V}_{k+1}\mathrm{x} = \mathbb{V}_{k+1}\mathrm{y} \Leftrightarrow \mathrm{x} = \mathrm{y}$: le minimum est unique.

2. Le minimum de f_{k+1} est atteint pour y^{k+1} tel que $\mathbf{grad}_{\mathrm{y}}\, f_{k+1}(\mathrm{y}^{k+1}) = \mathbf{0}_{k+1}$.
   C'est-à-dire que $\forall i \in \{0, \cdots, k\}$, $\dfrac{\partial f_{k+1}}{\partial \mathrm{y}_i}(\mathrm{y}^{k+1}) = 0$.
   Posons $w^i = \mathbb{A}v^i$. On a : $\mathbb{A}\mathbb{V}_{k+1} = \mathbb{A}\,(v^0, \cdots, v^k) = (\mathbb{A}\,v^0, \cdots, \mathbb{A}v^k) = (w^0, \cdots, w^k)$.
   Ainsi, pour $\mathrm{y} = (\mathrm{y}_0, \cdots, \mathrm{y}_k)^T \in \mathbb{K}^{k+1}$ on a : $\mathbb{A}\mathbb{V}_{k+1}\mathrm{y} = \displaystyle\sum_{i=0}^{k} \mathrm{y}_i w^i$, d'où :

$$\begin{aligned} f_{k+1}(\mathrm{y}) &= \left\| r^0 - \sum_{i=0}^{k} \mathrm{y}_i w^i \right\|^2 \\ &= \|r^0\|^2 - \sum_{j=0}^{k} \overline{\mathrm{y}}_j (r^0, w^j) - \sum_{i=0}^{k} \mathrm{y}_i (w^i, r^0) + \sum_{i=0}^{k} \mathrm{y}_i \sum_{j=0}^{k} \overline{\mathrm{y}}_j (w^i, w^j). \end{aligned}$$

   On obtient alors :

$$\forall i \in \{0, \cdots, k\},\ \frac{\partial f_{k+1}}{\partial \mathrm{y}_i}(\mathrm{y}) = 2\left(-(w^i, r^0) + \sum_{j=0}^{k} \overline{\mathrm{y}}_j (w^i, w^j) \right). \tag{7.10}$$

   Soit $\mathbb{Z}_{k+1} \in \mathbb{K}^{(k+1)\times(k+1)}$ telle que $\forall i, j \in \{0, \cdots, k\}$, $(\mathbb{Z}_{k+1})_{i,j} = (w^j, w^i)$.
   On a bien $\mathbb{Z}_{k+1}^* = \mathbb{Z}_{k+1}$.
   Soit $\mathrm{q}^{k+1} \in \mathbb{K}^{k+1}$ tel que $\forall i \in \{0, \cdots, k\}$, $\mathrm{q}_i^{k+1} = (r^0, w^i)$.
   Les équations (7.10) s'annulent pour $\mathrm{y} \in \mathbb{K}^{k+1}$ tel que $\overline{\dfrac{\partial f_{k+1}}{\partial \mathrm{y}_i}(\mathrm{y})} = 0$, ou $-\mathrm{q}_i^{k+1} + (\mathbb{Z}_{k+1}\mathrm{y})_i = 0, \forall i \in \{0, \cdots, k\}$. On en déduit que y^{k+1} est tel que : $\mathbb{Z}_{k+1}\mathrm{y}^{k+1} = \mathrm{q}^{k+1}$.
   On laisse le lecteur vérifier que $(\mathbb{V}_{k+1}^* \mathbb{A}^* \mathbb{A} \mathbb{V}_{k+1}) = \mathbb{Z}_{k+1}$ et que $\mathrm{q}^{k+1} = \mathbb{V}_{k+1}^* \mathbb{A}^* r^0$.
3. Soit $\mathrm{x} \in \mathbb{K}^{k+1} \backslash \{\mathbf{0}_{k+1}\}$. On a : $(\mathbb{Z}_{k+1}\mathrm{x}, \mathrm{x}) = (\mathbb{A}\mathbb{V}_{k+1}\mathrm{x}, \mathbb{A}\mathbb{V}_{k+1}\mathrm{x}) = \|\mathbb{A}\mathbb{V}_{k+1}\mathrm{x}\|^2 > 0$ car $\mathbb{V}_{k+1}$ est de rang maximal et $\mathbb{A}$ est inversible.

$\diamond$

L'algorithme (7.8) se réécrit, avec le changement de variable $y^{k+1} = \mathbb{V}_{k+1}\mathrm{y}^{k+1}$:

initialisation
$x^0 \in \mathbb{K}^n$
$r^0 = b - \mathbb{A}x^0$
construire v^0, $\mathbb{V}_1 = (v^0)$
$\mathrm{q}^0 = (r^0, \mathbb{A}v^0)$
itérations : pour $k = 0, 1, \cdots,$ **faire**
calculer $\mathbb{Z}_{k+1} = \mathbb{V}_{k+1}^* \mathbb{A}^* \mathbb{A} \mathbb{V}_{k+1}$
calculer $\mathrm{q}^{k+1} = \mathbb{V}_{k+1}^* \mathbb{A}^* r^0$
résoudre $\mathbb{Z}_{k+1}\mathrm{y}^{k+1} = \mathrm{q}^{k+1}$
$x^{k+1} = x^0 + \mathbb{V}_{k+1}\mathrm{y}^{k+1}$
$r^{k+1} = b - \mathbb{A}x^{k+1}$
tant que $\|r^{k+1}\| \geq \varepsilon \|r^0\|$
sinon
calculer v^{k+1} tel que $vect(v^0, \cdots, v^{k+1}) = \mathcal{K}_{k+2}(\mathbb{A}, r^0)$
poser $\mathbb{V}_{k+2} = (v^0, \cdots, v^{k+1})$
fin tant que
fin pour k
(7.11)

Si la base $(v^i)_{i=0,k}$ de $\mathcal{K}_{k+1}(\mathbb{A}, r^0)$ n'a pas de structure particulière, alors le système linéaire $\mathbb{Z}_{k+1}\mathrm{y}^{k+1} = \mathrm{q}^{k+1}$ peut être mal conditionné. Pour éviter la dégénérescence numérique de la base naturelle $(\mathbb{A}^i\, r^0)_{i=0,k}$ de $\mathcal{K}_{k+1}(\mathbb{A}, r^0)$ [25], on peut appliquer la procédure d'orthonormalisation de Gram-Schmidt (voir le §5.14.4).

7.3.2 Utilisation de l'algorithme d'Arnoldi

La procédure d'orthonormalisation de Gram-Schmidt appliquée à la base $(\mathbb{A}^i\, r^0)_{i=0,k}$ est connue sous le nom d'*algorithme d'Arnoldi* :

initialisation
$x^0 \in \mathbb{K}^n$
$r^0 = b - \mathbb{A}x^0$
$v^0 = r^0/\|r^0\|$
itérations : pour $k = 0, 1, \cdots,$ **faire**
initialiser v^{k+1}
$v^{k+1} = \mathbb{A}v^k$
orthogonaliser v^{k+1} **(procédure de Gram-Schmidt)**
itérations : pour $l = 0, \cdots, k,$ **faire**
$h_{l,k} = (\mathbb{A}v^k, v^l)$ ou bien $h_{l,k} = (v^{k+1}, v^l)$
$v^{k+1} = v^{k+1} - h_{l,k}v^l$
fin pour l
normaliser v^{k+1}
poser $h_{k+1,k} = \|v^{k+1}\|$
tant que $h_{k+1,k} \neq 0$
$v^{k+1} = v^{k+1}/h_{k+1,k}$
fin tant que
fin pour k
(7.12)

Dès que $h_{k+1,k} = 0$, on arrête l'algorithme car dans ce cas les espaces vectoriels $\mathcal{K}_{k+2}(\mathbb{A}, r^0)$ et $\mathcal{K}_{k+1}(\mathbb{A}, r^0)$ sont égaux, alors que $\mathcal{K}_k(\mathbb{A}, r^0) \subsetneq \mathcal{K}_{k+1}(\mathbb{A}, r^0)$. D'après le lemme 7.4, on a atteint la dimension maximale : $k+1 = m_{max}$.
La version $h_{l,k} = (\mathbb{A}v^k, v^l)$ correspond à la procédure de Gram-Schmidt classique, alors que la version $h_{l,k} = (v^{k+1}, v^l)$ correspond à la procédure de Gram-Schmidt modifiée. Ces procédures sont équivalentes lorsque les calculs sont exacts (voir le §4.1) : on a bien construit une base orthonormale.

Soit $\mathbb{H}_{k+2,k+1} \in \mathbb{K}^{(k+2)\times(k+1)}$ la matrice contenant les coefficients $(h_{l,m})_{l=0,k+1;\, m=0,k}$. Pour $l = 0, \cdots, k+1$ et $m = 0, \cdots, k$, on a : $(\mathbb{H}_{k+2,k+1})_{l+1,m+1} = \begin{cases} h_{l,m} & \text{si } l \leq m+1, \\ 0 & \text{si } l > m+1, \end{cases}$, c'est-à-dire :

$$\mathbb{H}_{k+2,k+1} = \begin{pmatrix} h_{0,0} & \cdots & \cdots & h_{0,k} \\ h_{1,0} & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \vdots \\ & \ddots & h_{k,k-1} & h_{k,k} \\ 0 & & 0 & h_{k+1,k} \end{pmatrix}.$$

Soit $\mathbb{H}_{k+1} \in \mathbb{K}^{(k+1)\times(k+1)}$ la **matrice de Hessenberg** supérieure[42] extraite de $\mathbb{H}_{k+2,k+1}$: pour tout $l, m = \{0, \cdots, k\}$, $(\mathbb{H}_{k+1})_{l+1,m+1} = (\mathbb{H}_{k+2,k+1})_{l+1,m+1}$. On a donc :

$$\mathbb{H}_{k+1} = \begin{pmatrix} h_{0,0} & \cdots & \cdots & h_{0,k} \\ h_{1,0} & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \vdots \\ & 0 & h_{k,k-1} & h_{k,k} \end{pmatrix}.$$

Dans la suite, on notera e_1^m le premier vecteur de la base canonique orthonormale de $\mathbb{K}^m$, respectivement I_m la matrice identité, pour $m \in \{1, \cdots, n-1\}$.

Proposition 7.19 *Soient $k+1 < m_{max}$ et $\mathbb{V}_{k+1} \in \mathbb{K}^{n\times(k+1)}$, la matrice dont les colonnes sont les vecteurs $(v^0, \cdots, v^k)$ de l'algorithme d'Arnoldi* (7.12).
On a l'égalité suivante : pour tout $k \in \{0, \cdots, m_{max} - 2\}$:

$$\mathbb{V}_{k+2}\mathbb{H}_{k+2,k+1} = \mathbb{A}\mathbb{V}_{k+1}. \tag{7.13}$$

On en déduit, par construction, que pour tout $k \in \{0, \cdots, m_{max} - 1\}$:

$$\mathbb{V}_{k+1}^*\mathbb{A}\mathbb{V}_{k+1} = \mathbb{H}_{k+1}. \tag{7.14}$$

D'autre part, la matrice $\mathbb{Z}_{k+1}$, construite dans l'algorithme (7.11) *en appliquant la procédure d'orthonormalisation de Gram-Schmidt, est telle que :*

$$\mathbb{Z}_{k+1} = \mathbb{H}_{k+2,k+1}^*\mathbb{H}_{k+2,k+1}. \tag{7.15}$$

Comme $\mathbb{A}$ est inversible et $\mathbb{V}_{k+1}$ de rang maximal, la matrice $\mathbb{H}_{k+1}$ est inversible. Elle correspond à la projection de l'application linéaire associée à $\mathbb{A}$ dans la base des vecteurs d'Arnoldi.
Démonstration :

42. une matrice de Hessenberg supérieure est une matrice carrée presque triangulaire : les termes sous la diagonale sont nuls, sauf ceux se trouvant immédiatement sous la diagonale (c'est-à-dire d'indice $(l+1, l)$).

1. Montrons l'égalité (7.13) par récurrence.
   On considère la procédure de Gram-Schmidt classique.
   Pour l'itération $k = 0$, d'après l'algorithme (7.12), on a : $\mathbb{V}_1 = (v^0)$, et $h_{1,0}v^1 = \mathbb{A}v^0 - h_{0,0}v^0 = \mathbb{A}\mathbb{V}_1 - h_{0,0}v^0$. On a donc : $h_{0,0}v^0 + h_{1,0}v^1 = \mathbb{A}\mathbb{V}_1$, d'où : $\mathbb{V}_2\mathbb{H}_{2,1} = \mathbb{A}\mathbb{V}_1$.
   On suppose qu'à l'itération $k > 0$, on a : $\mathbb{V}_{k+1}\mathbb{H}_{k+1,k} = \mathbb{A}\mathbb{V}_k$. Par construction de $\mathbb{V}_{k+2}$ et $\mathbb{H}_{k+2,k+1}$, on obtient :

$$(\mathbb{V}_{k+2}\mathbb{H}_{k+2,k+1})_{:,1:k} = \mathbb{V}_{k+1}\mathbb{H}_{k+1,k} = \mathbb{A}\mathbb{V}_k = (\mathbb{A}\mathbb{V}_{k+1})_{:,1:k}.$$

   D'après (7.12), v^{k+1} est construit de sorte que :

$$h_{k+1,k}v^{k+1} = \mathbb{A}v^k - \sum_{l=0}^{k} h_{l,k}v^l \Leftrightarrow \sum_{l=0}^{k+1} v^l h_{l,k} = \mathbb{A}v^k$$

   c'est-à-dire sous forme matricielle : $(\mathbb{V}_{k+2}\mathbb{H}_{k+2,k+1})_{:,k+1} = (\mathbb{A}\mathbb{V}_{k+1})_{:,k+1}$.
   D'où : $\mathbb{V}_{k+2}\mathbb{H}_{k+2,k+1} = \mathbb{A}\mathbb{V}_{k+1}$.
2. On déduit l'égalité (7.14) de l'égalité (7.13).
   En effet, on a : $\mathbb{V}_{k+1}^*\mathbb{A}\mathbb{V}_{k+1} = \mathbb{V}_{k+1}^*\mathbb{V}_{k+2}\mathbb{H}_{k+2,k+1}$.
   Pour $l = 1, \cdots, k+1$, les matrices $\mathbb{V}_l$ sont formées de vecteurs (colonne) unitaires, deux à deux orthogonaux, et sont construites de sorte que : $\mathbb{V}_{l+1} = (\mathbb{V}_l, v^l)$, avec $\forall m \in \{0, \cdots, l-1\}$, $(v^m, v^l) = 0$.
   On en déduit que : $(\mathbb{V}_{k+1}^*\mathbb{V}_{k+2})_{:,1:k+1} = I_{k+1}$, et $(\mathbb{V}_{k+1}^*\mathbb{V}_{k+2})_{:,k+2} = \mathbf{0}_{k+1}$.
   D'où : $\mathbb{V}_{k+1}^*\mathbb{A}\mathbb{V}_{k+1} = \mathbb{H}_{k+1}$.
3. Enfin, l'égalité (7.15) est déduite la définition de $\mathbb{Z}_{k+1}$ et de l'égalité (7.13).

◇

Proposition 7.20 *Le vecteur* q^{k+1} *construit dans l'algorithme* (7.11) *en appliquant la procédure d'orthonormalisation de Gram-Schmidt est tel que :*

$$\mathrm{q}^{k+1} = \|r^0\|\mathbb{H}_{k+2,k+1}^*\mathrm{e}_1, \quad \textit{où } \mathrm{e}_1 = \mathrm{e}_1^{k+2}.$$

Démonstration : On a $\mathrm{q}^{k+1} = \|r^0\|\mathbb{V}_{k+1}^*\mathbb{A}^*v^0$ par définition de v^0.
De plus, $\mathbb{V}_{k+1}^*\mathbb{A}^* = \mathbb{H}_{k+2,k+1}^*\mathbb{V}_{k+2}^*$. Comme $vect(v^0, \cdots, v^{k+1})$ est une famille libre orthonormée, alors $\mathbb{V}_{k+2}^*v^0 = ((v^0, v^0), (v^0, v^1), \cdots, (v^0, v^{k+1}))^T = (1, 0, \cdots, 0)^T = \mathrm{e}_1^{k+2} \in \mathbb{K}^{k+2}$. D'où $\mathbb{V}_{k+1}^*\mathbb{A}^*v^0 = \mathbb{H}_{k+2,k+1}^*\mathbb{V}_{k+2}^*v^0 = \mathbb{H}_{k+2,k+1}^*\mathrm{e}_1^{k+2}$ et la conclusion suit. ◇

Avec application de la procédure d'orthonormalisation de Gram-Schmidt, l'algorithme

(7.11) se réécrit :

$$
\begin{array}{l}
\textbf{initialisation} \\
\quad x^0 \in \mathbb{K}^n \\
\quad r^0 = b - \mathbb{A}x^0 \\
\quad v^0 = r^0/\|r^0\| \\
\quad \mathrm{q}^0 = (q_0) = (r^0, \mathbb{A}v^0) \\
\textbf{itérations : pour } k = 0, 1, \cdots, \textbf{ faire} \\
\quad v^{k+1} = \mathbb{A}v^k \\
\quad \textbf{itérations : pour } l = 0, \cdots, k, \textbf{ faire} \\
\quad\quad h_{l,k} = (\mathbb{A}v^k, v^l) \\
\quad\quad v^{k+1} = v^{k+1} - h_{l,k}v^l \\
\quad \textbf{fin pour } l \\
\quad h_{k+1,k} = \|v^{k+1}\| \\
\quad \textbf{tant que } h_{k+1,k} \neq 0 \\
\quad\quad v^{k+1} = v^{k+1}/h_{k+1,k} \\
\quad\quad \text{construire } \mathbb{H}_{k+2,k+1} \\
\quad\quad \text{calculer } \mathbb{Z}_{k+1} = \mathbb{H}^*_{k+2,k+1}\mathbb{H}_{k+2,k+1} \\
\quad\quad \text{calculer } \mathrm{q}^{k+1} = \|r^0\|\mathbb{H}^*_{k+2,k+1}\mathrm{e}_1 \\
\quad\quad \text{résoudre } \mathbb{Z}_{k+1}\mathrm{y}^{k+1} = \mathrm{q}^{k+1} \\
\quad\quad x^{k+1} = x^0 + \mathbb{V}_{k+1}\mathrm{y}^{k+1} \\
\quad\quad r^{k+1} = b - \mathbb{A}x^{k+1} \\
\quad\quad \textbf{tant que } \|r^{k+1}\| \geq \varepsilon\|r^0\| \\
\quad \textbf{fin tant que} \\
\textbf{fin pour } k
\end{array}
\tag{7.16}
$$

Soit $\mathbf{h}_{k+1} \in \mathbb{K}^{k+1}$ tel que : $\mathbf{h}_{k+1} := (h_{0,k}, \cdots, h_{k,k})^T$. On construit $\mathbb{H}_{k+2,k+1}$ à partir de $\mathbb{H}_{k+1,k}$ de la façon suivante :

$$
\mathbb{H}_{k+2,k+1} = \begin{pmatrix} \mathbb{H}_{k+1} \\ \mathbf{0}_k{}^T \quad h_{k+1,k} \end{pmatrix} = \begin{pmatrix} \mathbb{H}_{k+1,k} & \mathbf{h}_{k+1} \\ \mathbf{0}_k{}^T & h_{k+1,k} \end{pmatrix} \tag{7.17}
$$

Nous allons maintenant étudier le calcul de $\mathbb{Z}_{k+1}$ et la résolution de $\mathbb{Z}_{k+1}\mathrm{y}^{k+1} = \mathrm{q}^{k+1}$.

7.3.3 Factorisation QR de la matrice $\mathbb{H}_{k+2,k+1}$

D'après la proposition 5.28, il existe une matrice unitaire $\mathbb{Q}_{k+2} \in \mathbb{K}^{(k+2)\times(k+2)}$, et une matrice triangulaire supérieure $\mathbb{R}_{k+2,k+1} \in \mathbb{K}^{(k+2)\times(k+1)}$ telles que $\mathbb{H}_{k+2,k+1} = \mathbb{Q}_{k+2}\mathbb{R}_{k+2,k+1}$. La dernière ligne de $\mathbb{R}_{k+2,k+1}$ est nulle, puisque c'est une matrice triangulaire supérieure : $(\mathbb{R}_{k+2,k+1})_{l,m} = 0$ si $l > m$. On utilise explicitement cette propriété ci-dessous. On pose :

$$
\mathbb{R}_{k+1} = (\mathbb{R}_{k+2,k+1})_{1:k+1,:} \in \mathbb{K}^{(k+1)\times(k+1)},\ \widetilde{\mathbb{Q}}_{k+1} = (\mathbb{Q}_{k+2})_{1:k+1,1:k+1} \in \mathbb{K}^{(k+1)\times(k+1)}. \tag{7.18}
$$

On a donc : $\mathbb{R}_{k+2,k+1} = \begin{pmatrix} \mathbb{R}_{k+1} \\ \mathbf{0}_{k+1}{}^T \end{pmatrix}$.

Proposition 7.21 *La matrice $\mathbb{R}_{k+1}$ est inversible, et la résolution du système linéaire de l'algorithme* (7.16) *: Trouver* y^{k+1} *tel que* $\mathbb{Z}_{k+1}\mathrm{y}^{k+1} = \mathrm{q}^{k+1}$ *est équivalente à la résolution du système linéaire :*

$$\textit{Trouver } \mathrm{y}^{k+1} \textit{ tel que } \mathbb{R}_{k+1}\mathrm{y}^{k+1} = \|r^0\|\widetilde{\mathbb{Q}}_{k+1}^*\mathrm{e}_1, \quad \textit{où } \mathrm{e}_1 = \mathrm{e}_1^{k+1}. \tag{7.19}$$

Observons que la matrice $\mathbb{R}_{k+1}$ étant une matrice triangulaire supérieure, on utilise un algorithme de remontée pour résoudre le système linéaire (7.19), dont la complexité est d'ordre $(k+1)^2$ (voir la proposition 5.37).

Démonstration :

— Montrons que $\mathbb{R}_{k+1}$ est inversible. On a :

$$\begin{array}{rcll}
\mathbb{H}_{k+1} & = & (\mathbb{H}_{k+2,k+1})_{1:k+1,:} & \text{par définition,} \\
& = & (\mathbb{Q}_{k+2}\mathbb{R}_{k+2,k+1})_{1:k+1,:} & \text{factorisation QR,} \\
& = & (\mathbb{Q}_{k+2})_{1:k+1,1:k+1}(\mathbb{R}_{k+2,k+1})_{1:k+1,1:k+1} & \text{car } \mathbb{R}_{k+2,k+1} \text{ est triangulaire supérieure,} \\
& = & \widetilde{\mathbb{Q}}_{k+1}\mathbb{R}_{k+1} & \text{par définition.}
\end{array}$$

D'après (7.14), $\mathbb{H}_{k+1}$ est inversible. Les matrices $\widetilde{\mathbb{Q}}_{k+1}$ et $\mathbb{R}_{k+1}$ le sont donc aussi.

— Exprimons $\mathbb{Z}_{k+1}$ à l'aide de de la factorisation QR de la matrice $\mathbb{H}_{k+2,k+1}$:

$$\begin{array}{rcll}
\mathbb{Z}_{k+1} & = & \mathbb{H}_{k+2,k+1}^*\mathbb{H}_{k+2,k+1} & \text{d'après (7.15),} \\
& = & \mathbb{R}_{k+2,k+1}^*\mathbb{Q}_{k+2}^*\mathbb{Q}_{k+2}\mathbb{R}_{k+2,k+1} & \text{factorisation QR,} \\
& = & \mathbb{R}_{k+2,k+1}^*\mathbb{R}_{k+2,k+1} & \text{car } \mathbb{Q}_{k+2} \text{ est unitaire,} \\
& = & \mathbb{R}_{k+1}^*\mathbb{R}_{k+1} & \text{car } \mathbb{R}_{k+2,k+1} \text{ est triangulaire supérieure.}
\end{array}$$

— On note que :

$$\begin{array}{rcll}
\mathbb{H}_{k+2,k+1}^*\mathrm{e}_1^{k+2} & = & \begin{pmatrix} \mathbb{H}_{k+1}^* & \mathbf{0}_k \\ & h_{k+1,k} \end{pmatrix}\begin{pmatrix} \mathrm{e}_1^{k+1} \\ 0 \end{pmatrix} & \text{d'après (7.17),} \\
& = & \mathbb{H}_{k+1}^*\mathrm{e}_1^{k+1}, & \\
& = & \mathbb{R}_{k+1}^*\,\widetilde{\mathbb{Q}}_{k+1}^*\mathrm{e}_1^{k+1} & \text{voir la 1ère étape,}
\end{array}$$

d'où finalement $\mathrm{q}^{k+1} = \|r^0\|\mathbb{R}_{k+1}^*\,\widetilde{\mathbb{Q}}_{k+1}^*\mathrm{e}_1^{k+1}$ d'après la proposition 7.20.

Ainsi, résoudre $\mathbb{Z}_{k+1}\mathrm{y}^{k+1} = \mathrm{q}^{k+1}$ équivaut à résoudre : $\mathbb{R}_{k+1}^*\mathbb{R}_{k+1}\mathrm{y}^{k+1} = \|r^0\|\mathbb{R}_{k+1}^*\widetilde{\mathbb{Q}}_{k+1}^*\mathrm{e}_1^{k+1}$. Puisque $\mathbb{R}_{k+1}^*$ est inversible, c'est également équivalent à la résolution du système linéaire $\mathbb{R}_{k+1}\mathrm{y}^{k+1} = \|r^0\|\widetilde{\mathbb{Q}}_{k+1}^*\mathrm{e}_1^{k+1}$ comme annoncé. ◇

Proposition 7.22 *Posons* $q_{0,k+1} := (\mathbb{Q}_{k+2}^*)_{1,k+2}$. *Le résidu de l'algorithme* (7.16) *est tel que :*

$$\|r^{k+1}\| = \|r^0\|\,|(\mathbb{Q}_{k+2}^*)_{1,k+2}| = \|r^0\|\,|q_{0,k+1}|. \tag{7.20}$$

Démonstration : On a : $r^{k+1} = b - \mathbb{A}x^{k+1} = (b - \mathbb{A}x^0) - \mathbb{A}(x^{k+1} - x^0) = r^0 - \mathbb{A}\mathbb{V}_{k+1}\mathrm{y}^{k+1}$. De plus, on peut réécrire r^0 ainsi : $r^0 = \|r^0\|v^0 = \|r^0\|\mathbb{V}_{k+2}\mathrm{e}_1^{k+2}$. Finalement :

$$\begin{array}{rcll}
r^{k+1} & = & \|r^0\|\mathbb{V}_{k+2}\mathrm{e}_1^{k+2} - \mathbb{V}_{k+2}\mathbb{H}_{k+2,k+1}\mathrm{y}^k & \text{cf. (7.13),} \\
& = & \mathbb{V}_{k+2}\left(\|r^0\|\mathrm{e}_1^{k+2} - \mathbb{H}_{k+2,k+1}\mathrm{y}^{k+1}\right) & \\
& = & \mathbb{V}_{k+2}\left(\|r^0\|\mathrm{e}_1^{k+2} - \mathbb{Q}_{k+2}\mathbb{R}_{k+2,k+1}\mathrm{y}^{k+1}\right) & \text{factorisation QR,} \\
& = & \mathbb{V}_{k+2}\mathbb{Q}_{k+2}\left(\|r^0\|\mathbb{Q}_{k+2}^*\mathrm{e}_1^{k+2} - \mathbb{R}_{k+2,k+1}\mathrm{y}^{k+1}\right) & \mathbb{Q}_{k+2} \text{ est unitaire.}
\end{array}$$

Par ailleurs, en utilisant (7.19) :

$$\|r^0\|\mathbb{Q}_{k+2}^* \mathrm{e}_1^{k+2} - \mathbb{R}_{k+2,k+1}\mathrm{y}^{k+1} = \begin{pmatrix} \|r^0\|\widetilde{\mathbb{Q}}_{k+1}^* \mathrm{e}_1^{k+1} \\ \|r^0\|\overline{q}_{0,k+1} \end{pmatrix} - \begin{pmatrix} \mathbb{R}_{k+1}\mathrm{y}^{k+1} \\ 0 \end{pmatrix} = \begin{pmatrix} \mathbf{0}_{k+1} \\ \|r^0\|\overline{q}_{0,k+1} \end{pmatrix}$$

Or, la matrice $\mathbb{Q}_{k+2}$ est unitaire et on vérifie facilement par le calcul que la matrice $\mathbb{V}_{k+2} \in \mathbb{K}^{n\times(k+2)}$ est telle que $\mathbb{V}_{k+2}^*\mathbb{V}_{k+2} = I_{k+2}$: on en déduit donc (7.20). ◇

L'algorithme (7.16) se réécrit ainsi :

initialisation
  $x^0 \in \mathbb{K}^n$
  $r^0 = b - \mathbb{A}x^0$
  $v^0 = r^0/\|r^0\|$
  $\mathrm{q}^0 = (q_0) = (r^0, \mathbb{A}v^0)$
itérations : pour $k = 0, 1, \cdots,$ **faire**
  $v^{k+1} = \mathbb{A}v^k$
  itérations : pour $l = 0, \cdots, k,$ **faire**
    $h_{l,k} = (\mathbb{A}v^k, v^l)$
    $v^{k+1} = v^{k+1} - h_{l,k}v^l$
  fin pour l
  $h_{k+1,k} = \|v^{k+1}\|$
  tant que $h_{k+1,k} \neq 0$
    $v^{k+1} = v^{k+1}/h_{k+1,k}$
    construire et factoriser $\mathbb{H}_{k+2,k+1} = \mathbb{Q}_{k+2}\,\mathbb{R}_{k+2,k+1}$
    calculer $\mathrm{q}^{k+1} = \|r^0\|\widetilde{\mathbb{Q}}_{k+1}^*\mathrm{e}_1^{k+1}$ et $q_{0,k+1} := (\mathbb{Q}_{k+2})_{1,k+2}$
    résoudre $\mathbb{R}_{k+1}\mathrm{y}^{k+1} = \mathrm{q}^{k+1}$
    $x^{k+1} = x^0 + \mathbb{V}_{k+1}\mathrm{y}^{k+1}$
    tant que $|q_{0,k+1}| \geq \varepsilon$
  fin tant que
fin pour k (7.21)

Il existe plusieurs façons de procéder à la factorisation QR (voir la section 5.14). Nous allons utiliser les rotations de Givens décrites au §5.14.3.

7.3.4 Utilisation des rotations de Givens

Pour réaliser la factorisation de la matrice $\mathbb{H}_{k+2,k+1}$, et donc éliminer ses termes sous-diagonaux, on peut se servir des rotations dans un plan défini par deux vecteurs successifs de la base canonique. Considérons la matrice de rotation de Givens suivante, définie, pour $k > 1$, $l \in \{1, \cdots, k\}$:

$$\mathbb{G}_{k+1}^l = \begin{pmatrix} I_{l-1} & & & \\ & \overline{c_{l-1}} & \overline{s_{l-1}} & \\ & -s_{l-1} & c_{l-1} & \\ & & & I_{k-l} \end{pmatrix} \in \mathbb{K}^{(k+1)\times(k+1)}, \text{ avec } |c_{l-1}|^2 + |s_{l-1}|^2 = 1. \quad (7.22)$$

Par la suite, on note : $\widetilde{\mathbb{G}}_2^{l-1} := \begin{pmatrix} \overline{c_{l-1}} & \overline{s_{l-1}} \\ -s_{l-1} & c_{l-1} \end{pmatrix}$, de sorte que : $\mathbb{G}_{k+1}^{l} = \begin{pmatrix} I_{l-1} & & \\ & \widetilde{\mathbb{G}}_2^{l-1} & \\ & & I_{k-l} \end{pmatrix}$.

Rappelons que la matrice $\mathbb{G}_{k+1}^{l}$ est unitaire.

A l'itération 0 de l'algorithme (7.12), on a : $\mathbb{H}_{2,1} = \begin{pmatrix} h_{0,0} \\ h_{1,0} \end{pmatrix}$.

On considère la matrice $\mathbb{G}_2^1 = \begin{pmatrix} \overline{c_0} & \overline{s_0} \\ -s_0 & c_0 \end{pmatrix}$ telle que $\mathbb{G}_2^1 \, \mathbb{H}_{2,1} = \mathbb{R}_{2,1} := \begin{pmatrix} r_{0,0} \\ 0 \end{pmatrix}$, soit :

$$\left\{ \begin{array}{rcl} \overline{c_0} \, h_{0,0} + \overline{s_0} \, h_{1,0} & = & r_{0,0}, \\ -s_0 \, h_{0,0} + c_0 \, h_{1,0} & = & 0, \\ |c_0|^2 + |s_0|^2 & = & 1. \end{array} \right.$$

On fait le choix suivant : $\left\{ \begin{array}{rcl} c_0 & = & h_{0,0}/r_{0,0} \\ s_0 & = & h_{1,0}/r_{0,0} \end{array} \right.$ avec $r_{0,0} = (|h_{0,0}|^2 + |h_{1,0}|^2)^{1/2}$. On a donc : $\mathbb{H}_{2,1} = \mathbb{Q}_2 \, \mathbb{R}_{2,1}$, avec : $\mathbb{Q}_2 = (\mathbb{G}_2^1)^*$. Enfin, d'après la définition (7.18) pour $k = 0$, on a $\mathbb{R}_1 = r_{0,0}$.

A l'itération 1, on construit les matrices $\mathbb{H}_{3,2} = \begin{pmatrix} \mathbb{H}_{2,1} & \begin{pmatrix} h_{0,1} \\ h_{1,1} \end{pmatrix} \\ 0 & h_{2,1} \end{pmatrix}$ et $\mathbb{G}_3^1 := \begin{pmatrix} \mathbb{G}_2^1 & 0 \\ 0 & 1 \end{pmatrix}$.

Pour réaliser la factorisation de $\mathbb{H}_{3,2}$, on commence par calculer la matrice $\widetilde{\mathbb{H}}_{3,2}$ telle que :

$$\widetilde{\mathbb{H}}_{3,2} = \mathbb{G}_3^1 \, \mathbb{H}_{3,2} = \begin{pmatrix} r_{0,0} & r_{0,1} \\ 0 & \widetilde{h}_{1,1} \\ 0 & h_{2,1} \end{pmatrix} = \begin{pmatrix} \mathbb{R}_1 & r_{0,1} \\ 0 & \widetilde{h}_{1,1} \\ 0 & h_{2,1} \end{pmatrix},$$

avec : $\begin{pmatrix} r_{0,1} \\ \widetilde{h}_{1,1} \end{pmatrix} = \mathbb{G}_2^1 \begin{pmatrix} h_{0,1} \\ h_{1,1} \end{pmatrix}$, c'est-à-dire $\left\{ \begin{array}{rcl} r_{0,1} & = & \overline{c_0} \, h_{0,1} + \overline{s_0} \, h_{1,1} \\ \widetilde{h}_{1,1} & = & -s_0 \, h_{0,1} + c_0 \, h_{1,1} \end{array} \right.$. On remarque qu'il n'est pas nécessaire de stocker $\mathbb{H}_{2,1}$ pour calculer $\widetilde{\mathbb{H}}_{3,2}$, il suffit d'appliquer la rotation aux deux premiers termes de la dernière colonne de $\mathbb{H}_{3,2}$:

$$\begin{pmatrix} r_{0,1} \\ \widetilde{h}_{1,1} \\ h_{2,1} \end{pmatrix} = \begin{pmatrix} \mathbb{G}_2^1 \begin{pmatrix} h_{0,1} \\ h_{1,1} \end{pmatrix} \\ h_{2,1} \end{pmatrix}.$$

La matrice $\mathbb{G}_3^2$ est définie de la façon suivante $\mathbb{G}_3^2 := \begin{pmatrix} 1 & 0 & 0 \\ 0 & \overline{c_1} & \overline{s_1} \\ 0 & -s_1 & c_1 \end{pmatrix} = \begin{pmatrix} 1 & \mathbf{0}_2{}^T \\ \mathbf{0}_2 & \widetilde{\mathbb{G}}_2^1 \end{pmatrix}$, avec les coefficients $\left\{ \begin{array}{rcl} c_1 & = & \widetilde{h}_{1,1}/r_{1,1} \\ s_1 & = & h_{2,1}/r_{1,1} \end{array} \right.$, et $r_{1,1} = (|\widetilde{h}_{1,1}|^2 + |h_{2,1}|^2)^{1/2}$. La rotation s'applique aux deux derniers termes de la dernière colonne de $\widetilde{\mathbb{H}}_{3,2}$. La factorisation de la matrice $\mathbb{H}_{3,2} = \mathbb{Q}_3 \, \mathbb{R}_{3,2}$, avec $\mathbb{Q}_3^* = \mathbb{G}_3^2 \, \mathbb{G}_3^1$ est alors réalisée ainsi :

$$\mathbb{R}_{3,2} := \mathbb{Q}_3^* \, \mathbb{H}_{3,2} = \mathbb{G}_3^2 \, \widetilde{\mathbb{H}}_{3,2} = \begin{pmatrix} \mathbb{R}_1 & r_{0,1} \\ \begin{pmatrix} 0 \\ 0 \end{pmatrix} & \widetilde{\mathbb{G}}_2^1 \begin{pmatrix} \widetilde{h}_{1,1} \\ h_{2,1} \end{pmatrix} \end{pmatrix}.$$

Et on a $\mathbb{R}_2 = (\mathbb{R}_{3,2})_{1:2,:}$ par la définition (7.18) pour $k = 1$.
A l'itération $k > 1$, on doit construire la matrice $\mathbb{H}_{k+2,k+1}$ et les matrices $(\mathbb{G}^l_{k+2})^{k+1}_{l=1}$.
Tout d'abord, les matrices $(\mathbb{G}^l_{k+2})^k_{l=1}$ sont données par (7.22), puisqu'on a calculé les $(c_l)_{l=0,k-1}$ et $(s_l)_{l=0,k-1}$ aux itérations précédentes.
Pour réaliser la factorisation $\mathbb{H}_{k+2,k+1}$, on calcule la matrice $\widetilde{\mathbb{H}}_{k+2,k+1}$ telle que :

$$\widetilde{\mathbb{H}}_{k+2,k+1} = \mathbb{G}^k_{k+2} \cdots \mathbb{G}^1_{k+2}\, \mathbb{H}_{k+2,k+1} = \begin{pmatrix} \mathbb{R}_{k+1,k} & \widetilde{\mathbf{h}}_{k+1} \\ \mathbf{0}_k{}^T & h_{k+1,k} \end{pmatrix},$$

avec $\widetilde{\mathbf{h}}_{k+1} = \mathbb{G}^k_{k+2} \cdots \mathbb{G}^1_{k+2}\, \mathbf{h}_{k+1}$ où $\mathbf{h}_{k+1} = (h_{0,k}, \cdots, h_{k,k})^T$. La matrice $\mathbb{R}_{k+1,k}$ a été calculée à l'itération $k-1$. Le coefficient $h_{k+1,k}$ est calculé à la fin de la procédure d'orthonormalisation dans les algorithmes (7.16) ou (7.21). Il ne reste qu'à calculer le vecteur $\widetilde{\mathbf{h}}_{k+1}$ et la matrice $\mathbb{G}^{k+1}_{k+2}$. Le calcul de $\widetilde{\mathbf{h}}_{k+1}$ est effectué sans conserver les coefficients de la matrice $\mathbb{H}_{k+2,k+1}$. L'algorithme correspondant à ce calcul, et permettant de calculer les coefficients c_k et s_k de la matrice $\mathbb{G}^{k+1}_{k+2}$, s'écrit :

$$\left\| \begin{array}{l} \textbf{pour } l = 0, \cdots, k-1 \\ \quad \text{Mettre à jour } h_{l,k} \text{ et } h_{l+1,k} : \left\{ \begin{array}{rcl} h_{l,k} & = & \overline{c_l}\, h_{l,k} + \overline{s_l}\, h_{l+1,k} \\ h_{l+1,k} & = & -s_l\, h_{l,k} + c_l\, h_{l+1,k} \end{array} \right. . \\ \textbf{fin} \\ r_{k,k} = (|h_{k,k}|^2 + |h_{k+1,k}|^2)^{1/2} \text{ et } \left\{ \begin{array}{rcl} c_k & = & h_{k,k}/r_{k,k} \\ s_k & = & h_{k+1,k}/r_{k,k} \end{array} \right. . \end{array} \right. \tag{7.23}$$

On pose alors $(\widetilde{\mathbf{h}}_{k+1})_{l+1} = h_{l,k}$ pour $l = 0 : k$ et $\mathbb{G}^{k+1}_{k+2} := \begin{pmatrix} I_k & 0 \\ 0 & \widetilde{\mathbb{G}}^k_2 \end{pmatrix}$ avec $\widetilde{\mathbb{G}}^k_2 = \begin{pmatrix} \overline{c_k} & \overline{s_k} \\ -s_k & c_k \end{pmatrix}$.
On calcule $\mathbb{R}_{k+2,k+1} = \mathbb{G}^{k+1}_{k+2}\, \widetilde{\mathbb{H}}_{k+2,k+1}$, de la façon suivante :

$$\mathbb{R}_{k+2,k+1} = \begin{pmatrix} \mathbb{R}_k & (\widetilde{\mathbf{h}}_{k+1})_{1:k} \\ \begin{pmatrix} 0 \\ 0 \end{pmatrix} & \widetilde{\mathbb{G}}^k_2 \begin{pmatrix} (\widetilde{\mathbf{h}}_{k+1})_{k+1} \\ h_{k+1,k} \end{pmatrix} \end{pmatrix}.$$

A partir de là, on identifie la matrice $\mathbb{R}_{k+1}$ définie par (7.18).

D'après ce qui précède, la matrice $\mathbb{Q}_{k+2}$ est telle que $\mathbb{Q}^*_{k+2} = \mathbb{G}^{k+1}_{k+2} \cdots \mathbb{G}^1_{k+2}$.
Calculons $\widetilde{\mathbb{Q}}^*_{k+1} \mathrm{e}_1^{k+1} = \left(\mathbb{Q}^*_{k+2} \mathrm{e}_1^{k+2} \right)_{1:k+1}$.
Pour $l \in \{1, \cdots, k+1\}$, on pose : $\tilde{\mathrm{q}}^l := \mathbb{G}^l_{k+2} \cdots \mathbb{G}^1_{k+2} \mathrm{e}_1^{k+2}$.
Ainsi, pour $l \in \{2, \cdots, k+1\}$ on a : $\tilde{\mathrm{q}}^l = \mathbb{G}^l_{k+2}\, \tilde{\mathrm{q}}^{l-1}$. On en déduit que :

$$\text{Pour } l = 1 : \quad \tilde{\mathrm{q}}^1 = (\overline{c_0}, -s_0, 0, \cdots)^T.$$

$$\text{Pour } l = 2 : \quad \tilde{\mathrm{q}}^2 = (\overline{c_0}, -\overline{c_1}\, s_0, s_0\, s_1, 0, \cdots)^T.$$

On procède par récurrence sur $l \in \{3, \cdots, k+1\}$ pour montrer que

$$\tilde{\mathrm{q}}^l = \begin{pmatrix} (\tilde{\mathrm{q}}^{l-1})_{1:l-1} \\ (-1)^{l-1}\overline{c_{l-1}} \prod_{m=0}^{l-2} s_m \\ (-1)^l \prod_{m=0}^{l-1} s_m \\ \mathbf{0}_{k+1-l} \end{pmatrix}.$$

On suppose la formule vraie pour $\tilde{\mathrm{q}}^{l-1}$. Le calcul de $\tilde{\mathrm{q}}^l$ donne :

$$\tilde{\mathrm{q}}^l = \begin{pmatrix} (\tilde{\mathrm{q}}^{l-1})_{1:l-1} \\ \widetilde{\mathbb{G}}_2^{l-1}(\tilde{\mathrm{q}}^{l-1})_{l:l+1} \\ \mathbf{0}_{k+1-l} \end{pmatrix} \text{ et } \widetilde{\mathbb{G}}_2^{l-1}(\tilde{\mathrm{q}}^{l-1})_{l:l+1} = \begin{pmatrix} \overline{c_{l-1}} & \overline{s_{l-1}} \\ -s_{l-1} & c_{l-1} \end{pmatrix} \begin{pmatrix} (-1)^{l-1} \prod_{m=0}^{l-2} s_m \\ 0 \end{pmatrix},$$

d'où le résultat.
Pour $l = k+1$, on a donc :

$$(\tilde{\mathrm{q}}^{k+1})_{k+1} = (-1)^k \overline{c_k} \prod_{m=0}^{k-1} s_m \quad \text{et} \quad (\tilde{\mathrm{q}}^{k+1})_{k+2} = (-1)^{k+1} \prod_{m=0}^{k} s_m.$$

On en déduit que le second membre de l'équation (7.19), divisé par $\|r^0\|$, s'écrit ainsi :

$$\frac{1}{\|r^0\|}\mathrm{q}^{k+1} := \widetilde{\mathbb{Q}}_{k+1}^* \mathrm{e}_1^{k+1} = (\tilde{\mathrm{q}}^{k+1})_{1:k+1} = \left(\overline{c_0}, -\overline{c_1}s_0, \cdots, (-1)^k \overline{c_k} \prod_{m=0}^{k-1} s_m\right)^T. \tag{7.24}$$

D'après (7.20), on a : $\dfrac{\|r^{k+1}\|}{\|r^0\|} = |(\mathbb{Q}_{k+2})_{1,k+2}|$.

Notons enfin que : $(\mathbb{Q}_{k+2}^*)_{1,k+2} = \left(\mathbb{Q}_{k+2}^* \mathrm{e}_1^{k+2}\right)_{k+2} = (\tilde{\mathrm{q}}^{k+1})_{k+2} = (-1)^{k+1} \prod_{m=0}^{k} s_m$. On a donc établi une nouvelle expression du résidu.

Proposition 7.23 *Si on utilise l'algorithme de factorisation* (7.23)*, le résidu à l'itération k est tel que :*

$$\|r^{k+1}\| = \|r^0\| \prod_{m=0}^{k} |s_m|. \tag{7.25}$$

En utilisant les rotations de Givens pour factoriser la matrice $\mathbb{H}_{k+2,k+1}$ (rappelons qu'on détermine ici $\mathbb{H}_{k+2,k+1}$ à partir de $\mathbb{Q}_{k+2}$ et $\mathbb{R}_{k+2,k+1}$), l'algorithme GMRES (7.21) se ré-

écrit :

initialisation
$x^0 \in \mathbb{K}^n$
$r^0 = b - \mathbb{A}x^0$
$v^0 = r^0/\|r^0\|$
$q_0 = (r^0, \mathbb{A}v^0)$
itérations : pour $k = 0, \cdots,$ **faire**
$v^{k+1} = \mathbb{A}v^k$
itérations : pour $l = 0, \cdots, k,$ **faire**
$h_{l,k} = (\mathbb{A}v^k, v^l)$
$v^{k+1} = v^{k+1} - h_{l,k}v^l$
fin
$h_{k+1,k} = \|v^{k+1}\|$
tant que $h_{k+1,k} \neq 0$
$v^{k+1} = v^{k+1}/h_{k+1,k}$
Calcul de $\mathbb{H}_{k+2,k+1}$ (7.26)
itérations : pour $l = 0, \cdots, k-1,$ **faire**
$$\begin{pmatrix} h_{l,k} \\ h_{l+1,k} \end{pmatrix} = \begin{pmatrix} \overline{c_l} & \overline{s_l} \\ -s_l & c_l \end{pmatrix} \begin{pmatrix} h_{l,k} \\ h_{l+1,k} \end{pmatrix}$$
fin
$r_{k,k} = (|h_{k,k}|^2 + h_{k+1,k}^2)^{1/2}$
$c_k = h_{k,k}/r_{k,k}, \quad s_k = h_{k+1,k}/r_{k,k}$
$h_{k,k} = r_{k,k}, \quad h_{k+1,k} = 0$
Calculs de q^{k+1} et de q_{k+1}
$q_k = \overline{c_k}\, q_k, \quad q_{k+1} = -s_k\, q_k$
résoudre $\mathbb{H}_{k+1}\mathrm{y}^{k+1} = \mathrm{q}^{k+1}$ avec $\mathrm{q}^{k+1} = (q_0, \cdots, q_k)^T$
$x^{k+1} = x^0 + \mathbb{V}_{k+1}\mathrm{y}^{k+1}$
tant que $|q_{k+1}| \geq \varepsilon$
fin tant que
fin

L'algorithme (7.26) est une implémentation possible du GMRES, mais il existe d'autres méthodes de factorisation. Comme pour l'algorithme du gradient conjugué, le GMRES converge en au plus n itérations. A chaque itération k, il faut stocker les matrices $\mathbb{V}_{k+1}$ et $\mathbb{H}_{k+1}$, ce qui requiert de l'espace mémoire et peut être problématique quand k augmente. Une solution (appelée *restarted GMRES* en anglais) consiste à relancer l'algorithme au bout de k_* itérations, où $k_* > 0$ est un nombre d'itérations fixé a priori, en repartant de la dernière valeur de la solution approchée $x^0 = x^{k_*+1}$. Mais on perd la propriété de convergence en un nombre fini d'itérations.

Chapitre 8

Méthode de la puissance itérée

8.1 Introduction

Le but de ce chapitre est de construire des algorithmes de calcul effectif des valeurs propres et vecteurs propres d'une matrice. Dans ce chapitre, est présentée la méthode de la puissance itérée, ainsi que les méthodes dérivées : la puissance itérée avec translation, avec déflation, et la puissance itérée inverse.

8.2 Méthode de la puissance itérée

Tout d'abord, on propose un algorithme faisant intervenir les puissances successives d'une matrice $\mathbb{A}$ de $\mathbb{C}^{n\times n}$ pour calculer une valeur propre de plus grand module et un vecteur propre associé. Si on note $\|\cdot\|$ une norme quelconque de $\mathbb{C}^n$, soit l'algorithme :

1) **initialisation :**
   $v_0 \in \mathbb{C}^n$ tel que $\|v_0\| = 1$
2) **itérations : pour** $k = 1, 2, \ldots$ **faire**
   $v_k = \mathbb{A}v_{k-1}/\|\mathbb{A}v_{k-1}\|$
   fin

Par construction, on a pour tout $k \geq 1$, $\|v_k\| = 1$. On impose cette propriété pour éviter que la norme de ce vecteur tende vers l'infini... A partir des relations de cet algorithme et de la décomposition spectrale de la matrice $\mathbb{A}$ établie au §A.6, on vérifie que

$$v_0 = \sum_{i=1}^{d} \mathbb{P}_i v_0, \quad \mathbb{A}^k v_0 = \sum_{i=1}^{d} (\lambda_i \mathbb{P}_i + \underline{\mathbb{D}}_i)^k \mathbb{P}_i v_0, \ k \geq 1.$$

$$\text{Donc } v_k = \frac{1}{\alpha_k} \sum_{i=1}^{d} (\lambda_i \mathbb{P}_i + \underline{\mathbb{D}}_i)^k \mathbb{P}_i v_0, \text{ avec } \alpha_k = \|\sum_{i=1}^{d} (\lambda_i \mathbb{P}_i + \underline{\mathbb{D}}_i)^k \mathbb{P}_i v_0\|.$$

Théorème 8.1 *Soit $\mathbb{A} \in \mathbb{C}^{n\times n}$. On suppose qu'il n'existe qu'une seule valeur propre λ_1 de plus grand module et que cette valeur propre est semi–simple. Soit v_0 un choix initial possédant une composante non nulle sur $M_1 = Im(\mathbb{P}_1)$ ($\mathbb{P}_1 v_0 \neq 0$). Alors, en appelant r_1 (resp. θ_1) le module (resp. l'argument) de λ_1 ($\lambda_1 = r_1 e^{i\theta_1}$), on peut démontrer successivement que*

(i) $\lim_{k\to\infty} (e^{-\imath k\theta_1}\, v_k) = \dfrac{\mathbb{P}_1 v_0}{\|\mathbb{P}_1 v_0\|}$;
(ii) $\lim_{k\to\infty} \|\mathbb{A} v_k\| = r_1$;
(iii) Soit j telle que $(\mathbb{P}_1 v_0)_j \neq 0$: $\lim_{k\to\infty} \dfrac{v_{k+1}^j}{v_k^j} = e^{\imath\theta_1}$, *avec v^j la $j^{\text{ème}}$ composante de v.*

Remarque 8.2 *Dans le cas où $\lambda_1 \in \mathbb{R}_\star^+$, (ii) signifie que $\lim_{k\to\infty} \|\mathbb{A} v_k\| = \lambda_1$. C'est toujours le cas d'une matrice hermitienne (resp . symétrique) définie-positive de $\mathbb{C}^{n\times n}$ (resp. $\mathbb{R}^{n\times n}$).*

Démonstration : Puisque λ_1 est supposée semi–simple, $\underline{\mathbb{D}}_1 = [0]$ et on écrit

$$v_k = \frac{1}{\alpha_k}\left[\lambda_1^k \mathbb{P}_1 v_0 + \sum_{i=2}^{d} (\lambda_i \mathbb{P}_i + \underline{\mathbb{D}}_i)^k \mathbb{P}_i v_0\right] = \frac{\lambda_1^k}{\alpha_k}\left[\mathbb{P}_1 v_0 + e_k\right],$$
$$\text{avec } e_k = \sum_{i=2}^{d} \frac{1}{\lambda_1^k} (\lambda_i \mathbb{P}_i + \underline{\mathbb{D}}_i)^k \mathbb{P}_i v_0.$$

Le rayon spectral de la matrice

$$\mathbb{Q} = \sum_{i=2}^{d} \frac{1}{\lambda_1} (\lambda_i \mathbb{P}_i + \underline{\mathbb{D}}_i)\mathbb{P}_i,$$

égal à $|\lambda_2|/|\lambda_1|$, est strictement plus petit que 1 par hypothèse. Ainsi la suite de vecteurs $(e_k)_k$ tend vers 0 quand $k \to +\infty$, d'après le Théorème B.18. Par ailleurs,

$$\frac{\alpha_k}{r_1^k} = \frac{1}{r_1^k}\|\lambda_1^k(\mathbb{P}_1 v_0 + e_k)\| = \|\mathbb{P}_1 v_0 + e_k\| \;\to\; \|\mathbb{P}_1 v_0\|.$$

Notons que, d'après la Proposition A.32, puisque par hypothèse $\mathbb{P}_1 v_0 \neq 0$, $\mathbb{P}_1 v_0$ est un vecteur propre associé à λ_1. On introduit maintenant les vecteurs auxiliaires $w_k = e^{-\imath k\theta_1}\, v_k$. On trouve

$$w_k = e^{-\imath k\theta_1} \frac{\lambda_1^k}{\alpha_k}(\mathbb{P}_1 v_0 + e_k) = \frac{r_1^k}{\alpha_k}(\mathbb{P}_1 v_0 + e_k) \;\to\; \frac{\mathbb{P}_1 v_0}{\|\mathbb{P}_1 v_0\|},\ \text{c'est-à-dire (i).}$$

Pour prouver (ii), on remarque que

$$\mathbb{A} v_k = \mathbb{A}\left(e^{\imath k\theta_1} w_k\right) = e^{\imath k\theta_1}\mathbb{A} w_k,\ \text{d'où } \|\mathbb{A} v_k\| = \|\mathbb{A} w_k\|.$$

Comme $(w_k)_k$ est une suite convergente d'après (i), il en est de même pour $(\mathbb{A} w_k)_k$, et

$$\|\mathbb{A} v_k\| = \|\mathbb{A} w_k\| \;\to\; \frac{\|\mathbb{A}\mathbb{P}_1 v_0\|}{\|\mathbb{P}_1 v_0\|} = r_1,\ \text{c'est-à-dire (ii).}$$

Pour prouver (iii), nous considérons une coordonnée j telle que $(\mathbb{P}_1 v_0)_j \neq 0$, ou $(\mathbb{P}_1 v_0, e_j) \neq 0$, avec $(e_j)_j$ la base canonique de $\mathbb{C}^n$.
On a la relation $v_k^j = (v_k, e_j) = e^{\imath k\theta_1}(w_k, e_j)$. D'après (i), (w_k, e_j) tend vers $(\mathbb{P}_1 v_0)_j / \|\mathbb{P}_1 v_0\|$

qui est non nul par hypothèse. Ainsi, il existe k_0 tel que, pour tout $k \geq k_0$, $(w_k, e_j) \neq 0$, et donc $v_k^j \neq 0$. Par ailleurs,

$$v_{k+1}^j = (v_{k+1}, e_j) = \frac{(\mathbb{A}v_k, e_j)}{\|\mathbb{A}v_k\|} = e^{\imath k\theta_1} \frac{(\mathbb{A}w_k, e_j)}{\|\mathbb{A}w_k\|}.$$

Pour $k \geq k_0$, on a donc

$$\frac{v_{k+1}^j}{v_k^j} = \frac{1}{\|\mathbb{A}w_k\|} \frac{(\mathbb{A}w_k, e_j)}{(w_k, e_j)} \rightarrow \frac{\lambda_1}{r_1} = e^{\imath\theta_1}, \text{ c'est-à-dire (iii).}$$

$\diamond$

On note que

1) l'algorithme fournit une valeur propre et un vecteur propre associé. En effet, d'une part (ii-iii) fournissent λ_1 et, d'autre part, (i) fournit un vecteur propre de M_1, puisque $\mathbb{P}_1 v_0$ appartient toujours à ce sous-espace propre.

2) la vitesse de convergence de l'algorithme est lié au rapport $\rho_{1,2} = |\lambda_2|/|\lambda_1|$, où λ_2 est la deuxième valeur propre de plus grand module. De fait, la vitesse de convergence est liée à la façon dont $(e_k)_k$ tend vers 0, ce qui dépend du rayon spectral $\rho(\mathbb{Q})$, qui est lui-même inférieur ou égal à $\rho_{1,2}$ (cf. la discussion du §6.4.)

8.3 Méthode de la puissance inverse itérée

Si on suppose que la matrice $\mathbb{A} \in \mathbb{C}^{n\times n}$ est inversible, alors 0 n'est pas valeur propre. Rangeons les valeurs propres par ordre de module décroissant

$$Spe(\mathbb{A}) = \{\lambda_n, \lambda_{n-1}, \ldots, \lambda_2, \lambda_1\}$$

alors

$$Spe(\mathbb{A}^{-1}) = \{\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, \ldots, \frac{1}{\lambda_{n-1}}, \frac{1}{\lambda_n}\}$$

et les vecteurs propres de $\mathbb{A}$ sont aussi vecteurs propres de $\mathbb{A}^{-1}$: $\mathbb{A}u = \lambda u \Longleftrightarrow \mathbb{A}^{-1}u = \frac{1}{\lambda}u$. Donc si on veut calculer la valeur propre de $\mathbb{A}$ de plus petit module λ_n, on applique l'algorithme de la puissance itérée à la matrice inverse $\mathbb{A}^{-1}$. C'est la méthode de la puissance itérée inverse :

1) **initialisation :**
$v_0 \in \mathbb{C}^n$ tel que $\|v_0\| = 1$
2) **itérations : pour** $k = 1, 2, \ldots$ **faire**
$v_k = \mathbb{A}^{-1}v_{k-1}/\|\mathbb{A}^{-1}v_{k-1}\|$
fin

Cet algorithme fournit la valeur propre de plus grand module de $\mathbb{A}^{-1}$: soit $1/\lambda_n$. La vitesse de convergence est liée, cette fois, au rapport $\rho' = |\lambda_n|/|\lambda_{n-1}|$.

Dans la pratique pour calculer v_k, on peut réaliser une factorisation de la matrice $\mathbb{A}$ par la méthode de Gauss (resp. par la méthode de Cholesky si $\mathbb{A}$ est hermitienne définie-positive), et on résout le système linéaire $LUv_k = g$ (resp. $LL^*v_k = g$) par une technique de descente-remontée. Ou bien, on peut appliquer une méthode itérative pour résoudre le système linéaire $\mathbb{A}v_k = g$.

8.4 Technique de translation

Le problème qui se pose maintenant est comment obtenir les autres valeurs propres, une fois que l'on a calculé les valeurs propres extrêmes ? Une réponse est fournie par la technique de **translation** (**shift** en anglais), qui consiste à rechercher les valeurs propres de la matrice $\mathbb{A} - \sigma\mathbb{I}$. Si le spectre de $\mathbb{A}$ est

$$Spe(\mathbb{A}) = \{\lambda_n, \lambda_{n-1}, \ldots, \lambda_2, \lambda_1\}$$

alors le spectre de la matrice $\tilde{\mathbb{A}} = \mathbb{A} - \sigma\mathbb{I}$ est

$$Spe(\tilde{\mathbb{A}}) = \{\lambda_n - \sigma, \lambda_{n-1} - \sigma, \ldots, \lambda_2 - \sigma, \lambda_1 - \sigma\}.$$

Un choix judicieux de σ, c'est-à-dire tel que $\max_j(|\lambda_j - \sigma|) = |\lambda_i - \sigma|$ avec $\lambda_i \neq \lambda_1$ permet à la méthode de la puissance itérée appliquée à $\tilde{\mathbb{A}}$ de converger vers la valeur propre $\lambda_i - \sigma$ et donc de déterminer $\lambda_i \neq \lambda_1$.

Il faut être prudent dans le choix de σ car on n'obtient pas obligatoirement les valeurs propres dans l'ordre des modules décroissants par cette technique. Par exemple si $Spe(\mathbb{A}) = \{-2, 3, 5\}$, la méthode de la puissance itérée appliquée à $\mathbb{A}$ converge vers $\lambda_1 = 5$; si on l'applique à la matrice $\mathbb{A} - 2\mathbb{I}$, elle converge vers -4 car $Spe(\mathbb{A} - 2\mathbb{I}) = \{-4, 1, 3\}$; on a donc calculé la valeur propre $\lambda_3 = -4 + 2 = -2$ et non $\lambda_2 = 3$!

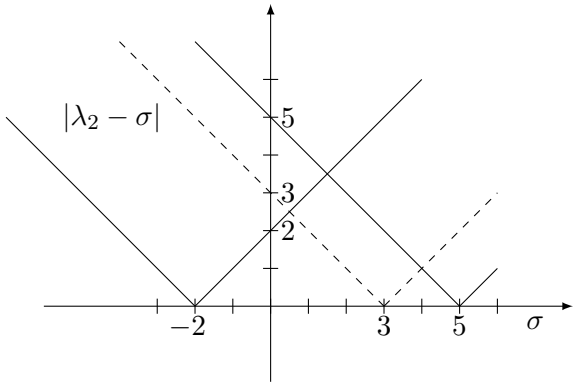


FIGURE 8.1 – Les variations de $|\lambda - \sigma|$

Qui plus est sur le graphique 8.1, on voit que la méthode de translation ne permet pas d'atteindre la valeur propre $\lambda_2 = 3$: pour toute valeur du paramètre σ, la courbe représentant les variations de $|\lambda_2 - \sigma|$ est toujours encadrée par les courbes $|\lambda_1 - \sigma|$ et $|\lambda_3 - \sigma|$.
Pour obtenir λ_2, il faut travailler sur le spectre de $\mathbb{A}^{-1}$, comme le montre la figure 8.2. Dans ce cas, on applique la technique de translation à l'algorithme de la puissance itérée inverse, en factorisant la matrice $\tilde{\mathbb{A}} = \mathbb{A} - \sigma\mathbb{I}$ pour le calcul des itérés successifs... Si λ est la valeur propre la plus proche de σ, alors $\dfrac{1}{\lambda - \sigma}$ est la valeur propre de plus grand module

de $(\mathbb{A} - \sigma\mathbb{I})^{-1}$. La convergence est liée cette fois au rapport

$$\frac{\dfrac{1}{|\lambda' - \sigma|}}{\dfrac{1}{|\lambda - \sigma|}} = \frac{|\lambda - \sigma|}{|\lambda' - \sigma|},$$

avec λ' telle que $|\lambda' - \sigma|$ est le deuxième plus petit module. Ce rapport peut être très petit si σ est proche de λ (et assez éloigné de λ'). La convergence de la méthode est donc très rapide (quelques itérations) si on dispose d'une bonne estimation de λ.

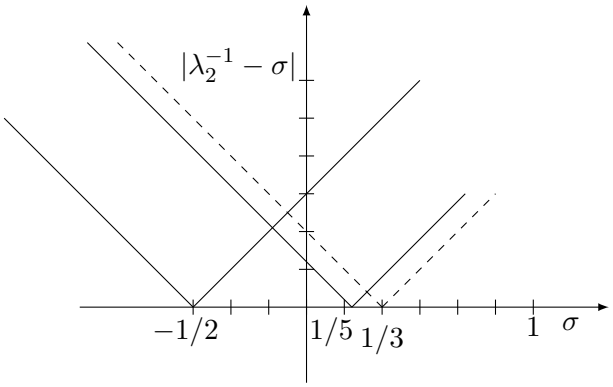


FIGURE 8.2 – Les variations de $|\lambda^{-1} - \sigma|$

Cette méthode est donc utilisée comme accélération de la méthode de la puissance itérée inverse, mais aussi pour le calcul des vecteurs propres lorsque l'on a obtenu une estimation des valeurs propres par un autre algorithme. On voit par ailleurs qu'il n'est pas nécessaire d'avoir une estimation fine de ces valeurs propres puisque la méthode de la puissance itérée inverse fournit des valeurs plus précises.

Remarque 8.3 *Quand la valeur σ est proche de la valeur exacte de λ, la matrice $\mathbb{A}-\sigma\mathbb{I}$ est presque singulière. A priori, ce phénomène pourrait introduire des problèmes numériques au cours de la factorisation de cette matrice. Pour des matrices symétriques, Parlett [31] a montré que les calculs restaient stables et qu'on pouvait donc utiliser cette méthode sans modification.*

8.5 Technique de déflation

Une autre façon de calculer différentes valeurs propres d'une matrice par la méthode de la puissance itérée, consiste à retirer les valeurs propres du spectre de $\mathbb{A}$ de la manière suivante, appelée technique de **déflation**.

On suppose connue une valeur propre λ_j de la matrice $\mathbb{A}$ et un vecteur propre associé u_j, on définit alors la matrice

$$\tilde{\mathbb{A}} = \mathbb{A} - \sigma u_j \cdot v^*$$

où σ est un paramètre complexe et $v^* \in \mathbb{C}^{1\times n}$ un vecteur ligne tel que $v^* u_j = 1$.

Théorème 8.4 *Soit une matrice* $\mathbb{A} \in \mathbb{C}^{n\times n}$ diagonalisable *possédant d valeurs propres distinctes, de spectre*

$$Spe(\mathbb{A}) = \{\lambda_d, \lambda_{d-1}, \ldots, \lambda_2, \lambda_1\}.$$

On suppose que la valeur propre λ_j *est* simple. *Alors la matrice* $\tilde{\mathbb{A}}$ *a pour spectre*

$$Spe(\tilde{\mathbb{A}}) = \{\lambda_d, \lambda_{d-1}, \ldots, \lambda_j - \sigma, \ldots, \lambda_2, \lambda_1\},$$

avec, le cas échéant, $\lambda_j - \sigma \in \{\lambda_i,\ i \neq j\}$.

Démonstration : On sait que les valeurs propres sont associées indifféremment à des vecteurs propres à gauche, ou à droite, cf. la Proposition A.12. Soit donc, pour $\lambda_i \neq \lambda_j$, $w_i^* \in \mathbb{C}^{1\times n}$ un vecteur propre à gauche de $\mathbb{A}$. D'après la Proposition A.13, comme λ_i est distincte de λ_j, on a $w_i^* u_j = 0$:

$$w_i^* \tilde{\mathbb{A}} = w_i^* \mathbb{A} - \sigma(w_i^* u_j) v^* = w_i^* \mathbb{A} = \lambda_i w_i^*.$$

Ainsi λ_i est valeur propre de $\tilde{\mathbb{A}}$, et w_i^* vecteur propre à gauche de $\mathbb{A}$ est aussi vecteur propre à gauche de $\tilde{\mathbb{A}}$. Ceci est valable pour tout vecteur propre à gauche associé à λ_i : les ordres de multiplicité de $\tilde{\mathbb{A}}$ et $\mathbb{A}$ vérifient donc

$$m_i(\tilde{\mathbb{A}}) \geq m_i(\mathbb{A}),\ \text{pour } i \neq j,\ \text{d'où } \sum_{i\neq j} m_i(\tilde{\mathbb{A}}) \geq n-1.$$

D'autre part

$$\tilde{\mathbb{A}} u_j = \mathbb{A} u_j - \sigma u_j (v^* u_j) = \mathbb{A} u_j - \sigma u_j = (\lambda_j - \sigma) u_j.$$

Donc u_j est un vecteur propre associé à la valeur propre $\lambda_j - \sigma$ de $\tilde{\mathbb{A}}$. Deux cas peuvent se présenter :

- $\lambda_j - \sigma \notin \{\lambda_i,\ i \neq j\}$: son ordre de multiplicité est de 1 pour $\tilde{\mathbb{A}}$, et on a bien retrouvé toutes les valeurs propres de $\tilde{\mathbb{A}}$ (avec le même ordre de multiplicité que pour $\mathbb{A}$.)
- $\lambda_j - \sigma = \lambda_i$, pour $i \neq j$. On note que, d'après la Proposition A.26, il existe m_i vecteurs propres à gauche indépendants $(w_{i,k}^*)_{1\leq k\leq m_i}$ associés à λ_i tels que $w_{i,k}^* u_j = 0$, ou $(u_j, w_{i,k}) = 0$. Ainsi la famille $(w_{i,1}, \cdots, w_{i,m_i}, u_j)$ est libre, et $m_i(\tilde{\mathbb{A}}) = m_i(\mathbb{A}) + 1$. On a également retrouvé toutes les valeurs propres de $\tilde{\mathbb{A}}$.

$\diamond$

Quels sont les autres vecteurs propres à droite de la matrice $\tilde{\mathbb{A}}$? On les cherche sous la forme $\tilde{u}_i = u_i - \gamma_i u_j$ pour $i \neq j$:

$$\tilde{\mathbb{A}} \tilde{u}_i = (\mathbb{A} - \sigma u_j \cdot v^*)(u_i - \gamma_i u_j) = \lambda_i u_i - (\gamma_i(\lambda_j - \sigma) + \sigma v^* u_i)\, u_j.$$

Pour que $\tilde{u}_i$ soit vecteur propre de $\tilde{\mathbb{A}}$ associé à λ_i, il faut et il suffit que

$$\lambda_i u_i - (\gamma_i(\lambda_j - \sigma) + \sigma v^* u_i)\, u_j = \lambda_i\, (u_i - \gamma_i u_j)$$

soit encore

$$\gamma_i(\lambda_j - \lambda_i - \sigma) = \sigma v^* u_i.$$

Finalement on a l'alternative

a) $\sigma \neq \lambda_j - \lambda_i \Longrightarrow \gamma_i = \dfrac{\sigma v^* u_i}{\lambda_j - \lambda_i - \sigma}$ et $u_i - \gamma_i u_j$ est aussi vecteur propre ;

b) $\sigma = \lambda_j - \lambda_i \Longrightarrow \lambda_i = \lambda_j - \sigma$ est alors valeur propre multiple de $\tilde{\mathbb{A}}$ et u_j est le seul vecteur propre connu.

Remarque 8.5 *1) le choix du vecteur v^* du théorème ne pose pas de difficulté a priori, on peut par exemple prendre v^* égal à w_j^*, vecteur propre à gauche associé à λ_j. Ce choix conduit à $\gamma_i = 0$, car dans ce cas on a automatiquement $v^* u_i = 0$.*

2) dans la pratique, il n'est pas nécessaire de calculer la matrice $\tilde{\mathbb{A}}$, car dans l'algorithme de la puissance itérée, il suffit de calculer le produit $\tilde{\mathbb{A}} v_k = \mathbb{A} v_k - \sigma u_j (v^* v_k)$.

Troisième partie

Méthodes de décomposition de domaine

Cette partie est une introduction aux méthodes de décompositions de domaines. On s'intéresse notamment au lien entre formulation variationnelle et formulation algébrique. Il existe une littérature abondante sur les méthodes de décompositions de domaines. Les ouvrages modernes de Dolean, Jolivet et Nataf [17], et Magoulès et Roux [25] proposent une étude approfondie de ces techniques et des algorithmes associés.

Les méthodes de décompositions de domaines se sont développées essentiellement pour satisfaire le besoin de calcul parallèle. Il s'agit de décomposer un problème global en plusieurs sous-problèmes locaux, qui seront couplés en leurs frontières par des conditions de transmission aux interfaces.

D'un point de vue algorithmique, on retrouvera les structures qui permettent notamment de paralléliser le produit matrice-vecteur, voir la section 4.5.

Chapitre 9

Introduction

Dans ce chapitre, nous allons aborder deux méthodes de décomposition de domaine pour un problème de type (3.1) d'un point de vue continu (formulation variationnelle) et discret (formulation algébrique). Ces deux méthodes se différencient par la façon dont on transmet l'information aux interfaces.

Soit Ω un domaine de $\mathbb{R}^d$. On s'intéresse au problème suivant :

$$\text{Trouver } u \in H^1_0(\Omega) \text{ tel que} \quad -\operatorname{div}(k\,\mathbf{grad}\,u) + q\,u = f \text{ sur } \Omega, \tag{9.1}$$

avec $f \in L^2(\Omega)$, k, $q \in L^\infty(\Omega)$ satisfaisant les hypothèses (3.2).

Le problème (9.1) peut également s'écrire sous la forme d'un problème mixte, à deux inconnues $(u, \boldsymbol{p}) \in H^1_0(\Omega) \times \mathbf{H}(\operatorname{div}, \Omega)$ (détaillé dans §3.2.2) :

$$\left\{ \begin{array}{l} \text{Trouver } (u, \boldsymbol{p}) \in H^1_0(\Omega) \times \mathbf{H}(\operatorname{div}, \Omega) \text{ tel que} \\ \operatorname{div} \boldsymbol{p} + q\,u = f \text{ et } k^{-1}\boldsymbol{p} + \mathbf{grad}\,u = 0 \text{ dans } \Omega. \end{array} \right. \tag{9.2}$$

L'existence et l'unicité d'une solution aux problèmes (9.1) et (9.2) sont établis au théorème 3.1.

9.1 Géométrie, espaces de Hilbert et et notations

On considère une partition de Ω en N_Ω **sous-domaines** (non-vides) $(\Omega_i)_{i\in\mathcal{I}_\Omega}$ [43], avec $\mathcal{I}_\Omega := \{1, \cdots, N_\Omega\}$: $\overline{\Omega} = \cup_{i\in\mathcal{I}_\Omega}\overline{\Omega_i}$. La partition est telle que $\forall (i,j) \in \mathcal{I}_\Omega \times \mathcal{I}_\Omega$, $i \neq j$, $\Omega_i \cap \Omega_j = \emptyset$. Pour $i \in \mathcal{I}_\Omega$, on appellera $\boldsymbol{n}_i$ la normale sortante unitaire à $\partial\Omega_i$ et, pour $p \in [1, \infty]$ et $z \in L^p(\Omega)$, on notera $z_i := z_{|\Omega_i}$.

On appelle Σ_{ij} l'interface entre Ω_i et Ω_j pour $i \neq j$, telle que $\Sigma_{ij} = int(\overline{\Omega_i} \cap \overline{\Omega_j})$ si la dimension d'Hausdorff de $\overline{\Omega_i} \cap \overline{\Omega_j}$ est égale à $d-1$: dans ce cas, on dit que les sous-domaines sont voisins ; et $\Sigma_{ij} = \emptyset$ sinon. Par construction, on a $\Sigma_{ij} = \Sigma_{ji}$, on utilisera indifféremment les deux notations. Notons que $\boldsymbol{n}_{i|\Sigma_{ij}} = -\boldsymbol{n}_{j|\Sigma_{ij}}$. Sur l'interface Σ_{ij}, on fixera $\boldsymbol{n}_{ij} = \boldsymbol{n}_{\min(i,j)|\Sigma_{ij}}$. On utilisera parfois la notation synthétique : $\partial_{\boldsymbol{n}_i} u = \mathbf{grad}\,u \cdot \boldsymbol{n}_i$ sur $\partial\Omega_i$ et $\partial_{\boldsymbol{n}_{ij}} u = \mathbf{grad}\,u \cdot \boldsymbol{n}_{ij}$ sur Σ_{ij}.

43. Chaque sous-domaine est de frontière "suffisamment régulière" (Annexe D).

On introduit alors les espaces d'indices et de couples suivants :

$$\begin{array}{rcll} \mathcal{I}_S & := & \{i \in \mathcal{I}_\Omega \,|\, \partial\Omega_i \cap \Sigma_S = \partial\Omega_i\}, & \text{(sous-dom. intérieurs)} \\ \mathcal{I}_{\Omega_i} & := & \{j \in \mathcal{I}_\Omega \,|\, \Sigma_{ij} \neq \emptyset\}, \quad \forall i & \text{(sous-dom. voisins de } \Omega_i) \\ \mathcal{I}\mathcal{J} & := & \{(i,j) \in \mathcal{I}_\Omega \times \mathcal{I}_\Omega,\, i<j \,|\, \Sigma_{ij} \neq \emptyset\}, & \text{(couples de sous-dom. voisins).} \end{array} \tag{9.3}$$

On note $N_{\mathcal{I}\mathcal{J}} = |\mathcal{I}\mathcal{J}|$ le nombre d'interfaces.
Pour $i \in \mathcal{I}_\Omega$, on appelle Σ_i la réunion des interfaces de $\partial\Omega_i$: $\Sigma_i = \cup_{j\in\mathcal{I}_{\Omega_i}} \Sigma_{ij}$. On définit l'interface Σ_S et le treillis [44] $\partial\Sigma_W$ par :

$$\Sigma_S = \bigcup_{i=1}^{N_\Omega} \bigcup_{j=i+1}^{N_\Omega} \overline{\Sigma_{ij}}, \quad \partial\Sigma_W = \bigcup_{i=1}^{N_\Omega} \bigcup_{j=i+1}^{N_\Omega} \partial\Sigma_{ij}.$$

Lorsque $d = 2$, le treillis est composé des sommets de l'interface, qui sont des points isolés.
Lorsque $d = 3$, le treillis est composé des sommets, et des arêtes (ouvertes), de l'interface.
Les points du treillis non-situés sur la frontière $\partial\Omega$ sont des points de croisement.
On introduit, pour $i \in \mathcal{I}_\Omega$, l'ouvert $\Gamma_i = \partial\Omega_i \backslash \overline{\Sigma_S}$. Par construction,

$$\partial\Omega_i = \Sigma_i \cup (\partial\Sigma_W \cap \partial\Omega_i) \cup \Gamma_i,$$

et la partition est disjointe.
On définit pour tout $i \in \mathcal{I}_\Omega$ le sous-espace des fonctions de $H^1(\Omega_i)$ s'annulant sur $\partial\Omega$:

$$V_i := \{v \in H^1(\Omega_i) \,|\, v_{|\Gamma_i} = 0\}. \tag{9.4}$$

On introduit les espaces produits, munis des normes produits suivants :

$$\begin{array}{rcll} \mathcal{P}V & = & \displaystyle\prod_{i\in\mathcal{I}_\Omega} V_i, & \|v\|_{\mathcal{P}V} = \left(\displaystyle\sum_{i\in\mathcal{I}_\Omega} \|v_i\|^2_{H^1(\Omega_i)}\right)^{1/2}, \\ \mathcal{P}\mathbf{H}(\mathrm{div},\Omega) & = & \displaystyle\prod_{i\in\mathcal{I}_\Omega} \mathbf{H}(\mathrm{div},\Omega_i), & \|\boldsymbol{q}\|_{\mathcal{P}\mathbf{H}(\mathrm{div},\Omega)} = \left(\displaystyle\sum_{i\in\mathcal{I}_\Omega} \|\boldsymbol{q}_i\|^2_{\mathbf{H}(\mathrm{div},\Omega_i)}\right)^{1/2}. \end{array} \tag{9.5}$$

Remarque 9.1 *On peut également écrire ces espaces sous la forme :*

$$\mathcal{P}V = \{v \in L^2(\Omega) \,|\, \forall i \in \mathcal{I}_\Omega,\, v_i \in V_i\}, \quad \mathcal{P}\mathbf{H}(\mathrm{div},\Omega) = \{\boldsymbol{q} \in \mathbf{L}^2(\Omega) \,|\, \forall i \in \mathcal{I}_\Omega,\, \boldsymbol{q}_i \in \mathbf{H}(\mathrm{div},\Omega_i)\}.$$

Pour $v \in \mathcal{P}V$, on introduit le saut de v sur Σ_{ij} :

$$\forall (i,j) \in \mathcal{I}\mathcal{J}, \quad [v]_{ij} := v_{i|\Sigma_{ij}} - v_{j|\Sigma_{ij}}. \tag{9.6}$$

Le saut $[v]_{ij}$ est bien défini dans $H^{1/2}(\Sigma_{ij})$. Le *saut global* $[v]$ de v sur Σ_S est défini par :

$$\forall (i,j) \in \mathcal{I}\mathcal{J}, \quad [v]_{|\Sigma_{ij}} := [v]_{ij}. \tag{9.7}$$

44. Appelé "wirebasket" en anglais.

Pour tout couple $(i,j) \in \mathcal{IJ}$, on note$H_{ij} := \widetilde{H}^{1/2}(\Sigma_{ij})$, l'espace des fonctions de $H^{1/2}(\Sigma_{ij})$ dont le prolongement par 0 dans $\partial\Omega_i$ appartient à $H^{1/2}(\partial\Omega_i)$ (voir l'annexe D.2). On peut montrer que $H_{ij} = H_{ji}$ (voir par exemple [3, chapitre 2]). On nommera H'_{ij} le dual de H_{ij}. Pour $\boldsymbol{q} \in \mathcal{P}\mathbf{H}(\mathrm{div},\Omega)$, on introduit le saut de la composante normale de $\boldsymbol{q}$ sur Σ_{ij} :

$$\forall (i,j) \in \mathcal{IJ}, \quad [\boldsymbol{q}\cdot\boldsymbol{n}]_{ij} := \boldsymbol{q}_i \cdot \boldsymbol{n}_{i|\Sigma_{ij}} + \boldsymbol{q}_j \cdot \boldsymbol{n}_{j|\Sigma_{ij}}. \tag{9.8}$$

Le saut $[\boldsymbol{q}\cdot\boldsymbol{n}]_{ij}$ est bien défini dans H'_{ij}, voir l'Annexe D.3.2. Le *saut global* $[\boldsymbol{q}\cdot\boldsymbol{n}]$ de la composante normale sur Σ_S est défini par :

$$\forall (i,j) \in \mathcal{IJ}, \quad [\boldsymbol{q}\cdot\boldsymbol{n}]_{|\Sigma_{ij}} := [\boldsymbol{q}\cdot\boldsymbol{n}]_{ij}. \tag{9.9}$$

Par définition, on a : $[\boldsymbol{q}\cdot\boldsymbol{n}] \in \widetilde{M}$, où $\widetilde{M}$ est l'espace produit $\widetilde{M} := \prod_{(i,j)\in\mathcal{IJ}} H'_{ij}$. Finalement, on introduit les espaces de Hilbert suivants (avec $\mu_{ij} = \mu_{|\Sigma_{ij}}$) :

$$\begin{aligned}
M &= \left\{ \mu \in \prod_{(i,j)\in\mathcal{IJ}} L^2(\Sigma_{ij}) \right\}, \quad \|\mu\|_M = \left(\sum_{(i,j)\in\mathcal{IJ}} \|\mu_{ij}\|^2_{L^2(\Sigma_{ij})} \right)^{1/2} ; \\
\widetilde{\mathbf{H}} &= \{\, \boldsymbol{q} \in \mathcal{P}\mathbf{H}(\mathrm{div},\Omega) \,|\quad [\boldsymbol{q}\cdot\boldsymbol{n}] \in M \,\}\,, \\
&\quad \|\boldsymbol{q}\|_{\widetilde{\mathbf{H}}} := \left(\|\boldsymbol{q}\|^2_{\mathcal{P}\mathbf{H}(\mathrm{div},\Omega)} + \|[\boldsymbol{q}\cdot\boldsymbol{n}]\|^2_M \right)^{1/2}.
\end{aligned} \tag{9.10}$$

Par construction, on a $M \subset \widetilde{M}$. Dans la définition des éléments $\boldsymbol{q} = (\boldsymbol{q}_i)_{i\in\mathcal{I}_\Omega}$ de $\widetilde{\mathbf{H}}$, il est important de noter que pour $(i,j) \in \mathcal{IJ}$, *il n'est pas requis que* $\boldsymbol{q}_i \cdot \boldsymbol{n}_{i|\Sigma_{ij}}, \boldsymbol{q}_j \cdot \boldsymbol{n}_{j|\Sigma_{ij}} \in L^2(\Sigma_{ij})$. Cette définition est fondée sur le fait que, concernant le problème (9.1), la trace normale du gradient sur l'interface, $\mathbf{grad}\, u \cdot \boldsymbol{n}_{|\Sigma_S}$, peut ne pas appartenir à M. En revanche, comme le saut $[k\,\mathbf{grad}\, u \cdot \boldsymbol{n}]$ est nul, alors il appartient systématiquement à M. Nous utiliserons cette observation pour bâtir une formulation variationnelle conforme dans $\widetilde{\mathbf{H}}$.

On rappelle ci-dessous des formules d'intégration par parties qui seront utiles pour construire les formulations variationnelles, en suivant les travaux de [9]. D'après la formule d'intégration par parties (D.2), on a $\forall i \in \mathcal{I}_\Omega$:

$$\forall (v_i\,,\, \boldsymbol{q}_i) \in H^1(\Omega_i) \times \mathbf{H}(\mathrm{div},\Omega_i),$$

$$\int_{\Omega_i} (v_i\ \mathrm{div}\, \boldsymbol{q}_i + \mathbf{grad}\, v_i \cdot \boldsymbol{q}_i)\, d\boldsymbol{x} = \langle \boldsymbol{q}_i \cdot \boldsymbol{n}_i, v_i \rangle_{H^{1/2}(\partial\Omega_i)}.$$

On introduit, pour $H \in \{H^1(\Omega), H^1_0(\Omega), (V_i)_{i\in\mathcal{I}_\Omega}, \mathcal{P}V\}$:

$$H_- := \{ v \in H \,|\, v = 0 \text{ dans un voisinage de } \partial\Sigma_W \}.$$

On rappelle maintenant des résultats de densité établis dans [14].

Théorème 9.2 *Soit $H \in \{H^1(\Omega), H^1_0(\Omega), (V_i)_{i\in\mathcal{I}_\Omega}, \mathcal{P}V\}$: H_- est dense dans H.*

A l'aide d'un de ces résultats de densité, on peut établir une nouvelle formule d'intégration par parties.

Corollaire 9.3 *On a la formule d'intégration par parties*

$$\begin{array}{l}\forall(v\,,\,\boldsymbol{q})\in H_0^1(\Omega)\times\widetilde{\mathbf{H}},\\ \displaystyle\int_\Omega(\mathbf{grad}\,v\cdot\boldsymbol{q}+v\ \mathrm{div}\,\boldsymbol{q})\,d\boldsymbol{x}=\int_{\Sigma_S}[\boldsymbol{q}\cdot\boldsymbol{n}]\,v_{|\Sigma_S}\,ds.\end{array}\tag{9.11}$$

Démonstration : Tout d'abord, pour $i\in\mathcal{I}_\Omega$ la formule d'intégration par parties (D.2-ii) (avec découpage des crochets de dualité) dans Ω_i devient, avec l'aide de l'espace des traces H_{ij} :

$$\begin{array}{l}\forall(v_i\,,\,\boldsymbol{q}_i,)\in(V_i)_-\times\mathbf{H}(\mathrm{div},\Omega_i),\\ \displaystyle\int_{\Omega_i}(\mathbf{grad}\,v_i\cdot\boldsymbol{q}_i+v_i\ \mathrm{div}\,\boldsymbol{q}_i)\,d\boldsymbol{x}=\sum_{j\in\mathcal{I}_{\Omega_i}}\langle\boldsymbol{q}_i\cdot\boldsymbol{n}_i,v_i\rangle_{H_{ij}}.\end{array}$$

Considérons $v\in\big(H_0^1(\Omega)\big)_-$ et $\boldsymbol{q}\in\widetilde{\mathbf{H}}$. Pour tout $(i,j)\in\mathcal{IJ}$, on a $v_{i|\Sigma_{ij}}=v_{j|\Sigma_{ij}}=v_{|\Sigma_{ij}}$. On peut donc en particulier définir $\mu=v_{|\Sigma_S}\in M$, et on peut écrire :[45]

$$\begin{array}{rcl}\displaystyle\int_\Omega(\mathbf{grad}\,v\cdot\boldsymbol{q}+v\ \mathrm{div}\,\boldsymbol{q})\,d\boldsymbol{x} &=& \displaystyle\sum_{i\in\mathcal{I}_\Omega}\int_{\Omega_i}(\mathbf{grad}\,v_i\cdot\boldsymbol{q}_i+v_i\ \mathrm{div}\,\boldsymbol{q}_i)\,d\boldsymbol{x}\\ &=& \displaystyle\sum_{i\in\mathcal{I}_\Omega}\sum_{j\in\mathcal{I}_{\Omega_i}}\langle\boldsymbol{q}_i\cdot\boldsymbol{n}_i,v_i\rangle_{H_{ij}},\\ &=& \displaystyle\sum_{(i,j)\in\mathcal{IJ}}\langle\boldsymbol{q}_i\cdot\boldsymbol{n}_i+\boldsymbol{q}_j\cdot\boldsymbol{n}_j,v_{|\Sigma_{ij}}\rangle_{H_{ij}},\\ &=& \displaystyle\sum_{(i,j)\in\mathcal{IJ}}\langle[\boldsymbol{q}\cdot\boldsymbol{n}]_{ij},\mu\rangle_{H_{ij}},\\ &=& \displaystyle\int_{\Sigma_S}[\boldsymbol{q}\cdot\boldsymbol{n}]\,\mu\,ds.\end{array}\tag{9.12}$$

D'après le théorème 9.2, le résultat est encore vrai pour $v\in H_0^1(\Omega)$ et $\boldsymbol{q}\in\widetilde{\mathbf{H}}$. ◇

9.2 Problèmes posés dans les sous-domaines

Proposition 9.4 *Le problème* (9.1) *est équivalent au problème :*
Trouver $u\in\mathcal{P}V$ tel que :

$$\left\{\begin{array}{lllrcll}(i) & \forall i\in\mathcal{I}_\Omega, & -\mathrm{div}(k_i\,\mathbf{grad}\,u_i)+q_iu_i &=& f_i & \textit{dans } \Omega_i,\\ (ii) & \forall(i,j)\in\mathcal{IJ}, & [u]_{ij} &=& 0 & \textit{sur } \Sigma_{ij},\\ (iii) & \forall(i,j)\in\mathcal{IJ}, & [k\,\mathbf{grad}\,u\cdot\boldsymbol{n}]_{ij} &=& 0 & \textit{sur } \Sigma_{ij},\end{array}\right.\tag{9.13}$$

avec l'identification $u_i:=u_{|\Omega_i}$ pour $i\in\mathcal{I}_\Omega$.

Les conditions aux limites sur les interfaces Σ_{ij}, (9.13)-(ii) et (iii), sont souvent appelées *conditions de transmission*. On peut les réécrire ainsi :

$$\forall(i,j)\in\mathcal{IJ},\ \left\{\begin{array}{lrcll}(ii) & u_i &=& u_j & \text{sur } \Sigma_{ij},\\ (iii) & k_i\,\mathbf{grad}\,u_i\cdot\boldsymbol{n}_{ij} &=& k_j\,\mathbf{grad}\,u_j\cdot\boldsymbol{n}_{ij} & \text{sur } \Sigma_{ij}.\end{array}\right.\tag{9.14}$$

45. En toute rigueur, l'intégrale sur Ω est valable pour les prolongements de $\boldsymbol{q}$ et $(\mathrm{div}\,\boldsymbol{q})$ dans Ω.

Montrons la Proposition 9.4 :

Démonstration :

(9.1) $\Rightarrow$ (9.13) Soit u solution du problème (9.1). Les fonctions locales $(u_i)_{i\in\mathcal{I}_\Omega}$ satisfont l'équation (9.13)-(i). Comme $u \in H_0^1(\Omega)$, alors $\forall (i,j) \in \mathcal{IJ}$, $u_i = u_j$ sur Σ_{ij} dans $H^{1/2}(\Sigma_{ij})$ (cf. [11], Prop. 1.2 p. 21), c'est-à-dire (9.13)-(ii). Montrons pour finir que $k_i \,\mathbf{grad}\, u_{|\Omega_i} \cdot \boldsymbol{n}_{ij} = k_j \,\mathbf{grad}\, u_{|\Omega_j} \cdot \boldsymbol{n}_{ij}$ dans H'_{ij}. Pour cela, on choisit $v \in \big(H_0^1(\Omega)\big)_-$ telle que $v_m = 0$ si $m \notin \{i,j\}$. En particulier, $v_{|\Sigma_{ij}} \in H_{ij}$ et v s'annule sur toutes les interfaces autres que Σ_{ij}. D'après (9.1), on a

$$\begin{aligned}
\sum_{m=i,j} \int_{\Omega_m} (f_m - q_m u_m) v_m \, d\boldsymbol{x} &= \int_\Omega (f - qu) v \, d\boldsymbol{x} = -\int_\Omega \operatorname{div}(k \,\mathbf{grad}\, u) v \, d\boldsymbol{x} \\
\text{ipp (D.1) dans } \Omega &= \int_\Omega k \,\mathbf{grad}\, u \cdot \mathbf{grad}\, v \, d\boldsymbol{x} \\
&= \sum_{m=i,j} \int_{\Omega_m} k_m \,\mathbf{grad}\, u_m \cdot \mathbf{grad}\, v_m \, d\boldsymbol{x} \\
\text{ipp (D.2-ii) dans } \Omega_m &= \sum_{m=i,j} \Big(-\int_{\Omega_m} \operatorname{div}(k_m \,\mathbf{grad}\, u_m) v_m \, d\boldsymbol{x} \\
&\qquad + \langle k_m \,\mathbf{grad}\, u_m \cdot \boldsymbol{n}_m, v_{|\Sigma_{ij}} \rangle_{H_{ij}} \Big).
\end{aligned}$$

Si on utilise (9.13)-(i), on en conclut que :

$$\begin{aligned}
&\forall v \in \big(H_0^1(\Omega)\big)_- \text{ telle que } v_m = 0 \text{ si } m \notin \{i,j\} \\
&\langle k_i \,\mathbf{grad}\, u_i \cdot \boldsymbol{n}_i + k_i \,\mathbf{grad}\, u_j \cdot \boldsymbol{n}_j, v_{|\Sigma_{ij}} \rangle_{H_{ij}} = 0.
\end{aligned}$$

A l'aide du théorème 9.2 appliqué dans V_i ou dans V_j (et de la surjectivité de l'application trace), on en déduit (9.13)-(iii).

(9.13) $\Rightarrow$ (9.1) Soit u solution du problème (9.13). On doit montrer que $u \in H_0^1(\Omega)$ et que u satisfait bien (9.1). Comme $u_i \in V_i$, alors $\forall j \in \Omega_i$, $u_{i|\Sigma_{ij}} \in H^{1/2}(\Sigma_{ij})$. La condition de continuité (9.13)-(ii) assure que $u \in H_0^1(\Omega)$ (cf. [11], Prop. 1.2 p. 21). Soit $\boldsymbol{p} \in \mathbf{L}^2(\Omega)$ tel que $\forall i \in \Omega_i : \boldsymbol{p}_i := -k_i \,\mathbf{grad}\, u_i$. D'après l'équation (9.13)-(i), on a $\forall i \in \Omega_i : \boldsymbol{p}_i \in \mathbf{H}(\operatorname{div}, \Omega_i)$. D'après l'équation (9.13)-(iii), on a : $\forall (i,j) \in \mathcal{IJ}$, $[\boldsymbol{p} \cdot \boldsymbol{n}]_{ij} = 0 \in L^2(\Sigma_{ij})$, c'est-à-dire que $\boldsymbol{p} \in \widetilde{\mathbf{H}}$.

En multipliant (9.13)-(i) par $v \in \mathcal{D}(\Omega)$, en intégrant sur Ω_i et en sommant les contributions, on a :

$$\sum_{i\in\mathcal{I}_\Omega} \int_{\Omega_i} \left(-\operatorname{div}(k_i \,\mathbf{grad}\, u_i)\, v + q_i u_i\, v\right) d\boldsymbol{x} = \sum_{i\in\mathcal{I}_\Omega} \int_{\Omega_i} f_i \, v \, d\boldsymbol{x}. \tag{9.15}$$

Comme $\boldsymbol{p} \in \widetilde{\mathbf{H}}$ et $v \in H_0^1(\Omega)$, on peut utiliser (9.11), ce qui donne :

$$\sum_{i\in\mathcal{I}_\Omega} \int_\Omega \left(\mathbf{grad}\, v_i \cdot \boldsymbol{p}_i + v_i \operatorname{div} \boldsymbol{p}_i\right) d\boldsymbol{x} = 0$$

puisque $[\boldsymbol{p} \cdot \boldsymbol{n}] = 0$. L'équation (9.15) s'écrit donc :

$$\begin{aligned}
\sum_{i\in\mathcal{I}_\Omega} \int_{\Omega_i} (k_i \,\mathbf{grad}\, u_i \cdot \mathbf{grad}\, v + q_i\, u_i\, v)\, d\boldsymbol{x} &= \sum_{i\in\mathcal{I}_\Omega} \int_{\Omega_i} f_i \, v \, d\boldsymbol{x}, \\
\Leftrightarrow \qquad \int_\Omega (k \,\mathbf{grad}\, u \cdot \mathbf{grad}\, v + q\, u\, v)\, d\boldsymbol{x} &= \int_\Omega f \, v \, d\boldsymbol{x}, \\
\Leftrightarrow \qquad \langle -\operatorname{div}(k \,\mathbf{grad}\, u) + qu, v \rangle &= \langle f, v \rangle.
\end{aligned}$$

Comme $f \in L^2(\Omega)$, on en déduit que $-\operatorname{div}(k\,\mathbf{grad}\,u) + qu = f$ est vrai au sens $L^2(\Omega)$, c'est-à-dire presque partout.

$\diamond$

Proposition 9.5 *Le problème* (9.2) *est équivalent au problème :*
Trouver $(u, \boldsymbol{p}) \in \mathcal{P}V \times \mathcal{P}\mathbf{H}(\operatorname{div}, \Omega)$ *tel que :*

$$\left\{\begin{array}{lllrcll} (0) & \forall i \in \mathcal{I}_\Omega, & \operatorname{div}\boldsymbol{p}_i + q_i u_i & = & f_i & \textit{dans } \Omega_i, \\ (i) & \forall i \in \mathcal{I}_\Omega, & k_i^{-1}\boldsymbol{p}_i + \mathbf{grad}\,u_i & = & 0 & \textit{dans } \Omega_i, \\ (ii) & \forall (i,j) \in \mathcal{IJ}, & [u]_{ij} & = & 0 & \textit{sur } \Sigma_{ij}, \\ (iii) & \forall (i,j) \in \mathcal{IJ}, & [\boldsymbol{p}\cdot\boldsymbol{n}]_{ij} & = & 0 & \textit{sur } \Sigma_{ij}, \end{array}\right. \tag{9.16}$$

avec l'identification $\forall i \in \mathcal{I}_\Omega$, $(u_i, \boldsymbol{p}_i) = (u_{|\Omega_i}, \boldsymbol{p}_{|\Omega_i})$.

On peut réécrire (9.16)-(ii) et (9.16)-(iii) ainsi :

$$\forall (i,j) \in \mathcal{IJ}, \left\{\begin{array}{llrcll} (ii) & & u_i & = & u_j & \text{sur } \Sigma_{ij}, \\ (iii) & & \boldsymbol{p}_i\cdot\boldsymbol{n}_{ij} & = & \boldsymbol{p}_j\cdot\boldsymbol{n}_{ij} & \text{sur } \Sigma_{ij}. \end{array}\right. \tag{9.17}$$

La démonstration, similaire à celle de la proposition 9.4, est laissée en exercice.

Chapitre 10

Problème à une inconnue, méthode de Schwarz

Nous allons maintenant étudier la résolution du problème (9.13) par la méthode de Schwarz itérative [34]. Pour cela, on va réécrire les conditions de transmission (9.14) sous une forme plus exploitable. A partir de là, on pourra définir un algorithme de résolution itératif. On suppose que la solution u est telle que :

$$\text{Pour tout } (i,j) \in \mathcal{I}_\Omega \times \mathcal{I}_{\Omega_i}, \quad k_i \, \mathbf{grad}\, u_i \cdot \boldsymbol{n}_{ij|\Sigma_{ij}} \in L^2(\Sigma_{ij}).$$

Remarque 10.1 *Si on revient à la proposition 3.19, ceci est garanti lorsque* $r_{\max} > 1/2$. *En effet, dans ce cas on a, pour tout* $i \in \mathcal{I}$, $\mathbf{grad}\, u_i \in \bigcap_{0 \le r < r_{\max}} \mathbf{H}^r(\Omega_i)$, *et on sait que la trace de* $\mathbf{grad}\, u_i$ *sur* $\partial\Omega_i$ *appartient à* $\mathbf{L}^2(\partial\Omega_i)$ *(voir le Chapitre 2 de [3] pour le théorème de trace correspondant).*

10.1 Approche continue

Pour tout $(i,j) \in \mathcal{I}_\Omega \times \mathcal{I}_{\Omega_i}$, on introduit les couples $(\alpha_{ij}^i, \alpha_{ij}^j)$ de coefficients positifs ou nuls tels que $\alpha_{ij}^i = \alpha_{ji}^i$; et $\alpha_{ij}^i + \alpha_{ij}^j > 0$: il n'y a qu'un seul couple de coefficients par interface, et pour un couple donné, les coefficients ne peuvent pas être tous les deux nuls. Les conditions de transmission de Dirichlet (9.14)-(ii) et de Neumann (9.14)-(iii) peuvent être combinées pour écrire les conditions de Robin suivantes : Pour tout $(i,j) \in \mathcal{I}\mathcal{J}$,

$$\left\{ \begin{array}{rcll} k_i \, \mathbf{grad}\, u_i \cdot \boldsymbol{n}_{ij} + \alpha_{ij}^i u_i & = & k_j \, \mathbf{grad}\, u_j \cdot \boldsymbol{n}_{ij} + \alpha_{ij}^i u_j & \text{sur } \Sigma_{ij}, \\ k_j \, \mathbf{grad}\, u_j \cdot \boldsymbol{n}_{ij} - \alpha_{ij}^j u_j & = & k_i \, \mathbf{grad}\, u_i \cdot \boldsymbol{n}_{ij} - \alpha_{ij}^j u_i & \text{sur } \Sigma_{ij}. \end{array} \right. \tag{10.1}$$

Ces conditions s'écrivent sous la forme synthétique suivante : Pour tout $(i,j) \in \mathcal{I}\mathcal{J}$,

$$\left\{ \begin{array}{rcll} k_i \partial_{\boldsymbol{n}_i} u_i + \alpha_{ij}^i u_i & = & -k_j \partial_{\boldsymbol{n}_j} u_j + \alpha_{ij}^i u_j & \text{sur } \Sigma_{ij}, \\ k_j \partial_{\boldsymbol{n}_j} u_j + \alpha_{ij}^j u_j & = & -k_i \partial_{\boldsymbol{n}_i} u_i + \alpha_{ij}^j u_i & \text{sur } \Sigma_{ij}. \end{array} \right. \tag{10.2}$$

On peut également les poser sous la forme d'une seule équation de la façon suivante :

$$\text{Pour tout } (i,j) \in \mathcal{I}_\Omega \times \mathcal{I}_{\Omega_i}, \quad k_i \partial_{\boldsymbol{n}_i} u_i + \alpha_{ij}^i u_i = -k_j \partial_{\boldsymbol{n}_j} u_j + \alpha_{ij}^i u_j \text{ sur } \Sigma_{ij}. \tag{10.3}$$

Le problème (9.13) est équivalent au problème :
Trouver $u \in \mathcal{P}V$ tel que : pour tout $i \in \mathcal{I}_\Omega$ et pour tout $j \in \mathcal{I}_{\Omega_i}$:

$$\left\{ \begin{array}{lrcll} (i) & -\operatorname{div}(k_i \, \mathbf{grad}\, u_i) + q_i u_i & = & f_i & \text{dans } \Omega_i, \\ (ii) & k_i \partial_{\boldsymbol{n}_i} u_i + \alpha_{ij}^i u_i & = & -k_j \partial_{\boldsymbol{n}_j} u_j + \alpha_{ij}^i u_j & \text{sur } \Sigma_{ij}. \end{array} \right. \tag{10.4}$$

Comme on ne connaît pas les valeurs de $u_{i|\Sigma}$ ni celles de $k_i \, \mathbf{grad}\, u_i \cdot \boldsymbol{n}_{i|\Sigma}$, on propose de résoudre les problèmes (10.4) de façon itérative dans les sous-domaines $(\Omega_i)_{i \in \mathcal{I}_\Omega}$, c'est-à-dire localement, et d'approcher les conditions de transmission. Afin de résoudre le problème (10.4), on propose donc l'algorithme itératif suivant [46] :

initialisation
$\forall i \in \mathcal{I}_\Omega,\ u_i^0 \in V_i$ **est donné.**
itérations : pour $\ell = 1, \ldots,$ $\quad \forall i \in \mathcal{I}_\Omega,$ **trouver** $u_i^\ell \in V_i$ **tel que**

$$\left\{ \begin{array}{lrcll} (i) & -\operatorname{div}(k_i \, \mathbf{grad}\, u_i^\ell) + q_i \, u_i^\ell & = & f_i & \text{dans } \Omega_i, \\ (ii) & \forall j \in \mathcal{I}_{\Omega_i},\ k_i \partial_{\boldsymbol{n}_i} u_i^\ell + \alpha_{ij}^i u_i^\ell & = & -k_j \partial_{\boldsymbol{n}_j} u_j^{\ell-1} + \alpha_{ij}^i u_j^{\ell-1} & \text{sur } \Sigma_{ij}, \end{array} \right. \tag{10.5}$$

jusqu'à convergence.

Le critère de convergence peut s'établir en mesurant la différence entre les solutions locales à l'interface, entre deux itérations consécutives. Pour le calcul parallèle, l'intérêt de cet algorithme (qui correspond à un algorithme de Jacobi) est de pouvoir résoudre simultanément les problèmes locaux (pour $i \in \mathcal{I}_\Omega$). Qu'en est-il de la convergence de l'algorithme ? Pour l'établir *dans le cas où* $\forall (i,j) \in \mathcal{IJ},\ \alpha_{ij}^i = \alpha_{ij}^j$, P.-L. Lions dans [24] étudie la convergence de la suite des erreurs $\left((e_i^\ell)_{i \in \mathcal{I}_\Omega} \right)_{\ell \in \mathbb{N}}$, où pour tout $i \in \mathcal{I}_\Omega$, $\forall \ell \in \mathbb{N}$, $e_i^\ell := u_i^\ell - u_i \in V_i$. On fait l'hypothèse que le choix initial est tel que

$$\text{Pour tout } (i,j) \in \mathcal{I}_\Omega \times \mathcal{I}_{\Omega_i}, \quad k_i \, \mathbf{grad}\, u_i^0 \cdot \boldsymbol{n}_{ij|\Sigma_{ij}} \in L^2(\Sigma_{ij}).$$

Théorème 10.2 *Supposons que* $\forall (i,j) \in \mathcal{IJ},\ \alpha_{ij}^i = \alpha_{ij}^j$. *Alors, la suite* $\left((u_i^\ell)_{i \in \mathcal{I}_\Omega} \right)_\ell$ *converge vers* $(u_i)_{i \in \mathcal{I}_\Omega}$ *dans* $\mathcal{P}V$.

Démonstration : On note $\forall (i,j) \in \mathcal{IJ},\ \alpha_{ij} := \alpha_{ij}^i = \alpha_{ij}^j > 0$. Nous adaptons ici la preuve écrite dans l'ouvrage de Quarteroni et Valli [32] pour l'équation de Poisson pour deux sous-domaines. Les suites $\left((e_i^\ell)_{\ell \in \mathbb{N}} \right)_{i \in \mathcal{I}_\Omega}$ satisfont pour tout $i \in \mathcal{I}_\Omega$, et pour tout $j \in \mathcal{I}_{\Omega_i}$:

$$\left\{ \begin{array}{lrcll} (i) & -\operatorname{div}(k_i \, \mathbf{grad}\, e_i^\ell) + q_i \, e_i^\ell & = & 0 & \text{dans } \Omega_i, \\ (ii) & k_i \partial_{\boldsymbol{n}_i} e_i^\ell + \alpha_{ij} e_i^\ell & = & -k_j \partial_{\boldsymbol{n}_j} e_j^{\ell-1} + \alpha_{ij} e_j^{\ell-1} & \text{sur } \Sigma_{ij}. \end{array} \right. \tag{10.6}$$

On sait que, pour tout ℓ et pour tout $i \in \mathcal{I}_\Omega$, $e_{i|\partial\Omega_i}^\ell \in H^{1/2}(\partial\Omega_i) \subset L^2(\partial\Omega_i)$: en particulier, pour tout $j \in \mathcal{I}_{\Omega_i}$, $e_{i|\Sigma_{ij}}^\ell \in L^2(\Sigma_{ij})$. Par récurrence sur ℓ, on déduit de (10.6)-(ii) que, pour

46. Cet algorithme fut initialement proposé par Schwarz [34] pour déterminer la solution de l'équation de Laplace dans un domaine égal à l'union d'un disque et d'un carré non-disjoints (avec des sous-domaines recouvrants, et la condition de transmission de Dirichlet).

tout $i \in \mathcal{I}_\Omega$ et pour tout $j \in \mathcal{I}_{\Omega_i}$, on a $k_i \partial_{\boldsymbol{n}_i} e_{i|\Sigma_{ij}}^\ell \in L^2(\Sigma_{ij})$. On multiplie (10.6)-(i) par e_i^ℓ et on intègre sur Ω_i, ce qui donne :

$$\int_{\Omega_i} \Big(-\operatorname{div}(k_i \,\mathbf{grad}\, e_i^\ell) + q_i e_i^\ell \Big)\; e_i^\ell d\boldsymbol{x} = 0.$$

On utilise la formule d'intégration par parties (D.1) pour obtenir :

$$\int_{\Omega_i} \Big(-\operatorname{div}(k_i \,\mathbf{grad}\, e_i^\ell)\, e_i^\ell \Big)\, d\boldsymbol{x} = \int_{\Omega_i} k_i |\,\mathbf{grad}\, e_i^\ell|^2 d\boldsymbol{x} - \sum_{j \in \mathcal{I}_{\Omega_i}} \int_{\Sigma_{ij}} k_i \partial_{\boldsymbol{n}_i} e_i^\ell\, e_i^\ell\, ds.$$

Les suites $(\,(e_i^\ell)_\ell\,)_{i \in \mathcal{I}_\Omega}$ satisfont donc : pour tout $i \in \mathcal{I}_\Omega$

$$\int_{\Omega_i} \Big(k_i |\,\mathbf{grad}\, e_i^\ell|^2 + q_i\, (e_i^\ell)^2 \Big)\; d\boldsymbol{x} = \sum_{j \in \mathcal{I}_{\Omega_i}} \int_{\Sigma_{ij}} k_i \partial_{\boldsymbol{n}_i} e_i^\ell\, e_i^\ell\, ds. \tag{10.7}$$

Pour traiter les intégrales sur les interfaces Σ_{ij}, on utilise l'égalité

$$AB = \big(\,(\alpha A + B)^2 - (\alpha A - B)^2\,\big)\,/(4\,\alpha) \tag{10.8}$$

avec $\alpha = \alpha_{ij}$; $A = e_i^\ell$ et $B = k_i \partial_{\boldsymbol{n}_i} e_i^\ell$ presque partout sur Σ_{ij}. D'où :

$$k_i \partial_{\boldsymbol{n}_i} e_i^\ell\, e_i^\ell = \Big(\,(\alpha_{ij} e_i^\ell + k_i \partial_{\boldsymbol{n}_i} e_i^\ell)^2 - (\alpha_{ij} e_i^\ell - k_i \partial_{\boldsymbol{n}_i} e_i^\ell)^2\,\Big)\,/(4\,\alpha_{ij}) \text{ presque partout.}$$

On en déduit que les suites $(e_i^\ell)_\ell$, $i \in \mathcal{I}_\Omega$ satisfont :

$$\begin{aligned} &\|\sqrt{k_i}\,\mathbf{grad}\, e_i^\ell\|_{L^2(\Omega_i)}^2 + \|\sqrt{q_i} e_i^\ell\|_{L^2(\Omega_i)}^2 \\ &= \sum_{j \in \mathcal{I}_{\Omega_i}} \Big[\Big(\|\alpha_{ij} e_i^\ell + k_i \partial_{\boldsymbol{n}_i} e_i^\ell\|_{L^2(\Sigma_{ij})}^2 - \|\alpha_{ij} e_i^\ell - k_i \partial_{\boldsymbol{n}_i} e_i^\ell\|_{L^2(\Sigma_{ij})}^2 \Big)\,/(4\,\alpha_{ij}) \Big]. \end{aligned}$$

Les conditions de transmission sur Σ_{ij} s'écrivent pour tout $i \in \mathcal{I}_\Omega$, et pour tout $j \in \mathcal{I}_{\Omega_i}$ (voir (10.6)-(ii)) :

$$\forall \ell \in \mathbb{N}^*, \quad \alpha_{ij} e_i^\ell + k_i \partial_{\boldsymbol{n}_i} e_i^\ell = \alpha_{ij} e_j^{\ell-1} - k_j \partial_{\boldsymbol{n}_j} e_j^{\ell-1}.$$

Ainsi les suites $(e_i^\ell)_{\ell \in \mathbb{N}}$ satisfont pour $\ell \in \mathbb{N}^*$, pour $(i,j) \in \mathcal{I}\mathcal{J}$:

$$\begin{aligned} &\|\sqrt{k_i}\,\mathbf{grad}\, e_i^\ell\|_{L^2(\Omega_i)}^2 + \|\sqrt{q_i} e_i^\ell\|_{L^2(\Omega_i)}^2 \\ &= \sum_{j \in \mathcal{I}_{\Omega_i}} \Big[\Big(\|\alpha_{ij} e_j^{\ell-1} - k_j \partial_{\boldsymbol{n}_j} e_j^{\ell-1}\|_{L^2(\Sigma_{ij})}^2 - \|\alpha_{ij} e_i^\ell - k_i \partial_{\boldsymbol{n}_i} e_i^\ell\|_{L^2(\Sigma_{ij})}^2 \Big)\,/(4\,\alpha_{ij}) \Big]. \end{aligned}$$

En sommant les contributions de chaque sous-domaine, on obtient pour $\ell \in \mathbb{N}^*$:

$$\begin{aligned} &\sum_{i \in \mathcal{I}_\Omega} \Big(\|\sqrt{k_i}\,\mathbf{grad}\, e_i^\ell\|_{L^2(\Omega_i)}^2 + \|\sqrt{q_i} e_i^\ell\|_{L^2(\Omega_i)}^2 \Big) \\ &= \sum_{i \in \mathcal{I}_\Omega} \sum_{j \in \mathcal{I}_{\Omega_i}} \Big[\Big(\|\alpha_{ij} e_j^{\ell-1} - k_j \partial_{\boldsymbol{n}_j} e_j^{\ell-1}\|_{L^2(\Sigma_{ij})}^2 - \|\alpha_{ij} e_i^\ell - k_i \partial_{\boldsymbol{n}_i} e_i^\ell\|_{L^2(\Sigma_{ij})}^2 \Big)\,/(4\,\alpha_{ij}) \Big] \\ &= \sum_{(i,j) \in \mathcal{I}\mathcal{J}} \Big(\|\alpha_{ij} e_j^{\ell-1} - k_j \partial_{\boldsymbol{n}_j} e_j^{\ell-1}\|_{L^2(\Sigma_{ij})}^2 - \|\alpha_{ij} e_j^\ell - k_j \partial_{\boldsymbol{n}_j} e_j^\ell\|_{L^2(\Sigma_{ij})}^2 \\ &\qquad + \|\alpha_{ij} e_i^{\ell-1} - k_i \partial_{\boldsymbol{n}_i} e_i^{\ell-1}\|_{L^2(\Sigma_{ij})}^2 - \|\alpha_{ij} e_i^\ell - k_i \partial_{\boldsymbol{n}_i} e_i^\ell\|_{L^2(\Sigma_{ij})}^2 \Big)\,/(4\,\alpha_{ij}) \end{aligned}$$

En effet, ci-dessus chaque couple $(i,j) \in \mathcal{IJ}$ regroupe deux contributions dans la double somme précédente. En sommant cette fois sur les itérations $\ell \in \{1, \cdots, N\}$, les contributions aux itérations successives $\ell - 1$ et ℓ se compensent pour tout $\ell \in \{2, \cdots, N-1\}$ et il reste finalement :

$$\begin{aligned}
&\sum_{\ell=1}^{N} \sum_{i \in \mathcal{I}_\Omega} \left[\left(\| \sqrt{k_i}\, \mathbf{grad}\, e_i^\ell \|^2_{L^2(\Omega_i)} + \| \sqrt{q_i} e_i^\ell \|^2_{L^2(\Omega_i)} \right) \right] \\
+ &\sum_{(i,j) \in \mathcal{IJ}} \left[\left(\| \alpha_{ij} e_i^N - k_i \partial_{\boldsymbol{n}_i} e_i^N \|^2_{L^2(\Sigma_{ij})} + \| \alpha_{ij} e_j^N - k_j \partial_{\boldsymbol{n}_j} e_j^N \|^2_{L^2(\Sigma_{ij})} \right) / (4\, \alpha_{ij}) \right] \\
= &\sum_{(i,j) \in \mathcal{IJ}} \left[\left(\| \alpha_{ij} e_i^0 - k_i \partial_{\boldsymbol{n}_i} e_i^0 \|^2_{L^2(\Sigma_{ij})} + \| \alpha_{ij} e_j^0 - k_j \partial_{\boldsymbol{n}_j} e_2^0 \|^2_{L^2(\Sigma_{ij})} \right) / (4\, \alpha_{ij}) \right]
\end{aligned}$$

On a donc montré que la série $\displaystyle\sum_{\ell=1}^{\infty} \sum_{i \in \mathcal{I}_\Omega} \left(\| \sqrt{k_i}\, \mathbf{grad}\, e_i^\ell \|^2_{L^2(\Omega_i)} + \| \sqrt{q_i} e_i^\ell \|^2_{L^2(\Omega_i)} \right)$, qui est croissante, est en outre bornée : la suite $\left(\displaystyle\sum_{i \in \mathcal{I}_\Omega} \left(\| \sqrt{k_i}\, \mathbf{grad}\, e_i^\ell \|^2_{L^2(\Omega_i)} + \| \sqrt{q_i} e_i^\ell \|^2_{L^2(\Omega_i)} \right) \right)_{\ell \in \mathbb{N}}$ converge donc vers 0, c'est-à-dire que la suite $(e_i^\ell)_\ell$ tend vers 0 dans $H^1(\Omega_i)$, $i \in \mathcal{I}_\Omega$.
Par hypothèse, on a également que $(e_i^\ell)_{\ell \in \mathbb{N}}$ tend vers 0 dans V_i, $i \in \mathcal{I}_\Omega$. $\diamond$

La preuve est établie dans le cas où $\alpha_{ij}^i = \alpha_{ij}^j$. En pratique, on choisit souvent $\alpha_{ij}^i \neq \alpha_{ij}^j$.

Remarque 10.3 *On indique ci-dessous quelques points à développer :*

- *La preuve de la convergence de l'algorithme* (10.5) *a d'abord été établie par P.-L. Lions dans [24] pour le problème de Poisson.*
- *On peut écrire une variante de l'algorithme* (10.5) *sous la forme d'un algorithme de Gauss-Seidel, la preuve de convergence se faisant avec les mêmes outils.*
- *Si deux sous-domaines Ω_i et Ω_j, $i \neq j$ se recouvrent (c'est-à-dire $\int(\Omega_i) \cap \int(\Omega_j) \neq \emptyset$), on définit deux interfaces $\Sigma_{ij} = int(\Omega_i \cap \partial\Omega_j)$ et $\Sigma_{ji} = int(\Omega_j \cap \partial\Omega_i)$. Sur chacune de ces interfaces, on peut utiliser au choix la condition de transmission de Dirichlet, celle de Neumann ou celle de Robin. La preuve de la convergence de l'algorithme de Schwarz avec des sous-domaines recouvrants et des conditions d'interface de Dirichlet a été réalisée par P.-L. Lions dans [23].*

10.2 Optimisation des paramètres de Robin

Comment choisir les paramètres $(\alpha_{ij}^i, \alpha_{ij}^j)$? Dans l'article [28], Nataf et Nier proposent l'étude asymptotique qui suit. Supposons que $\Omega = \mathbb{R}^2$, $N_\Omega = 2$ avec $\Omega_1 =]-\infty, 0[\times \mathbb{R}$, $\Omega_2 =]0, +\infty[\times \mathbb{R}$. L'interface est telle que $\Sigma = \{(x,y) \,|\, x = 0,\ y \in \mathbb{R}\}$, et l'on a : $\boldsymbol{n}_1 = \boldsymbol{e}_1$ et $\boldsymbol{n}_2 = -\boldsymbol{e}_1$. On note $\alpha_1 = \alpha_{12}^1$ et $\alpha_2 = \alpha_{12}^2$. Dans chaque sous-domaine Ω_i, on suppose que les coefficients k_i et q_i sont constants. La suite d'erreurs $(e_1^\ell, e_2^\ell)_{\ell \in \mathbb{N}}$ satisfait les équations (voir (10.6)) :

$$\forall \ell \in \mathbb{N}^*, \quad \left\{ \begin{array}{rcll} -\Delta e_i^\ell + \dfrac{q_i}{k_i}\, e_i^\ell &=& 0, & \text{dans } \Omega_i, \\ k_1\, \partial_x e_1^\ell + \alpha_1\, e_1^\ell &=& k_2\, \partial_x e_2^{\ell-1} + \alpha_1\, e_2^{\ell-1}, & \text{sur } \Sigma, \\ -k_2\, \partial_x e_2^\ell + \alpha_2\, e_2^\ell &=& -k_1\, \partial_x e_1^{\ell-1} + \alpha_2\, e_1^{\ell-1}, & \text{sur } \Sigma. \end{array} \right. \tag{10.9}$$

Soit $\hat{e}_i^\ell$ la transformée de Fourier partielle le long de l'axe y de e_i^ℓ (on a $e_i^\ell(x,.) \in L^1(\mathbb{R})$) :

$$\forall\, (x,\kappa) \in \hat{\Omega}_i, \quad \hat{e}_i^\ell(x,\kappa) := \int_{\mathbb{R}} \exp(-\imath\kappa y)\, e_i^\ell(x,y)\, dy, \tag{10.10}$$

avec $\hat{\Omega}_1 =]-\infty, 0[\times\mathbb{R}$, $\hat{\Omega}_2 =]0, +\infty[\times\mathbb{R}$. Après une double intégration par parties, on trouve que :

$$\widehat{\Delta e_i^\ell}(x,\kappa) = \int_{\mathbb{R}} \exp(-\imath\kappa y)\Delta e_i^\ell(x,y)dy = -(\partial_{xx}^2 \hat{e}_i^\ell + \kappa^2\, \hat{e}_i^\ell)(x,\kappa). \tag{10.11}$$

Après transformation de Fourier partielle des équations (10.9), on obtient ainsi que la suite $(\hat{e}_1^\ell, \hat{e}_2^\ell)_{\ell\in\mathbb{N}}$ est telle que :

$$\forall \ell \in \mathbb{N}^*, \quad \left\{ \begin{array}{rcll} -\partial_{xx}^2 \hat{e}_i^\ell + \left(\kappa^2 + \dfrac{q_i}{k_i}\right) \hat{e}_i^\ell &=& 0, & \text{dans } \hat{\Omega}_i, \\ k_1\, \partial_x \hat{e}_1^\ell + \alpha_1\, \hat{e}_1^\ell &=& k_2\, \partial_x \hat{e}_2^{\ell-1} + \alpha_1\, \hat{e}_2^{\ell-1}, & \text{sur } \Sigma, \\ -k_2\, \partial_x \hat{e}_2^\ell + \alpha_2\, \hat{e}_2^\ell &=& -k_1\, \partial_x \hat{e}_1^{\ell-1} + \alpha_2\, \hat{e}_1^{\ell-1}, & \text{sur } \Sigma. \end{array} \right. \tag{10.12}$$

Posons : $\lambda_i := \lambda_i(\kappa) = \sqrt{\kappa^2 + \dfrac{q_i}{k_i}}$. D'après la première équation de (10.12) et comme les erreurs tendent vers 0 en $\pm\infty$ ($\hat{e}_i^\ell \in L^2(\hat{\Omega}_i)$), on a :

$$\forall \ell \in \mathbb{N}, \quad \hat{e}_1^\ell(x,\kappa) = b_1^\ell\, \exp(\, \lambda_1(\kappa)\, x\,) \text{ et } \hat{e}_2^\ell(x,\kappa) = b_2^\ell\, \exp(\, -\lambda_2(\kappa)\, x\,). \tag{10.13}$$

Ci-dessus, les coefficients b_i^ℓ dépendent des λ_i (donc de κ), et sont déterminés à l'aide des conditions de transmission (deuxième et troisième équations de (10.12)). On obtient : $\forall \ell \in \mathbb{N}$, $b_i^{\ell+1}(\, \alpha_i + k_i\, \lambda_i\,) = b_j^\ell(\, \alpha_i - k_j\, \lambda_j\,)$ avec $(i,j) \in \{(1,2),(2,1)\}$. On en déduit par récurrence que :

$$\text{Pour } i \in \{1,2\},\ \forall \ell \in \mathbb{N},\ b_i^{\ell+2} = \rho\, b_i^\ell, \text{ avec } \rho = \frac{(\alpha_1 - k_2\,\lambda_2)}{(\alpha_1 + k_1\,\lambda_1)} \frac{(\alpha_2 - k_1\,\lambda_1)}{(\alpha_2 + k_2\,\lambda_2)}. \tag{10.14}$$

Comme u^0 n'est (presque sûrement) pas la solution exacte, observons que les premières erreurs e_1^0 et e_2^0 ne sont pas nulles, et $b_i^0 \neq 0$, cf. (10.13).

Par ailleurs, on a : $b_i^{\ell+2} = \rho\, b_i^\ell$ pour $\ell \geq 0$, d'où : $b_i^{\ell+2} = \left\{ \begin{array}{ll} \rho^{\frac{\ell+2}{2}} b_i^0 & \text{si } \ell \text{ est pair,} \\ \rho^{\frac{\ell+1}{2}} b_i^1 & \text{si } \ell \text{ est impair.} \end{array} \right.$

Comme $\alpha_i \geq 0$, $k_i > 0$ et $\lambda_i > 0$, on a $|\rho| \leq 1$.

Dans le cas où $\alpha_1 = \alpha_2 = 0$ (condition d'interface de Neumann), on trouve $\rho = 1$: les coefficients $(b_i^\ell)_\ell$ ne tendent pas vers 0, l'algorithme ne converge pas.

Si au contraire $\alpha_1 \neq 0$ ou $\alpha_2 \neq 0$, on a $|\rho| < 1$ et l'algorithme converge. Plus précisément, si les α_i sont très grands (condition d'interface de "type Dirichlet", avec une petite perturbation), on trouve $\rho = 1^-$: il y a convergence, mais les coefficients $(b_i^\ell)_\ell$ tendent très lentement vers 0. Notons que, plus $|\rho|$ est petit, meilleure est la vitesse de convergence (les coefficients $(b_i^\ell)_\ell$ tendent plus rapidement vers 0) ; et que, si $\rho = 0$, la convergence est

obtenue dès la deuxième itération [47]...
Afin de réduire ρ, il faut choisir $\alpha_i \approx k_j\,\lambda_j$. A l'ordre 0 en κ, les paramètres optimaux sont tels que :

$$\alpha_i = \sqrt{q_j\,k_j}. \tag{10.15}$$

Remarque 10.4 — *Un étude approfondie du choix des paramètres se trouve dans [19].*
- *Ce résultat théorique peut également être utilisé efficacement en pratique dans le cas où les coefficients k_i et q_i ne sont pas constants.*
- *Si la solution est suffisamment régulière, on peut utiliser des conditions d'interface dites conditions de Ventcell, qui correspondant aux conditions d'interface de Robin auxquelles on a ajouté des termes dépendant des dérivées secondes tangentielles des u_i.*

Les méthodes utilisant l'algorithme itératif de Schwarz avec conditions de transmission optimisées sont appelées les *méthodes de Schwarz optimisées*, ce qui se dit en anglais optimized Schwarz methods (OSM). Dans le cas où $q_i = 0$, on ne peut pas faire cette étude asymptotique. Néanmoins, on verra que l'approche algébrique de l'algorithme nous donne une autre technique pour optimiser la convergence.

10.3 Formulation variationnelle

Nous allons maintenant construire une formulation variationnelle associée au problème (10.5) (pour ℓ donné). On considère une fonction-test $v_i \in (V_i)_-$ que l'on multiplie à l'équation (10.5)-(i), et on intègre sur Ω_i. Rappelons que la formule d'intégration par parties (D.2-i) (avec découpage des crochets de dualité) dans Ω_i donne :

$$-\int_{\Omega_i} \mathrm{div}(k_i\,\mathbf{grad}\,u_i^\ell)\,v_i d\boldsymbol{x} = \int_{\Omega_i} k_i\,\mathbf{grad}\,u_i^\ell \cdot \mathbf{grad}\,v_i d\boldsymbol{x} - \sum_{j\in\mathcal{I}_{\Omega_i}} \int_{\Sigma_{ij}} k_i \partial_{\boldsymbol{n}_i} u_i^\ell\,v_i ds.$$

Par ailleurs, d'après la condition de transmission (10.5)-(ii) sur Σ_{ij}, on a :

$$-k_i\partial_{\boldsymbol{n}_i} u_i^\ell = \alpha_{ij}^i u_i^\ell + k_j \partial_{\boldsymbol{n}_j} u_j^{\ell-1} - \alpha_{ij}^i u_j^{\ell-1}.$$

Une formulation variationnelle d'une itération du problème (10.5) s'écrit donc :
Pour $i \in \mathcal{I}_\Omega$, trouver $u_i^\ell \in V_i$ tel que $\forall v_i \in (V_i)_-$,

$$\begin{aligned} &\int_{\Omega_i} k_i\,\mathbf{grad}\,u_i^\ell \cdot \mathbf{grad}\,v_i d\boldsymbol{x} + \int_{\Omega_i} q_i u_i^\ell\,v_i d\boldsymbol{x} + \sum_{j\in\mathcal{I}_{\Omega_i}} \int_{\Sigma_{ij}} \alpha_{ij}^i u_i^\ell v_i\,ds \\ &\qquad = \\ &\int_{\Omega_i} f_i\,v_i d\boldsymbol{x} - \sum_{j\in\mathcal{I}_{\Omega_i}} \int_{\Sigma_{ij}} k_j \partial_{\boldsymbol{n}_j} u_j^{\ell-1}\,v_i ds + \sum_{j\in\mathcal{I}_{\Omega_i}} \int_{\Sigma_{ij}} \alpha_{ij}^i u_j^{\ell-1} v_i\,ds \end{aligned} \tag{10.16}$$

47. Dans le cas plus général de N_Ω sous-domaines découpés en tranches, on peut vérifier que la convergence se fera en N_Ω itérations au mieux (voir [28]).

Les gradients des éléments finis P_1 ne sont pas conformes dans $\mathbf{H}(\mathrm{div},\cdot)$, et ne permettent donc pas de discrétiser $\boldsymbol{p}_j \cdot \boldsymbol{n}_j$ sur Σ_{ij}, avec $\boldsymbol{p}_j = k_j \,\mathbf{grad}\, u_j^\ell \in \mathbf{H}(\mathrm{div},\Omega_j)$. Afin d'obtenir un schéma numérique correct, l'intégrale sur Σ_{ij} pour $j \in \mathcal{I}_{\Omega_i}$ contenant le terme $k_j \partial_{\boldsymbol{n}_j} u_j^{ell-1}$ doit être réécrite sous la forme d'une intégrale volumique. Pour cela, on choisit un relèvement $v_j^\star$ dans V_j de la trace sur Σ_{ij} de la fonction-test v_i. En effet la trace $v_{i|\Sigma_{ij}} \in H_{ij}$, et on peut la prolonger par 0 sur $\partial\Omega_j$ pour obtenir un élément de $H^{1/2}(\partial\Omega_j)$. D'après la surjectivité de l'application trace (voir le théorème D.1), il existe un relèvement du prolongement dans V_j, qui appartient par définition à l'espace V_j^i tel que :

$$V_j^i := \{v \in V_j \,|\, v_{|\partial\Omega_j \setminus \Sigma_{ij}} = 0\}. \tag{10.17}$$

Par construction, $v^\star_{j\,|\Sigma_{ij}} = v_{i|\Sigma_{ij}}$, et on a : $\displaystyle\int_{\Sigma_{ij}} k_j \partial_{\boldsymbol{n}_j} u_j^{\ell-1}\, v_i ds = \int_{\partial\Omega_j} k_j \partial_{\boldsymbol{n}_j} u_j^{\ell-1}\, v_j^\star\, ds$. La formule d'intégration par parties (D.1) dans Ω_j, puis l'équation (10.5)-(i) permettent d'écrire :

$$\begin{aligned} -\int_{\Sigma_{ij}} k_j \partial_{\boldsymbol{n}_j} u_j^{\ell-1}\, v_j^\star\, ds &= -\int_{\Omega_j} \mathrm{div}(k_j \,\mathbf{grad}\, u_j^{\ell-1})\, v_j^\star\, d\boldsymbol{x} - \int_{\Omega_j} k_j \,\mathbf{grad}\, u_j^{\ell-1} \cdot \mathbf{grad}\, v_j^\star\, d\boldsymbol{x}, \\ &= \int_{\Omega_j} f_j\, v_j^\star\, d\boldsymbol{x} - \int_{\Omega_j} q_j u_j^{\ell-1}\, v_j^\star\, d\boldsymbol{x} - \int_{\Omega_j} k_j \,\mathbf{grad}\, u_j^{\ell-1} \cdot \mathbf{grad}\, v_j^\star\, d\boldsymbol{x}. \end{aligned}$$

La formulation variationnelle (10.16) se réécrit alors :
Pour $i \in \mathcal{I}_\Omega$, trouver $u_i^\ell \in V_i$ tel que $\forall v_i \in (V_i)_-$,

$$\begin{aligned} &\int_{\Omega_i} k_i \,\mathbf{grad}\, u_i^\ell \cdot \mathbf{grad}\, v_i d\boldsymbol{x} + \int_{\Omega_i} q_i u_i^\ell\, v_i d\boldsymbol{x} + \sum_{j\in\mathcal{I}_{\Omega_i}} \int_{\Sigma_{ij}} \alpha_{ij}^i u_i^\ell\, v_i\, ds \\ &\quad = \int_{\Omega_i} f_i\, v_i d\boldsymbol{x} + \sum_{j\in\mathcal{I}_{\Omega_i}} \int_{\Omega_j} f_j\, v_j^\star\, d\boldsymbol{x} - \sum_{j\in\mathcal{I}_{\Omega_i}} \int_{\Omega_j} q_j u_j^{\ell-1}\, v_j^\star\, d\boldsymbol{x} \\ &\quad - \sum_{j\in\mathcal{I}_{\Omega_i}} \int_{\Omega_j} k_j \,\mathbf{grad}\, u_j^{\ell-1} \cdot \mathbf{grad}\, v_j^\star\, d\boldsymbol{x} + \sum_{j\in\mathcal{I}_{\Omega_i}} \int_{\Sigma_{ij}} \alpha_{ij}^i u_j^{\ell-1} v_j^\star\, ds. \end{aligned} \tag{10.18}$$

Remarque 10.5 *Dans (10.16), il reste possible de choisir une fonction-test $v_i \in V_i$. La difficulté est qu'on peut avoir $v_{i|\Sigma_{ij}} \in H^{1/2}(\Sigma_{ij}) \setminus H_{ij}$. Dans ce cas,* on ne peut pas prolonger cette trace par 0 sur $\partial\Omega_j$ *pour obtenir un élément de $H^{1/2}(\partial\Omega_j)$, et on ne peut plus transformer l'intégrale sur Σ_{ij} en une intégrale sur $\partial\Omega_j$ pour aboutir à (10.18).*

Par la suite, on va utiliser, pour $i \in \mathcal{I}_\Omega$, les formes bilinéaires a_1^i, continues et coercitives sur $V_i \times V_i$ (voir la preuve du théorème 3.1) :

$$\left\{ \begin{array}{rcl} a_1^i : V_i \times V_i & \to & \mathbb{R} \\ (v,w) & \mapsto & \displaystyle\int_{\Omega_i} (k_i \,\mathbf{grad}\, v \cdot \mathbf{grad}\, w + q_i v\, w)\, d\boldsymbol{x} \quad . \end{array} \right. \tag{10.19}$$

La forme bilinéaire $\widetilde{a}_1^i$ associée à la formulation variationnelle (10.18) s'écrit :

$$\left\{ \begin{array}{rcl} \widetilde{a}_1^i : V_i \times V_i & \to & \mathbb{R} \\ (v,w) & \mapsto & \displaystyle a_1^i(v,w) + \sum_{j\in\mathcal{I}_{\Omega_i}} \int_{\Sigma_{ij}} \alpha_{ij}^i v\, w\, ds \quad . \end{array} \right. \tag{10.20}$$

On peut montrer que si $\alpha_{ij}^i \geq 0$, la forme bilinéaire $\widetilde{a}_1^i$ est continue (utiliser le théorème de trace D.1) et coercitive.

Dans la formulation variationnelle (10.18), la solution appartient à V_i, et les fonctions-test à $(V_i)_-$. Lors de la discrétisation on peut choisir de rester dans $(V_i)_-$ en annulant tous les degrés de liberté associés aux sommets situé sur $\partial\Sigma_W$ à la fois pour les fonctions-test discrètes mais aussi pour la solution discrète (de façon à obtenir un système linéaire carré). Sinon on se place dans V_i, et dans ce cas il reste à prendre en compte parmi les fonctions-test discrètes celles qui sont associées aux degrés de liberté situés sur $\partial\Sigma_W$ (encore une fois pour obtenir un système linéaire carré). C'est cette seconde approche qu'on retient par la suite.

10.4 Discrétisation

Considérons comme discrétisation des espaces V_i les éléments finis de Lagrange P_1 décrits au §3.3.1, avec notamment la prise en compte de la condition aux limites de Dirichlet. On définit une famille de triangulations $(\mathcal{T}_{i,h})_h$ de Ω_i constituées de L_i triangles tels que : $\overline{\Omega_i} = \cup_{l\in\mathcal{I}_i^T} T_{i,l}$ avec $\mathcal{I}_i^T = \{1, \cdots, L_i\}$. Parmi les sommets, on ne conserve que ceux qui n'appartiennent pas à Γ_i ; après renumérotation, ce sont les N_i^+ sommets $(M_{i,m})_{m\in\mathcal{I}_i^+}$ tels que $\mathcal{I}_i^+ = \{1, \cdots, N_i^+\}$. Pour $i \in \mathcal{I}_\Omega$, il y a donc trois types de sommets : ceux qui sont à l'intérieur de Ω_i, ceux qui se trouvent sur Σ_i, et enfin ceux qui se trouvent sur le treillis $\partial\Sigma_W$; en particulier, on définit $W_i = \{M_{i,m} \in \partial\Sigma_W,\, m \in \mathcal{I}_i^+\}$ les sommets de la triangulation $\mathcal{T}_{i,h}$ appartenant à $\partial\Sigma_W$.
On décompose $\mathcal{I}_i^+$ en trois sous-ensembles disjoints d'indices : $\mathcal{I}_i^+ = \mathcal{I}_i \cup \mathcal{I}_{i,\Sigma_i} \cup \mathcal{I}_{i,W_i}$ tels que :

$$\left\{\begin{array}{rclrcl} \mathcal{I}_i &=& \{m \in \mathcal{I}_i^+ \,|\, M_{i,m} \in \int(\Omega_i)\}, & N_i &=& |\mathcal{I}_i|, \\ \mathcal{I}_{i,\Sigma_i} &=& \{m \in \mathcal{I}_i^+ \,|\, M_{i,m} \in \Sigma_i\}, & N_{\Sigma_i} &=& |\mathcal{I}_{i,\Sigma_i}|, \\ \mathcal{I}_{i,W_i} &=& \{m \in \mathcal{I}_i^+ \,|\, M_{i,m} \in W_i\}, & N_{W_i} &=& |\mathcal{I}_{i,W_i}|. \end{array}\right. \tag{10.21}$$

On a donc : $N_i^+ = |\mathcal{I}_i^+| = N_i + N_{\Sigma_i} + N_{W_i}$.
On définit enfin les sous-ensembles d'indices suivants, pour $j \in \mathcal{I}_{\Omega_i}$:

$$\left\{\begin{array}{rclrcl} \mathcal{I}_{i,\Sigma_i}^{ij} &=& \{m \in \mathcal{I}_{i,\Sigma_i} \,|\, M_{i,m} \in \Sigma_{ij}\}, & N_{\Sigma_{ij}}^i &=& |\mathcal{I}_{i,\Sigma_i}^{ij}|, \\ \mathcal{I}_{i,W_i}^{ij} &=& \{m \in \mathcal{I}_{i,W_i} \,|\, M_{i,m} \in \partial\Sigma_{ij}\}, & N_{W_i}^{ij} &=& |\mathcal{I}_{i,W_i}^{ij}|, \\ \mathcal{I}_i^{ij} &=& \mathcal{I}_{i,\Sigma_i}^{ij} \cup \mathcal{I}_{i,W_i}^{ij}. & & & \end{array}\right. \tag{10.22}$$

Les sous-ensembles $(\mathcal{I}_{i,W_i}^{ij})_{j\in\mathcal{I}_{\Omega_i}}$ ne sont pas disjoints en général, chaque sommet de W_i pouvant appartenir à deux interfaces parmi $(\partial\Sigma_{ij})_{j\in\mathcal{I}_{\Omega_i}}$. Par contre, $\mathcal{I}_{i,\Sigma_i} = \cup_{j\in\mathcal{I}_{\Omega_i}} \mathcal{I}_{i,\Sigma_i}^{ij}$ et $N_{\Sigma_i} = \sum_{j\in\Omega_i} N_{\Sigma_{ij}}^i$.
On suppose que les couples de triangulations $(\mathcal{T}_{i,h}, \mathcal{T}_{j,h})_{(i,j)\in\mathcal{IJ}}$ partagent :

— leurs $N_{W_{ij}}$ sommets sur le treillis, avec : $N_{W_{ij}} = N_{W_i}^{ij} = N_{W_j}^{ji}$,

— leurs $N_{\Sigma_{ij}}$ sommets sur l'interface Σ_{ij}, avec : $N_{\Sigma_{ij}} = N_{\Sigma_{ij}}^i = N_{\Sigma_{ij}}^j$.

Il peut arriver que le maillage soit tel que, pour un sous-domaine Ω_i, un triangle repose sur deux interfaces distinctes Σ_{ij} et $\Sigma_{ij'}$ $(j \neq j')$. Plus précisément, dans le cas $d = 2$, étudions la Figure 10.4, qui est un zoom sur un point de croisement M_W, partagé entre quatre sous-domaines $(\Omega_i)_{i=1,4}$:

— Dans le sous-domaine Ω_4, délimité par les interfaces Σ_{14} et Σ_{34}, le point de croisement M_W appartient au seul triangle T_1^4. On dit que les sommets M_{14} de l'interface Σ_{14} et M_{34} de l'interface Σ_{34} sont couplés car ils appartiennent au même triangle T_1^4 et par conséquent $a_1^4(w_{1,M_{14}}, w_{1,M_{34}}) \neq 0$.

— Dans le sous-domaine Ω_3, délimité par les interfaces Σ_{23} et Σ_{34}, le point de croisement M_W appartient aux deux triangles T_1^3 et T_2^3. On dit que les sommets M_{23} de l'interface Σ_{23} et M_{34} de l'interface Σ_{34} ne sont pas couplés car ils n'appartiennent pas au même triangle et par conséquent $a_1^3(w_{3,M_{23}}, w_{3,M_{34}}) = 0$.

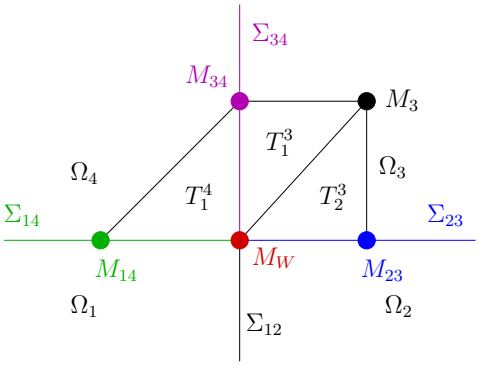


FIGURE 10.1 – Point de croisement

L'espace de discrétisation de V_i est noté $V_{i,h}^+$ et est tel que :

$$V_{i,h}^+ := \left\{ v_{i,h} \in C^0(\overline{\Omega_i}) \ : \ \forall l \in \mathcal{I}_i^T, \ v_{i,h|T_l} \in P_1(T_l), \ v_{i,h|\Gamma_i} = 0 \right\}. \tag{10.23}$$

On appelle $(w_{i,m})_{m\in\mathcal{I}_i^+}$ les fonctions de base "chapeau" associées à $V_{i,h}^+ := vect\left((w_{i,m})_{m\in\mathcal{I}_i^+}\right)$. Pour tout $(i,j) \in \mathcal{IJ}$, les sommets des triangulations $\mathcal{T}_{i,h}$ et $\mathcal{T}_{j,h}$ étant identiques sur l'interface Σ_{ij} et son bord $\partial\Sigma_{ij}$, les espaces des traces $\mathfrak{T}_{i,h}^{ij}$ et $\mathfrak{T}_{j,h}^{ij}$ tels que

$$\mathfrak{T}_{i,h}^{ij} := vect\left((w_{i,m|\Sigma_{ij}})_{m\in\mathcal{I}_i^{ij}} \right), \quad \mathfrak{T}_{j,h}^{ij} := vect\left((w_{j,m|\Sigma_{ij}})_{m\in\mathcal{I}_j^{ij}} \right) \tag{10.24}$$

sont égaux : pour tout $m_i \in \mathcal{I}_i^{ij}$ (resp. $m_j \in \mathcal{I}_j^{ij}$) il existe un unique $m_j \in \mathcal{I}_j^{ij}$ (resp. $m_i \in \mathcal{I}_i^{ij}$) tel que $w_{i,m_i|\Sigma_{ij}} = w_{j,m_j|\Sigma_{ij}}$ (les fonctions de base des éléments finis de Lagrange P_1 sont continues).
On précise maintenant les choix de relèvement (discret).

$$\begin{aligned} &\forall (i,j) \in \mathcal{IJ}, \ \forall (m_i, m_j) \in \mathcal{I}_i^{ij} \times \mathcal{I}_j^{ij} \text{ tel que } M_{m_i} = M_{m_j} \ : \\ &\text{- la fonction de base } w_{j,m_j} \text{ est un relèvement de } w_{i,m_i|\Sigma_{ij}} \text{ dans } \Omega_j \, ; \\ &\text{- la fonction de base } w_{i,m_i} \text{ est un relèvement de } w_{j,m_j|\Sigma_{ij}} \text{ dans } \Omega_i. \end{aligned} \tag{10.25}$$

Pour $m \in \mathcal{I}_{i,W_i}$, on note $\mathcal{I}_{\Sigma_i}^m = \{j \in \mathcal{I}_{\Omega_i} \,|\, M_{i,m} \in \overline{\Sigma_{ij}}\}$.
On cherche $(u_{i,h}^\ell)_{i\in\mathcal{I}_\Omega}$ appartenant à l'espace produit $\prod_{i\in\mathcal{I}_\Omega} V_{i,h}^+$ une approximation de $(u_i^\ell)_{i\in\mathcal{I}_\Omega}$ telle que pour tout $i \in \mathcal{I}_\Omega$, $u_{i,h}^\ell = \sum_{m'\in\mathcal{I}_i^+} U_{i,m'}^\ell w_{i,m'}$, où $U_{i,m'}^\ell := u_{i,h}^\ell(M_{i,m'})$.
On raisonne d'abord par analogie avec (10.18). En effet, si on considère $w_{i,m}$ avec $i \in \mathcal{I}_i \cup \mathcal{I}_{i,\Sigma_i}$, alors on a $w_{i,m|\Sigma_{ij}} \in H_{ij}$ pour tout $j \in \mathcal{I}_{\Omega_i}$. On obtient alors :
Trouver $(u_{i,h}^\ell)_{i\in\mathcal{I}_\Omega} \in \prod_{i\in\mathcal{I}_\Omega} V_{i,h}^+$ tel que $\forall i \in \mathcal{I}_\Omega$, $\forall j \in \mathcal{I}_{\Omega_i}$:

$$\left\{\begin{array}{ll} \forall m \in \mathcal{I}_i, & a_1^i(u_{i,h}^\ell, w_{i,m}) = \displaystyle\int_{\Omega_i} f_i\, w_{i,m} d\boldsymbol{x}, \\[2ex] \forall m_i \in \mathcal{I}_{i,\Sigma_i}^{ij}, & a_1^i(u_{i,h}^\ell, w_{i,m_i}) + \displaystyle\int_{\Sigma_{ij}} \alpha_{ij}^i u_{i,h}^\ell w_{i,m_i}\, ds \\ & \quad = \displaystyle\int_{\Omega_i} f_i\, w_{i,m_i} d\boldsymbol{x} + \int_{\Omega_j} f_j\, w_{j,m_j} d\boldsymbol{x} \\ & \quad -a_1^j(u_{j,h}^{\ell-1}, w_{j,m_j}) + \displaystyle\int_{\Sigma_{ij}} \alpha_{ij}^i u_{j,h}^{\ell-1} w_{j,m_j}\, ds. \end{array}\right. \tag{10.26}$$

Si maintenant $m \in \mathcal{I}_{i,W_i}$, alors, pour $j \in \mathcal{I}_{\Omega_i}$, soit $w_{i,m|\Sigma_{ij}} = 0$, soit $w_{i,m|\Sigma_{ij}} \neq 0$ et par définition $w_{i,m|\Sigma_{ij}} \notin H_{ij}$. Dans le premier cas, l'intégrale sur Σ_{ij} s'annule. Dans le second cas, on ne peut plus procéder comme avant. Néanmoins, on sait qu'un relèvement (discret) existe d'après (10.25). Il suffit alors de reprendre le même raisonnement qu'avant avec ce relèvement. On obtient alors :
Trouver $(u_{i,h}^\ell)_{i\in\mathcal{I}_\Omega} \in \prod_{i\in\mathcal{I}_\Omega} V_{i,h}^+$ tel que $\forall i \in \mathcal{I}_\Omega$, $\forall j \in \mathcal{I}_{\Omega_i}$:

$$\left\{\begin{array}{ll} \forall m_i \in \mathcal{I}_{i,W_i}, & a_1^i(u_{i,h}^\ell, w_{i,m_i}) + \displaystyle\sum_{j\in\mathcal{I}_{\Sigma_i}^m} \int_{\Sigma_{ij}} \alpha_{ij}^i u_{i,h}^\ell w_{i,m_i}\, ds \\[2ex] (\text{déf. } m_j \text{ cf. } (10.25)) & \quad = \displaystyle\int_{\Omega_i} f_i\, w_{i,m_i} d\boldsymbol{x} + \sum_{j\in\mathcal{I}_{\Sigma_i}^{m_i}} \int_{\Omega_j} f_j\, w_{j,m_j} d\boldsymbol{x} \\ & \quad + \displaystyle\sum_{j\in\mathcal{I}_{\Sigma_i}^{m_i}} \left(-a_1^j(u_{j,h}^{\ell-1}, w_{j,m_j}) + \int_{\Sigma_{ij}} \alpha_{ij}^i u_{j,h}^{\ell-1} w_{j,m_j}\, ds \right). \end{array}\right. \tag{10.27}$$

Après avoir décomposé $u_{i,h}^{\ell}$ sur la base $(w_{i,m})_{m\in\mathcal{I}_i^+}$, (10.26)-(10.27) se réécrit :
Trouver $(U_{i,m'}^{\ell})_{i\in\mathcal{I}_\Omega, m'\in\mathcal{I}_i^+}$ tel que $\forall i\in\mathcal{I}_\Omega$, $\forall j\in\mathcal{I}_{\Omega_i}$:

$$
\left\{
\begin{array}{ll}
\forall m\in\mathcal{I}_i, & \displaystyle\sum_{m'\in\mathcal{I}_i^+} a_1^i(w_{i,m'},w_{i,m})U_{i,m'}^{\ell} = \int_{\Omega_i} f_i\, w_{i,m} d\boldsymbol{x},\\[2ex]
\forall m_i\in\mathcal{I}_{i,\Sigma_i}^{ij}, & \displaystyle\sum_{m'\in\mathcal{I}_i^+}\left(a_1^i(w_{i,m'},w_{i,m_i}) + \int_{\Sigma_{ij}} \alpha_{ij}^i w_{i,m'}w_{i,m_i}\,ds\right)U_{i,m'}^{\ell}\\
& \qquad = \displaystyle\int_{\Omega_i} f_i\, w_{i,m_i} d\boldsymbol{x} + \int_{\Omega_j} f_j\, w_{j,m_j} d\boldsymbol{x}\\
& \quad + \displaystyle\sum_{m'\in\mathcal{I}_j^+}\left(-a_1^j(w_{j,m'},w_{j,m_j}) + \int_{\Sigma_{ij}} \alpha_{ij}^i w_{j,m'}w_{j,m_j}\,ds\right)U_{j,m'}^{\ell-1},\\[2ex]
\forall m_i\in\mathcal{I}_{i,W_i}, & \displaystyle\sum_{m'\in\mathcal{I}_i^+}\left(a_1^i(w_{i,m'},w_{i,m_i}) + \sum_{j\in\mathcal{I}_{\Sigma_i}^{m_i}}\int_{\Sigma_{ij}} \alpha_{ij}^i w_{i,m'}w_{i,m_i}\,ds\right)U_{i,m'}^{\ell}\\
& \qquad = \displaystyle\int_{\Omega_i} f_i\, w_{i,m_i} d\boldsymbol{x} + \sum_{j\in\mathcal{I}_{\Sigma_i}^{m_i}}\int_{\Omega_j} f_j\, w_{j,m_j} d\boldsymbol{x}\\
& \quad + \displaystyle\sum_{j\in\mathcal{I}_{\Sigma_i}^{m_i}}\sum_{m'\in\mathcal{I}_j^+}\left(-a_1^j(w_{j,m'},w_{j,m_j}) + \int_{\Sigma_{ij}} \alpha_{ij}^i w_{j,m'}w_{j,m_j}\,ds\right)U_{j,m'}^{\ell-1}.
\end{array}
\right. \tag{10.28}
$$

On peut mettre les équations (10.28) sous forme matricielle (voir §3.3).
Soit $i\in\mathcal{I}_\Omega$. Soit $\mathbb{A}_{i,i}\in\mathbb{R}^{N_i^+\times N_i^+}$ la matrice symétrique telle que :

$$\forall (m,m')\in\mathcal{I}_i^+\times\mathcal{I}_i^+, (\mathbb{A}_{i,i})_{m,m'} = a_1^i(w_{i,m'},w_{i,m}).$$

Soit $\widetilde{\mathbb{A}}_{i,i}\in\mathbb{R}^{N_i^+\times N_i^+}$ la matrice symétrique telle que $\forall j\in\mathcal{I}_{\Omega_i}$:

$$
\begin{array}{lll}
\forall (m,m')\in\mathcal{I}_i\times\mathcal{I}_i^+, & (\widetilde{\mathbb{A}}_{i,i})_{m,m'} &= (\mathbb{A}_{i,i})_{m,m'},\\
\forall (m,m')\in\mathcal{I}_{i,\Sigma_i}^{ij}\times\mathcal{I}_i^+, & (\widetilde{\mathbb{A}}_{i,i})_{m,m'} &= (\mathbb{A}_{i,i})_{m,m'} + \displaystyle\int_{\Sigma_{ij}} \alpha_{ij}^i w_{i,m'}w_{i,m}\,ds,\\
\forall (m,m')\in\mathcal{I}_{i,W_i}\times\mathcal{I}_i^+, & (\widetilde{\mathbb{A}}_{i,i})_{m,m'} &= (\mathbb{A}_{i,i})_{m,m'} + \displaystyle\sum_{j'\in\mathcal{I}_{\Sigma_i}^m}\int_{\Sigma_{ij'}} \alpha_{ij'}^i w_{i,m'}w_{i,m}\,ds.
\end{array}
$$

Soit $i\in\mathcal{I}_\Omega$ et $j\in\mathcal{I}_{\Omega_i}$. Soit $\widetilde{\mathbb{A}}_{i,j}\in\mathbb{R}^{N_i^+\times N_j^+}$ la matrice de couplage entre Ω_i et Ω_j :

$$
\begin{array}{lll}
\forall (m,m')\in\mathcal{I}_i\times\mathcal{I}_j^+, & (\widetilde{\mathbb{A}}_{i,j})_{m,m'} &= 0,\\
\forall (m_i,m')\in\mathcal{I}_i^{ij}\times\mathcal{I}_j^+, & (\widetilde{\mathbb{A}}_{i,j})_{m_i,m'} &= a_1^j(w_{j,m'},w_{j,m_j}) - \displaystyle\int_{\Sigma_{ij}} \alpha_{ij}^i w_{j,m'}w_{j,m_j}\,ds,
\end{array}
$$

où l'expression de $(\widetilde{\mathbb{A}}_{i,j})_{(m_i,m')\in\mathcal{I}_i^{ij}\times\mathcal{I}_j^+}$ est obtenue suite au choix du relèvement discret sur l'interface Σ_{ij}.

Soit $\widetilde{U}_i^\ell \in \mathbb{R}^{N_i^+}$ le vecteur tel que pour $m \in \mathcal{I}_i^+$: $\left(\widetilde{U}_i^\ell\right)_m := U_{i,m}^\ell$.
Soit $\widetilde{F}_i \in \mathbb{R}^{N_i^+}$ le vecteur tel que $\forall j \in \mathcal{I}_{\Omega_i}$:

$$\begin{aligned}
\forall m \in \mathcal{I}_i, \qquad \left(\widetilde{F}_i\right)_m &= \int_{\Omega_i} f w_{i,m}\, d\boldsymbol{x}, \\
\forall m_i \in \mathcal{I}_{i,\Sigma_i}^{ij}, \quad \left(\widetilde{F}_i\right)_{m_i} &= \int_{\Omega_i} f w_{i,m_i}\, d\boldsymbol{x} + \int_{\Omega_j} f w_{j,m_j}\, d\boldsymbol{x}, \\
\forall m_i \in \mathcal{I}_{W_i}, \quad \left(\widetilde{F}_i\right)_{m_i} &= \int_{\Omega_i} f w_{i,m_i}\, d\boldsymbol{x} + \sum_{j \in \mathcal{I}_{\Sigma_i}^{m_i}} \int_{\Omega_j} f w_{j,m_j}\, d\boldsymbol{x}
\end{aligned}$$

où l'expression de $\left(\widetilde{F}_i\right)_{m_i \in \mathcal{I}_i^{ij}}$ est également obtenue suite au choix des relèvements discrets sur les interfaces $(\,\Sigma_{ij}\,)_{j \in \mathcal{I}_{\Sigma_i}^{m_i}}$.
La discrétisation de l'algorithme (10.5) s'écrit finalement :

initialisation : $\forall i \in \mathcal{I}_\Omega$, $\widetilde{U}_i^0 \in \mathbb{R}^{N_i^+}$ **est donné.**
itérations : pour $\ell = 1, \ldots,$ $\forall i \in \mathcal{I}_\Omega$, **trouver** $\widetilde{U}_i^\ell \in \mathbb{R}^{N_i^+}$ **tel que :**

$$\widetilde{\mathbb{A}}_{i,i}\widetilde{U}_i^\ell = \widetilde{F}_i - \sum_{j \in \mathcal{I}_{\Omega_i}} \widetilde{\mathbb{A}}_{i,j}\widetilde{U}_j^{\ell-1}, \tag{10.29}$$

jusqu'à convergence.

10.5 Interprétation algébrique

On pose $N^+ = \sum_{i \in \mathcal{I}_\Omega} N_i^+$.
Soit $\widetilde{\mathbb{A}} \in \mathbb{R}^{N^+ \times N^+}$ la matrice définie par $N_\Omega \times N_\Omega$ blocs (voir le §5.3). Les blocs non nuls sont les blocs diagonaux : $\forall i \in \mathcal{I}_\Omega$, $[\widetilde{\mathbb{A}}]_{i,i} = \widetilde{\mathbb{A}}_{i,i}$, et les blocs extra-diagonaux correspondants à des sous-domaines voisins : $\forall i \in \mathcal{I}_\Omega$, $\forall j \in \mathcal{I}_{\Omega_i}$, $[\widetilde{\mathbb{A}}]_{i,j} = \widetilde{\mathbb{A}}_{i,j}$.
Soit le vecteur $\widetilde{F} \in \mathbb{R}^{N^+}$ constitué de N_Ω blocs et tel que : $\forall i \in \mathcal{I}_\Omega$, $[\widetilde{F}]_i = \widetilde{F}_i$.
L'algorithme (10.29) s'interprète comme un *algorithme de Jacobi par blocs* (décrit en (6.16)) appliqué au système linéaire suivant :

$$\text{Trouver } \widetilde{U} \in \mathbb{R}^{N^+} \text{ tel que : } \widetilde{\mathbb{A}}\widetilde{U} = \widetilde{F}, \tag{10.30}$$

pour lequel la matrice $D_B \in \mathbb{R}^{N^+ \times N^+}$ est diagonale par blocs, avec $\forall i \in \mathcal{I}_\Omega$, $[D_B]_{i,i} = \widetilde{\mathbb{A}}_{i,i}$, et $E_B + F_B = -(\widetilde{\mathbb{A}} - D_B)$ est la matrice contenant les sous-blocs extra-diagonaux.

Remarque 10.6 *La matrice D_B est symétrique car ses blocs diagonaux le sont, mais la matrice $\widetilde{\mathbb{A}}$ n'est pas symétrique car $\forall i \in \mathcal{I}_\Omega$, $\forall \in \mathcal{I}_{\Omega_i}$, $\widetilde{\mathbb{A}}_{i,j} \neq \widetilde{\mathbb{A}}_{j,i}$. Ainsi, si on veut utiliser un algorithme (parallélisable) plus performant que l'algorithme de Jacobi (10.29) pour la résolution de* (10.30)*, on peut choisir l'algorithme GMRES, préconditionné par D_B (§7.3).*

Le système linéaire (10.30), issu d'une méthode de décomposition de domaine, s'obtient également de façon algébrique, en manipulant les sous-blocs du système linéaire obtenu

après discrétisation du problème *sans décomposition de domaine*.
Plus précisément, étudions le cas de deux sous-domaines. Dans cette configuration simplifiée (sans point de croisement), on note que : soit Σ_{12} est fermée ($\partial\Omega_1$ ou $\partial\Omega_2$ n'intersecte pas $\partial\Omega$), et dans ce cas le treillis $\partial\Sigma_W$ est vide ; soit $\partial\Omega_1$ et $\partial\Omega_2$ intersectent $\partial\Omega$, et dans ce cas $\partial\Sigma_W \subset \partial\Omega$. Dans les deux cas, il n'y a aucune inconnue associée aux sommets de $\partial\Sigma_W$ (dans le second cas, à cause de la condition aux limites). On note $\Sigma = \Sigma_{12}$ dans la suite.
La discrétisation P_1 du problème (9.1) dans $V_h \subset H_0^1(\Omega)$ défini sur la triangulation $\mathcal{T}_h = \mathcal{T}_{1,h} \cup \mathcal{T}_{2,h}$ s'écrit : [48]

$$\text{Trouver } U \in \mathbb{R}^{N_1+N_\Sigma+N_2} \text{ tel que } AU = F, \text{ avec :}$$
$$A := \begin{pmatrix} \mathring{\mathbb{A}}_{1,1} & \mathbb{A}_{1,\Sigma} & 0 \\ (\mathbb{A}_{1,\Sigma})^T & \mathbb{A}^1_{\Sigma,\Sigma} + \mathbb{A}^2_{\Sigma,\Sigma} & (\mathbb{A}_{2,\Sigma})^T \\ 0 & \mathbb{A}_{2,\Sigma} & \mathring{\mathbb{A}}_{2,2} \end{pmatrix}, \quad U := \begin{pmatrix} U_1 \\ U_\Sigma \\ U_2 \end{pmatrix}, \quad F := \begin{pmatrix} F_1 \\ F_\Sigma \\ F_2 \end{pmatrix}. \tag{10.31}$$

Ci-dessus, $u_{h|\Sigma}$ et donc U_Σ sont définis de manière univoque.

En introduisant $U_{1,\Sigma}$ et $U_{2,\Sigma}$: $U_{1,\Sigma} = U_{2,\Sigma} = U_\Sigma$, on peut réécrire le système linéaire $AU = F$ ainsi :

$$\begin{pmatrix} \mathring{\mathbb{A}}_{1,1} & \mathbb{A}_{1,\Sigma} & 0 & 0 \\ (\mathbb{A}_{1,\Sigma})^T & \mathbb{A}^1_{\Sigma,\Sigma} & (\mathbb{A}_{2,\Sigma})^T & \mathbb{A}^2_{\Sigma,\Sigma} \\ 0 & 0 & \mathring{\mathbb{A}}_{2,2} & \mathbb{A}_{2,\Sigma} \\ (\mathbb{A}_{1,\Sigma})^T & \mathbb{A}^1_{\Sigma,\Sigma} & (\mathbb{A}_{2,\Sigma})^T & \mathbb{A}^2_{\Sigma,\Sigma} \end{pmatrix} \begin{pmatrix} U_1 \\ U_{1,\Sigma} \\ U_2 \\ U_{2,\Sigma} \end{pmatrix} = \begin{pmatrix} F_1 \\ F_\Sigma \\ F_2 \\ F_\Sigma \end{pmatrix}. \tag{10.32}$$

Si, pour $i \in \{1,2\}$, les blocs diagonaux $\begin{pmatrix} \mathring{\mathbb{A}}_{i,i} & \mathbb{A}_{i,\Sigma} \\ (\mathbb{A}_{i,\Sigma})^T & \mathbb{A}^i_{\Sigma,\Sigma} \end{pmatrix}$ sont bien inversibles, en revanche la matrice globale de ce système linéaire ne l'est pas, puisque les deuxième et dernier blocs de lignes sont identiques.
En fait, il manque l'information que $U_{1,\Sigma} = U_{2,\Sigma}$. Soient $\widetilde{\mathbb{M}}^i_{\Sigma,\Sigma} \in \mathbb{R}^{N_\Sigma \times N_\Sigma}$, $i \in \{1,2\}$, deux matrices inversibles, telles que les matrices $\mathbb{A}^i_{\Sigma,\Sigma} + \widetilde{\mathbb{M}}^i_{\Sigma,\Sigma}$, $i \in \{1,2\}$, et $\widetilde{\mathbb{M}}^1_{\Sigma,\Sigma} + \widetilde{\mathbb{M}}^2_{\Sigma,\Sigma}$ soient également inversibles. Considérons maintenant la réécriture de (10.31) via le système linéaire *inversible* suivant :

$$\begin{pmatrix} \mathring{\mathbb{A}}_{1,1} & \mathbb{A}_{1,\Sigma} & 0 & 0 \\ (\mathbb{A}_{1,\Sigma})^T & \mathbb{A}^1_{\Sigma,\Sigma} + \widetilde{\mathbb{M}}^1_{\Sigma,\Sigma} & (\mathbb{A}_{2,\Sigma})^T & \mathbb{A}^2_{\Sigma,\Sigma} - \widetilde{\mathbb{M}}^1_{\Sigma,\Sigma} \\ 0 & 0 & \mathring{\mathbb{A}}_{2,2} & \mathbb{A}_{2,\Sigma} \\ (\mathbb{A}_{1,\Sigma})^T & \mathbb{A}^1_{\Sigma,\Sigma} - \widetilde{\mathbb{M}}^2_{\Sigma,\Sigma} & (\mathbb{A}_{2,\Sigma})^T & \mathbb{A}^2_{\Sigma,\Sigma} + \widetilde{\mathbb{M}}^2_{\Sigma,\Sigma} \end{pmatrix} \begin{pmatrix} U_1 \\ U_{1,\Sigma} \\ U_2 \\ U_{2,\Sigma} \end{pmatrix} = \begin{pmatrix} F_1 \\ F_\Sigma \\ F_2 \\ F_\Sigma \end{pmatrix} \tag{10.33}$$

48. Pour $i = 1, 2$, on utilise la décomposition

$$\mathbb{A}_{i,i} = \begin{pmatrix} \mathring{\mathbb{A}}_{i,i} & \mathbb{A}_{i,\Sigma} \\ (\mathbb{A}_{i,\Sigma})^T & \mathbb{A}^i_{\Sigma,\Sigma} \end{pmatrix},$$

avec $\mathring{\mathbb{A}}_{i,i}$ correspondant aux interactions entre les N_i sommets situés dans Ω_i, $\mathbb{A}_{i,\Sigma}$ correspondant aux interactions entre les sommets situés dans Ω_i et les N^i_Σ sommets situés sur $\partial\Omega_i \cap \Sigma$, et enfin $\mathbb{A}^i_{\Sigma,\Sigma}$ correspondant aux contributions venant de Ω_i des interactions entre les sommets situés sur l'interface. Dans le cas particulier de deux sous-domaines, on a noté que $\partial\Omega_1 \cap \Sigma = \partial\Omega_2 \cap \Sigma$.

Pour $i \in \{1,2\}$, les blocs diagonaux $\begin{pmatrix} \overset{\circ}{\mathbb{A}}_{i,i} & \mathbb{A}_{i,\Sigma} \\ (\mathbb{A}_{i,\Sigma})^T & \mathbb{A}^i_{\Sigma,\Sigma} + \widetilde{\mathbb{M}}^i_{\Sigma,\Sigma} \end{pmatrix}$ sont toujours inversibles, alors que les deuxième et dernier blocs de lignes ne sont plus identiques. En effet, on a :

$$\left\{ \begin{array}{rcl} (\mathbb{A}_{1,\Sigma})^T U_1 + (\mathbb{A}^1_{\Sigma,\Sigma} + \widetilde{\mathbb{M}}^1_{\Sigma,\Sigma})\, U_{1,\Sigma} + (\mathbb{A}_{2,\Sigma})^T U_2 + (\mathbb{A}^2_{\Sigma,\Sigma} - \widetilde{\mathbb{M}}^1_{\Sigma,\Sigma})\, U_{2,\Sigma} & = & F_\Sigma, \\ (\mathbb{A}_{1,\Sigma})^T U_1 + (\mathbb{A}^1_{\Sigma,\Sigma} - \widetilde{\mathbb{M}}^2_{\Sigma,\Sigma})\, U_{1,\Sigma} + (\mathbb{A}_{2,\Sigma})^T U_2 + (\mathbb{A}^2_{\Sigma,\Sigma} + \widetilde{\mathbb{M}}^2_{\Sigma,\Sigma})\, U_{2,\Sigma} & = & F_\Sigma, \end{array} \right.$$

et la différence entre ces deux équations redonne $U_{1,\Sigma} = U_{2,\Sigma}$, puisque $\widetilde{\mathbb{M}}^1_{\Sigma,\Sigma} + \widetilde{\mathbb{M}}^2_{\Sigma,\Sigma}$ est inversible.
La matrice du système linéaire (10.33) est similaire à la matrice $\widetilde{\mathbb{A}}$ du système linéaire (10.30). Cette approche algébrique permet d'exhiber un choix optimal des matrices $\widetilde{\mathbb{M}}^i_{\Sigma,\Sigma}$. On peut en effet montrer que si $\widetilde{\mathbb{M}}^i_{\Sigma,\Sigma}$ est choisie égale à $\mathbb{A}^j_{\Sigma,\Sigma} - (\mathbb{A}_{j,\Sigma})^T (\overset{\circ}{\mathbb{A}}_{j,j})^{-1} \mathbb{A}_{j,\Sigma}$ pour $(i,j) \in \{(1,2),(2,1)\}$, alors la convergence de l'algorithme de Jacobi est optimale (en 1 itération). Mais ces matrices sont coûteuses à calculer du fait de la présence de $(\overset{\circ}{\mathbb{A}}_{j,j})^{-1}$. En pratique lorsque $q_i \neq 0$, il est plus simple de prendre $\widetilde{\mathbb{M}}^i_{\Sigma,\Sigma} = \mathbb{M}^i_{\Sigma,\Sigma}$, avec un choix optimisé des paramètres α_i.

Chapitre 11

Problème à deux inconnues, méthode de Schwarz

11.1 Approche continue

On suppose que la solution $(u, \boldsymbol{p})$ est telle que :

$$\text{Pour tout } (i,j) \in \mathcal{I}_\Omega \times \mathcal{I}_{\Omega_i}, \quad \boldsymbol{p}_i \cdot \boldsymbol{n}_{ij|\Sigma_{ij}} \in L^2(\Sigma_{ij}),$$

et on introduit les espaces $\mathbf{Q}_i$, pour $i \in \mathcal{I}_\Omega$ munis de la norme canonique :

$$\mathbf{Q}_i := \left\{ \boldsymbol{q} \in \mathbf{H}(\mathrm{div}, \Omega_i) \,:\, \boldsymbol{q} \cdot \boldsymbol{n}_{i|\Sigma_i} \in L^2(\Sigma_i) \right\}, \ \|\boldsymbol{q}\|^2_{\mathbf{Q}_i} := \|\boldsymbol{q}\|^2_{\mathbf{H}(\mathrm{div},\Omega_i)} + \|\boldsymbol{q}\|^2_{L^2(\Sigma_i)}. \quad (11.1)$$

Pour tout $(i,j) \in \mathcal{I}_\Omega \times \mathcal{I}_{\Omega_i}$, on introduit les couples $(\alpha^i_{ij}, \alpha^j_{ij})$ de coefficients positifs ou nuls tels que $\alpha^i_{ij} + \alpha^j_{ij} > 0$. Les conditions de transmission de Dirichlet (9.17)-(ii) et de Neumann (9.17)-(iii) peuvent être combinées pour écrire les conditions de Robin suivantes :

$$\text{Pour tout } (i,j) \in \mathcal{IJ}, \left\{ \begin{array}{rcll} \boldsymbol{p}_i \cdot \boldsymbol{n}_{ij} + \alpha^i_{ij} u_i & = & \boldsymbol{p}_j \cdot \boldsymbol{n}_{ij} + \alpha^i_{ij} u_j & \text{sur } \Sigma_{ij}, \\ \boldsymbol{p}_j \cdot \boldsymbol{n}_{ij} - \alpha^j_{ij} u_j & = & \boldsymbol{p}_i \cdot \boldsymbol{n}_{ij} - \alpha^j_{ij} u_i & \text{sur } \Sigma_{ij}. \end{array} \right. \quad (11.2)$$

On peut également les poser sous la forme d'une seule équation de la façon suivante :

$$\text{Pour tout } (i,j) \in \mathcal{I} \times \mathcal{I}_{\Omega_i}, \quad \boldsymbol{p}_i \cdot \boldsymbol{n}_i + \alpha^i_{ij} u_i = -\boldsymbol{p}_j \cdot \boldsymbol{n}_j + \alpha^i_{ij} u_j \text{ sur } \Sigma_{ij}. \quad (11.3)$$

Le problème (9.16) est équivalent au problème :
Trouver $(u, \boldsymbol{p}) \in \mathcal{P}V \times \prod_{i \in \mathcal{I}_\Omega} \mathbf{Q}_i$ tel que pour tout $i \in \mathcal{I}_\Omega$ et pour tout $j \in \mathcal{I}_{\Omega_i}$:

$$\left\{ \begin{array}{lrcll} (0) & \mathrm{div}\, \boldsymbol{p}_i + q_i u_i & = & f_i & \text{dans } \Omega_i, \\ (i) & k_i^{-1} \boldsymbol{p}_i + \mathbf{grad}\, u_i & = & 0 & \text{dans } \Omega_i, \\ (ii) & \boldsymbol{p}_i \cdot \boldsymbol{n}_i + \alpha^i_{ij} u_i & = & -\boldsymbol{p}_j \cdot \boldsymbol{n}_j + \alpha^i_{ij} u_j & \text{sur } \Sigma_{ij}. \end{array} \right. \quad (11.4)$$

Comme on ne connaît pas les valeurs de $u_{i|\Sigma}$ ni celles de $\boldsymbol{p}_i \cdot \boldsymbol{n}_{i|\Sigma}$, on propose de résoudre les problèmes (11.4) localement de façon itérative dans les $(\Omega_i)_{i \in \mathcal{I}_\Omega}$, et d'approcher les

conditions de transmission. Afin de résoudre le problème (11.4), on propose maintenant l'algorithme de Schwarz itératif suivant :

initialisation
$\forall i \in \mathcal{I}_\Omega, \quad (u_i^0, \boldsymbol{p}_i^0) \in V_i \times \mathbf{Q}_i$ **est donné.**
itérations : pour $\ell = 1, \ldots, \quad \forall i \in \mathcal{I}_\Omega,$ **trouver** $(u_i^\ell, \boldsymbol{p}_i^\ell) \in V_i \times \mathbf{Q}_i$ **tel que**

$$\left\{ \begin{array}{lrcll} (0) & \operatorname{div} \boldsymbol{p}_i^\ell + q_i\, u_i^\ell & = & f_i & \text{dans } \Omega_i, \\ (i) & k_i^{-1} \boldsymbol{p}_i^\ell + \mathbf{grad}\, u_i^\ell & = & 0 & \text{dans } \Omega_i, \\ (ii) & \forall j \in \mathcal{I}_{\Omega_i},\ -\boldsymbol{p}_i^\ell \cdot \boldsymbol{n}_i + \alpha_{ij}^i u_i^\ell & = & \boldsymbol{p}_j^{\ell-1} \cdot \boldsymbol{n}_j + \alpha_{ij}^i u_j^{\ell-1} & \text{sur } \Sigma_{ij}, \end{array} \right. \tag{11.5}$$

jusqu'à convergence.

On fait l'hypothèse que le choix initial est tel que

$$\text{Pour tout } (i,j) \in \mathcal{I}_\Omega \times \mathcal{I}_{\Omega_i}, \quad \boldsymbol{p}_i^0 \cdot \boldsymbol{n}_{ij|\Sigma_{ij}} \in L^2(\Sigma_{ij}).$$

Théorème 11.1 *Supposons que* $\forall (i,j) \in \mathcal{IJ}$, $\alpha_{ij}^i = \alpha_{ij}^j$. *Alors la suite* $((u_i^\ell, \boldsymbol{p}_i^\ell)_{i \in \mathcal{I}_\Omega})_\ell$ *converge vers* $(u, \boldsymbol{p})$ *dans* $\mathcal{P}V \times \mathcal{P}\mathbf{H}(\operatorname{div}, \Omega)$.

Démonstration : La preuve est similaire à celle du théorème 10.2. On note $\forall (i,j) \in \mathcal{IJ}$, $\alpha_{ij} := \alpha_{ij}^i = \alpha_{ij}^j > 0$. On introduit la suite de couple d'erreurs $\left((e_i^\ell, \boldsymbol{e}_i^\ell)_{\ell \in \mathbb{N}} \right)_{i \in \mathcal{I}_\Omega}$, où : $e_i^\ell := u_i^\ell - u_i \in V_i$ et $\boldsymbol{e}_i^\ell := \boldsymbol{p}_i^\ell - \boldsymbol{p}_i \in \mathbf{H}(\operatorname{div}, \Omega_i)$. Les suites $\left((e_i^\ell, \boldsymbol{e}_i^\ell)_{\ell \in \mathbb{N}} \right)_{i \in \mathcal{I}_\Omega}$ satisfont pour tout $i \in \mathcal{I}_{\Omega_i}$, et pour tout $j \in \mathcal{I}_{\Omega_i}$:

$$\left\{ \begin{array}{lrcll} (0) & \operatorname{div} \boldsymbol{e}_i^\ell + q_i e_i^\ell & = & 0 & \text{dans } \Omega_i, \\ (i) & k_i^{-1} \boldsymbol{e}_i^\ell + \mathbf{grad}\, e_i^\ell & = & 0 & \text{dans } \Omega_i, \\ (ii) & -\boldsymbol{e}_i^\ell \cdot \boldsymbol{n}_i + \alpha_{ij} e_i^\ell & = & \boldsymbol{e}_j^{\ell-1} \cdot \boldsymbol{n}_j + \alpha_{ij} e_j^{\ell-1} & \text{sur } \Sigma_{ij}. \end{array} \right. \tag{11.6}$$

Par récurrence sur ℓ, on vérifie facilement que $\boldsymbol{e}_i^\ell \cdot \boldsymbol{n}_{i|\Sigma_{ij}} \in L^2(\Sigma_{ij})$. On multiplie (11.6)-(0) par e_i^ℓ et on intègre sur Ω_i, ce qui donne :

$$\int_{\Omega_i} \operatorname{div} \boldsymbol{e}_i^\ell \, e_i^\ell \, d\boldsymbol{x} + \| \sqrt{q_i} e_i^\ell \|^2_{L^2(\Omega_i)} = 0. \tag{11.7}$$

On multiplie (11.6)-(i) par $\boldsymbol{e}_i^\ell$, on intègre sur Ω_i, puis on utilise la formule d'intégration par parties (D.1) pour obtenir :

$$\| \sqrt{k_i^{-1}} \boldsymbol{e}_i^\ell \|^2_{L^2(\Omega_i)} - \int_{\Omega_i} \operatorname{div} \boldsymbol{e}_i^\ell \, e_i^\ell \, d\boldsymbol{x} = - \sum_{j \in \mathcal{I}_{\Omega_i}} \int_{\Sigma_{ij}} \boldsymbol{e}_i^\ell \cdot \boldsymbol{n}_i \, e_i^\ell \, ds. \tag{11.8}$$

Pour traiter les intégrales sur Σ_{ij}, on utilise l'égalité (10.8) avec $\alpha = \alpha_{ij}$; $A = e_i^\ell$ et $B = \boldsymbol{e}_i^\ell \cdot \boldsymbol{n}_i$ presque partout sur Σ_{ij} ; puis les conditions de transmission sur Σ_{ij} (11.6)-(ii). On a pour tout $i \in \mathcal{I}_\Omega$ et pour tout $j \in \mathcal{I}_{\Omega_i}$:

$$\int_{\Sigma_{ij}} \boldsymbol{e}_i^\ell \cdot \boldsymbol{n}_i \, e_i^\ell \, ds = \left(\| \boldsymbol{e}_i^\ell \cdot \boldsymbol{n}_i + \alpha_{ij} e_i^\ell \|^2_{L^2(\Sigma_{ij})} - \| \boldsymbol{e}_j^{\ell-1} \cdot \boldsymbol{n}_j + \alpha_{ij} e_j^{\ell-1} \|^2_{L^2(\Sigma_{ij})} \right) / (4\, \alpha_{ij}). \tag{11.9}$$

En sommant (11.7) et (11.8), et en utilisant (11.9) on obtient alors pour $i \in \mathcal{I}_\Omega$:

$$\begin{aligned}&\|\sqrt{q_i}e_i^\ell\|^2_{L^2(\Omega_i)} + \|\sqrt{k_i^{-1}}\boldsymbol{e}_i^\ell\|^2_{L^2(\Omega_i)}\\ &= \sum_{j\in\mathcal{I}_{\Omega_i}} \left(\|\boldsymbol{e}_j^{\ell-1}\cdot\boldsymbol{n}_j + \alpha_{ij}e_j^{\ell-1}\|^2_{L^2(\Sigma_{ij})} - \|\boldsymbol{e}_i^\ell\cdot\boldsymbol{n}_i + \alpha_{ij}e_i^\ell\|^2_{L^2(\Sigma_{ij})}\right)/(4\,\alpha_{ij}).\end{aligned}$$

En sommant les contributions de chaque sous-domaine, on obtient pour $\ell \in \mathbb{N}^*$:

$$\begin{aligned}&\sum_{i\in\mathcal{I}_\Omega}\left(\|\sqrt{q_i}e_i^\ell\|^2_{L^2(\Omega_i)} + \|\sqrt{k_i^{-1}}\boldsymbol{e}_i^\ell\|^2_{L^2(\Omega_i)}\right)\\ &= \sum_{i\in\mathcal{I}_\Omega}\sum_{j\in\mathcal{I}_{\Omega_i}}\left(\|\boldsymbol{e}_j^{\ell-1}\cdot\boldsymbol{n}_j + \alpha_{ij}e_j^{\ell-1}\|^2_{L^2(\Sigma_{ij})} - \|\boldsymbol{e}_i^\ell\cdot\boldsymbol{n}_i + \alpha_{ij}e_i^\ell\|^2_{L^2(\Sigma_{ij})}\right)/(4\,\alpha_{ij}),\\ &= \sum_{(i,j)\in\mathcal{IJ}}\Big(\|\boldsymbol{e}_j^{\ell-1}\cdot\boldsymbol{n}_j + \alpha_{ij}e_j^{\ell-1}\|^2_{L^2(\Sigma_{ij})} - \|\boldsymbol{e}_j^\ell\cdot\boldsymbol{n}_j + \alpha_{ij}e_j^\ell\|^2_{L^2(\Sigma_{ij})}\cdot\\ &\qquad\qquad \|\boldsymbol{e}_i^{\ell-1}\cdot\boldsymbol{n}_i + \alpha_{ij}e_i^{\ell-1}\|^2_{L^2(\Sigma_{ij})} - \|\boldsymbol{e}_i^\ell\cdot\boldsymbol{n}_i + \alpha_{ij}e_i^\ell\|^2_{L^2(\Sigma_{ij})}\Big)/(4\,\alpha_{ij})\end{aligned}$$

En sommant cette fois sur les itérations $\ell \in \{1,\cdots,N\}$, les contributions aux itérations successives $\ell-1$, ℓ se compensent pour $\ell \in \{1,\cdots,N-1\}$ et on a :

$$\begin{aligned}&\sum_{\ell=1}^{N}\sum_{i\in\mathcal{I}_\Omega}\left(\|\sqrt{q_i}e_i^\ell\|^2_{L^2(\Omega_i)} + \|\sqrt{k_i^{-1}}\boldsymbol{e}_i^\ell\|^2_{L^2(\Omega_i)}\right) + \sum_{(i,j)\in\mathcal{IJ}}\|\boldsymbol{e}_i^N\cdot\boldsymbol{n}_i + \alpha_{ij}e_i^N\|^2_{L^2(\Sigma_{ij})}/(4\,\alpha_{ij})\\ &\qquad = \sum_{(i,j)\in\mathcal{IJ}}\|\boldsymbol{e}_i^0\cdot\boldsymbol{n}_i + \alpha_{ij}e_i^0\|^2_{L^2(\Sigma_{ij})}/(4\,\alpha_{ij})\end{aligned}$$

On a donc montré que la série $\displaystyle\sum_{\ell=1}^{\infty}\sum_{i\in\mathcal{I}_\Omega}\left(\|\sqrt{q_i}e_i^\ell\|^2_{L^2(\Omega_i)} + \|\sqrt{k_i^{-1}}\boldsymbol{e}_i^\ell\|^2_{L^2(\Omega_i)}\right)$ converge. On en déduit que la suite $\displaystyle\left(\sum_{i\in\mathcal{I}_\Omega}\left(\|\sqrt{q_i}e_i^\ell\|^2_{L^2(\Omega_i)} + \|\sqrt{k_i^{-1}}\boldsymbol{e}_i^\ell\|^2_{L^2(\Omega_i)}\right)\right)_{\ell\in\mathbb{N}}$ converge vers 0, c'est-à-dire que pour $i \in \mathcal{I}_\Omega$, la suite $(e_i^\ell, \boldsymbol{e}_i^\ell)_{\ell\in\mathbb{N}}$ tend vers 0 dans $L^2(\Omega_i)\times\mathbf{L}^2(\Omega_i)$. En utilisant finalement les relations (11.6)-(0) et (11.6)-(i), on obtient la convergence de la suite $(e_i^\ell, \boldsymbol{e}_i^\ell)_{\ell\in\mathbb{N}}$ vers 0 dans $V_i \times \mathbf{H}(\mathrm{div},\Omega_i)$. ◇

L'étude asymptotique de Nataf et Nier [28] permet d'optimiser le choix des paramètres α_{ij}^i (10.15).

11.2 Formulation variationnelle

Pour construire une formulation variationnelle associée au problème (11.5), comme dans le cas d'un seul domaine (voir le §3.3.2), nous procédons en deux étapes (il y a maintenant deux équations à prendre en compte).

On considère une fonction-test $\boldsymbol{q}_i \in \mathbf{Q}_i$ que l'on multiplie à l'équation (11.5)-(i), et on effectue le produit scalaire dans $\mathbf{L}^2(\Omega_i)$. On obtient :

$$\int_{\Omega_i} k_i^{-1}\boldsymbol{p}_i^\ell\cdot\boldsymbol{q}_i d\boldsymbol{x} + \int_{\Omega_i}\mathbf{grad}\,u_i^\ell\cdot\boldsymbol{q}_i d\boldsymbol{x} = 0. \tag{11.10}$$

La formule d'intégration par parties (D.2-i) donne :

$$\int_{\Omega_i} \mathbf{grad}\, u_i^\ell \cdot \boldsymbol{q}_i d\boldsymbol{x} = -\int_{\Omega_i} u_i^\ell \operatorname{div} \boldsymbol{q}_i d\boldsymbol{x} + \sum_{j\in\mathcal{I}_{\Omega_i}} \int_{\Sigma_{ij}} u_i^\ell\, \boldsymbol{q}_i \cdot \boldsymbol{n}_i\, ds,$$

donc l'équation (11.10) se réécrit :

$$\int_{\Omega_i} \left(k_i^{-1}\boldsymbol{p}_i^\ell \cdot \boldsymbol{q}_i - u_i^\ell \operatorname{div} \boldsymbol{q}_i\right) d\boldsymbol{x} + \sum_{j\in\mathcal{I}_{\Omega_i}} \int_{\Sigma_{ij}} u_i^\ell\, \boldsymbol{q}_i \cdot \boldsymbol{n}_i\, ds = 0. \tag{11.11}$$

Par ailleurs, d'après la condition de transmission (11.5)-(ii), on a :

$$u_i^\ell = (\alpha_{ij}^i)^{-1}\boldsymbol{p}_i^\ell \cdot \boldsymbol{n}_i + (\alpha_{ij}^i)^{-1}\boldsymbol{p}_j^{\ell-1} \cdot \boldsymbol{n}_j + u_j^{\ell-1},$$

donc l'équation (11.11) se réécrit :

$$\begin{aligned}&\int_{\Omega_i} \left(k_i^{-1}\boldsymbol{p}_i^\ell \cdot \boldsymbol{q}_i - u_i^\ell \operatorname{div} \boldsymbol{q}_i\right) d\boldsymbol{x} + \sum_{j\in\mathcal{I}_{\Omega_i}} \int_{\Sigma_{ij}} (\alpha_{ij}^i)^{-1}\boldsymbol{p}_i^\ell \cdot \boldsymbol{n}_i\, \boldsymbol{q}_i \cdot \boldsymbol{n}_i\, ds \\ &= -\sum_{j\in\mathcal{I}_{\Omega_i}} \int_{\Sigma_{ij}} (\alpha_{ij}^i)^{-1}\boldsymbol{p}_j^{\ell-1} \cdot \boldsymbol{n}_j\, \boldsymbol{q}_i \cdot \boldsymbol{n}_i\, ds - \sum_{j\in\mathcal{I}_{\Omega_i}} \int_{\Sigma_{ij}} u_j^{\ell-1}\, \boldsymbol{q}_i \cdot \boldsymbol{n}_i\, ds.\end{aligned} \tag{11.12}$$

On se place dans les espaces produits $\mathcal{V}_i^+ = L^2(\Omega_i) \times \mathbf{Q}_i$ pour $i \in \mathcal{I}_\Omega$, munis de la norme produit

$$\|(v,\boldsymbol{q})\|_{\mathcal{V}_i^+} = \left(\|v\|^2_{L^2(\Omega_i)} + \|\boldsymbol{q}\|^2_{\mathbf{H}(\operatorname{div},\Omega_i)} + \|\boldsymbol{q}\cdot\boldsymbol{n}_i\|^2_{L^2(\Sigma_i)}\right)^{1/2}.$$

Dans ce cas, on ne peut pas conserver la trace de $u_j^{\ell-1}$ sur Σ_{ij}, il faut donc faire une intégration par parties du terme $\displaystyle\int_{\Sigma_{ij}} u_j^{\ell-1}\, \boldsymbol{q}_i \cdot \boldsymbol{n}_i\, ds$.
On introduit $\boldsymbol{q}_j^\star \in \mathbf{Q}_j$ un relèvement de $\boldsymbol{q}_i \cdot \boldsymbol{n}_{i|\Sigma_{ij}}$ dans $\mathbf{H}(\operatorname{div},\Omega_j)$.
On peut choisir par exemple $\boldsymbol{q}_j^\star = \mathbf{grad}\, v_j^\star$ où $v_j^\star \in V_j^i$ (l'espace V_j^i est défini en (10.17)) est la solution du Laplacien avec condition aux limites de Neumann sur $\partial\Omega_j$:
Trouver $v_j^\star \in V_j^i$ tel que

$$\left\{\begin{array}{rcll} \Delta v_j^\star &=& c(\boldsymbol{q}_i) & \text{dans } \Omega_j, \\ \partial_{\boldsymbol{n}_j} v_j^\star &=& -\boldsymbol{q}_i \cdot \boldsymbol{n}_{i|\Sigma_{ij}} & \text{sur } \Sigma_{ij}, \\ \partial_{\boldsymbol{n}_j} v_j^\star &=& 0 & \text{sur } \partial\Omega_j \setminus \overline{\Sigma_{ij}}. \end{array}\right.$$

Ci-dessus, la constante $c(\boldsymbol{q}_i)$ est choisie de façon à assurer la condition de compatibilité par le problème de Neumann, à savoir (cf. [11]),

$$\int_{\Omega_j} c(\boldsymbol{q}_i)\, d\boldsymbol{x} + \int_{\Sigma_{ij}} \boldsymbol{q}_i \cdot \boldsymbol{n}_i\, ds = 0.$$

Par construction, on a $\operatorname{div} \boldsymbol{q}_j^\star = c(\boldsymbol{q}_i)$ dans Ω_j, et $\boldsymbol{q}_j^\star \cdot \boldsymbol{n}_j = -\boldsymbol{q}_i \cdot \boldsymbol{n}_i$ sur Σ_{ij}, $\boldsymbol{q}_j^\star \cdot \boldsymbol{n}_j = 0$ sur $\partial\Omega_j \setminus \overline{\Sigma_{ij}}$. On proposera dans la suite un choix "naturel" de relèvement pour la

discrétisation. La formule d'intégration par parties (D.2-i) dans Ω_j permet d'écrire :

$$\begin{aligned}
-\int_{\Sigma_{ij}} u_j^{\ell-1}\,\boldsymbol{q}_i\cdot\boldsymbol{n}_i\,ds &= \int_{\Sigma_{ij}} u_j^{\ell-1}\,\boldsymbol{q}_j^\star\cdot\boldsymbol{n}_j\,ds = \int_{\partial\Omega_j} u_j^{\ell-1}\,\boldsymbol{q}_j^\star\cdot\boldsymbol{n}_j\,ds \\
&= \int_{\Omega_j} \mathbf{grad}\,u_j^{\ell-1}\cdot\boldsymbol{q}_j^\star\,d\boldsymbol{x} + \int_{\Omega_j} u_j^{\ell-1}\,\operatorname{div}\boldsymbol{q}_j^\star\,d\boldsymbol{x}, \\
&= \int_{\Omega_j}\left(-k_j^{-1}\boldsymbol{p}_j^{\ell-1}\cdot\boldsymbol{q}_j^\star + u_j^{\ell-1}\,\operatorname{div}\boldsymbol{q}_j^\star\right)d\boldsymbol{x}, \quad \text{d'après (11.5)-(i).}
\end{aligned}$$

L'équation (11.12) se réécrit :

$$\begin{aligned}
&\int_{\Omega_i}\left(k_i^{-1}\boldsymbol{p}_i^\ell\cdot\boldsymbol{q}_i - u_i^\ell\operatorname{div}\boldsymbol{q}_i\right)d\boldsymbol{x} + \sum_{j\in\mathcal{I}_{\Omega_i}}\int_{\Sigma_{ij}}(\alpha_{ij}^i)^{-1}\boldsymbol{p}_i^\ell\cdot\boldsymbol{n}_i\,\boldsymbol{q}_i\cdot\boldsymbol{n}_i\,ds \\
&\qquad = \sum_{j\in\mathcal{I}_{\Omega_i}}\int_{\Sigma_{ij}}(\alpha_{ij}^i)^{-1}\boldsymbol{p}_j^{\ell-1}\cdot\boldsymbol{n}_j\,\boldsymbol{q}_j^\star\cdot\boldsymbol{n}_j\,ds \\
&\qquad\quad + \sum_{j\in\mathcal{I}_{\Omega_i}}\int_{\Omega_j}\left(-k_j^{-1}\boldsymbol{p}_j^{\ell-1}\cdot\boldsymbol{q}_j^\star + u_j^{\ell-1}\,\operatorname{div}\boldsymbol{q}_j^\star\right)d\boldsymbol{x}.
\end{aligned} \tag{11.13}$$

Pour prendre en compte (11.5)-(0), on utilise simplement l'égalité variationnelle équivalente :

$$\forall v_i\in L^2(\Omega_i),\quad 0 = \int_{\Omega_i}(\operatorname{div}\boldsymbol{p}_i^\ell + q_i\,u_i^\ell - f_i)v_i\,d\boldsymbol{x}. \tag{11.14}$$

Pour $i\in\mathcal{I}_\Omega$, par soustraction des égalités variationnelles (11.13) et (11.14), on en conclut que si $(u_i^\ell,\boldsymbol{p}_i^\ell)$ est solution de (11.5), alors $(u_i^\ell,\boldsymbol{p}_i^\ell)$ est solution de la **formulation variationnelle** ci-dessous :
Trouver $(u_i^\ell,\boldsymbol{p}_i^\ell)\in\mathcal{V}_i^+$ tel que $\forall\,(v_i,\boldsymbol{q}_i)\in\mathcal{V}_i^+$

$$\left\{\begin{aligned}
&\int_{\Omega_i}\left(-k_i^{-1}\boldsymbol{p}_i^\ell\cdot\boldsymbol{q}_i + u_i^\ell\operatorname{div}\boldsymbol{q}_i + v_i\operatorname{div}\boldsymbol{p}_i^\ell + q_i\,u_i^\ell v_i\right)d\boldsymbol{x} \\
&\qquad\qquad - \sum_{j\in\mathcal{I}_{\Omega_i}}\int_{\Sigma_{ij}}(\alpha_{ij}^i)^{-1}\boldsymbol{p}_i^\ell\cdot\boldsymbol{n}_i\,\boldsymbol{q}_i\cdot\boldsymbol{n}_i\,ds \\
&= \int_{\Omega_i} f_i v_i\,d\boldsymbol{x} - \sum_{j\in\mathcal{I}_{\Omega_i}}\int_{\Omega_j}\left(-k_j^{-1}\boldsymbol{p}_j^{\ell-1}\cdot\boldsymbol{q}_j^\star + u_j^{\ell-1}\,\operatorname{div}\boldsymbol{q}_j^\star\right)d\boldsymbol{x} \\
&\qquad\qquad - \sum_{j\in\mathcal{I}_{\Omega_i}}\int_{\Sigma_{ij}}(\alpha_{ij}^i)^{-1}\boldsymbol{p}_j^{\ell-1}\cdot\boldsymbol{n}_j\,\boldsymbol{q}_j^\star\cdot\boldsymbol{n}_j\,ds.
\end{aligned}\right. \tag{11.15}$$

Par la suite, on va utiliser, pour $i\in\mathcal{I}_\Omega$, les formes bilinéaires a_2^i, continues sur $\mathcal{V}_i^+\times\mathcal{V}_i^+$:

$$a_2^i\ :\ ((v,\boldsymbol{q}),(w,\boldsymbol{r}))\mapsto\int_{\Omega_i}\left(-k_i^{-1}\boldsymbol{q}\cdot\boldsymbol{r} + v\operatorname{div}\boldsymbol{r} + w\operatorname{div}\boldsymbol{q} + q_i\,vw\right)d\boldsymbol{x}. \tag{11.16}$$

On a vu au §3.2.2 que cette forme bilinéaire était continue et satisfaisait la condition de stabilité sur $\mathcal{V}_i$ (voir la démonstration du théorème 3.2).

La forme bilinéaire $\widetilde{a}_2^i$ associée à la formulation variationnelle (11.15) s'écrit :

$$\left\{\begin{array}{rcl} \widetilde{a}_2^i : \mathcal{V}_i^+ \times \mathcal{V}_i^+ & \to & \mathbb{R} \\ ((v,\boldsymbol{q}),(w,\boldsymbol{r})) & \mapsto & a_2^i((v,\boldsymbol{q}),(w,\boldsymbol{r})) - \displaystyle\sum_{j\in\mathcal{I}_{\Omega_i}} \int_{\Sigma_{ij}} (\alpha_{ij}^i)^{-1} \boldsymbol{q}\cdot\boldsymbol{n}_i\, \boldsymbol{r}\cdot\boldsymbol{n}_i\, ds \end{array}\right. \tag{11.17}$$

Si $\alpha_{ij}^i > 0$, la forme bilinéaire $\widetilde{a}_2^i$ satisfait également la condition de stabilité sur $\mathcal{V}_i^+$ (il suffit de reprendre la démonstration du théorème 3.2), ce qui permet de montrer que la formulation variationnelle (11.15) est bien posée.

11.3 Discrétisation

On se place dans le cas $2D$. Considérons comme discrétisation produit des espaces $\mathcal{V}_i^+$ les éléments finis de Raviart-Thomas $P_0 - \mathbf{RT}_0$ décrits au §3.3.2. On définit une famille de triangulations $(\mathcal{T}_{i,h})_h$ de Ω_i constituées de L_i éléments tels que : $\overline{\Omega_i} = \cup_{l\in\mathcal{I}_i^T} T_{i,l}$ avec $\mathcal{I}_i^T = \{1,\cdots,L_i\}$; et de A_i^+ arêtes $(a_{i,m})_{m\in\mathcal{I}_i^{a+}}$ avec $\mathcal{I}_i^{a+} = \{1,\cdots,A_i^+\}$. On décompose $\mathcal{I}_i^{a+}$ ainsi : $\mathcal{I}_i^{a+} = \mathcal{I}_i^a \cup \mathcal{I}_{i,\Sigma_i}^a$ de sorte que pour $\left\{\begin{array}{ll} m\in\mathcal{I}_i^a & : \quad a_{i,m}\notin\Sigma_i \\ m\in\mathcal{I}_{i,\Sigma_i}^a & : \quad a_{i,m}\in\Sigma_i \end{array}\right.$.
On décompose de plus $\mathcal{I}_{i,\Sigma_i}^a$ ainsi : $\mathcal{I}_{i,\Sigma_i}^a = \cup_{j\in\mathcal{I}_{\Omega_i}} \mathcal{I}_{i,\Sigma_i}^{ij,a}$, où pour tout $m\in\mathcal{I}_{i,\Sigma_i}^{ij,a}$, $a_{i,m}\in\Sigma_{ij}$.

On fait l'hypothèse suivante :

Pour tout $(i,j)\in\mathcal{IJ}$, les triangulations $\mathcal{T}_{i,h}$ et $\mathcal{T}_{j,h}$ partagent leurs arêtes sur Σ_{ij}.

Soient $A_{\Sigma_{ij}}^i = |\mathcal{I}_{i,\Sigma_i}^{ij,a}|$ et $A_{\Sigma_{ij}}^j = |\mathcal{I}_{j,\Sigma_j}^{ij,a}|$, alors : $A_{\Sigma_{ij}}^i = A_{\Sigma_{ij}}^j = A_{\Sigma_{ij}}$.
L'espace de discrétisation de $\mathcal{V}_i^+$ est noté $\mathcal{V}_{i,h}$ et est tel que : $\mathcal{V}_{i,h} := M_{i,h}\times\mathbf{Q}_{i,h}$ avec :

$$\begin{array}{rcl} M_{i,h} & = & \{\, v_{i,h}\in L^2(\Omega_i)\,|\,\forall l\in\mathcal{I}_i^T,\ v_{i,h|T_l}\in P_0(T_l)\,\}, \\ \mathbf{Q}_{i,h} & = & \left\{\boldsymbol{q}_{i,h}\in\mathbf{Q}_i\,|\,\forall l\in\mathcal{I}_i^T,\ \boldsymbol{q}_{i,h|T_l}\in\mathbf{RT}_0(T_l)\right\}. \end{array} \tag{11.18}$$

On rappelle que (voir (3.44)) :

$$\begin{array}{rcl} \mathbf{RT}_0(T) & := & \{\,\boldsymbol{q}\in(P(T))^2\ :\ \exists\boldsymbol{a}\in(P_0(T))^2,\exists b\in P_0(T),\ \forall\boldsymbol{x}\in T,\ \boldsymbol{q}(\boldsymbol{x}) = \boldsymbol{a}+b\boldsymbol{x} \\ & & \qquad\qquad \text{et } \forall e\in\{1,2,3\},\ (\boldsymbol{q}\cdot\boldsymbol{n})_{|a_e^T}\in P_0(a_e^T)\,\}. \end{array} \tag{11.19}$$

On appelle $(\underline{w}_{i,l})_{l\in\mathcal{I}_i^T}$ les fonctions de base de $M_{i,h}$, telles que : $\underline{w}_{i,l|T_{l'}} = \delta_{l,l'}$.
On appelle $(\underline{\boldsymbol{\omega}}_{i,m})_{m\in\mathcal{I}_i^{a+}}$ les fonctions de base associées à $\mathbf{Q}_{i,h}$.
Pour tout $(i,j)\in\mathcal{IJ}$, les triangulations $\mathcal{T}_{i,h}$ et $\mathcal{T}_{j,h}$ étant identiques sur l'interface Σ_{ij}, les espaces des traces normales aux faces sur Σ_{ij},

$$\mathfrak{T}_{i,h}^{ij,a} := vect\left(\left(\underline{\boldsymbol{\omega}}_{i,m}\cdot\boldsymbol{n}_{i|\Sigma_{ij}}\right)_{m\in\mathcal{I}_{i,\Sigma_i}^{ij,a}}\right),\quad \mathfrak{T}_{j,h}^{ij,a} := vect\left(\left(\underline{\boldsymbol{\omega}}_{j,m}\cdot\boldsymbol{n}_{j|\Sigma_{ij}}\right)_{m\in\mathcal{I}_{j,\Sigma_j}^{ij,a}}\right) \tag{11.20}$$

sont égaux : pour tout $m_i\in\mathcal{I}_{i,\Sigma_i}^{ij,a}$ il existe un unique $m_j\in\mathcal{I}_{j,\Sigma_j}^{ij,a}$ tel que $\underline{\boldsymbol{\omega}}_{i,m_i}\cdot\boldsymbol{n}_{i|\Sigma_{ij}} = -\underline{\boldsymbol{\omega}}_{j,m_j}\cdot\boldsymbol{n}_{j|\Sigma_{ij}}$ (voir la démonstration de la proposition 3.28, avec m_j tel que $a_{i,m_i} = a_{j,m_j}$).

Ceci permet de proposer un choix de relèvement (discret).

$$\begin{array}{l}\forall (i,j) \in \mathcal{IJ},\ \forall (m_i, m_j) \in \mathcal{I}^{ij,a}_{i,\Sigma_i} \times \mathcal{I}^{ij,a}_{j,\Sigma_i} \text{ tel que } a_{i,m_i} = a_{j,m_j}\ : \\ \text{- la fonction de base } \underline{\boldsymbol{\omega}}_{j,m_j} \text{ est un relèvement de } \underline{\boldsymbol{\omega}}_{i,m_i} \cdot \boldsymbol{n}_{i|\Sigma_{ij}} \text{ dans } \Omega_j\,; \\ \text{- la fonction de base } \underline{\boldsymbol{\omega}}_{i,m_i} \text{ est un relèvement de } \underline{\boldsymbol{\omega}}_{j,m_j} \cdot \boldsymbol{n}_{j|\Sigma_{ij}} \text{ dans } \Omega_i.\end{array} \tag{11.21}$$

On cherche $(u^\ell_{i,h}, \boldsymbol{p}^\ell_{i,h})_{i\in\mathcal{I}_\Omega}$, appartenant à l'espace produit $\displaystyle\prod_{i\in\mathcal{I}_\Omega} \mathcal{V}_{i,h}$, une approximation de $(u^\ell_i, \boldsymbol{p}^\ell_i)_{i\in\mathcal{I}_\Omega}$ telle que pour tout $i \in \mathcal{I}_\Omega$: $u^\ell_{i,h} = \displaystyle\sum_{l\in\mathcal{I}^T_i} U^\ell_{i,l}\, w_{i,l}$ et $\boldsymbol{p}^\ell_{i,h} = \displaystyle\sum_{m\in\mathcal{I}^{a+}_i} P^\ell_{i,m}\, \underline{\boldsymbol{\omega}}_{i,m}$.
La discrétisation de la formulation variationnelle (11.15) s'écrit :
Trouver $(u^\ell_{i,h}, \boldsymbol{p}^\ell_{i,h})_{i\in\mathcal{I}_\Omega} \in \displaystyle\prod_{i\in\mathcal{I}_\Omega} \mathcal{V}_{i,h}$ tel que $\forall i \in \mathcal{I}_\Omega$, $\forall j \in \mathcal{I}_{\Omega_i}$:

$$\left\{\begin{array}{ll}
\forall\,(l,m) \in \mathcal{I}^T_i \times \mathcal{I}^a_i & a^i_2((u^\ell_{i,h}, \boldsymbol{p}^\ell_{i,h})\,,\,(\underline{w}_{i,l}, \underline{\boldsymbol{\omega}}_{i,m})) = \displaystyle\int_{\Omega_i} f_i\, \underline{w}_{i,l}\, d\boldsymbol{x}, \\[2ex]
\forall\,(l,m_i) \in \mathcal{I}^T_i \times \mathcal{I}^{ij,a}_{i,\Sigma_i} & a^i_2((u^\ell_{i,h}, \boldsymbol{p}^\ell_{i,h})\,,\,(\underline{w}_{i,l}, \underline{\boldsymbol{\omega}}_{i,m_i})) - \displaystyle\int_{\Sigma_{ij}} (\alpha^i_{ij})^{-1} \boldsymbol{p}^\ell_{i,h} \cdot \boldsymbol{n}_i\, \underline{\boldsymbol{\omega}}_{i,m_i} \cdot \boldsymbol{n}_i\, ds \\
& = \displaystyle\int_{\Omega_i} f_i\, \underline{w}_{i,l} d\boldsymbol{x} - \int_{\Omega_j} \left(-k_j^{-1} \boldsymbol{p}^{\ell-1}_{j,h} \cdot \underline{\boldsymbol{\omega}}_{j,m_j} + u^{\ell-1}_j\ \mathrm{div}\, \underline{\boldsymbol{\omega}}_{j,m_j}\right) d\boldsymbol{x} \\
& - \displaystyle\int_{\Sigma_{ij}} (\alpha^i_{ij})^{-1} \boldsymbol{p}^{\ell-1}_{j,h} \cdot \boldsymbol{n}_j\ \underline{\boldsymbol{\omega}}_{j,m_j} \cdot \boldsymbol{n}_j\, ds.
\end{array}\right.$$

Après avoir décomposé $(u^\ell_{i,h}, \boldsymbol{p}^\ell_{i,h})$ dans leurs bases respectives, on obtient :
Trouver $\left((U^\ell_{i,l'})_{l'\in\mathcal{I}^T_i}, (P^\ell_{i,m'})_{m'\in\mathcal{I}^{a+}_i} \right)$ tels que $\forall i \in \mathcal{I}_\Omega$, $\forall j \in \mathcal{I}_{\Omega_i}$:

$$\left\{\begin{array}{l}
\forall\, l \in \mathcal{I}^T_i\ : \\
\displaystyle\sum_{l'\in\mathcal{I}^T_i} U^\ell_{i,l'} \int_{\Omega_i} q_i\, \underline{w}_{i,l}\, \underline{w}_{i,l'}\, d\boldsymbol{x} + \sum_{m'\in\mathcal{I}^{a+}_i} P^\ell_{i,m'} \int_{\Omega_i} \underline{w}_{i,l}\ \mathrm{div}\, \underline{\boldsymbol{\omega}}_{i,m'}\, d\boldsymbol{x} \\
\qquad\qquad = \displaystyle\int_{\Omega_i} f_i\, \underline{w}_{i,l}\, d\boldsymbol{x},
\end{array}\right. \tag{11.22}$$

ainsi que

$$
\left\{
\begin{array}{l}
\forall\, m \in \mathcal{I}_i^a : \\
-\displaystyle\sum_{m' \in \mathcal{I}_i^{a+}} P_{i,m'}^{\ell} \int_{\Omega_i} k_i^{-1} \underline{\boldsymbol{\omega}}_{i,m} \cdot \underline{\boldsymbol{\omega}}_{i,m'}\, d\boldsymbol{x} + \sum_{l' \in \mathcal{I}_i^T} U_{i,l'}^{\ell} \int_{\Omega_i} \operatorname{div} \underline{\boldsymbol{\omega}}_{i,m}\, \underline{w}_{i,l'}\, d\boldsymbol{x} = 0, \\
\forall\, m_i \in \mathcal{I}_{i,\Sigma_{ij}}^{ij,a},\ \text{avec (11.21) pour définir la correspondance } m_i \text{ à } m_j) : \\
-\displaystyle\sum_{m' \in \mathcal{I}_i^{a+}} P_{i,m'}^{\ell} \int_{\Omega_i} k_i^{-1} \underline{\boldsymbol{\omega}}_{i,m_i} \cdot \underline{\boldsymbol{\omega}}_{i,m'}\, d\boldsymbol{x} \\
-\displaystyle\sum_{m' \in \mathcal{I}_{i,\Sigma_i}^{ij,a}} P_{i,m'}^{\ell} \int_{\Sigma_{ij}} (\alpha_{ij}^i)^{-1} \underline{\boldsymbol{\omega}}_{i,m_i} \cdot \boldsymbol{n}_i\, \underline{\boldsymbol{\omega}}_{i,m'} \cdot \boldsymbol{n}_i\, ds \\
+\displaystyle\sum_{l' \in \mathcal{I}_i^T} U_{i,l'}^{\ell} \int_{\Omega_i} \operatorname{div} \underline{\boldsymbol{\omega}}_{i,m_i}\, \underline{w}_{i,l'}\, d\boldsymbol{x} = \sum_{m' \in \mathcal{I}_j^{a+}} P_{j,m'}^{\ell-1} \int_{\Omega_j} k_j^{-1} \underline{\boldsymbol{\omega}}_{j,m_j} \cdot \underline{\boldsymbol{\omega}}_{j,m'} d\boldsymbol{x} \\
-\displaystyle\sum_{m' \in \mathcal{I}_{j,\Sigma_{ij}}^{ij,a}} P_{j,m'}^{\ell-1} \int_{\Sigma_{ij}} (\alpha_{ij}^i)^{-1} \underline{\boldsymbol{\omega}}_{j,m_j} \cdot \boldsymbol{n}_j\, \underline{\boldsymbol{\omega}}_{j,m'} \cdot \boldsymbol{n}_j\, ds \\
-\displaystyle\sum_{l' \in \mathcal{I}_j^T} U_{j,l'}^{\ell-1} \int_{\Omega_j} \operatorname{div} \underline{\boldsymbol{\omega}}_{j,m_j}\, \underline{w}_{j,l'}\, d\boldsymbol{x}.
\end{array}
\right. \tag{11.23}
$$

On peut mettre les équations (11.22)-(11.23) sous forme matricielle (on utilise des notations similaires à celles du paragraphe 3.3.2).
Pour tout $i \in \mathcal{I}_\Omega$, $j \in \mathcal{I}_{\Omega_i}$, on définit les matrices et vecteurs associés à l'inconnue scalaire et au second membre :

— $U_i^\ell \in \mathbb{R}^{L_i}$ est le vecteur tel que pour $l \in \mathcal{I}_i^T$: $\left(U_i^\ell\right)_l := U_{i,l}^\ell$.

— $\mathbb{M}_u^i \in \mathbb{R}^{L_i \times L_i} \mid \quad \forall\, (l,l') \in \mathcal{I}_i^T \times \mathcal{I}_i^T,\ (\mathbb{M}_u^i)_{l,l'} = \displaystyle\int_{\Omega_i} q_i \underline{w}_{i,l}\, \underline{w}_{i,l'}\, d\boldsymbol{x}$.

— $F_i \in \mathbb{R}^{L_i}$ est le vecteur tel que pour $l \in \mathcal{I}_i^T$: $(F_i)_l := F_{i,l} = \displaystyle\int_{\Omega_i} f_i \underline{w}_{i,l}\, d\boldsymbol{x}$.

Les matrices $\mathbb{M}_u^i$ sont diagonales et inversibles.
On définit les matrices et vecteurs associés à l'inconnue vectorielle : Soit $P_i^\ell \in \mathbb{R}^{A_i^+}$ le vecteur tel que pour $m \in \mathcal{I}_i^{a+}$: $\left(P_i^\ell\right)_m := P_{i,m}^\ell$.
Soit $\mathbb{M}_{\boldsymbol{p}}^{i,i} \in \mathbb{R}^{A_i^+ \times A_i^+}$ la matrice symétrique telle que :

$$
\forall\, (m,m') \in \mathcal{I}_i^{a+} \times \mathcal{I}_i^{a+},\ (\mathbb{M}_{\boldsymbol{p}}^{i,i})_{m,m'} = \int_{\Omega_i} k_i^{-1} \underline{\boldsymbol{\omega}}_{i,m'} \cdot \underline{\boldsymbol{\omega}}_{i,m}\, d\boldsymbol{x}.
$$

Soit $\widetilde{\mathbb{M}}_{\boldsymbol{p}}^{i,i} \in \mathbb{R}^{A_i^+ \times A_i^+}$ la matrice symétrique telle que pour tout $j \in \mathcal{I}_{\Omega_i}$:

$$
\begin{array}{rcl}
\forall (m,m') \in \mathcal{I}_i^a \times \mathcal{I}_i^{a+}, & (\widetilde{\mathbb{M}}_{\boldsymbol{p}}^{i,i})_{m,m'} & = (\mathbb{M}_{\boldsymbol{p}}^{i,i})_{m,m'}, \\
\forall (m,m') \in \mathcal{I}_{i,\Sigma_i}^{ij,a} \times \mathcal{I}_i^{a+}, & (\widetilde{\mathbb{M}}_{\boldsymbol{p}}^{i,i})_{m,m'} & = (\mathbb{M}_{\boldsymbol{p}}^{i,i})_{m,m'} + \displaystyle\int_{\Sigma_{ij}} (\alpha_{ij}^i)^{-1} \underline{\boldsymbol{\omega}}_{i,m'} \cdot \boldsymbol{n}_i\, \underline{\boldsymbol{\omega}}_{i,m} \cdot \boldsymbol{n}_i\, ds.
\end{array}
$$

Soit $i \in \mathcal{I}_\Omega$ et $j \in \mathcal{I}_{\Omega_i}$. Soit $\widetilde{\mathbb{M}}_{\boldsymbol{p}}^{i,j} \in \mathbb{R}^{A_i^+ \times A_j^+}$ la matrice de couplage entre Ω_i et Ω_j :

$$\begin{aligned}
\forall (m,m') \in \mathcal{I}_i^a \times \mathcal{I}_j^{a+}, \quad & (\widetilde{\mathbb{M}}_{\boldsymbol{p}}^{i,j})_{m,m'} &= 0, \\
\forall (m_i,m') \in \mathcal{I}_{i,\Sigma_i}^{ij,a} \times \mathcal{I}_j^{a+}, \quad & (\widetilde{\mathbb{M}}_{\boldsymbol{p}}^{i,j})_{m_i,m'} &= \int_{\Omega_j} k_j^{-1} \underline{\boldsymbol{\omega}}_{j,m'} \cdot \underline{\boldsymbol{\omega}}_{j,m_j} \, d\boldsymbol{x} \\
& & \quad - \int_{\Sigma_{ij}} (\alpha_{ij}^i)^{-1} \underline{\boldsymbol{\omega}}_{j,m'} \cdot \boldsymbol{n}_j \, \underline{\boldsymbol{\omega}}_{j,m_j} \cdot \boldsymbol{n}_j \, ds,
\end{aligned}$$

où l'expression de $(\widetilde{\mathbb{M}}_{\boldsymbol{p}}^{i,j})_{(m_i,m') \in \mathcal{I}_i^{ij,a} \times \mathcal{I}_j^{a+}}$ est obtenue suite au choix du relèvement discret 11.21) sur l'interface Σ_{ij}.

Soit $\mathbb{B}_{i,i} \in \mathbb{R}^{A_i^+ \times L_i}$ la matrice de couplage entre les deux inconnues au sein d'un même sous-domaine :

$$\forall\, (m,l) \in \mathcal{I}_i^{a+} \times \mathcal{I}_i^T, \quad (\mathbb{B}_{i,i})_{m,l} = \int_{\Omega_i} \operatorname{div} \underline{\boldsymbol{\omega}}_{i,m} \, \underline{w}_{i,l} \, d\boldsymbol{x}.$$

Soit $\widetilde{\mathbb{B}}_{i,j} \in \mathbb{R}^{A_i^+ \times L_j}$ la matrice de couplage entre les deux inconnues dans des sous-domaines voisins (voir encore une fois (11.21) pour la correspondance de m_i à m_j) :

$$\forall\, (m,l) \in \mathcal{I}_i^{a+} \times \mathcal{I}_j^T, \quad (\widetilde{\mathbb{B}}_{i,j})_{m,l} = \begin{cases} \displaystyle\int_{\Omega_j} \operatorname{div} \underline{\boldsymbol{\omega}}_{j,m_j} \, \underline{w}_{j,l} \, d\boldsymbol{x} & \text{si} \quad m = m_i \in \mathcal{I}_{i,\Sigma_i}^{ij,a}, \\ 0 & \text{si} \quad m \in \mathcal{I}_i^a. \end{cases}$$

Les équations (11.22)-(11.23) se mettent finalement sous la forme du système linéaire suivant :
Pour tout $i \in \mathcal{I}_\Omega$, trouver (U_i^ℓ, P_i^ℓ) tel que :

$$\left\{ \begin{array}{rcl} \mathbb{M}_u^i U_i^\ell + \mathbb{B}_{i,i}^T P_i^\ell &=& F_i, \\ \\ \mathbb{B}_{i,i} U_i^\ell - \widetilde{\mathbb{M}}_{\boldsymbol{p}}^{i,i} P_i^\ell &=& \displaystyle\sum_{j \in \mathcal{I}_{\Omega_i}} \widetilde{\mathbb{M}}_{\boldsymbol{p}}^{i,j} P_j^{\ell-1} - \widetilde{\mathbb{B}}_{i,j} U_j^{\ell-1}. \end{array} \right. \tag{11.24}$$

La discrétisation de l'algorithme (11.5) s'écrit :

initialisation : $\forall i \in \mathcal{I}_\Omega$, $(U_i^0, P_i^0) \in \mathbb{R}^{L_i} \times \mathbb{R}^{A_i^+}$ **est donné.**
itérations : pour $\ell = 1, \ldots,$ $\forall i \in \mathcal{I}_\Omega$, **trouver** $(U_i^\ell, P_i^\ell) \in \mathbb{R}^{L_i} \times \mathbb{R}^{A_i^+}$ **tel que :**

$$\begin{pmatrix} \mathbb{M}_u^i & \mathbb{B}_{i,i}^T \\ \mathbb{B}_{i,i} & -\widetilde{\mathbb{M}}_{\boldsymbol{p}}^{i,i} \end{pmatrix} \begin{pmatrix} U_i^\ell \\ P_i^\ell \end{pmatrix} = \begin{pmatrix} F_i \\ 0 \end{pmatrix} - \sum_{j \in \mathcal{I}_{\Omega_i}} \begin{pmatrix} 0 & 0 \\ \widetilde{\mathbb{B}}_{i,j} & -\widetilde{\mathbb{M}}_{\boldsymbol{p}}^{i,j} \end{pmatrix} \begin{pmatrix} U_j^{\ell-1} \\ P_j^{\ell-1} \end{pmatrix}, \tag{11.25}$$

jusqu'à convergence.

L'algorithme (11.25) correspond à un algorithme de Jacobi par blocs pour résoudre le système linéaire (11.24). La matrice globale n'étant pas symétrique, on peut aussi choisir d'utiliser l'algorithme GMRES pour le résoudre.

Remarque 11.2 *On peut vérifier, comme au chapitre 10, que l'algorithme* (11.25) *peut être obtenu par des manipulations algébriques.*

Chapitre 12

Problème à une inconnue, méthode avec contrainte

On part du problème (9.1) ou (9.2), dont on sait qu'il est bien posé d'après le théorème 3.1 du chapitre 3. On a établi au chapitre 9 qu'une reformulation équivalente par sous-domaine est donnée par le problème (9.13). On en propose une version équivalente, qu'on exprime variationnellement avant de la discrétiser. Encore une fois, on suppose que la solution u est telle que :

$$\text{Pour tout } (i,j) \in \mathcal{I}_\Omega \times \mathcal{I}_{\Omega_i}, \quad k\,\mathbf{grad}\,u \cdot \boldsymbol{n}_{ij|\Sigma_{ij}} \in L^2(\Sigma_{ij}).$$

12.1 Approche continue

On considère le problème (9.13) pour lequel on ajoute comme inconnues les fonctions $(\lambda_{ij})_{(i,j)\in\mathcal{IJ}}$ représentant les traces $k\,\mathbf{grad}\,u \cdot \boldsymbol{n}_{ij|\Sigma_{ij}}$ sur Σ_{ij}, et on note $\lambda \in M$ tel que $\lambda_{|\Sigma_{ij}} = \lambda_{ij}$ pour tout $(i,j) \in \mathcal{IJ}$. Pour un élément $\mu \in M$ (voir (9.10)), on note $\mu_{ij} = \mu_{|\Sigma_{ij}}$ pour $(i,j) \in \mathcal{IJ}$. Par la suite, pour le couple (j,i) tel que $(i,j) \in \mathcal{IJ}$, on pourra utiliser la notation μ_{ji}, en ayant posé : $\mu_{ji} = \mu_{ij}$. Le problème (9.13) est équivalent au problème : Trouver $(u,\lambda) \in \mathcal{P}V \times M$ tel que :

$$\left\{\begin{array}{lllrcll} (i) & \forall i \in \mathcal{I}_\Omega & & -\operatorname{div}(k_i\,\mathbf{grad}\,u_i) + q_i u_i &=& f_i & \text{dans } \Omega_i, \\ (ii) & \forall (i,j) \in \mathcal{IJ} & & u_i &=& u_j & \text{sur } \Sigma_{ij}, \\ (iii) & \forall (i,j) \in \mathcal{I}_\Omega \times \mathcal{I}_{\Omega_i} & & k_i\,\mathbf{grad}\,u_i \cdot \boldsymbol{n}_{ij} &=& \lambda_{ij} & \text{sur } \Sigma_{ij}. \end{array}\right. \tag{12.1}$$

Pour $(i,j) \in \mathcal{I}_\Omega \times \mathcal{I}_\Omega$, $i \neq j$, on pose : $\operatorname{sgn}(j-i) = 1$ si $i < j$ ou -1 si $i > j$. On peut réécrire (12.1)-(iii) sous la forme :

$$\forall (i,j) \in \mathcal{I}_\Omega \times \mathcal{I}_{\Omega_i}, \quad k_i\,\mathbf{grad}\,u_i \cdot \boldsymbol{n}_{i|\Sigma_{ij}} = \operatorname{sgn}(j-i)\lambda_{ij}.$$

Par exemple, pour $(i,j) = (1,2)$, on a défini $\boldsymbol{n}_{21} = \boldsymbol{n}_{12} = \boldsymbol{n}_{1|\Sigma_{12}} = -\boldsymbol{n}_{2|\Sigma_{12}}$, d'où :

$k_1\,\mathbf{grad}\,u_1 \cdot \boldsymbol{n}_{1|\Sigma_{12}} = \operatorname{sgn}(2-1)\lambda_{12} = \lambda_{12}$ et $k_2\,\mathbf{grad}\,u_2 \cdot \boldsymbol{n}_{2|\Sigma_{12}} = \operatorname{sgn}(1-2)\lambda_{21} = -\lambda_{21} = -\lambda_{12}$.

On a donc bien : $k_1\,\mathbf{grad}\,u_1 \cdot \boldsymbol{n}_{12} = k_2\,\mathbf{grad}\,u_2 \cdot \boldsymbol{n}_{21} = \lambda_{12} = \lambda_{21}$.
Nous allons voir que l'intérêt de cette méthode de décomposition de domaine est qu'on peut

choisir pour un couple $(i,j) \in \mathcal{I}_\Omega \times \mathcal{I}_{\Omega_i}$ des triangulations $\mathcal{T}_{i,h}$ et $\mathcal{T}_{j,h}$ non coïncidentes sur Σ_{ij} (c'est-à-dire que $\mathfrak{T}_{i,h} \neq \mathfrak{T}_{j,h}$, voir la définition (10.24)). Les multiplicateurs de Lagrange discrets permettent dans ce cas plus général de minimiser l'écart entre les solutions discrètes $u_{i,h}$ et $u_{j,h}$ restreintes à Σ_{ij}.

12.2 Formulation variationnelle

La formule d'intégration par parties (D.1) donne, pour $v_i \in V_i$:

$$\begin{aligned} -\int_{\Omega_i} \operatorname{div}(k_i \,\mathbf{grad}\, u_i)\, v_i \, d\boldsymbol{x} &= \int_{\Omega_i} k_i \,\mathbf{grad}\, u_i \cdot \mathbf{grad}\, v_i \, d\boldsymbol{x} - \sum_{j \in \mathcal{I}_{\Omega_i}} \int_{\Sigma_{ij}} k_i \partial_{\boldsymbol{n}_i} u_i \, v_i \, ds, \\ &= \int_{\Omega_i} k_i \,\mathbf{grad}\, u_i \cdot \mathbf{grad}\, v_i \, d\boldsymbol{x} - \sum_{j \in \mathcal{I}_{\Omega_i}} \operatorname{sgn}(j-i) \int_{\Sigma_{ij}} \lambda_{ij} \, v_i \, ds. \end{aligned}$$

Le problème (12.1) implique que (u,λ) est solution de la formulation variationnelle :
Trouver $(u,\lambda) \in \mathcal{P}V \times M$ tel que $\forall\, (v,\mu) \in \mathcal{P}V \times M$:

$$\left\{ \begin{array}{rrcl} \forall i \in \mathcal{I}_\Omega \quad a_1^i(u_i,v_i) - \displaystyle\sum_{j \in \mathcal{I}_{\Omega_i}} \operatorname{sgn}(j-i) \int_{\Sigma_{ij}} \lambda_{ij} \, v_i \, ds &=& \displaystyle\int_{\Omega_i} f_i \, v_i d\boldsymbol{x}, \\ \forall (i,j) \in \mathcal{I}\mathcal{J} \qquad \displaystyle\int_{\Sigma_{ij}} \mu_{ij} \, (u_i - u_j) \, ds &=& 0. \end{array} \right. \tag{12.2}$$

Les formes bilinéaires a_1^i ont été définies en (10.19). En faisant la somme sur $i \in \mathcal{I}_\Omega$, on obtient :
Trouver $(u,\lambda) \in \mathcal{P}V \times M$ tel que $\forall\, (v,\mu) \in \mathcal{P}V \times M$:

$$\left\{ \begin{array}{rcl} \displaystyle\sum_{i \in \mathcal{I}_\Omega} a_1^i(u_i,v_i) - \sum_{(i,j) \in \mathcal{I}\mathcal{J}} \int_{\Sigma_{ij}} \lambda_{ij} \, (v_i - v_j) \, ds &=& \displaystyle\int_\Omega f \, v d\boldsymbol{x}, \\ \displaystyle\sum_{(i,j) \in \mathcal{I}\mathcal{J}} \int_{\Sigma_{ij}} \mu_{ij} \, (u_i - u_j) \, ds &=& 0. \end{array} \right.$$

Soient $a_{\mathcal{P}V}$ et b les formes bilinéaires associées à la formulation variationnelle (12.2) :

$$\left\{ \begin{array}{rcl} a_{\mathcal{P}V} : \mathcal{P}V \times \mathcal{P}V &\to& \mathbb{R} \\ (u,v) &\mapsto& \displaystyle\sum_{i \in \mathcal{I}_\Omega} a_1^i(u_i,v_i) \end{array} \right. ,$$

$$\left\{ \begin{array}{rcl} b : \mathcal{P}V \times M &\to& \mathbb{R} \\ (v,\mu) &\mapsto& - \displaystyle\sum_{(i,j) \in \mathcal{I}\mathcal{J}} \int_{\Sigma_{ij}} \mu_{ij} \, (v_i - v_j) \, ds \end{array} \right. ,$$

La formulation variationnelle (12.2) s'écrit alors sous la forme du problème mixte suivant :
Trouver $(u,\lambda) \in \mathcal{P}V \times M$ tel que :

$$\left\{ \begin{array}{rcll} a_{\mathcal{P}V}(u,v) + b(v,\lambda) &=& \displaystyle\int_\Omega f \, v d\boldsymbol{x} & \forall v \in \mathcal{P}V, \\ b(u,\mu) &=& 0 & \forall \mu \in \times M. \end{array} \right. \tag{12.3}$$

Théorème 12.1 *Le problème* (12.3) *est équivalent au problème* (12.1).

Démonstration : On a vu précédemment que si (u, λ) est solution du problème (12.1), alors (u, λ) est solution du problème (12.3).
Réciproquement, soit (u, λ) une solution du problème (12.3).
Soit $i \in \mathcal{I}_\Omega$ fixé. En prenant des fonctions-test v telles que $v_i \in \mathcal{D}(\Omega_i)$ et $v_m = 0$ si $m \neq i$, on retrouve (12.1)-(i) au sens des distributions dans $\mathcal{D}'(\Omega_i)$, et donc dans $L^2(\Omega_i)$ puisque $f_i \in L^2(\Omega_i)$.
Ensuite, soient $i \in \mathcal{I}_\Omega$ et $j \in \mathcal{I}_{\Omega_i}$ fixés tels que $(i, j) \in \mathcal{IJ}$. En prenant des fonctions-test v telles que

$$v_i \in C^\infty_{\partial\Omega_i \setminus \overline{\Sigma_{ij}}}(\overline{\Omega_i}) := \{v \in C^\infty(\overline{\Omega_i}) \,:\, v = 0 \text{ dans un voisinage de } \partial\Omega_i \setminus \overline{\Sigma_{ij}}\},$$

et $v_m = 0$ si $m \neq i$, après intégration par parties (D.2-i) et à l'aide de (12.1)-(i), on trouve

$$\int_{\Sigma_{ij}} (\lambda_{ij} - k_i \partial_{\boldsymbol{n}_i} u_i)\, v_i \, ds = 0.$$

Or, l'ensemble des traces sur Σ_{ij} de $C^\infty_{\partial\Omega_i \setminus \overline{\Sigma_{ij}}}(\overline{\Omega_i})$ est dense dans $L^2(\Sigma_{ij})$ (voir le §D.3.1). On retrouve donc (12.1)-(iii) dans $L^2(\Sigma_{ij})$.
Enfin, on sait que $[u] \in M$. En prenant $\mu = [u]$, on aboutit à (12.1)-(ii). ◇

12.3 Discrétisation

On se place dans le cas $2D$. Considérons comme discrétisation des espaces V_i les éléments finis de Lagrange P_1 décrits au §3.3.1. Comme au paragraphe 10.4, l'espace de discrétisation de V_i est l'espace $V^+_{i,h}$ (10.23). Pour tout $(i, j) \in \mathcal{IJ}$, on définit une famille de triangulations $(T_{\Sigma_{ij},h})_h$ de Σ_{ij}, constituées de $A^{ij,a}_\lambda$ arêtes $(a_{ij,m})_{m \in \mathcal{I}^{ij}_\lambda}$, $\mathcal{I}^{ij,a}_\lambda = \{1, \cdots, A^{ij,a}_\lambda\}$. On a donc :

$$\overline{\Sigma_{ij}} = \cup_{m \in \mathcal{I}^{ij,a}_\lambda} a_{ij,m}.$$

L'espace de discrétisation de $L^2(\Sigma_{ij})$ est noté $W_{ij,h}$ et est tel que :

$$W_{ij,h} := \left\{ \mu_h \in C^0(\overline{\Sigma_{ij}}) \,:\, \forall m \in \mathcal{I}^{ij,a}_\lambda,\, \mu_{h|a_{ij,m}} \in P_1(a_{ij,m}),\, \mu_{h|\partial\Sigma_{ij} \cap \partial\Omega} = 0 \right\}. \tag{12.4}$$

On note $(M_{ij,m})_{m \in \mathcal{I}^{ij}_\lambda}$ les sommets de la triangulation $T_{\Sigma_{ij},h}$ (h donné) avec $\mathcal{I}^{ij}_\lambda = \{1, \cdots, N^{ij}_\lambda\}$ et $N^{ij}_\lambda = |\mathcal{I}^{ij}_\lambda| = A^{ij,a}_\lambda + 1$. Pour construire une base de $W_{ij,h}$, on choisit une famille de N^{ij}_λ fonctions $(\mu_{ij,m})_{m \in \mathcal{I}^{ij}_\lambda}$ telles que :

$$\text{Pour tout } m \in \mathcal{I}^{ij}_\lambda,\ \mu_{ij,m} \in W_{ij,h},\ \text{et } \mu_{ij,m}(M_{ij,m'}) = \delta_{mm'} \ \text{ pour tout } m' \in \mathcal{I}^{ij}_\lambda.$$

Remarque 12.2 *Par définition, il y a deux degrés de liberté pour toute extrémité commune à deux interfaces distinctes Σ_{ij_1} et Σ_{ij_2}. Si m_1 et m_2 sont tels que $M_{ij_1,m_1} = M_{ij_2,m_2}$ est l'extrémité $\partial\Sigma_{ij_1} \cap \partial\Sigma_{ij_2}$, alors $\mu_{ij_1,m_1} \in W_{ij_1,h}$ et $\mu_{ij_2,m_2} \in W_{ij_2,h}$ sont les deux degrés de liberté correspondants.*

On pose $N_\lambda = \sum_{(i,j)\in\mathcal{IJ}} N_\lambda^{ij}$, $\mathcal{I}_\lambda = \{1, \cdots, N_\lambda\}$ et $(\mu_m)_{m\in\mathcal{I}_\lambda}$ la base de $\prod_{(i,j)\in\mathcal{IJ}} W_{ij,h}$ telle que, pour tout $m \in \mathcal{I}_\lambda$, il existe $(i,j) \in \mathcal{IJ}$ et $m_{ij} \in \mathcal{I}_\lambda^{ij}$ uniques vérifiant $\mu_m = \mu_{ij,m_{ij}}$.
On cherche $\left((u_{i,h})_{i\in\mathcal{I}_\Omega}\,,\, (\lambda_{ij,h})_{(i,j)\in\mathcal{IJ}} \right) \in \left(\prod_{i\in\mathcal{I}_\Omega} V_{i,h}^+\right) \times \left(\prod_{(i,j)\in\mathcal{IJ}} W_{ij,h}\right)$ une approximation de $\left((u_i)_{i\in\mathcal{I}_\Omega}, (\lambda_{ij})_{(i,j)\in\mathcal{IJ}} \right)$ telle que :

$$\begin{cases} \forall i \in \mathcal{I}_\Omega, \quad u_{i,h} = \displaystyle\sum_{m\in\mathcal{I}_i^+} U_{i,m}\, w_{i,m} \text{ où } U_{i,m} := u_{i,h}(M_{i,m}), \\ \forall (i,j) \in \mathcal{IJ}, \quad \lambda_{ij,h} = \displaystyle\sum_{m\in\mathcal{I}_i^+} \Lambda_{ij,m}\, \mu_{ij,m} \text{ où } \Lambda_{ij,m} := \lambda_{ij,h}(M_{ij,m}). \end{cases}$$

La discrétisation de la formulation variationnelle (12.2) s'écrit :
Trouver $\left((u_{i,h})_{i\in\mathcal{I}_\Omega}\,,\, (\lambda_{ij,h})_{(i,j)\in\mathcal{IJ}} \right) \in \left(\prod_{i\in\mathcal{I}_\Omega} V_{i,h}^+\right) \times \left(\prod_{(i,j)\in\mathcal{IJ}} W_{ij,h}\right)$ tel que :

$$\begin{cases} \forall i \in \mathcal{I}_\Omega, \quad \forall m \in \mathcal{I}_i^+, \\ \qquad a_1^i(u_{i,h}, w_{i,m}) - \displaystyle\sum_{j\in\mathcal{I}_{\Omega_i}} \mathrm{sgn}(j-i)(\lambda_{ij,h}\,,\, w_{i,m}\,)_{L^2(\Sigma_{ij})} = \int_{\Omega_i} f_i\, w_{i,m} d\boldsymbol{x}, \\ \\ \forall (i,j) \in \mathcal{IJ}, \quad \forall m \in \mathcal{I}_\lambda^{ij} \qquad (\,\mu_m\,,\, u_{i,h} - u_{j,h}\,)_{L^2(\Sigma_{ij})} = 0. \end{cases} \tag{12.5}$$

Pour $(i,j) \in \mathcal{IJ}$, l'utilisation d'un multiplicateur de Lagrange permet de choisir $V_{i,h}$ et $V_{j,h}$ de sorte que $\mathfrak{T}_{i,h} \neq \mathfrak{T}_{j,h}$ (voir la définition (10.24)). On a alors $N^i_{\Sigma_{ij}} \neq N^j_{\Sigma_{ij}}$ (voir la définition (10.22)). Ce type de méthode de décomposition de domaine est appelée souvent méthode de décomposition de domaine non-conforme.
Il faut prouver que le problème (12.5) est bien posé, en particulier montrer la condition de stabilité (C.7). On peut établir (exercice) qu'elle est bien vérifiée lorsque $W_{ij,h} = \mathfrak{T}_{i,h} + \mathfrak{T}_{j,h}$ (voir [36]). En particulier, lorsque les triangulations à l'interface sont égales ou si l'une est une sous-triangulation de l'autre, on peut choisir la plus fine des deux, soit : $W_{ij,h} = \mathfrak{T}_{j,h}$ lorsque $N^i_{\Sigma_{ij}} \leq N^j_{\Sigma_{ij}}$. Sinon, on peut créer une intersection des arêtes des deux triangulations et utiliser les éléments finis P_1 sur cette nouvelle triangulation.
Après avoir décomposé $u_{i,h}$ sur la base $(w_{i,m})_{m\in\mathcal{I}_i^+}$ pour $i \in \mathcal{I}_\Omega$ et $\lambda_{ij,h}$ sur la base $(\mu_{ij,m})_{m\in\mathcal{I}_\lambda^{ij}}$ pour $(i,j) \in \mathcal{IJ}$, la formulation variationnelle (12.5) se réécrit :

$$\begin{cases} \forall i \in \mathcal{I}_\Omega, \quad \forall m \in \mathcal{I}_i^+, \\ \displaystyle\sum_{m'\in\mathcal{I}_i^+} U_{i,m'}\, a_1^i(w_{i,m'}, w_{i,m}) - \sum_{j\in\mathcal{I}_{\Omega_i}} \mathrm{sgn}(j-i) \sum_{m'\in\mathcal{I}_\lambda^{ij}} \Lambda_{ij,m'}(\,\mu_{ij,m'}\,,\, w_{i,m}\,)_{L^2(\Sigma_{ij})} \\ \qquad = \displaystyle\int_{\Omega_i} f_i\, w_{i,m} d\boldsymbol{x}, \\ \\ \forall (i,j) \in \mathcal{IJ}, \quad \forall m \in \mathcal{I}_\lambda^{ij} \\ \displaystyle\sum_{m'\in\mathcal{I}_i^+} U_{i,m'}(\,\mu_{ij,m}\,,\, w_{i,m'})_{L^2(\Sigma_{ij})} - \sum_{m'\in\mathcal{I}_j^+} U_{j,m'}(\,\mu_{ij,m}\,,\, w_{j,m'}\,)_{L^2(\Sigma_{ij})} = 0. \end{cases} \tag{12.6}$$

On peut mettre les équations (12.6) sous forme matricielle.

Soit $i \in \mathcal{I}_\Omega$. Soit $\mathbb{A}_{i,i} \in \mathbb{R}^{N_i^+ \times N_i^+}$ la matrice symétrique telle que :

$$\forall (m,m') \in \mathcal{I}_i^+ \times \mathcal{I}_i^+, (\mathbb{A}_{i,i})_{m,m'} = a_1^i(w_{i,m'}, w_{i,m}).$$

Soit $\mathbb{M}_{i,\lambda} \in \mathbb{R}^{N_i^+ \times N_\lambda}$ la matrice telle que :

$$\forall\, (m,m') \in \mathcal{I}_i^+ \times \mathcal{I}_\lambda, \quad (\mathbb{M}_{i,\lambda})_{m,m'} = \left\{ \begin{array}{rcl} 0 & \text{si} & m \in \mathcal{I}_i \\ (\,\mu_{m',h}\,,\, w_{i,m}\,)_{L^2(\Sigma_{ij})} & \text{si} & m \in \mathcal{I}_{i,\Sigma_i}^{ij} \end{array} \right. .$$

Soient $\widetilde{U}_i$ et $\widetilde{F}_i \in \mathbb{R}^{N_i^+}$ les vecteurs tels que : $\left(\widetilde{U}_i\right)_m = U_{i,m}$ et $\left(\widetilde{F}_i\right)_m = F_{i,m}$.
Pour $(i,j) \in \mathcal{IJ}$, on définit $\Lambda_{ij} \in \mathbb{R}^{N_\lambda^{ij}}$ le vecteur tel que $\forall m \in \mathcal{I}_\lambda^{ij}$, $(\Lambda_{ij})_m = \Lambda_{ij,m}$.
Le système linéaire correspondant à (12.6) s'écrit :
Trouver $\left(\left(\widetilde{U}_i\right)_{i\in\mathcal{I}_\Omega}, (\Lambda_{ij})_{(i,j)\in\mathcal{IJ}} \right) \in \prod_{i\in\mathcal{I}_\Omega} \mathbb{R}^{N_i^+} \times \prod_{(i,j)\in\mathcal{IJ}} \mathbb{R}^{N_\lambda^{ij}}$ tel que :

$$\left\{ \begin{array}{l} \forall i \in \mathcal{I}_\Omega, \quad \mathbb{A}_{i,i} U_i - \displaystyle\sum_{j\in\mathcal{I}_{\Omega_i}} \mathrm{sgn}(j-i)\mathbb{M}_{i,\lambda}\Lambda_{ij} = \widetilde{F}_i \\ \\ \forall (i,j) \in \mathcal{IJ}, \quad \mathbb{M}_{i,\lambda}^T\, U_i - \mathbb{M}_{j,\lambda}^T\, U_j = 0. \end{array} \right. \tag{12.7}$$

On rappelle que $N^+ = \displaystyle\sum_{i\in\mathcal{I}_\Omega} N_i^+$. Soit $\mathbb{A}_\lambda \in \mathbb{R}^{N^+} \times \mathbb{R}^{N^+}$ la matrice diagonale par blocs, dont les blocs diagonaux sont tels que : $\forall i \in \mathcal{I}_\Omega$, $[\mathbb{A}_\lambda]_{i,i} = \mathbb{A}_{i,i}$ (voir le §5.3).
Soit $\mathbb{M}_\lambda \in \mathbb{R}^{N^+} \times \mathbb{R}^{N_\lambda}$ la matrice composée de $N_\Omega \times N_{\mathcal{IJ}}$ blocs, où $N_{\mathcal{IJ}}$ est le nombre d'interfaces. Les blocs non nuls sont tels que : $\forall (i,j) \in \mathcal{IJ}$, $[\mathbb{M}_\lambda]_{i,(i,j)} = -\mathbb{M}_{i,\lambda}$ et $[\mathbb{M}_\lambda]_{j,(i,j)} = \mathbb{M}_{j,\lambda}$.
Soient $\widetilde{U}$ et $\widetilde{F} \in \mathbb{R}^{N^+}$ les vecteurs composés de N_Ω blocs tels que : $\forall i \in N_\Omega$, $[\widetilde{U}]_i = \widetilde{U}_i$, $[\widetilde{F}]_i = \widetilde{F}_i$.
Soit $\Lambda \in \mathbb{R}^{N_\lambda}$ le vecteur composé de $N_{\mathcal{IJ}}$ blocs tel que : $\forall (i,j) \in N_{\mathcal{IJ}}$, $[\Lambda]_{(i,j)} = \Lambda_{ij}$.
Le système linéaire (12.7) se réécrit :
Trouver $(\widetilde{U}, \Lambda) \in \mathbb{R}^{N^+} \times \mathbb{R}^{N_\lambda}$ tel que :

$$\left(\begin{array}{cc} \mathbb{A}_\lambda & \mathbb{M}_\lambda \\ (\mathbb{M}_\lambda)^T & 0 \end{array} \right) \left(\begin{array}{c} \widetilde{U} \\ \Lambda \end{array} \right) = \left(\begin{array}{c} \widetilde{F} \\ 0 \end{array} \right). \tag{12.8}$$

Pour résoudre ce système linéaire, on peut utiliser l'algorithme suivant (dit **algorithme d'Uzawa**), en calculant d'abord Λ puis $\widetilde{U}$:

$$\left\{ \begin{array}{rcl} (\mathbb{M}_\lambda)^T \mathbb{A}_\lambda^{-1} \mathbb{M}_\lambda \Lambda & = & (\mathbb{M}_\lambda)^T \mathbb{A}_\lambda^{-1} \widetilde{F}, \\ \mathbb{A}_\lambda \widetilde{U} & = & \widetilde{F} - \mathbb{M}_\lambda \Lambda. \end{array} \right. \tag{12.9}$$

On reconnaît à la première ligne le complément de Schur.
Les matrices $\mathbb{A}_{i,i}$ sont symétriques, définies-positives, et les matrices $\mathbb{M}_{i,\lambda}$ sont de rang maximal. La matrice $(\mathbb{M}_\lambda)^T \mathbb{A}_\lambda^{-1} \mathbb{M}_\lambda$ est donc symétrique définie-positive : on peut utiliser l'algorithme du gradient conjugué pour calculer Λ. La matrice $\mathbb{A}_\lambda$ étant diagonale par blocs, la résolution $\widetilde{U}$ peut se faire en parallèle sur chaque bloc.

12.4 Interprétation algébrique

Le système linéaire (12.8), issu d'une méthode de décomposition de domaine, s'obtient également de façon algébrique, en manipulant les sous-blocs du système linéaire obtenu après discrétisation du problème *sans décomposition de domaine*. Plus précisément, étudions le cas de deux sous-domaines. [49]
En partant du problème (10.31) et en manipulant les blocs d'équations, on peut obtenir un système linéaire proche du système linéaire (12.7). Remarquons que l'on a :

$$(\mathbb{A}_{1,\Sigma})^T U_1 + \mathbb{A}^1_{\Sigma,\Sigma} U_\Sigma + \mathbb{A}^2_{\Sigma,\Sigma} U_\Sigma + (\mathbb{A}_{2,\Sigma})^T U_2 = F_\Sigma = F_{1,\Sigma} + F_{2,\Sigma}.$$

On introduit alors $\Lambda' \in \mathbb{R}^{N_\Sigma}$ tel que :

$$\Lambda' := (\mathbb{A}_{1,\Sigma})^T U_1 + \mathbb{A}^1_{\Sigma,\Sigma} U_\Sigma - F_{1,\Sigma} = -(\mathbb{A}_{2,\Sigma})^T U_2 - \mathbb{A}^2_{\Sigma,\Sigma} U_\Sigma + F_{2,\Sigma},$$

de sorte que :

$$(\mathbb{A}_{1,\Sigma})^T U_1 + \mathbb{A}^1_{\Sigma,\Sigma} U_\Sigma - \Lambda' = F_{1,\Sigma}, \quad \mathbb{A}^2_{\Sigma,\Sigma} U_\Sigma + (\mathbb{A}_{2,\Sigma})^T U_2 + \Lambda' = F_{2,\Sigma}.$$

Pour traiter le vecteur Λ' comme une nouvelle inconnue, il faut introduire les $\mathbb{R}^{N_\Sigma}$ équations indépendantes données par la relation $U_{1,\Sigma} = U_{2,\Sigma}$. Soit $\mathbb{M}_\Sigma \in \mathbb{R}^{N_\Sigma \times N_\Sigma}$ une matrice inversible, si on introduit $\widetilde{\Lambda} := (\mathbb{M}_\Sigma)^{-1}\Lambda'$, on peut finalement réécrire le système linéaire (10.31) sous la forme :

$$\begin{pmatrix} \mathring{\mathbb{A}}_{1,1} & \mathbb{A}_{1,\Sigma} & 0 & 0 & 0 \\ (\mathbb{A}_{1,\Sigma})^T & \mathbb{A}^1_{\Sigma,\Sigma} & 0 & 0 & -\mathbb{M}_\Sigma \\ 0 & 0 & \mathring{\mathbb{A}}_{2,2} & \mathbb{A}_{2,\Sigma} & 0 \\ 0 & 0 & (\mathbb{A}_{2,\Sigma})^T & \mathbb{A}^2_{\Sigma,\Sigma} & \mathbb{M}_\Sigma \\ 0 & -\mathbb{M}_\Sigma & 0 & \mathbb{M}_\Sigma & 0 \end{pmatrix} \begin{pmatrix} U_1 \\ U_{1,\Sigma} \\ U_2 \\ U_{2,\Sigma} \\ \widetilde{\Lambda} \end{pmatrix} = \begin{pmatrix} F_1 \\ F_{1,\Sigma} \\ F_2 \\ F_{2,\Sigma} \\ 0 \end{pmatrix}. \tag{12.10}$$

Soient

$$\widetilde{U}_1 = \begin{pmatrix} U_1 \\ U_{1,\Sigma} \end{pmatrix} \in \mathbb{R}^{N_1^+}, \ \widetilde{U}_2 = \begin{pmatrix} U_2 \\ U_{2,\Sigma} \end{pmatrix} \in \mathbb{R}^{N_2^+} \text{ et } \widetilde{U} = \begin{pmatrix} \widetilde{U}_1 \\ \widetilde{U}_2 \end{pmatrix} \in \mathbb{R}^{N^+}.$$

De même pour définir $\widetilde{F} \in \mathbb{R}^{N^+}$. Soient $\widetilde{\mathbb{M}}_{i,\lambda} := \begin{pmatrix} 0 \\ \mathbb{M}_\Sigma \end{pmatrix} \in \mathbb{R}^{N_i^+ \times N_\Sigma}$ et $\widetilde{\mathbb{M}}_\lambda := \begin{pmatrix} -\widetilde{\mathbb{M}}_{1,\lambda} \\ \widetilde{\mathbb{M}}_{2,\lambda} \end{pmatrix}$.
Le système linéaire (12.10) se met sous la forme :
Trouver $(\widetilde{U}, \widetilde{\Lambda}) \in \mathbb{R}^{N^+} \times \mathbb{R}^{N_\lambda}$ tel que :

$$\begin{pmatrix} \mathbb{A}_\lambda & \widetilde{\mathbb{M}}_\lambda \\ (\widetilde{\mathbb{M}}_\lambda)^T & 0 \end{pmatrix} \begin{pmatrix} \widetilde{U} \\ \widetilde{\Lambda} \end{pmatrix} = \begin{pmatrix} \widetilde{F} \\ 0 \end{pmatrix}. \tag{12.11}$$

Cette approche donne une méthode systématique pour résoudre en parallèle le système linéaire (10.31), similaire à la méthode du complément de Schur.

49. Voir la note de bas de page[48] page 205. En outre, on décompose F_Σ en deux contributions $F_{i,\Sigma}$, $i \in \{1,2\}$ telles que $(F_{i,\Sigma})_m = \int_{\Omega_i} f\, w_{i,m} d\boldsymbol{x}$.

Si $\mathbb{M}_\Sigma$ est égale à la matrice identité d'ordre N_Σ, on remarque que la matrice de couplage $\widetilde{\mathbb{M}}_\lambda$ ne dépend pas des volumes des éléments d'interface, ce qui pourrait poser des problèmes de stabilité numérique.
Si $\mathbb{M}_\Sigma$ est la matrice de masse associée aux degrés de liberté de l'interface, on trouve $\widetilde{\mathbb{M}}_\lambda = \mathbb{M}_\lambda$.
La généralisation aux maillages non conformes est possible en adaptant le nombre de degrés de liberté de $\widetilde{\Lambda}$, et en modifiant astucieusement la matrice $\mathbb{M}_\Sigma$ dans la matrice $\widetilde{\mathbb{M}}_{i,\lambda}$ (qui ne sera plus nécessairement une matrice carrée).

Chapitre 13

Problème à deux inconnues, méthode avec contrainte

On souhaite à nouveau résoudre le problème (9.1) ou (9.2). Comme au chapitre 11, on va s'appuyer sur les inconnues $(u_i, \boldsymbol{p}_i)_{i\in\mathcal{I}_\Omega}$. Par contre, et contrairement à l'hypothèse faite aux chapitres précédents, il n'est plus nécessaire de supposer que la trace normale de $\boldsymbol{p}_i = k_i \,\mathbf{grad}\, u_i$ sur les interfaces Σ_{ij} appartient à $L^2(\Sigma_{ij})$ pour $(i,j) \in \mathcal{I}_\Omega \times \mathcal{I}_{\Omega_i}$. L'idée développée ci-dessous consiste, plutôt que de considérer séparément la trace normale de chaque côté des interfaces, à s'intéresser directement à la différence de ces traces sur les interfaces (ce qu'on a appelé le saut, défini en (9.8)).

13.1 Approche continue

Si on part du problème à deux inconnues (9.16), le multiplicateur de Lagrange λ représente la trace de u sur Σ_S et appartient à l'espace $H_-^{1/2}(\Sigma_S)$

$$H_-^{1/2}(\Sigma_S) = \left\{ \mu \in M \,|\, \mu_{ij} \in H^{1/2}(\Sigma_{ij}),\, \forall (i,j) \in \mathcal{IJ} \right\}, \tag{13.1}$$

où on rappelle que $M = \left\{ \mu \in \prod_{(i,j)\in\mathcal{IJ}} L^2(\Sigma_{ij}) \right\}$. Par la suite, pour tout $\mu \in M$, on choisit la notation "naturelle" $\mu_{ij} := \mu_{|\Sigma_{ij}}$. Le problème (9.16) est équivalent au problème :
Trouver $(u, \boldsymbol{p}, \lambda) \in \mathcal{P}V \times \mathcal{P}\mathbf{H}(\mathrm{div}, \Omega) \times H_-^{1/2}(\Sigma_S)$ tel que :

$$\left\{ \begin{array}{lllrcll} (0) & \forall i \in \mathcal{I}_\Omega, & & \mathrm{div}\, \boldsymbol{p}_i + q_i u_i &=& f_i & \text{dans } \Omega_i, \\ (i) & \forall i \in \mathcal{I}_\Omega, & & k_i^{-1}\boldsymbol{p}_i + \mathbf{grad}\, u_i &=& 0 & \text{dans } \Omega_i, \\ (ii) & \forall i \in \mathcal{I}_\Omega,\, \forall j \in \mathcal{I}_{\Omega_i}, & & u_i &=& -\lambda_{ij} & \text{sur } \Sigma_{ij}, \\ (iii) & \forall (i,j) \in \mathcal{IJ}, & & [\boldsymbol{p}\cdot\boldsymbol{n}]_{ij} &=& 0 & \text{sur } \Sigma_{ij}, \end{array} \right. \tag{13.2}$$

13.2 Formulation variationnelle

Bien sûr, $u \in L^2(\Omega)$. D'autre part, on a vu à la section 9.1 que $[\boldsymbol{p}\cdot\boldsymbol{n}]_{ij} \in H'_{ij}$ pour tout $(i,j) \in \mathcal{I}_\Omega \times \mathcal{I}_{\Omega_i}$. Par ailleurs, $L^2(\Sigma_{ij}) \subset H'_{ij}$ (voir §D.2). Comme $0 \in L^2(\Sigma_{ij})$, on en conclut que, d'après (13.2)-(iii), on a en particulier $[\boldsymbol{p}\cdot\boldsymbol{n}]_{ij} \in L^2(\Sigma_{ij})$ pour tout $(i,j) \in \mathcal{I}_\Omega \times \mathcal{I}_{\Omega_i}$,

et $[\boldsymbol{p}\cdot\boldsymbol{n}] \in M$. Ainsi, $\boldsymbol{p} \in \widetilde{\mathbf{H}}$, où $\widetilde{\mathbf{H}}$ est défini en (9.10). On va maintenant chercher $(u,\boldsymbol{p}) \in L^2(\Omega)\times\widetilde{\mathbf{H}}$. Pour cela, on construit une nouvelle formulation variationnelle.
On rappelle que, par équivalence entre (9.16) et le problème initial (9.1) ou (9.2), on sait que $u \in H_0^1(\Omega)$. Ceci nous permet d'utiliser la formule d'intégration par parties (9.11), pour $\boldsymbol{q}\in\widetilde{\mathbf{H}}$, puis (13.2)-(ii) comme caractérisation de la trace de u sur Σ_S, pour trouver :

$$\begin{aligned}\int_\Omega \mathbf{grad}\, u\cdot\boldsymbol{q}\,d\boldsymbol{x} &= -\int_\Omega u\ \mathrm{div}\,\boldsymbol{q}\,d\boldsymbol{x} + \int_{\Sigma_S}[\boldsymbol{q}\cdot\boldsymbol{n}]\,u_{|\Sigma_S}\,ds\\ &= -\int_\Omega u\ \mathrm{div}\,\boldsymbol{q}\,d\boldsymbol{x} - \int_{\Sigma_S}[\boldsymbol{q}\cdot\boldsymbol{n}]\,\lambda\,ds.\end{aligned}$$

(Voir la note de bas de page [45] page 190 pour l'écriture d'intégrales sur Ω).
Si maintenant on utilise (13.2)-(i), on trouve cette fois

$$\begin{aligned}\int_\Omega \mathbf{grad}\, u\cdot\boldsymbol{q}\,d\boldsymbol{x} &= \sum_{i\in\mathcal{I}_\Omega}\int_{\Omega_i}\mathbf{grad}\,u_i\cdot\boldsymbol{q}_i\,d\boldsymbol{x} = -\sum_{i\in\mathcal{I}_\Omega}\int_{\Omega_i}k_i^{-1}\boldsymbol{p}_i\cdot\boldsymbol{q}_i\,d\boldsymbol{x}\\ &= -\int_\Omega k^{-1}\boldsymbol{p}\cdot\boldsymbol{q}\,d\boldsymbol{x}.\end{aligned}$$

Ainsi, pour tout $\boldsymbol{q}\in\widetilde{\mathbf{H}}$, on a :

$$-\int_\Omega k^{-1}\boldsymbol{p}\cdot\boldsymbol{q}\,d\boldsymbol{x} + \int_\Omega u\ \mathrm{div}\,\boldsymbol{q}\,d\boldsymbol{x} + \int_{\Sigma_S}[\boldsymbol{q}\cdot\boldsymbol{n}]\,\lambda\,ds = 0.$$

Ci-dessus, le gradient de u a disparu.
Par ailleurs, pour tout $v\in L^2(\Omega)$, on a d'après (13.2)-(0) que :

$$\int_\Omega(\mathrm{div}\,\boldsymbol{p}+qu)v\,d\boldsymbol{x} = \int_\Omega f v\,d\boldsymbol{x}.$$

Enfin, pour tout $\mu\in M$, d'après (13.2)-(iii) il suit :

$$\int_{\Sigma_S}[\boldsymbol{p}\cdot\boldsymbol{n}]\,\mu\,ds = 0.$$

Pour résumer, si le triplet $(u,\boldsymbol{p},\lambda)$ est solution de (13.2), alors c'est une solution de la formulation variationnelle :
Trouver $(u,\boldsymbol{p},\lambda)\in L^2(\Omega)\times\widetilde{\mathbf{H}}\times M$ tel que $\forall(v,\boldsymbol{q},\mu)\in L^2(\Omega)\times\widetilde{\mathbf{H}}\times M$:

$$\begin{aligned}&\int_\Omega\left(-k^{-1}\boldsymbol{p}\cdot\boldsymbol{q} + u\ \mathrm{div}\,\boldsymbol{q} + v\,\mathrm{div}\,\boldsymbol{p} + quv\right)d\boldsymbol{x}\\ &\qquad+\int_{\Sigma_S}\left([\boldsymbol{q}\cdot\boldsymbol{n}]\,\lambda + [\boldsymbol{p}\cdot\boldsymbol{n}]\,\mu\right)ds = \int_\Omega f v\,d\boldsymbol{x}.\end{aligned}\tag{13.3}$$

Théorème 13.1 *Le problème* (13.2) *est équivalent au problème* (13.3).

Démonstration : On vient d'établir que si $(u, \boldsymbol{p}, \lambda)$ est solution du problème (13.2), alors $(u, \boldsymbol{p}, \lambda)$ est solution du problème (13.3).
Réciproquement, soit $(u, \boldsymbol{p}, \lambda)$ une solution du problème (13.3). Puisque $\widetilde{\mathbf{H}} \subset \mathcal{P}\mathbf{H}(\mathrm{div}, \Omega)$, on sait déjà que $\boldsymbol{p} \in \mathcal{P}\mathbf{H}(\mathrm{div}, \Omega)$. Ci-dessous, $(v, \boldsymbol{q}, \mu)$ sont des fonctions-test bien choisies. Soit $i \in \mathcal{I}_\Omega$ fixé. On choisit v avec $v_m = 0$ si $m \neq i$, et $\boldsymbol{q} = 0$, $\mu = 0 : \forall v_i \in L^2(\Omega_i)$,

$$\int_{\Omega_i} (\mathrm{div}\, \boldsymbol{p}_i + q u_i - f_i) v_i \, d\boldsymbol{x} = 0,$$

c'est-à-dire (13.2)-(0).
On choisit maintenant $v = 0$, $\boldsymbol{q}$ avec $\boldsymbol{q}_i \in \boldsymbol{D}(\Omega_i)$ et $\boldsymbol{q}_m = 0$ si $m \neq i$, et $\mu = 0 : \forall \boldsymbol{q}_i \in \boldsymbol{D}(\Omega_i)$,

$$\langle -k_i^{-1} \boldsymbol{p}_i - \mathbf{grad}\, u_i, \boldsymbol{q}_i \rangle = 0,$$

c'est-à-dire (13.2)-(i) au sens des distributions dans $\boldsymbol{D}'(\Omega_i)$. Ainsi, $u_i \in H^1(\Omega_i)$ pour tout $i \in \mathcal{I}_\Omega$. En particulier, la trace de u_i sur $\partial\Omega_i$ est bien définie. Montrons que $u_i = 0$ sur $\Gamma_i = \partial\Omega_i \cap \partial\Omega$. On introduit $\nu \in L^2(\partial\Omega_i)$ telle que $\nu = u_i$ sur Γ_i, et 0 ailleurs (c'est-à-dire sur $\partial\Omega_i \cap \Sigma_S$). On rappelle que $L^2(\partial\Omega_i) \subset (H^{1/2}(\partial\Omega_i))'$. Par surjectivité de l'application trace normale (voir le théorème D.3), il existe $\boldsymbol{q}_i \in \mathbf{H}(\mathrm{div}, \Omega_i)$ tel que $\boldsymbol{q}_i \cdot \boldsymbol{n}_i = \nu$ sur $\partial\Omega_i$. Si on annule v, $\boldsymbol{q}_m$ si $m \neq i$ et μ, on trouve

$$\int_{\Omega_i} (-k_i^{-1} \boldsymbol{p}_i \cdot \boldsymbol{q}_i + u_i \ \mathrm{div}\, \boldsymbol{q}_i) \, d\boldsymbol{x} = 0.$$

Or, on sait déjà que $-k_i^{-1} \boldsymbol{p}_i = \mathbf{grad}\, u_i$ dans Ω_i : on remplace $-k_i^{-1} \boldsymbol{p}_i$ par $\mathbf{grad}\, u_i$ dans l'intégrale volumique et on intègre par parties (D.2) pour trouver

$$\int_{\partial\Omega_i} \boldsymbol{q}_i \cdot \boldsymbol{n}_i \, u_i \, ds = 0.$$

Soit, en remplaçant la trace normale de $\boldsymbol{q}_i$ par sa valeur, que $\|u_i\|^2_{L^2(\Gamma_i)} = 0$. En d'autres termes, on a obtenu que $u_i \in V_i$, et donc que $u \in \mathcal{P}V$.
Ensuite, soient $i \in \mathcal{I}_\Omega$ et $j \in \mathcal{I}_{\Omega_i}$ fixés tels que $(i, j) \in \mathcal{I}\mathcal{J}$. On choisit maintenant $v = 0$, $\boldsymbol{q} = 0$, et μ avec $\mu_{ij} = [\boldsymbol{p} \cdot \boldsymbol{n}]_{ij}$ et $\mu_{mn} = 0$ si $(m, n) \neq (i, j)$: on a $\|[\boldsymbol{p} \cdot \boldsymbol{n}]_{ij}\|^2_{L^2(\Sigma_{ij})} = 0$, c'est-à-dire (13.2)-(iii).
Soit une dernière fois $i \in \mathcal{I}_\Omega$ fixé. On définit $\lambda_i \in L^2(\partial\Omega_i)$ tel que pour tout $j \in \mathcal{I}_{\Omega_i}$, $\lambda_{i|\Sigma_{ij}} = \lambda_{ij}$, et $\lambda_{i|\Gamma_i} = 0$. On sait que $u_i + \lambda_i \in L^2(\partial\Omega_i) \subset (H^{1/2}(\partial\Omega_i))'$. Par surjectivité de l'application trace normale (voir le théorème D.3), il existe $\boldsymbol{q}_i \in \mathbf{H}(\mathrm{div}, \Omega_i)$ tel que $\boldsymbol{q}_i \cdot \boldsymbol{n}_i = u_i + \lambda_i$ sur $\partial\Omega_i$. Si on annule v, $\boldsymbol{q}_m$ si $m \neq i$ et μ, on trouve

$$\int_{\Omega_i} (-k_i^{-1} \boldsymbol{p}_i \cdot \boldsymbol{q}_i + u_i \ \mathrm{div}\, \boldsymbol{q}_i) \, d\boldsymbol{x} + \int_{\partial\Omega_i} \boldsymbol{q}_i \cdot \boldsymbol{n}_i \lambda_i \, ds = 0.$$

On remplace $-k_i^{-1} \boldsymbol{p}_i$ par $\mathbf{grad}\, u_i$ dans l'intégrale volumique et on intègre par parties (D.2) pour trouver

$$\int_{\partial\Omega_i} \boldsymbol{q}_i \cdot \boldsymbol{n}_i (u_i + \lambda_i) \, ds = 0.$$

Soit, en remplaçant la trace normale de $\boldsymbol{q}_i$ par sa valeur $u_i + \lambda_i$ qui s'annule sur Γ_i, que $\|u_i + \lambda_i\|^2_{L^2(\partial\Omega_i \cap \Sigma_S)} = 0$. En d'autres termes, on a obtenu (13.2)-(ii). Et il suit finalement que $\lambda \in H^{1/2}_-(\Sigma_S)$. ◇

On va réécrire cette formulation variationnelle sous une forme plus compacte. Considérons les espaces produits suivants, munis de leur norme produit :

$$\widetilde{\mathcal{V}} = L^2(\Omega) \times \widetilde{\mathbf{H}} \; ; \quad \|(v, \boldsymbol{q})\|_{\widetilde{\mathcal{V}}} = \left(\|v\|^2_{L^2(\Omega)} + \|\boldsymbol{q}\|^2_{\widetilde{\mathbf{H}}} \right)^{1/2}, \tag{13.4}$$

et $\widetilde{a}_2$ la forme bilinéaire suivante :

$$\left\{ \begin{array}{rcl} \widetilde{a}_2 : \widetilde{\mathcal{V}} \times \widetilde{\mathcal{V}} & \to & \mathbb{R} \\ ((v, \boldsymbol{q}), (w, \boldsymbol{r})) & \mapsto & \displaystyle\sum_{i \in \mathcal{I}_\Omega} \int_{\Omega_i} \left(-k_i^{-1} \boldsymbol{q} \cdot \boldsymbol{r} + v \operatorname{div} \boldsymbol{r} + w \operatorname{div} \boldsymbol{q} + q_i \, vw \right) d\boldsymbol{x} \end{array} \right. . \tag{13.5}$$

On peut alors réécrire la formulation variationnelle (13.3) sous la forme :
Trouver $((u, \boldsymbol{p}), \lambda) \in \widetilde{\mathcal{V}} \times M$ tel que $\forall ((v, \boldsymbol{q}), \mu) \in \widetilde{\mathcal{V}} \times M$:

$$\widetilde{a}_2((u, \boldsymbol{p}), (v, \boldsymbol{q})) + ([\boldsymbol{q} \cdot \boldsymbol{n}], \lambda)_{L^2(\Sigma_S)} + ([\boldsymbol{p} \cdot \boldsymbol{n}], \mu)_{L^2(\Sigma_S)} = (f, v)_{L^2(\Omega)}. \tag{13.6}$$

On peut alors appliquer la théorie de Banach-Necas-Babuska ou la théorie de la T-coercivité pour établir le caractère bien posé de (13.6), voir [9].

Théorème 13.2 *Sous les hypothèses (3.2), la formulation variationnelle (13.6) est bien posée. En particulier il existe une constante $C > 0$ ne dépendant que de Ω et des coefficients telle que :*

$$\begin{array}{c} \forall f \in L^2(\Omega), \; \exists ! (u, \boldsymbol{p}, \lambda) \text{ solution de (13.6) tel que} \\ \|u\|_{L^2(\Omega)} + \|\boldsymbol{p}\|_{\widetilde{\mathbf{H}}} + \|\lambda\|_{L^2(\Sigma_S)} \leq C \, \|f\|_{L^2(\Omega)}. \end{array}$$

13.3 Discrétisation

On se place dans le cas $2D$. Considérons comme discrétisation produit de l'espace $L^2(\Omega) \times \widetilde{\mathbf{H}}$ les éléments finis de Raviart-Thomas $P_0 - \mathbf{RT}_0$ décrits au §3.3.2. L'espace de discrétisation de $L^2(\Omega)$ est $M_h := \prod_{i \in \mathcal{I}_\Omega} M_{i,h}$ et l'espace de discrétisation de $\widetilde{\mathbf{H}}$ est $\widetilde{\mathbf{H}}_h := \prod_{i \in \mathcal{I}_\Omega} \mathbf{Q}_{i,h}$,voir (11.18). [50] Pour tout $(i, j) \in \mathcal{I}\mathcal{J}$, on définit une famille de triangulations $(T_{\Sigma_{ij},h})_h$ de Σ_{ij}, constituées de $A_\lambda^{ij,a}$ arêtes $(a_{ij,m})_{m \in \mathcal{I}_\lambda^{ij,a}}$, $\mathcal{I}_\lambda^{ij,a} = \{1, \cdots, A_\lambda^{ij,a}\}$. On a donc :

$$\overline{\Sigma_{ij}} = \bigcup_{m \in \mathcal{I}_\lambda^{ij,a}} a_{ij,m}.$$

L'espace de discrétisation de $L^2(\Sigma_{ij})$ est noté $M_{\Sigma_{ij},h}$ et est tel que :

$$M_{\Sigma_{ij},h} := \left\{ \mu_h \in L^2(\Sigma_{ij}) \; : \; \forall m \in \mathcal{I}_\lambda^{ij,a}, \; \mu_{h|a_{ij,m}} \in P_0(a_{ij,m}) \right\}. \tag{13.7}$$

Pour construire une base de $M_{\Sigma_{ij},h}$, on choisit les fonctions caractéristiques $(\mu_m)_{m \in \mathcal{I}_\lambda^{ij,a}}$:

$$\forall m, \, m' \in \mathcal{I}_\lambda^{ij,a}, \; \mu_{m|a_{ij,m'}} = \delta_{m,m'}.$$

50. Les espaces abstraits sont différents de ceux du paragraphe §11.3, et en particulier ils sont munis de normes différentes, comparer $\|\cdot\|_{\mathcal{V}_i^+}$ à $\|\cdot\|_{\widetilde{\mathbf{H}}}$. Néanmoins on choisit la même discrétisation.

Posons $M_{\Sigma,h} = \prod_{(i,j)\in\mathcal{IJ}} M_{\Sigma_{ij},h}$. La formulation discrète de (13.6) s'écrit :
Trouver $(u_h, \boldsymbol{p}_h, \lambda_h) \in M_h \times \widetilde{\mathbf{H}}_h \times M_{\Sigma,h}$ tel que :

$$\forall (v_h, \boldsymbol{p}_h, \mu_h) \in M_h \times \widetilde{\mathbf{H}}_h \times M_{\Sigma,h},$$

$$\widetilde{a}_2((u_h, \boldsymbol{p}_h), (v_h, \boldsymbol{p}_h)) + ([\boldsymbol{q}_h \cdot \boldsymbol{n}], \lambda_h)_{L^2(\Sigma_S)} + ([\boldsymbol{p}_h \cdot \boldsymbol{n}], \mu_h)_{L^2(\Sigma_S)} = (f, v_h)_{L^2(\Omega)}. \tag{13.8}$$

On reprend les notations du §11. Comme pour le cas à une inconnue, l'utilisation d'un multiplicateur de Lagrange permet de choisir, pour chaque couple $(i,j) \in \mathcal{IJ}$, des espaces $\widetilde{\mathbf{H}}_{i,h}$ et $\widetilde{\mathbf{H}}_{j,h}$ tels que $\mathfrak{T}^{ij,a}_{i,h} \neq \mathfrak{T}^{ij,a}_{j,h}$ (voir la définition (11.20)) [9]. Il faut prouver que le problème (13.8) est bien posé, en particulier montrer la condition de stabilité (C.7). On peut montrer que celle-ci est vérifiée lorsque $M_{\Sigma_{ij},h} = \mathfrak{T}^{ij,a}_{i,h} + \mathfrak{T}^{ij,a}_{j,h}$. Ainsi lorsque les triangulations à l'interface sont égales, on peut choisir : $M_{\Sigma_{ij},h} = \mathfrak{T}^{ij,a}_{i,h} = \mathfrak{T}^{ij,a}_{j,h}$. Si l'une des triangulation est une sous-triangulation de l'autre, on peut choisir la plus fine des deux. Sinon, on peut calculer l'intersection des arêtes des deux triangulations et utiliser les éléments finis P_0 sur cette nouvelle triangulation. Dans les cas énumérés ci-dessus, on dit que les discrétisations entre $\widetilde{\mathbf{H}}_{i,h}$ et $\widetilde{\mathbf{H}}_{j,h}$ d'une part, et $M_{\Sigma_{ij},h}$ d'autre part, sont compatibles. On a le résultat suivant, voir [9].

Théorème 13.3 *Sous les hypothèses (3.2) et si les discrétisations sont compatibles, la formulation variationnelle discrète (13.8) est bien posée, et elle vérifie une condition de stabilité uniforme. En particulier :*

$$\lim_{h\to 0} \left(\|u - u_h\|_{L^2(\Omega)} + \|\boldsymbol{p} - \boldsymbol{p}_h\|_{\widetilde{\mathbf{H}}} + \|\lambda - \lambda_h\|_{L^2(\Sigma_S)} \right) = 0.$$

La formulation discrète (13.8) s'écrit (les espaces d'indices sont définis au §11.3) :
Trouver $((u_{i,h}, \boldsymbol{p}_{i,h})_{i\in\mathcal{I}_\Omega}, (\lambda_{ij,h})_{(i,j)\in\mathcal{IJ}})$ tel que :

$$\forall i \in \mathcal{I}_\Omega,\ \forall\, (l,m) \in \mathcal{I}_i^T \times \mathcal{I}_i^{a+} \ :\ a_2^i((u_{i,h}, \boldsymbol{p}_{i,h})\,,\, (\underline{w}_{i,l}, \underline{\boldsymbol{\omega}}_{i,m})) + \sum_{j\in\mathcal{I}_{\Omega_i}} (\underline{\boldsymbol{\omega}}_{i,m} \cdot \boldsymbol{n}_i, \lambda_{ij,h})_{L^2(\Sigma_{ij})} = (f, \underline{w}_{i,l})_{L^2(\Omega_i)},$$

$$\forall (i,j) \in \mathcal{IJ},\ \forall\, m \in \mathcal{I}_\lambda^{ij,a} \ :\ (\boldsymbol{p}_{i,h} \cdot \boldsymbol{n}_i + \boldsymbol{p}_{j,h} \cdot \boldsymbol{n}_j, \mu_m)_{L^2(\Sigma_{ij})} = 0.$$

Après avoir décomposé $u_{i,h}$, $\boldsymbol{p}_{i,h}$ et $\lambda_{ij,h}$ dans leurs bases respectives, on obtient :
Trouver $\Big(((U_{i,l'})_{l'\in\mathcal{I}_i^T}, (P_{i,m'})_{m'\in\mathcal{I}_i^{a+}})_{i\in\mathcal{I}_\Omega}, ((\lambda_{ij,m'})_{m'\in\mathcal{I}_\lambda^{ij,a}})_{(i,j)\in\mathcal{IJ}} \Big)$ tel que $\forall i \in \mathcal{I}_\Omega$, et

$\forall j \in \mathcal{I}_{\Omega_i}$:

$$\left\{\begin{array}{rcl}
\forall\, l \in \mathcal{I}_i^T & : & \displaystyle\sum_{l' \in \mathcal{I}_i^T} U_{i,l'} \int_{\Omega_i} q_i\, \underline{w}_{i,l}\, \underline{w}_{i,l'}\, d\boldsymbol{x} + \sum_{m' \in \mathcal{I}_i^{a+}} P_{i,m'} \int_{\Omega_i} \underline{w}_{i,l}\, \mathrm{div}\, \underline{\boldsymbol{\omega}}_{i,m'}\, d\boldsymbol{x} \\
& & \qquad = (f_i, \underline{w}_{i,l})_{L^2(\Omega_i)}, \\
\forall\, m \in \mathcal{I}_i^a & : & -\displaystyle\sum_{m' \in \mathcal{I}_i^{a+}} P_{i,m'} \int_{\Omega_i} k_i^{-1} \underline{\boldsymbol{\omega}}_{i,m} \cdot \underline{\boldsymbol{\omega}}_{i,m'}\, d\boldsymbol{x} \\
& & \qquad + \displaystyle\sum_{l' \in \mathcal{I}_i^T} U_{i,l'} \int_{\Omega_i} \mathrm{div}\, \underline{\boldsymbol{\omega}}_{i,m}\ \underline{w}_{i,l'}\, d\boldsymbol{x} = 0, \\
\forall\, m \in \mathcal{I}_{i,\Sigma_i}^{ij,a} & : & -\displaystyle\sum_{m' \in \mathcal{I}_i^{a+}} P_{i,m'} \int_{\Omega_i} k_i^{-1} \underline{\boldsymbol{\omega}}_{i,m} \cdot \underline{\boldsymbol{\omega}}_{i,m'}\, d\boldsymbol{x} \\
& & + \displaystyle\sum_{l' \in \mathcal{I}_i^T} U_{i,l'} \int_{\Omega_i} \mathrm{div}\, \underline{\boldsymbol{\omega}}_{i,m}\ \underline{w}_{i,l'}\, d\boldsymbol{x} + \sum_{m' \in \mathcal{I}_\lambda^{ij,a}} \lambda_{ij,m'} \int_{\Sigma_S} \underline{\boldsymbol{\omega}}_{i,m} \cdot \boldsymbol{n}_i\, \mu_{m'}\, ds = 0,
\end{array}\right. \tag{13.9}$$

et $\forall (i,j) \in \mathcal{IJ}$, $\forall\, m \in \mathcal{I}_\lambda^{ij,a}$:

$$\sum_{m' \in \mathcal{I}_i^{a+}} P_{i,m'} \int_{\Sigma_S} \mu_m\, \underline{\boldsymbol{\omega}}_{i,m'} \cdot \boldsymbol{n}_i\, ds + \sum_{m' \in \mathcal{I}_j^{a+}} P_{j,m'} \int_{\Sigma_S} \mu_m\, \underline{\boldsymbol{\omega}}_{j,m'} \cdot \boldsymbol{n}_j\, ds = 0. \tag{13.10}$$

On peut mettre les équations (13.9)-(13.10) sous forme matricielle. On reprend les définitions des matrices $\mathbb{M}_u^i$, $\mathbb{M}_{\boldsymbol{p}}^{i,i}$, $\mathbb{B}_{i,i}$ données au chapitre 11. Pour $i \in \mathcal{I}_\Omega$, $j \in \mathcal{I}_{\Omega_i}$, on appelle $\mathbb{C}_{i,ij} \in \mathbb{R}^{A_i^+ \times A_{\Sigma_{ij}}^i}$ la matrice de couplage entre $\boldsymbol{p}_{i,h}$ et $\lambda_{ij,h}$, telle que :

$$\forall\, (m,m') \in \mathcal{I}_i^{a+} \times \mathcal{I}_\lambda^{ij,a} \ : \quad (\mathbb{C}_{i,ij})_{m,m'} = \left\{\begin{array}{rl} 0 & \text{si} \quad m \in \mathcal{I}_i^a, \\ \displaystyle\int_{\Sigma_{ij}} \underline{\boldsymbol{\omega}}_{i,m} \cdot \boldsymbol{n}_i\, \mu_{m'}\, ds & \text{si} \quad m \in \mathcal{I}_{i,\Sigma_i}^{ij,a}. \end{array}\right.$$

Le système linéaire correspondant à (13.9)-(13.10) se met sous la forme :
Trouver $((U_i, P_i)_{i \in \mathcal{I}_\Omega}, (\Lambda_{ij})_{(i,j) \in \mathcal{IJ}})$ tel que :

$$\left\{\begin{array}{lrcl}
\forall i \in \mathcal{I}_\Omega, & \mathbb{M}_u^i U_i + \mathbb{B}_{i,i}^T P_i & = & F_i, \\
\forall i \in \mathcal{I}_\Omega, & \mathbb{B}_{i,i} U_i - \mathbb{M}_{\boldsymbol{p}}^i P_i + \displaystyle\sum_{j \in \mathcal{I}_{\Omega_i}} \mathbb{C}_{i,ij} \Lambda_{ij} & = & 0, \\
\forall (i,j) \in \mathcal{IJ} & \mathbb{C}_{i,ij}^T U_i + \mathbb{C}_{j,ij}^T U_j & = & 0.
\end{array}\right. \tag{13.11}$$

Les matrices $\mathbb{M}_u^i$, $\mathbb{B}_{i,i}$ et $\mathbb{M}_{\boldsymbol{p}}^i$ (pour $i \in \mathcal{I}_\Omega$) ont été introduites au chapitre 11. On rappelle que les matrices $\mathbb{M}_u^i$ sont diagonales, et inversibles.
En utilisant la première équation de (13.11), on obtient : $U_i = (\mathbb{M}_u^i)^{-1}(F_i - \mathbb{B}_{i,i}^T P_i)$. En utilisant ce résultat dans la seconde équation de (13.11), on a :

$$\mathbb{S}_{\boldsymbol{p}}^i P_i - \sum_{j \in \mathcal{I}_{\Omega_i}} \mathbb{C}_{i,ij} \Lambda_{ij} = \widetilde{F}_i, \text{ avec : } \left\{\begin{array}{rcl} \mathbb{S}_{\boldsymbol{p}}^i & := & \mathbb{B}_{i,i} (\mathbb{M}_u^i)^{-1} \mathbb{B}_{i,i}^T + \mathbb{M}_{\boldsymbol{p}}^i, \\ \widetilde{F}_i & := & \mathbb{B}_{i,i} (\mathbb{M}_u^i)^{-1} F_i. \end{array}\right.$$

Le complément de Schur $\mathbb{S}_{\boldsymbol{p}}^i$ est une matrice symétrique définie-positive car $(\mathbb{M}_u^i)^{-1}$ et $\mathbb{M}_{\boldsymbol{p}}^i$ sont symétriques définies-positives.

On peut réécrire (13.11) ainsi :
Trouver $((P_i)_{i\in\mathcal{I}_\Omega}, (\Lambda_{ij})_{(i,j)\in\mathcal{IJ}})$ tel que :

$$\left\{\begin{array}{lll} \forall i \in \mathcal{I}_\Omega, & \mathbb{S}^i_{\boldsymbol{p}} P_i + \sum_{j\in\mathcal{I}_{\Omega_i}} \mathbb{C}_{i,ij}\Lambda_{ij} &= \widetilde{F}_i \\ \forall (i,j) \in \mathcal{IJ}, & \mathbb{C}^T_{i,ij} P_i + \mathbb{C}^T_{j,ij} P_j &= 0. \end{array}\right. \tag{13.12}$$

De même que pour le cas à une inconnue, on peut résoudre ce système linéaire en calculant d'abord $(\Lambda_{ij})_{(i,j)\in\mathcal{IJ}}$ puis $(P_i)_{i\in\mathcal{I}_\Omega}$.
Pour finir, on en déduit $(U_i)_{i\in\mathcal{I}_\Omega}$ à l'aide de

$$\forall i \in \mathcal{I}_\Omega, \quad U_i = (\mathbb{M}^i_u)^{-1} \left(F_i - \mathbb{B}^T_{i,i} P_i\right),$$

ce qui est peu coûteux car les matrices $\mathbb{M}^i_u$ sont diagonales et inversibles.

Quatrième partie

Annexes

Annexe A

Valeurs propres et vecteurs propres

A.1 Introduction

Dans ce chapitre, on donne quelques résultats théoriques fondamentaux qui sont utilisés pour construire des algorithmes de calcul des éléments propres des matrices. La présentation de la forme de Jordan, puis de la décomposition spectrale d'une matrice permet d'établir une distinction entre matrices diagonalisables et matrices défectives.

A.2 Rappels

Avant de commencer l'étude des propriétés spectrales des matrices il faut rappeler quelques notions utiles : tout d'abord il est nécessaire de se placer dans le corps $\mathbb{C}$ des nombres complexes, car les valeurs propres et vecteurs propres d'une matrice à éléments réels peuvent être complexes. A l'exception de certains cas particuliers, tous les calculs présentés dans ce chapitre sont donc effectués avec des nombres complexes.

Soit $B = \{b_1, b_2, \ldots, b_n\}$ une base de $\mathbb{C}^n$. On peut écrire tout vecteur x sous la forme $x = x_1 b_1 + x_2 b_2 + \cdots + x_n b_n$, avec $(x_i)_{i=1,n}$ les n coordonnées de x dans la base B. On peut alors définir le produit scalaire complexe associé [51] de $\mathbb{C}^n$

$$(u, v) = v^* u = \sum_{i=1}^{n} u_i \, \overline{v_i},$$

$\overline{x}$ désignant le complexe conjugué de x. Si besoin, on note ce produit scalaire $(\cdot, \cdot)_{\mathbb{C}^n}$.

51. Dans un $\mathbb{C}$-espace vectoriel V, le produit scalaire complexe $(\cdot, \cdot)_V$ possède les propriétés suivantes :
— Il est *linéaire* par rapport à la première variable :
$\forall a_1, a_2 \in \mathbb{C},\ \forall v_1, v_2, w \in V,\ (a_1 v_1 + a_2 v_2, w)_V = a_1 (v_1, w)_V + a_2 (v_2, w)_V$.
— Il est *anti-linéaire* par rapport à la deuxième variable :
$\forall a_1, a_2 \in \mathbb{C},\ \forall v, w_1, w_2 \in V,\ (v, a_1 w_1 + a_2 w_2)_V = \overline{a_1} (v, w_1)_V + \overline{a_2} (v, w_2)_V$.
— Il est *hermitien* :
$\forall v, w \in V,\ (v, w)_V = \overline{(w, v)_V}$.
— Il est *défini-positif* :
$\forall v \in V \setminus \{0\},\ (v, v)_V > 0$.
Dans ce cas, $\|v\|_V : V \to \mathbb{R}$, défini par $\|v\|_V = ((v, v)_V)^{1/2}$, est une norme sur V. De plus, l'inégalité de Cauchy-Schwarz est vérifiée, voir la Proposition B.3 : $\forall v, w \in V,\ |(v, w)_V| \leq \|v\|_V \, \|w\|_V$.

On associe à toute matrice $A \in \mathbb{C}^{n\times m}$, la **matrice adjointe** de A, notée $A^* \in \mathbb{C}^{m\times n}$, définie par

$$\forall u \in \mathbb{C}^m,\, v \in \mathbb{C}^n \qquad (Au, v)_{\mathbb{C}^n} = (u, A^* v)_{\mathbb{C}^m}.$$

Ceci implique que

$$(A^*)_{i,j} = \overline{A_{j,i}} \quad 1 \le i \le m, \quad 1 \le j \le n.$$

De la même façon, si on se place dans $\mathbb{R}^n$ et qu'on choisit une base, on peut définir le produit scalaire associé de $\mathbb{R}^n$ par

$$(u, v) = v^T u = \sum_{i=1}^{n} u_i\, v_i$$

Si besoin, on note ce produit scalaire $(\cdot,\cdot)_{\mathbb{R}^n}$. On associe à toute matrice $A \in \mathbb{R}^{n\times m}$, la **matrice transposée** de A, notée $A^T \in \mathbb{R}^{m\times n}$, définie par

$$\forall u \in \mathbb{R}^m,\, v \in \mathbb{R}^n \qquad (Au, v)_{\mathbb{R}^n} = (u, A^T v)_{\mathbb{R}^m}.$$

Ceci implique que

$$(A^T)_{i,j} = A_{j,i} \quad 1 \le i \le m, \quad 1 \le j \le n.$$

Définition A.1 — *On appelle* **valeur propre** *d'une matrice A, toute racine complexe λ_i, ou $\lambda_i(A)$, du* **polynôme caractéristique** $p(\lambda) = \det(A - \lambda I)$. *A ce titre on associe à chaque valeur propre sa* **multiplicité algébrique** m_i, *qui est l'ordre de multiplicité de λ_i en tant que racine de p. Si on note d le nombre de racines distinctes de p, on peut donc écrire*

$$p(\lambda) = \prod_{i=1}^{d} (\lambda - \lambda_i)^{m_i}.$$

— *On définit aussi la* **valeur propre** λ_i *et un* **vecteur propre** *associé u_i comme un couple (λ_i, u_i) solution du problème $Au_i = \lambda_i u_i$,* avec $u_i \neq 0$, *ce qui peut encore s'exprimer par la relation d'appartenance $u_i \in \mathrm{Ker}\,(A - \lambda_i I) \setminus \{0\}$. On introduit donc naturellement la notion de* **multiplicité géométrique** *de λ_i par $g_i = dim\,(\mathrm{Ker}\,(A - \lambda_i I))$, pour i compris entre 1 et d.*

Proposition A.2 *Les multiplicités $(m_i)_{i=1,d}$ et $(g_i)_{i=1,d}$ vérifient les relations :*

(i) $\sum_{i=1}^{d} m_i = n$;

(ii) $g_i \le m_i$, *pour* $i \in \{1, \cdots, d\}$;

(iii) $\sum_{i=1}^{d} g_i \le n$.

Démonstration : Comme tout polynôme à coefficients complexes est scindé dans $\mathbb{C}$, on en déduit immédiatement la relation (i).

Pour prouver (ii), on introduit, pour i fixé, $B_i = (e_1, \cdots, e_{g_i})$ une base de $\mathrm{Ker}\,(A-\lambda_i I)$. On la complète en une base B' de $\mathbb{C}^n$. La matrice A est alors semblable à la matrice par blocs ci-dessous,

$$A' = \begin{bmatrix} \lambda_i I_{g_i} & X \\ 0 & Y \end{bmatrix},$$

qui représente l'application linéaire asssociée, exprimée cette fois dans la base B'. On a alors

$$p(\lambda) = \det\,(A' - \lambda I_n) = (\lambda_i - \lambda)^{g_i} \det\,(Y - \lambda I_{n-g_i}).$$

Ainsi λ_i est racine de p d'ordre au moins g_i, ce qui prouve (ii).

Pour finir, (iii) est une conséquence immédiate des deux points précédents. ◇

Avant de détailler plus avant la présentation, nous rappelons le résultat bien connu

Proposition A.3 *Soient $(v_k)_k$ d' vecteurs propres de A, associés à des valeurs propres deux à deux distinctes. Alors $(v_k)_k$ est une famille libre de $\mathbb{C}^n$.*

Démonstration : Raisonnons par récurrence sur d'.
Pour $d' = 1$, on note que, par définition des vecteurs propres, $v_1 \neq 0$. Ainsi, (v_1) est bien une famille libre.
Supposons le résultat vrai $d' - 1$. Considérons une famille $(v_k)_k$ de d' vecteurs propres de A, associés à des valeurs propres deux à deux distinctes. Si la famille $(v_k)_k$ était liée, on pourrait par exemple écrire

$$v_1 = \sum_{k=2}^{d'} \alpha_k v_k.$$

Or, par application de A (resp. par multiplication par λ_1), on trouve

$$\lambda_1 v_1 = \sum_{k=2}^{d'} \lambda_k \alpha_k v_k \text{ (resp. } \lambda_1 v_1 = \sum_{k=2}^{d'} \lambda_1 \alpha_k v_k).$$

Par différence, on en déduit que

$$\sum_{k=2}^{d'} (\lambda_k - \lambda_1)\alpha_k v_k = 0.$$

Si on applique l'hypothèse de récurrence, on trouve $(\lambda_k - \lambda_1)\alpha_k = 0$, pour $k \in \{2, \cdots, d'\}$: Comme $\lambda_k \neq \lambda_1$, on a en fait $\alpha_k = 0$, pour $k \in \{2, \cdots, d'\}$, ce qui entraîne $v_1 = 0$ et aboutit à une contradiction. La famille $(v_k)_k$ est donc libre. ◇

Définition A.4 *Une matrice A de $\mathbb{C}^{n\times n}$ est dite* **diagonalisable** *lorsqu'elle est semblable à une matrice diagonale.*

On commence par le résultat ci-dessous.

Proposition A.5 *Une matrice $A \in \mathbb{C}^{n\times n}$ est diagonalisable si, et seulement si, il existe une base de $\mathbb{C}^n$ formée de vecteurs propres de A.*

Démonstration : Si A est diagonalisable, on peut écrire $A = U\Lambda U^{-1}$, avec U inversible et Λ diagonale,

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \lambda_n \end{bmatrix}.$$

On note $(u_i)_{1\leq i\leq n}$ les vecteurs colonnes de U. Comme U est inversible, ils sont linéairement indépendants : ils forment donc une base de $\mathbb{C}^n$. Qui plus est, si on écrit $AU = U\Lambda$ colonne par colonne, on trouve $Au_i = \lambda_i u_i$, pour i variant de 1 à n.

Réciproquement s'il existe n vecteurs propres $u_1, u_2, \cdots, u_n$ linéairement indépendants, alors la matrice

$$U = \begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix} \in \mathbb{C}^{n\times n},$$

est inversible. Des relations $Au_i = \lambda_i u_i$, pour $1 \leq i \leq n$, on tire successivement $AU = U\Lambda$, puis $A = U\Lambda U^{-1}$. ◇

Proposition A.6 *Une matrice A de $\mathbb{C}^{n\times n}$ est diagonalisable si, et seulement si,*

$$\sum_{i=1}^{d} g_i = n.$$

Démonstration : Par définition, si A est diagonalisable, alors elle est semblable à la matrice par blocs

$$A' = \begin{bmatrix} \lambda_1 I_{g_1} & 0 & \dots & 0 \\ 0 & \lambda_2 I_{g_2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \lambda_d I_{g_d} \end{bmatrix},$$

écrite dans la base de vecteurs propres $(e_1^1, \cdots, e_{g_1}^1, e_1^2, \cdots, e_{g_2}^2, \cdots, e_1^d, \cdots, e_{g_d}^d)$. On constate alors que les racines du polynôme caractéristique p sont $\lambda_1, \cdots, \lambda_d$ et que, de plus, λ_i est exactement racine d'ordre g_i. Dans ce cas, $\sum_i g_i = n$.

Réciproquement, soit $(e_i^1, \cdots, e_{g_i}^1)$ une base de $\mathrm{Ker}\,(A - \lambda_i I)$, pour chaque i. Par hypothèse, $(e_k)_k = (e_1^1, \cdots, e_{g_1}^1, e_1^2, \cdots, e_{g_2}^2, \cdots, e_1^d, \cdots, e_{g_d}^d)$ est une famille à n éléments de $\mathbb{C}^n$. Pour prouver que c'est une base, il suffit de vérifier qu'elle est libre. Soit donc $(\alpha_k)_k$ tels que

$$\sum_{k=1}^{n} \alpha_k e_k = 0.$$

On regroupe alors les éléments de chaque $\mathrm{Ker}\,(A - \lambda_i I)$, pour trouver $\sum_{i=1}^{d} v_i = 0$, avec

$$v_i = \sum_{m=1}^{g_i} \alpha_m^i e_m^i \in \mathrm{Ker}\,(A - \lambda_i I).$$

D'après la Proposition A.3, chaque $v_i = 0$. Finalement, comme $(e_i^1, \cdots, e_{g_i}^1)$ est une base de $\mathrm{Ker}\,(A - \lambda_i I)$, on a de plus $\alpha_m^i = 0$, pour $m \in \{1, \cdots, g_i\}$. En conclusion, tous les $(\alpha_k)_k$ sont nuls, et $(e_k)_k$ est une base de $\mathbb{C}^n$ formée de vecteurs propres de A. $\diamond$

Corollaire A.7 *Si toutes les valeurs propres d'une matrice sont distinctes, elle est diagonalisable.*

Démonstration : Dans ce cas, on a $g_i = 1$, pour i variant de 1 à n. Leur somme vaut donc n. $\diamond$

Définition A.8 *Une matrice A de $\mathbb{C}^{n\times n}$ est dite* **défective** *lorsqu'elle n'est pas diagonalisable.*

On introduit ensuite les notions suivantes :

Définition A.9 *L'ensemble des valeurs propres d'une matrice A s'appelle le* **spectre** *de A. On le note $Spe(A)$.*

- *λ_i est dite valeur propre* **simple** *si, et seulement si, $m_i = 1$; sinon λ_i est valeur propre* **multiple**.
- *λ_i valeur propre* **multiple**, *est dite* **semi-simple** *si, et seulement si, $m_i = g_i > 1$; sinon λ_i est valeur propre* **défective** *(on a alors $m_i > g_i$).*

Remarque A.10 *D'après ce que l'on a vu ci-dessous, une matrice A admet au moins une valeur propre défective si, et seulement si, elle est défective.*
Par ailleurs, seule une valeur propre multiple peut être défective, puisque pour une valeur propre simple λ_i, on a $m_i = g_i = 1$!

Enfin pour en terminer avec les définitions, rappelons encore que

Définition A.11 *On dit que $v \in \mathbb{C}^{1\times n} \setminus \{0\}$ est* **vecteur propre à gauche** *de la matrice A, si et seulement si il existe $\mu \in \mathbb{C}$ tel que $v^*A = \mu v^*$. Par cohérence les vecteurs propres usuels sont appelés* **vecteurs propres à droite**.

Un vecteur propre à gauche est un *vecteur ligne* de $\mathbb{C}^{1\times n}$. Un vecteur propre à droite est un *vecteur colonne* de $\mathbb{C}^{n\times 1}$, que l'on identifie à $\mathbb{C}^n$. On ne distingue pas valeur propre à droite de valeur propre à gauche. De fait, ces notions coïncident.

Proposition A.12 *A chaque valeur propre correspond un vecteur propre à droite, et un vecteur propre à gauche.*

Démonstration : 1) Pour commencer, on a la série d'équivalences sur les valeurs propres (à droite)

$$\begin{aligned}\lambda_i \text{ v. p. de } A &\iff \det(A - \lambda_i I_n) = 0 \iff \det(A - \lambda_i I_n)^* = 0 \\ &\iff \det(A^* - \overline{\lambda_i} I_n) = 0 \iff \overline{\lambda_i} \text{ v. p. de } A^*.\end{aligned}$$

2) Ensuite, on vérifie par transposition et passage au complexe conjugué que, pour $u \in \mathbb{C}^n$, $u \neq 0$, $Au = \mu u$ équivaut à $vA^* = \overline{\mu} v$, avec $v = u^* = (\overline{u}_1, \cdots, \overline{u}_n) \in \mathbb{C}^{1\times n} \setminus \{0\}$: λ_i est

valeur propre à droite de A si, et seulement si, $\overline{\lambda_i}$ est valeur propre à gauche de A^*.
3) On conclut que

$$\lambda_i \text{ v. p. à gauche de } A \overset{2)}{\Longleftrightarrow} \overline{\lambda_i} \text{ v. p. à droite de } A^* \overset{1)}{\Longleftrightarrow} \lambda_i = \overline{\overline{\lambda_i}} \text{ v. p. à droite de } A.$$

◇

Proposition A.13 *Soit u_i un vecteur propre à droite de la matrice $A \in \mathbb{C}^{n\times n}$: $Au_i = \lambda_i u_i$, et soit v_j un vecteur propre à gauche de la matrice A : $v_j^* A = \lambda_j v_j^*$. Si $\lambda_i \neq \lambda_j$, alors $v_j^* u_i = 0$.*

Démonstration : On note que, comme $v_j^* \in \mathbb{C}^{1\times n}$ et $u_i \in \mathbb{C}^{n\times 1}$, leur produit $v_j^* u_i$ appartient à $\mathbb{C}$.
Soient λ_i et u_i tels que $Au_i = \lambda_i u_i$, λ_j et v_j tels que $v_j^* A = \lambda_j v_j^*$, alors

$$\left.\begin{array}{r} Au_i = \lambda_i u_i \Longrightarrow v_j^* A u_i = \lambda_i v_j^* u_i \\ A^* v_j = \overline{\lambda}_j v_j \Longrightarrow u_i^* A^* v_j = \overline{\lambda}_j u_i^* v_j \end{array}\right\} \Longrightarrow \lambda_i v_j^* u_i = v_j^* A u_i = (u_i^* A^* v_j)^* = \lambda_j v_j^* u_i,$$

soit finalement $(\lambda_i - \lambda_j) v_j^* u_i = 0$. ◇

Pour conclure ce paragraphe, considérons les deux matrices suivantes :

$$A_1 = \begin{pmatrix} 1 & 2 & -4 \\ 0 & 2 & 2 \\ 0 & 0 & 3 \end{pmatrix} \quad \text{et} \quad A_2 = \begin{pmatrix} 1 & 2 & -4 \\ 0 & 2 & 2 \\ 0 & 0 & 2 \end{pmatrix};$$

elles ne différent que par le dernier élément, mais ont des propriétés spectrales distinctes

$$\mathrm{Spe}(A_1) = \{1, 2, 3\} \quad \text{et} \quad \mathrm{Spe}(A_2) = \{1, 2\}.$$

La matrice A_1 ayant ses valeurs propres réelles distinctes, ses vecteurs propres forment une base de $\mathbb{C}^3$ (ou de $\mathbb{R}^3$). A_1 est donc *diagonalisable* :

$$A_1 = \begin{pmatrix} 1 & 2 & -4 \\ 0 & 2 & 2 \\ 0 & 0 & 3 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix} \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix}^{-1}.$$

La matrice A_2 est *défective* : la valeur propre $\lambda(A_2) = 2$ a une multiplicité algébrique $m = 2$ car le polynôme caractéristique est divisible par $(\lambda-2)^2$. Sa multiplicité géométrique est $g = 1$: en effet tout vecteur propre $v = (v_1, v_2, v_3)^T$ associé à la valeur propre $\lambda = 2$ vérifie nécessairement les relations

$$\begin{array}{ll} v_1 + 2v_2 - 4v_3 & = 2v_1 \\ 2v_2 + 2v_3 & = 2v_2 \\ 2v_3 & = 2v_3 \end{array}$$

on en déduit que $v_3 = 0$ et $v_1 = 2v_2$; le sous–espace propre relatif à la valeur propre $\lambda = 2$ est donc engendré par le vecteur $v = (2, 1, 0)^T$. La matrice A_2 est défective, et on peut seulement écrire

$$A_2 = \begin{pmatrix} 1 & 2 & -4 \\ 0 & 2 & 2 \\ 0 & 0 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 1/2 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{pmatrix} \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 1/2 \end{pmatrix}^{-1}.$$

Cette différence peut s'avérer très importante dans la pratique, en particulier lorsque l'on doit évaluer A_1^k et A_2^k.

A.3 Localisation des valeurs propres

Théorème A.14 *[Gerschgorin–Hadamard] Le spectre de la matrice $A \in \mathbb{C}^{n\times n}$ est contenu dans l'ensemble $\mathcal{D}$ réunion des disques D_i du plan complexe définis par*

$$D_i = \{z \in \mathbb{C},\ |z - A_{i,i}| \leq \sum_{j\neq i} |A_{i,j}|\}.$$

Démonstration : Soit λ une valeur propre de A et u un vecteur propre associé, et soit $|u_i| = \max_j |u_j|$, alors $|u_i| \neq 0$ et

$$\sum_j A_{i,j}\ u_j = \lambda u_i \Longleftrightarrow \lambda - A_{i,i} = \sum_{j\neq i} A_{i,j}\frac{u_j}{u_i} \Longleftrightarrow |\lambda - A_{i,i}| \leq \sum_{j\neq i} |A_{i,j}|.$$

Ainsi, λ appartient au disque D_i de rayon $\sum_{j\neq i} |A_{i,j}|$ centré en $A_{i,i}$. A toute valeur propre λ, on peut ainsi associer un disque D_i, et le spectre de la matrice A est donc contenu dans l'ensemble

$$\mathcal{D} = \cup_{i=1}^n D_i = \cup_{i=1}^n \{z \in \mathbb{C},\ |z - A_{i,i}| \leq \sum_{j\neq i} |A_{i,j}|\}.$$

◇

Ce résultat permet de localiser simplement les valeurs propres d'une matrice dans le plan complexe. Dans le cas d'une matrice non hermitienne N, on peut appliquer le Théorème A.14 aux matrices N et N^* qui ont les mêmes valeurs propres, mais des disques de Gerschgorin différents, ce qui permet une localisation plus précise du spectre de N.

Remarque A.15 *Il existe de nombreuses configurations. Par exemple, soit λ une valeur propre de la matrice irréductible*[52] *$A \in \mathbb{C}^{n\times n}$: on peut montrer que si λ appartient à la frontière de l'ensemble $\mathcal{D}$, alors tous les cercles de Gerschgorin passent nécessairement par λ. En conséquence tout point de $\partial\mathcal{D}$, et tel qu'il existe au moins un cercle ∂D_i qui ne le contienne pas, ne peut correspondre à une valeur propre de la matrice.*

Enfin il existe une variante du Théorème de Gerschgorin-Hadamard, qui permet de mieux localiser les valeurs propres :

Théorème A.16 *Soit une matrice $A \in \mathbb{C}^{n\times n}$ et soient les n disques D_i du plan complexe définis par*

$$D_i = \{z \in \mathbb{C},\ |z - A_{i,i}| \leq \sum_{j\neq i} |A_{i,j}|\}.$$

S'il existe p disques D_i formant un ensemble connexe E, sans intersection avec les $n - p$ disques restant, alors E contient exactement p valeurs propres de la matrice A.

52. Pour rappel, une matrice $A \in \mathbb{C}^{n\times n}$ est **réductible** s'il existe une matrice de permutation P telle que $P^T AP$ soit triangulaire supérieure par blocs. Sinon, on dit que A est **irréductible**.

Démonstration : On écrit A sous la forme $A = Dia + R$ où Dia est la partie diagonale de A, et on définit les n rayons

$$r_i = \sum_{j \neq i} |A_{i,j}| = \sum_j |R_{i,j}| \qquad 1 \leq i \leq n,$$

ainsi que les n disques

$$D_i = \{z \in \mathbb{C},\ |z - Dia_{i,i}| \leq r_i\}.$$

Puis pour tout $\varepsilon \geq 0$ on définit de manière cohérente la matrice $A(\varepsilon) = Dia + \varepsilon R$, et les n disques associés

$$D_i(\varepsilon) = \{z \in \mathbb{C},\ |z - Dia_{i,i}| \leq \varepsilon\ r_i\}.$$

Par définition $A(0) = Dia$, $A(1) = A$ et pour tout i et tout ε inférieur à 1, $D_i(\varepsilon) \subset D_i(1) = D_i$. Par application du Théorème A.14, on sait que le spectre de $A(\varepsilon)$ est contenu dans l'ensemble $\cup_{i=1}^n D_i(\varepsilon)$ pour tout ε. Sans nuire à la généralité on suppose que l'ensemble connexe E est constitué par l'union des p premiers disques : $E = \cup_{i=1}^p D_i$. On définit alors l'ensemble $E(\varepsilon) = \cup_{i=1}^p D_i(\varepsilon)$. L'hypothèse

$$\forall j > p \quad D_j \cap E = \emptyset$$

entraine

$$\forall j > p, \forall \varepsilon \leq 1 \quad D_j(\varepsilon) \cap E(\varepsilon) = \emptyset.$$

Pour $\varepsilon = 0$, chaque disque $D_i(0)$ est réduit à un point et

$$E(0) = \cup_{i=1}^p D_i(0) = \{\lambda_1, \lambda_2, \ldots, \lambda_p\}$$

quand ε tend vers 1, $E(\varepsilon) \subset E$ contient toujours exactement p valeurs propres : $\lambda_1, \lambda_2, \ldots, \lambda_p$, les autres valeurs propres restant dans leurs disques. Cette configuration reste vraie à la limite, puisque les valeurs propres de $A(\varepsilon)$ dépendent continûment de ε. ◇

A.4 Matrices diagonalisables

On se place dans un espace vectoriel sur $\mathbb{C}$.

Définition A.17 *Une matrice $A \in \mathbb{C}^{n \times n}$ est dite* **hermitienne** *lorsque $A^* = A$.*
Une matrice $A \in \mathbb{C}^{n \times n}$ est dite **normale** *lorsque $A^*A = AA^*$.*
Une matrice $U \in \mathbb{C}^{n \times n}$ est dite **unitaire** *lorsque $UU^* = U^*U = I_n$.*

Les mêmes résultats restent valables dans un espace vectoriel sur $\mathbb{R}$, sous réserve que l'on remplace partout

— complexe par réelle ($A \in \mathbb{R}^{n \times n}$),
— adjointe par transposée (* par ^{T}),
— hermitienne par symétrique ($A^T = A$),
— normale par normale ($A^TA = AA^T$),
— unitaire par orthogonale ($OO^T = O^TO = I_n$).

Les résultats s'appliquent donc en particulier aux matrices issues de la discrétisation par différences finies du Laplacien (voir le chapitre 2).

Proposition A.18 *Les valeurs propres d'une matrice hermitienne sont réelles.*

Démonstration : En effet, pour un couple vecteur-valeur propres (u, λ), on a la suite d'égalités,

$$\lambda(u,u) = (Au,u) = (u,A^*u) \overset{A^*=A}{=} (u,Au) = (u,\lambda u) = \overline{\lambda}(u,u).$$

On en déduit que $\lambda \in \mathbb{R}$ puisque $u \neq 0$. ◇

Proposition A.19 *Les vecteurs propres d'une matrice hermitienne correspondant à des valeurs propres distinctes sont orthogonaux.*

Démonstration : En effet, pour deux couples vecteur-valeur propres (u, λ) et (v, μ), on peut écrire,

$$\left.\begin{array}{c}(Au,v) = (\lambda u, v) = \lambda(u,v) \\ (Au,v) \overset{A^*=A}{=} (u,Av) = (u,\mu v) \overset{\overline{\mu}=\mu}{=} \mu(u,v)\end{array}\right\} \Longrightarrow (\lambda-\mu)(u,v) = 0 \Longrightarrow (u,v) = 0 \text{ si } \lambda \neq \mu.$$

◇

Proposition A.20 *Toute matrice hermitienne est diagonalisable. Qui plus est, on peut choisir ses vecteurs propres de sorte qu'ils forment une base orthonormale.*
En d'autres termes, il existe Q unitaire et D diagonale telles que $A = QDQ^*$.

Cette propriété découle d'un résultat plus général

Proposition A.21 *[Forme de Schur] Soit $A \in \mathbb{C}^{n\times n}$ il existe une matrice unitaire Q telle que $T = Q^*AQ$ soit une matrice triangulaire supérieure avec pour éléments diagonaux les valeurs propres de la matrice A.*

Démonstration : La démontration est effectuée par récurrence : la propriété est évidente à l'ordre $n = 1$. Supposons-la vraie jusqu'à l'ordre $n-1$ inclus. Soit $A \in \mathbb{C}^{n\times n}$ et λ une valeur propre de A, u un vecteur propre associé de norme 1 ; d'après le théorème de la base incomplète, il existe une matrice $U \in \mathbb{C}^{n-1\times n-1}$ unitaire ($U^*U = UU^* = I$) telle que la matrice $[u, U] \in \mathbb{C}^{n\times n}$ soit aussi unitaire (et en particulier $U^*u = 0$), car on peut toujours construire une base orthogonale de $\mathbb{C}^n$ dont $u \neq 0$ soit le premier vecteur de base.

Ainsi par construction $U^*u = 0$ et $A[u, U] = [\lambda u, AU]$, soit encore

$$[u,U]^*\ A\ [u,U] = \begin{bmatrix}u^*\\U^*\end{bmatrix}\ [\lambda u, A\ U] \overset{U^*u=0}{=} \begin{bmatrix}\lambda & u^*AU\\0 & U^*AU\end{bmatrix}.$$

Comme $U^*AU \in \mathbb{C}^{n-1\times n-1}$, on peut lui appliquer l'hypothèse de récurrence : il existe $\tilde{Q} \in \mathbb{C}^{n-1\times n-1}$ unitaire telle que $\tilde{Q}^*U^*AU\tilde{Q} = \tilde{T}$; alors

$$[u,U\tilde{Q}]^*\ A\ [u,U\tilde{Q}] = \begin{bmatrix}u^*\\\tilde{Q}^*U^*\end{bmatrix}\ [\lambda u, A\ U\tilde{Q}] = \begin{bmatrix}\lambda & u^*AU\tilde{Q}\\0 & \tilde{T}\end{bmatrix} = T.$$

Enfin puisque $Q = [u, U\tilde{Q}]$ est unitaire, les matrices A et T sont semblables, et possèdent de ce fait les mêmes valeurs propres. Plus précisément, si μ est une valeur propre de U^*AU associée au vecteur propre $v \in \mathbb{C}^{n-1}$, μ est aussi une valeur propre de A puisque

$$U^*AU\ v = \mu v \quad \Longrightarrow A(Uv) = \mu(Uv).$$

On obtient donc la propriété à l'ordre n avec $Q = [u, U]$, et dans cette écriture les termes diagonaux sont bien les valeurs propres de A.

Les vecteurs colonnes de la matrice Q sont appelés **vecteurs de Schur** ; ils vérifient la relation $AQ = QT$. ◇

Dans le cas où la matrice A est hermitienne, la matrice triangulaire supérieure $T = Q^*AQ$ est aussi hermitienne, car elle vérifie $T^* = Q^*A^*Q = Q^*AQ$. Elle est donc diagonale, à éléments réels, ce qui démontre la Proposition A.20.

De plus dans ce cas particulier, la relation $AQ = QT$ avec T diagonale, montre que les vecteurs de Schur sont les vecteurs propres de la matrice hermitienne A. La matrice Q dont les colonnes sont les vecteurs propres de A, est unitaire par construction. On en déduit que les vecteurs propres de A forment une base orthogonale de $\mathbb{C}^n$. Ce résultat étant vrai, que les valeurs propres soient distinctes ou non, constitue donc une extension de la Proposition A.19.

Proposition A.22 *Une matrice $A \in \mathbb{C}^{n\times n}$ est normale si, et seulement si, elle est diagonalisable dans une base orthonormale.*

Démonstration : Supposons d'abord que A soit diagonalisable dans une base orthonormale : il existe Q unitaire et D diagonale telles que $A = QDQ^*$. On a alors la suite d'égalités

$$AA^* = QDQ^*QD^*Q^* \overset{Q^*Q=I_n}{=} QDD^*Q^* \overset{D^*D=DD^*}{=} QD^*DQ^* \overset{Q^*Q=I_n}{=} QD^*Q^*QDQ^* = A^*A.$$

Réciproquement, soit A une matrice normale. On écrit $A = QTQ^*$, avec Q matrice unitaire et T matrice triangulaire supérieure. De l'égalité $A^*A = AA^*$ on tire $T^*T = TT^*$. Mais la matrice T étant triangulaire supérieure, on peut écrire pour tout i

$$\sum_{j\geq i} |T_{i,j}|^2 = (TT^*)_{i,i} = (T^*T)_{i,i} = \sum_{j\leq i} |T_{j,i}|^2.$$

Pour $i = 1$ on trouve donc que

$$\sum_{j\geq 1} |T_{1,j}|^2 = (TT^*)_{1,1} = (T^*T)_{1,1} = |T_{1,1}|^2,$$

ce qui entraine que tous les éléments extra-diagonaux $T_{1,j}$ sont nuls. En appliquant le même raisonnement à la ligne $i = 2$, on trouve

$$\sum_{j\geq 2} |T_{2,j}|^2 = (TT^*)_{2,2} = (T^*T)_{2,2} = |T_{1,2}|^2 + |T_{2,2}|^2 = |T_{2,2}|^2,$$

puisque $T_{1,2} = 0$. Tous les éléments extra-diagonaux $T_{2,j}$ sont nuls... En réitérant ce procédé, on montre que la matrice T est diagonale. ◇

Remarque A.23 *Des relations $AQ = QD$ et $A^*Q = QD^*$ on déduit que si A est normale, alors les matrices A et A^* admettent la même base de vecteurs propres, qui sont les vecteurs colonnes de la matrice Q.*

Il s'agit bien d'une *généralisation* des résultats concernant les matrices hermitiennes, car la matrice A définie par ($\imath^2 = -1$)

$$A = \begin{bmatrix} \imath & 0 & \dots & 0 \\ 0 & \imath & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \imath \end{bmatrix}$$

est diagonalisable (car normale) *sans être hermitienne.*

Un exemple moins trivial est fourni par la matrice de permutation $P \in \mathbb{R}^{n\times n}$ de rang n

$$P = \begin{bmatrix} 0 & \dots & \dots & 0 & 1 \\ 1 & 0 & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 & 0 \end{bmatrix}$$

qui est normale, donc diagonalisable toujours sans être hermitienne :

$$Q^*PQ = \Lambda.$$

Les matrices Q et Λ sont définies en posant $z = e^{\imath\pi/n}$:

$$Q = \frac{1}{\sqrt{n}} \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & \overline{z} & \overline{z}^2 & \vdots & \overline{z}^{(n-1)} \\ 1 & \overline{z}^2 & \overline{z}^4 & \vdots & \overline{z}^{2(n-1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \overline{z}^{n-1} & \overline{z}^{2(n-1)} & \dots & \overline{z}^{(n-1)(n-1)} \end{bmatrix} \quad \text{et} \quad \Lambda = \begin{bmatrix} 1 & 0 & \dots & \dots & 0 \\ 0 & z & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & z^{(n-2)} & 0 \\ 0 & \dots & \dots & 0 & z^{(n-1)} \end{bmatrix}.$$

Enfin, toute matrice A diagonalisable n'est pas nécessairement normale [53].

Complétons ce paragraphe par l'énoncé de quelques résultats utiles sur les matrices hermitiennes. Pour une matrice A et $x \neq 0$, on appelle $\rho_A(x) = \dfrac{(Ax, x)}{(x, x)}$ le **quotient de Rayleigh** de la matrice A. L'ensemble

$$F(A) = \{\rho_A(x), x \in \mathbb{C}^n, x \neq 0\}$$

est appelé le **champ des valeurs** de la matrice A. Le champ des valeurs de A contient le spectre de A.

53. La matrice $A = \begin{pmatrix} 0 & -1 \\ 2 & 3 \end{pmatrix}$ est diagonalisable, mais $AA^* \neq A^*A$

Théorème A.24 (Courant–Fisher) *Soit $A \in \mathbb{C}^{n\times n}$ une matrice hermitienne dont les valeurs propres (réelles) sont rangées suivant*

$$\lambda_n \leq \lambda_{n-1} \leq \ldots \leq \lambda_2 \leq \lambda_1$$

alors

$$(i) \quad \lambda_k = \max_{V,\, dim V=k} \; \min_{x\in V-\{0\}} \frac{(Ax,x)}{(x,x)},$$

$$(ii) \quad \lambda_k = \min_{W,\, dim W=n-k+1} \; \max_{x\in W-\{0\}} \frac{(Ax,x)}{(x,x)}.$$

Démonstration : Pour démontrer le premier point du théorème, on considère le sous–espace engendré par les vecteurs propres u_i associés aux $n-k+1$ valeurs propres λ_i ($k \leq i \leq n$). Soit $V \subset \mathbb{C}^n$ un sous–espace quelconque de dimension k : $W_k = vect(u_n, u_{n-1}, \ldots, u_k)$; puisque $\dim W_k = n-k+1$, $W_k \cap V \neq \{0\}$, il existe donc au moins un vecteur commun non nul $x \in W_k \cap V$, que l'on écrit $x = \sum_{i=k}^{n} \alpha_i u_i$. Alors

$$\rho_A(x) = \frac{(Ax,x)}{(x,x)} = \frac{\sum_{i=k}^{n} \lambda_i \alpha_i^2 (u_i,u_i)}{\sum_{i=k}^{n} \alpha_i^2 (u_i,u_i)} \leq \lambda_k.$$

Par conséquent $m(V) = \min_{x\in V-\{0\}} \rho_A(x) \leq \lambda_k$, et on en déduit que le maximum de $m(V)$ sur tous les sous-espaces V de dimension k est plus petit que λ_k. Si on prend en particulier $V = vect(u_1, u_2, \cdots, u_k)$, alors $\dim V = k$ et $m(V)$ atteint la valeur maximale λ_k pour $x = u_k \in V$.

On procède de même pour le second point du théorème, en introduisant cette fois $V_k = vect(u_1, u_2, \ldots, u_k)$ le sous–espace de dimension k. Alors pour tout sous–espace W de dimension $n-k+1$, $W \cap V_k \neq \{0\}$ et par le même raisonnement, on en déduit que $M(V) = \max_{x\in W-\{0\}} \frac{(Ax,x)}{(x,x)} \geq \lambda_k$, puis que

$$\min_{\substack{W \\ \dim W=n-k+1}} \; \max_{x\in W-\{0\}} \frac{(Ax,x)}{(x,x)} \geq \lambda_k.$$

La valeur minimale λ_k est atteinte en prenant pour sous–espace $W =< u_n, u_{n-1}, \ldots, u_k >$ et pour vecteur $x = u_k$. ◇

Théorème A.25 *Soit $B = A + E$ la somme de deux matrices hermitiennes de $\mathbb{C}^{n\times n}$, on range les valeurs propres par ordre croissant :*

$$A \; : \quad \lambda_n \leq \lambda_{n-1} \leq \ldots \leq \lambda_2 \leq \lambda_1$$

$$E \; : \quad \varepsilon_n \leq \varepsilon_{n-1} \leq \ldots \leq \varepsilon_2 \leq \varepsilon_1$$

$$B \ : \quad \mu_n \leq \mu_{n-1} \leq \ldots \leq \mu_2 \leq \mu_1.$$

Alors pour tout $k = 1, 2, \ldots, n$

$$\begin{array}{ll} (i) & \lambda_k + \varepsilon_n \leq \mu_k \leq \lambda_k + \varepsilon_1 \\ (ii) & |\mu_k - \lambda_k \ | \leq \|E\|. \end{array}$$

Démonstration : Soit $u_1,\ u_2,\ \ldots, u_n$ une base orthonormée des vecteurs propres de la matrice A, et μ_k une valeur propre de B : on pose $W_k = vect(u_n, u_{n-1}, \ldots, u_k)$, d'après le théorème de Courant–Fisher

$$\mu_k \leq \max_{x \in W_k - \{0\}} \rho_B(x) \leq \max_{x \in W_k - \{0\}} \rho_A(x) + \max_{x \in W_k - \{0\}} \rho_E(x)$$

soit encore $\mu_k \leq \lambda_k + \varepsilon_1$.

Pour la minoration, on écrit $A = B - E = B + E'$ et le résultat précédent appliqué à la matrice $B + E'$ devient $\lambda_k \leq \mu_k - \varepsilon_n$.

Finalement, pour tout $k = 1, 2, \ldots, n$, $\lambda_k + \varepsilon_n \leq \mu_k \leq \lambda_k + \varepsilon_1$, soit encore $\varepsilon_n \leq \mu_k - \lambda_k \leq \varepsilon_1$. On en déduit (ii) puisque pour toute matrice E et toute norme matricielle $\|E\|$, $|\varepsilon_k| \leq \|E\| \quad \forall k = 1, 2, \ldots, n$. ◇

Ce résultat semble prometteur sur le plan numérique, car il montre que la recherche des valeurs propres d'une matrice A hermitienne est théoriquement stable. Les valeurs propres $\lambda(A)$ dépendent continûment des éléments de A de la manière suivante : si on pose $A_{\mathcal{E}} = A + \mathcal{E}$ avec $\mathcal{E} \in \mathbb{C}^{n \times n}$ matrice de perturbation **hermitienne**, alors

$$\max_{\lambda} |\lambda(A_{\mathcal{E}}) - \lambda(A)| = \max_{\lambda} |\lambda(\mathcal{E})| \|\mathcal{E}\|_2 \leq \|\mathcal{E}\|_F.$$

Cette majoration donne une borne maximale de variation des valeurs propres de A en fonction des éléments de $\mathcal{E}$. Malheureusement dans la pratique, les erreurs commises sur les éléments de la matrice A sont dues soit à la représentation machine des nombres (erreur de troncature ou d'arrondi), soit aux erreurs de calcul qui en découlent. En conséquence, bien qu'il soit souvent possible d'estimer $\|\mathcal{E}\|_F$, le Théorème A.25 n'est pas utilisable pour le calcul numérique car la matrice $\mathcal{E}$ n'est jamais hermitienne !

Pour finir ce paragraphe, nous introduisons la décomposition spectrale d'une matrice diagonalisable. Pour cela, on se souvient que, par définition, lorsqu'une matrice A est diagonalisable, il existe U inversible et Λ diagonale, telles que $A = U \Lambda U^{-1}$. On a vu à la Proposition A.5 que les vecteurs colonnes de U sont des vecteurs propres à droite de A. De la même façon, les vecteurs ligne de U^{-1}, appartenant à $\mathbb{C}^{1 \times n}$ et notés $(v_i^*)_{1 \leq i \leq n}$,

$$U^{-1} = \begin{bmatrix} v_1^* \\ v_2^* \\ \vdots \\ v_n^* \end{bmatrix},$$

sont des vecteurs propres à gauche de A. En effet, la relation $U^{-1} A = \Lambda U^{-1}$ peut s'écrire, ligne par ligne, sous la forme $v_i^* A = \lambda_i v_i^*$, pour i variant de 1 à n. En particulier, les ordres

de multiplicité de λ_i à gauche et à droite sont *identiques*.

Au final, on peut résumer la relation $A = U\Lambda U^{-1}$ dans les formules

$$A = \begin{bmatrix} u_1 & u_2 & \dots & u_n \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \lambda_n \end{bmatrix} \begin{bmatrix} v_1^* \\ v_2^* \\ \dots \\ v_n^* \end{bmatrix}, \text{ ou bien } A = \sum_{i=1}^{n} \lambda_i \, u_i \cdot v_i^*,$$

en se souvenant que $u_i \cdot v_i^*$ est une matrice de $\mathbb{C}^{n\times n}$. Si on regroupe les $u_i \cdot v_i^*$ correspondant à une même valeur propre λ_k, on peut écrire A sous la forme

$$A = \sum_{k=1}^{d} \lambda_k P_k, \text{ avec } P_k = \sum_{i,\, \lambda_i = \lambda_k} u_i \cdot v_i^*. \tag{A.1}$$

Cette relation est appelée **décomposition spectrale** de la matrice diagonalisable A. Par construction, $d \leq n$. Qui plus est, les $(P_k)_{1\leq k\leq d}$, appelées les **projections spectrales**, vérifient les relations suivantes.

Proposition A.26 *Pour k variant de 1 à d, P_k est la matrice d'un projecteur sur le sous-espace propre* $\mathrm{Ker}\,(A - \lambda_k I_n)$. *De plus, la somme de ces projecteurs est égale à l'identité. En d'autres termes :*

(i) pour $1 \leq i \leq n$, $1 \leq k \leq d$: $P_k u_i = u_i$ si $\lambda_i = \lambda_k$, $P_k u_i = 0$ sinon ;
(ii) $\sum_{1\leq k\leq d} P_k = I_n$;
(iii) pour $1 \leq k \leq d$: $P_k^2 = P_k$;
(iii) pour $1 \leq k \neq l \leq d$: $P_k P_l = 0$.

Démonstration : La démonstration est relativement aisée, sous réserve que l'on se souvienne de la définition des vecteurs $(u_i)_i$ et $(v_i^*)_i$. En effet, de la relation $U^{-1}U = I_n$, on tire immédiatement

$$v_i^* u_j = \delta_{ij}, \text{ pour } 1 \leq i, j \leq n.$$

Pour prouver (i), on écrit simplement

$$P_k u_i = \sum_{j,\, \lambda_j = \lambda_k} (u_j \cdot v_j^*) u_i = \sum_{j,\, \lambda_j = \lambda_k} u_j (v_j^* u_i) = \sum_{j,\, \lambda_j = \lambda_k} \delta_{ij} u_j = \begin{cases} u_i & \text{si } i \in \{j,\, \lambda_j = \lambda_k\} \\ 0 & \text{sinon} \end{cases}.$$

Les points (ii) et (iii) sont des conséquences simples de (i) car, $(u_i)_i$ étant une base de $\mathbb{C}^n$, les P_k sont complètement déterminées par leur action sur ceux-ci. ◇

Pour conclure le cas des matrices diagonalisables, on déduit de (A.1) et de la Proposition A.26 que

$$AP_k = P_k A = \lambda_k P_k, \text{ pour } 1 \leq k \leq d.$$

A.5 Matrices défectives et forme de Jordan

Par définition, les matrices défectives ne sont pas diagonalisables. Il faut donc construire d'autres vecteurs associés aux vecteurs propres, appelés **vecteurs principaux**. Cette construction conduit à la **forme de Jordan** dans laquelle la matrice est écrite sous une forme "presque diagonale" (voir Chatelin [7]).

Théorème A.27 *Soit $A \in \mathbb{C}^{n\times n}$ une matrice admettant d valeurs propres distinctes λ_i de multiplicité algébrique m_i et de multiplicité géométrique g_i ($g_i \leq m_i$). Il existe une matrice $X \in \mathbb{C}^{n\times n}$ telle que*

$$A = X \begin{bmatrix} J_1 & 0 & \dots & 0 \\ 0 & J_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & J_d \end{bmatrix} X^{-1}.$$

Dans cette écriture $J_i \in \mathbb{C}^{m_i\times m_i}$ est la boîte de Jordan *associée à la valeur propre λ_i ; chaque boîte de Jordan se décompose elle-même en une matrice diagonale par blocs $g_i \times g_i$, dont les blocs diagonaux $J_{i,j}$ sont appelés* blocs de Jordan *:*

$$J_i = \begin{bmatrix} J_{i,1} & 0 & \dots & 0 \\ 0 & J_{i,2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & J_{i,g_i} \end{bmatrix} \text{ avec } J_{i,j} = [\lambda_i] \text{ ou } \begin{bmatrix} \lambda_i & 1 & \dots & \dots & 0 \\ 0 & \lambda_i & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \lambda_i & 1 \\ 0 & \dots & \dots & \dots & \lambda_i \end{bmatrix}.$$

Remarque A.28 *Pour toute valeur propre* semi-simple *λ_i, comme on a $m_i = g_i$ blocs diagonaux $J_{i,j}$ sur la diagonale de J_i, une matrice de $\mathbb{C}^{m_i\times m_i}$, on a nécessairement $J_{i,j} = [\lambda_i]$ pour $1 \leq j \leq g_i$, et J_i est une matrice* diagonale *égale à $J_i = \lambda_i I_{m_i}$. On retrouve donc la définition* non défective équivaut à diagonalisable*, dans le sous-espace* $\mathrm{Ker}\,(A - \lambda_i I_n)$.

La démonstration de ce Théorème est effectuée par étapes et requiert deux résultats intermédiaires que nous admettons.

Proposition A.29 *Soit $R \in \mathbb{C}^{n\times n}$ une matrice triangulaire supérieure admettant d valeurs propres distinctes λ_i, alors il existe une matrice $Z \in \mathbb{C}^{n\times n}$ telle que*

$$R = Z^{-1} \begin{bmatrix} \tilde{R}_1 & 0 & \dots & 0 \\ 0 & \tilde{R}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \tilde{R}_d \end{bmatrix} Z,$$

où $\tilde{R}_i = \lambda_i I + U_i$, avec U_i une matrice strictement triangulaire supérieure ($U_{i,j} = 0$ si $j \leq i$).

Proposition A.30 *Soit $U \in \mathbb{C}^{n\times n}$ une matrice strictement triangulaire supérieure. Alors il existe une matrice inversible Y et g matrices $E_j \in \mathbb{C}^{k_j\times k_j}$ avec $k_1 \geq k_2 \geq \ldots \geq k_g \geq 1$ telles que*

$$Y^{-1}UY = \begin{bmatrix} E_1 & 0 & \ldots & 0 \\ 0 & E_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & E_g \end{bmatrix} \quad \textit{avec } E_j = [0] \textit{ ou } \begin{bmatrix} 0 & I_{k_j-1} \\ 0 & 0 \end{bmatrix}.$$

Démonstration : (du Théorème A.27) La forme de Jordan d'une matrice $A \in \mathbb{C}^{n\times n}$ est obtenue de la façon suivante :

— On commence par mettre A sous forme triangulaire supérieure (forme de Schur de la Proposition A.21) $Q^*AQ = R$.

— On applique le résultat de la Proposition A.29 à la matrice R, et on obtient la matrice

$$\tilde{R} = \begin{bmatrix} \tilde{R}_1 & 0 & \ldots & 0 \\ 0 & \tilde{R}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \tilde{R}_d \end{bmatrix}.$$

Les d blocs diagonaux $\tilde{R}_i$ correspondent aux d valeurs propres distinctes de R qui est semblable à A par construction.

— On applique ensuite, pour $j = 1, 2, \ldots, g_i$, le résultat de la Proposition A.30 à chaque bloc $U_i = \lambda_i I - \tilde{R}_i$:

$$Y_i^{-1}\left(\lambda_i I + U_i\right) Y_i = \lambda_i I + \begin{bmatrix} E_{i,1} & 0 & \ldots & 0 \\ 0 & E_{i,2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & E_{i,g_i} \end{bmatrix}, \; E_{i,j} = [0] \text{ ou } \begin{bmatrix} 0 & 1 & \ldots & \ldots & 0 \\ 0 & 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 & 1 \\ 0 & \ldots & \ldots & \ldots & 0 \end{bmatrix}$$

et les blocs $E_{i,j}$ sont rangés par ordre de rang croissant.

— On pose maintenant

$$\tilde{Y} = \begin{bmatrix} Y_1 & 0 & \ldots & 0 \\ 0 & Y_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & Y_d \end{bmatrix},$$

et on obtient

$$\left(\tilde{Y}^{-1}Z^{-1}Q^*\right) A \left(QZ\tilde{Y}\right) = J = \begin{bmatrix} J_1 & 0 & \ldots & 0 \\ 0 & J_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & J_d \end{bmatrix}.$$

Dans cette écriture, J_i est la **boîte de Jordan** associée à λ_i :

$$J_i = \begin{bmatrix} J_{i,1} & 0 & \dots & 0 \\ 0 & J_{i,2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & J_{i,g_i} \end{bmatrix} \text{ avec } J_{i,j} = [\lambda_i] \text{ ou } \begin{bmatrix} \lambda_i & 1 & \dots & \dots & 0 \\ 0 & \lambda_i & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \lambda_i & 1 \\ 0 & \dots & \dots & \dots & \lambda_i \end{bmatrix}.$$

$\diamond$

Proposition A.31 *Pour $k \geq 2$, la matrice $E_k \in \mathbb{C}^{k\times k}$ définie par*

$$E_k = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 1 \\ 0 & \dots & \dots & \dots & 0 \end{bmatrix}.$$

possède les propriétés suivantes

$$\begin{array}{ll} (i) & E_k^k = [0] \\ (ii) & E_k^* E_k = \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & I_{k-1} \end{bmatrix} \\ (iii) & I_k - E_k^* E_k^k = e_1 \cdot e_1^* \\ (iv) & E_k e_{j+1} = e_j, \quad j = 1, 2, \dots, k-1 \end{array}$$

où e_j est le $j^{ième}$ vecteur de base de $\mathbb{C}^k$.

Dans l'énoncé du Théorème A.27, les matrices A et J sont semblables et les λ_i sont les d valeurs propres distinctes de A. De plus par construction, le rang du bloc J_i est égal à la multiplicité algébrique m_i de λ_i dans J, donc dans A. Ainsi si on note M_i le sous–espace de $\mathbb{C}^n$ associé au bloc J_i, comme $\dim M_i = m_i$ et $\sum_{i=1}^d m_i = n$, on a

$$\bigoplus_{i=1,d} M_i = \mathbb{C}^n.$$

A chacun des g_i sous–blocs $J_{i,j}$ de rang $m_{i,j} \geq 1$, est associé un sous–espace $M_{i,j}$ de M_i. Cherchons un vecteur propre u dans ce sous–espace. Il doit vérifier les $m_{i,j}$ relations :

$$\lambda_i u_1 + u_2 = \lambda_i u_1,\ \lambda_i u_2 + u_3 = \lambda_i u_2,\ \cdots,\ \lambda_i u_{m_{i,j}-1} + u_{m_{i,j}} = \lambda_i u_{m_{i,j}-1},\ \lambda_i u_{m_{i,j}} = \lambda_i u_{m_{i,j}}.$$

On en déduit que nécessairement $u_2 = u_3 = \ldots = u_{m_{i,j}} = 0$! Le seul vecteur propre possible dans $M_{i,j}$ s'écrit donc $u = (1, 0, \dots, 0)^T$. Or, il ne peut y avoir que g_i vecteurs propres linéairement indépendants dans M_i (autant que de sous–espaces $M_{i,j}$) et g_i correspond

donc bien à la *multiplicité géométrique* de λ_i.

Soit maintenant $z_0 \in \mathrm{Ker}\,(A-\lambda_i I)$ un vecteur propre de A, existe-t-il un vecteur $z_1 \neq 0$ tel que $(A-\lambda_i I)z_1 = z_0$? Un tel vecteur satisfait nécessairement la relation

$$(A-\lambda_i I)^2 z_1 = (A-\lambda_i I)z_0 = 0 \quad \text{soit} \quad z_1 \in \mathrm{Ker}\,(A-\lambda_i I)^2.$$

On voit donc que pour que z_1 existe, il faut et il suffit que $\mathrm{Ker}\,(A-\lambda_i I)^2 \neq \{0\}$. On peut poursuivre en définissant une suite de vecteurs z_k par

$$(A-\lambda_i I)z_k = z_{k-1},$$

et l'on doit chercher z_k dans $\mathrm{Ker}\,(A-\lambda_i I)^{k+1}$. Mais puisque

$$\mathrm{Ker}\,(A-\lambda_i I) \subset \mathrm{Ker}\,(A-\lambda_i I)^2 \subset \ldots \subset \mathrm{Ker}\,(A-\lambda_i I)^k \subset \ldots \subset \mathbb{C}^n,$$

il existe nécessairement un entier $l_i \leq n$ tel que

$$\mathrm{Ker}\,(A-\lambda_i I)^{l_i} = \mathrm{Ker}\,(A-\lambda_i I)^l \quad \forall l \geq l_i.$$

Cet entier est tel que $\mathrm{Ker}\,(A-\lambda_i I)^{l_i} = M_i$: on l'appelle **indice** de la valeur propre λ_i. Les vecteurs z_k sont appelés **vecteurs principaux** associés à z_0 dans M_i. Ils vérifient les relations

$$Az_0 = \lambda_i z_0,\ Az_1 = \lambda_i z_1 + z_0,\ Az_2 = \lambda_i z_2 + z_1,\ \cdots,\ Az_{l_i} = \lambda_i z_{l_i} + z_{l_{i-1}}.$$

Soit encore

$$A\begin{bmatrix} z_0 & z_1 & \ldots & z_{l_i}\end{bmatrix} = \begin{bmatrix} z_0 & z_1 & \ldots & z_{l_i}\end{bmatrix}\begin{bmatrix} \lambda_i & 1 & \ldots & 0 \\ 0 & \lambda_i & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 0 & \ldots & \ldots & \lambda_i \end{bmatrix}.$$

On reconnaît à droite le bloc de Jordan $J_{i,j}$ associé au vecteur z_0, et ceci montre que le rang de $J_{i,j}$ est inférieur ou égal à l'indice l_i.

On en déduit que la représentation de Jordan peut ne pas être unique, car la décomposition de la boîte J_i en blocs $J_{i,j}$ dépend du choix du vecteur z_0 dans chaque $M_{i,j}$. Par exemple, pour une valeur propre λ_i de multiplicité algébrique $m_i = 7$, de multiplicité géométrique $g_i = 3$ et d'indice $l_i = 3$, on obtient deux formes de Jordan différentes :

$$\left[\begin{array}{c|ccc|ccc} \lambda_i & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & \lambda_i & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda_i & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda_i & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & \lambda_i & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & \lambda_i & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & \lambda_i \end{array}\right] \quad \text{ou} \quad \left[\begin{array}{cc|cc|ccc} \lambda_i & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda_i & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & \lambda_i & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda_i & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & \lambda_i & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & \lambda_i & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & \lambda_i \end{array}\right].$$

Soit encore comment écrire 7 ($= m_i$) comme somme de 3 ($= g_i$) entiers naturels non nuls, chaque entier étant inférieur ou égal à 3 ($= l_i$) :

$$7 = 1 + 3 + 3 = 2 + 2 + 3.$$

Dans le cas général, cette écriture contient le cas A diagonalisable pour lequel $l_i = 1$ et $m_i = g_i$ pour tout i.

A.6 Décomposition spectrale d'une matrice quelconque

D'après le Théorème A.27, on peut écrire toute matrice $A \in \mathbb{C}^{n\times n}$ ayant d valeurs propres distinctes selon $A = XJX^{-1}$, avec

$$X = \begin{bmatrix} X_1 & X_2 & \dots & X_d \end{bmatrix}.$$

Ci-dessus, chaque bloc $X_i \in \mathbb{C}^{n\times m_i}$ correspond aux m_i colonnes de X associées au sous–espace M_i, et on introduit de manière cohérente

$$X^{-1} = \begin{bmatrix} Y_1^* \\ Y_2^* \\ \dots \\ Y_d^* \end{bmatrix}.$$

Les vecteurs colonnes de X_i forment une base de M_i, et les vecteurs lignes $Y_i^* \in \mathbb{C}^{m_i\times n}$ forment une base adjointe. De la relation $X^{-1}X = I_n$, on déduit comme précédemment

$$Y_i^* \cdot X_i = I_{m_i} \text{ et } Y_i^* \cdot X_j = [0] \text{ si } i \neq j.$$

La matrice $P_i = X_i \cdot Y_i^* \in \mathbb{C}^{n\times n}$ est la matrice représentant dans $\mathbb{C}^n$ la projection sur M_i le long de l'ensemble $\{z \in \mathbb{C}^n,\ X_i^* z = 0\} = \bigoplus_{j\neq i} M_j$. On l'appelle **projection spectrale** associée à la valeur propre λ_i. En particulier on vérifie que

$$Y_i^* \cdot X_i = I_{m_i} \Longrightarrow P_i^2 = P_i\,; \quad Y_i^* \cdot X_j = [0] \Longrightarrow P_iP_j = 0,\ i \neq j,$$

à comparer à la Proposition A.26. Finalement on résume ces résultats dans la formule

$$A = \sum_{i=1}^{d} (\lambda_i P_i + D_i)\,,\ \text{avec } J_i = \lambda_i I_{m_i} + N_i,\ D_i = X_i N_i Y_i^* \text{ et } \sum_{i=1}^{d} P_i = I_n,$$

qui est la **décomposition spectrale** d'une matrice quelconque. Une forme des N_i est par exemple, pour $m_i = 7$, $g_i = 3$ et $l_i = 3$,

$$N_i = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

On rappelle que

— le rang de J_i est m_i ;
— le nombre de blocs présents est g_i ;
— le rang maximum d'un bloc est l_i.

On en déduit que si on itère l_i fois, ${N_i}^{l_i} = [0]$ soit ${D_i}^{l_i} = [0]$.

Dans le cas particulier d'une valeur propre λ_i *semi-simple*, on a $m_i = g_i$ et $l_i = 1$. La matrice N_i est nulle, et donc $D_i = [0]$.

Enfin, de la formule générale on déduit que

$$AP_j = \sum_{i=1}^{d} (\lambda_i P_i + D_i) P_j = \lambda_j P_j^2 + D_j P_j = P_j (\lambda_j P_j + D_j)$$

(P_j et D_j commutent par définition), et plus généralement

$$A^k P_j = P_j (\lambda_j P_j + D_j)^k .$$

Proposition A.32 *Soit $v \in \mathbb{C}^n$ un vecteur quelconque, pour toute valeur propre λ_i semi–simple, $P_i v$ est vecteur propre associé à λ_i si, et seulement si, $P_i v \neq 0$.*

Démonstration : La démonstration découle de l'étude précédente puisque pour tout vecteur $v \in \mathbb{C}^n$, le vecteur $P_i v$ appartient au sous–espace $M_i = \mathrm{Ker}\,(A - \lambda_i)^{l_i}$. Donc

$$AP_i v = \lambda_i P_i^2 v + D_i P_i v = \lambda_i P_i v + D_i P_i v.$$

Mais dire que λ_i est semi–simple est équivalent à $D_i = [0]$, soit

$$AP_i v = \lambda_i P_i v.$$

◇

Cette forme générale de la décomposition spectrale d'une matrice est utilisée au chapitre 8 pour étudier la convergence d'algorithmes de calcul de valeurs propres. Le lecteur intéressé par cet aspect de l'algèbre linéaire peut se reporter aux livres de F. Chatelin [7], B. N. Parlett [31] et Y. Saad [33].

Annexe B

Normes vectorielles et matricielles

B.1 Introduction

On rappelle dans ce chapitre quelques notions indispensables à l'étude des propriétés des matrices en vue de la résolution de systèmes linéaires par des méthodes itératives, mais aussi pour le calcul des valeurs propres et vecteurs propres. L'outil principal introduit dans ce chapitre est la norme (de vecteur ou de matrice) qui permet de définir la notion de convergence de suites (de vecteurs ou de matrices).

On raisonne en général dans des espaces vectoriels sur $\mathbb{C}$, en suivant [13]. Les résultats se transposent sans difficulté au cas d'espaces vectoriels définis sur $\mathbb{R}$.

B.2 Normes de vecteurs

Définition B.1 *soit* $\mathtt{E}$ *un espace vectoriel sur* $\mathbb{C}$*, on appelle* **norme** *une application de* $\mathtt{E}$ *dans* $\mathbb{R}^+$*, notée* $\|\cdot\|$*, qui vérifie les trois propriétés suivantes :*

- $\forall x \in \mathtt{E}$, $\forall \lambda \in \mathbb{C}$, $\|\lambda x\| = |\lambda| \ \|x\|$
- $\forall x \in \mathtt{E}$, $\forall y \in \mathtt{E}$, $\|x + y\| \leq \|x\| + \|y\|$
- $\|x\| = 0 \quad \Longleftrightarrow \quad x = 0.$

La seconde propriété est appelée inégalité triangulaire ; un espace vectoriel muni d'une norme est dit espace vectoriel **normé**. Il est immédiat de vérifier que l'application "module" est une norme sur l'espace vectoriel $\mathbb{C}$.

D'une façon plus générale, si $\mathtt{E}$ est un espace vectoriel sur $\mathbb{C}$ de dimension finie n, et si $B = \{b_1, b_2, \ldots, b_n\}$ est une base de $\mathtt{E}$, tout vecteur de $\mathtt{E}$ s'écrit de manière unique

$$x = \alpha_1 b_1 + \alpha_2 b_2 + \ldots + \alpha_n b_n.$$

À l'aide des coordonnées $\alpha_i \in \mathbb{C}$, on définit l'application $\|.\|$ de $\mathtt{E}$ dans $\mathbb{R}^+$ par

$$\|x\| = \left(|\alpha_1|^2 + |\alpha_2|^2 + \ldots + |\alpha_n|^2\right)^{1/2}.$$

Cette application est une norme, appelée **norme euclidienne** ou encore **norme canonique** associée à la base B.

Si $\mathtt{E}$ est un espace vectoriel sur $\mathbb{C}$ de dimension finie n, muni d'un produit scalaire $(\cdot,\cdot)$, on peut définir une norme par la relation

$$\|x\| = (x,x)^{1/2}.$$

Remarque B.2 *La réciproque n'est pas vraie, il existe des espaces vectoriels normés qui ne sont pas euclidiens, car la norme doit posséder des propriétés supplémentaires pour permettre de définir un produit scalaire.*

Proposition B.3 *[Inégalité de Cauchy-Schwarz] Pour tout couple de vecteurs x, y d'un espace vectoriel euclidien $\mathtt{E}$ sur $\mathbb{C}$*

$$|(x,y)| \leq \|x\| \times \|y\|.$$

Démonstration : Soient $x, y \in \mathtt{E}$, on écrit $(x,y) = |(x,y)|\, e^{\imath\theta}$, pour $\theta \in [0, 2\pi[$:

$$\begin{aligned} \forall \lambda \in \mathbb{R} \qquad \|\lambda\, e^{-\imath\theta} x + y\|^2 &= (\lambda\, e^{-\imath\theta} x + y, \lambda\, e^{-\imath\theta} x + y) \\ &= \lambda^2 \|x\|^2 + \lambda e^{-\imath\theta}\,(x,y) + \lambda e^{\imath\theta}\,\overline{(y,x)} + \|y\|^2 \\ &= \lambda^2 \|x\|^2 + \lambda e^{-\imath\theta}\,(x,y) + \lambda e^{\imath\theta}\,\overline{(x,y)} + \|y\|^2 \\ &= \lambda^2 \|x\|^2 + 2\lambda |(x,y)| + \|y\|^2 \end{aligned}$$

en prenant $x \neq 0$, le trinome du second degré en λ garde un signe constant quel que soit la valeur de $\lambda \in \mathbb{R}$, ce qui implique que le discriminant est négatif, soit

$$|(x,y)|^2 - \|x\|^2\, \|y\|^2 \leq 0.$$

$\diamond$

D'une manière générale, à l'aide des coordonnées α_i d'un vecteur x dans la base B, on peut lui associer, pour tout entier $p > 0$ fini, les normes suivantes appelées **normes de Hölder**

$$\|x\|_p = \left(|\alpha_1|^p + |\alpha_2|^p + \ldots + |\alpha_n|^p\right)^{1/p}.$$

Cette définition comprend les cas particuliers

$$\|x\|_1 = \sum_{i=1}^{n} |\alpha_i| \quad \text{et} \quad \|x\|_2 = (\sum_{i=1}^{n} |\alpha_i|^2)^{1/2}$$

et s'étend au cas $p = \infty$ avec la norme $\|x\|_\infty = \max_i |\alpha_i|$.

En utilisant la convexité de la fonction $x \mapsto e^x$, on établit alors la majoration suivante, appelée **inégalité de Hölder**

$$\forall x \in \mathtt{E} \qquad |(x,y)| \leq \|x\|_p\, \|y\|_q$$

pour tout couple d'entiers $p > 0$ et $q > 0$ liés par la relation

$$\frac{1}{p} + \frac{1}{q} = 1,$$

dont l'inégalité de Cauchy-Schwarz est un cas particulier, avec $p = q = 2$.

Définition B.4 *on dit que deux normes* $\|\cdot\|_\diamond$ *et* $\|\cdot\|_\star$ *définies sur un espace vectoriel* $\mathtt{E}$ *sont* **équivalentes** *s'il existe deux constantes* $C_m > 0$ *et* $C_M > 0$ *telles que :*

$$\forall x \in \mathtt{E} \qquad C_m \, \|x\|_\diamond \leq \|x\|_\star \leq C_M \, \|x\|_\diamond.$$

En utilisant les propriétés selon lesquelles d'une part toute fonction continue sur un ensemble compact atteint ses extrema, et d'autre part tout ensemble fermé et borné dans un espace vectoriel de dimension finie est compact, on en déduit facilement le résultat ci-dessous.

Théorème B.5 *Dans un espace vectoriel* $\mathtt{E}$ *sur* $\mathbb{C}$ *de dimension finie toutes les normes sont équivalentes.*

Démonstration : Soient $\|\cdot\|_\star$ et $\|\cdot\|_\diamond$ deux normes sur $\mathtt{E}$. Les fonctions $x \mapsto \|x\|_\star$ et $x \mapsto \|x\|_\diamond$ sont continues sur $\mathtt{E}$. La sphère unité $S_\star = \{x \in \mathtt{E} \ : \ \|x\|_\star = 1\}$ est donc compacte dans $\mathtt{E}$ (puisque fermée et bornée). En outre, la fonction de $S_\star$ dans $\mathbb{R}$, $x \mapsto \|x\|_\diamond$, atteint ses extrema. Soient C_{min} et C_{max} respectivement le minimum et le maximum de la fonction : il existe x_{min} et x_{max} deux éléments de $S_\star$ tels que $C_{min} = \|x_{min}\|_\diamond$ et $C_{max} = \|x_{max}\|_\diamond$. Ainsi, $C_{min} > 0$ puisque $x_{min} \neq 0$, et de même C_{max} est finie (et non nulle). Soit maintenant $x \in \mathtt{E} \setminus \{0\}$ quelconque, on a $(\|x\|_\star)^{-1} x \in S_\star$ et par conséquent :

$$C_{min} \leq (\|x\|_\star)^{-1} \|x\|_\diamond \leq C_{max} \quad \Longrightarrow \quad C_{min} \, \|x\|_\star \leq \|x\|_\diamond \leq C_{max} \, \|x\|_\star$$

La même inégalité est bien sûr vérifiée lorsque $x = 0$. $\diamond$

Ce résultat s'exprime plus précisément sous les formes particulières suivantes, dans un espace vectoriel $\mathtt{E}$ de dimension finie n :

$$\begin{array}{ll} \|x\|_2 & \leq \|x\|_1 \leq \sqrt{n} \, \|x\|_2 \\ \|x\|_\infty & \leq \|x\|_2 \leq \sqrt{n} \, \|x\|_\infty \\ \|x\|_\infty & \leq \|x\|_1 \leq n \, \|x\|_\infty. \end{array}$$

B.3 Normes de matrices

D'après l'étude des applications linéaires l'ensemble $\mathbb{C}^{m\times n}$ des matrices à m lignes et n colonnes est un espace vectoriel sur $\mathbb{C}$ de dimension $m \times n$; on peut donc considérer toute matrice A de $\mathbb{C}^{m\times n}$ comme un vecteur à $m \times n$ composantes et ainsi utiliser une des normes vectorielles précédentes pour définir $\|A\|_{m,n}$. Il est cependant nécessaire d'introduire une condition supplémentaire pour obtenir un outil de démonstration de la convergence de suites et de séries de matrices.
Pour éviter toute confusion les vecteurs sont représentés par des lettres minuscules et les matrices par des majuscules.

Définition B.6 *On dit que la norme vectorielle* $\|\cdot\|$ *définie sur* $\mathbb{C}^{n\times n}$ *est une* **norme matricielle***, si et seulement si elle vérifie pour tout couple* (A, B) *de matrices de* $\mathbb{C}^{n\times n}$

$$\|AB\| \leq \|A\| \, \|B\|.$$

Mais si $\mathtt{E}$ et $\mathtt{F}$ sont deux espaces vectoriels sur $\mathbb{C}$ de dimension finie, resp. $dim(\mathtt{E}) = n$, $dim(\mathtt{F}) = m$, il existe une bijection entre l'ensemble $\mathcal{L}(\mathtt{E}, \mathtt{F})$ des applications linéaires de $\mathtt{E}$ dans $\mathtt{F}$ et l'ensemble $\mathbb{C}^{m\times n}$, *une fois que* l'on a choisi une base de $\mathtt{E}$ et une base de $\mathtt{F}$; on peut alors définir une norme à partir de l'application linéaire associée : si A est la matrice associée à l'application f

$$\|A\| = \max_{x\neq 0} \frac{\|f(x)\|_{\mathtt{F}}}{\|x\|_{\mathtt{E}}}$$

De manière équivalente, $\|A\|$ peut être définie par

$$\|A\| = \inf\{\alpha \in \mathbb{R} \ : \ \|Ax\|_m \leq \alpha \, \|x\|_n, \ \forall x \in \mathtt{E}\}$$

De plus, comme la sphère unité est compacte dans l'espace vectoriel de dimension finie $\mathtt{E}$, il existe $x_0 \in \mathtt{E} \setminus \{0\}$ tel que $\|Ax_0\| = \|A\| \, \|x_0\|$.

Dans le cas particulier où $\mathtt{E} = \mathbb{C}^n$ et $\mathtt{F} = \mathbb{C}^m$, la matrice A est rectangulaire $A \in \mathbb{C}^{m\times n}$ et on note

$$\|A\|_{\mathtt{E},\mathtt{F}} = \max_{x\neq 0} \frac{\|Ax\|_{\mathtt{F}}}{\|x\|_{\mathtt{E}}}.$$

Proposition B.7 *Soient* $\mathtt{D} = \mathbb{C}^p$, $\mathtt{E} = \mathbb{C}^n$, $\mathtt{F} = \mathbb{C}^m$. *Les normes induites vérifient*

$$\forall A \in \mathbb{C}^{m\times n}, B \in \mathbb{C}^{n\times p} \qquad \|AB\|_{\mathtt{D},\mathtt{F}} \leq \|A\|_{\mathtt{E},\mathtt{F}} \, \|B\|_{\mathtt{D},\mathtt{E}}.$$

Démonstration : Pour $A \in \mathbb{C}^{m\times n}$ et $B \in \mathbb{C}^{n\times p}$, le produit AB est une matrice de $\mathbb{C}^{m\times p}$ qui vérifie

$$\|AB\|_{\mathtt{D},\mathtt{F}} = \max_{x\neq 0} \frac{\|ABx\|_{\mathtt{F}}}{\|x\|_{\mathtt{D}}} = \max_{x\neq 0, Bx\neq 0} \frac{\|ABx\|_{\mathtt{F}}}{\|Bx\|_{\mathtt{E}}} \frac{\|Bx\|_{\mathtt{E}}}{\|x\|_{\mathtt{D}}} \leq \max_{y\neq 0} \frac{\|Ay\|_{\mathtt{F}}}{\|y\|_{\mathtt{E}}} \max_{x\neq 0} \frac{\|Bx\|_{\mathtt{E}}}{\|x\|_{\mathtt{D}}}.$$

◇

Lorsque $m = n$, cette application satisfait aux axiomes de la Définition B.1 et aussi à la Définition B.6 (voir la Proposition B.7) ; on dit que cette norme est **associée** à la norme vectorielle $\|\cdot\|$, ou **induite** par la norme vectorielle, ou encore **subordonnée** à la norme vectorielle. Dans ce cas, il est d'usage de noter de la même façon la norme vectorielle et la norme matricielle associée :

$$\|A\| = \max_{x\neq 0} \frac{\|Ax\|}{\|x\|}.$$

Proposition B.8 *Pour toute matrice carrée* $A \in \mathbb{C}^{n\times n}$ *les normes matricielles de Hölder vérifient*

(i) $\|A\|_1 = \max_{x\neq 0} \frac{\|Ax\|_1}{\|x\|_1} = \max_j \sum_i |A_{i,j}|.$

(ii) $\|A\|_\infty = \max_{x\neq 0} \frac{\|Ax\|_\infty}{\|x\|_\infty} = \max_i \sum_j |A_{i,j}|.$

Démonstration : Pour tout vecteur $v \in \mathbb{C}^n$

$$\|Av\|_1 = \sum_i |\sum_j A_{i,j} v_j| \leq \sum_i \sum_j |A_{i,j}| \, |v_j| \leq (\max_j \sum_i |A_{i,j}|)\|v\|_1;$$

pour obtenir l'égalité, on construit un vecteur v particulier : soit j_0 un indice pour lequel

$$\sum_i |A_{i,j_0}| = \max_j \sum_i |A_{i,j}| \ ;$$

le vecteur e_{j_0} dont toutes les composantes sont nulles à l'exception de $e_{j_0} = 1$ répond à la question.

De même

$$\|Av\|_\infty = \max_i |\sum_j A_{i,j} v_j| \leq (\max_i \sum_j |A_{i,j}|)\|v\|_\infty;$$

soit i_0 un indice tel que

$$\sum_j |A_{i_0,j}| = \max_i \sum_j |A_{i,j}| \ ;$$

le vecteur v dont les composantes sont

$$v_j = 1 \quad \text{si } A_{i_0,j} = 0 \qquad \text{et } v_j = \frac{A_{i_0,j}}{|A_{i_0,j}|} \quad \text{si } A_{i_0,j} \neq 0$$

permet d'atteindre l'égalité. ◇

Il est aussi possible de définir une norme à partir des éléments d'une matrice carrée d'ordre n. C'est la cas de la norme de **Schur-Frobenius**

$$\|A\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^n |A_{i,j}|^2} = \sqrt{tr(A^* A)},$$

où tr désigne la trace.

Proposition B.9 *La norme de Schur-Frobenius $\|\cdot\|_F$ est une norme matricielle, mais elle n'est pas une norme matricielle induite.*

Démonstration : Pour cette norme, on peut écrire

$$\begin{aligned}
\|AB\|_F^2 &= \sum_{i=1}^n \sum_{j=1}^n |(AB)_{i,j}|^2 = \sum_{i=1}^n \sum_{j=1}^n \Big|\sum_{k=1}^n A_{i,k} B_{k,j}\Big|^2 \\
\text{(Cauchy-Schwarz)} \quad &\leq \sum_{i=1}^n \sum_{j=1}^n \big((\sum_{k=1}^n |A_{i,k}|^2)(\sum_{k=1}^n |B_{k,j}|^2\big) = \big(\sum_{i=1}^n \sum_{k=1}^n |A_{i,k}|^2\big)\big(\sum_{j=1}^n \sum_{k=1}^n |B_{k,j}|^2\big) \\
&= \|A\|_F^2 \, \|B\|_F^2,
\end{aligned}$$

ainsi la norme de Schur-Frobenius est bien une norme matricielle.
On vérifie directement que $\|I_n\|_F = \sqrt{n}$. Or, pour toute norme induite, on a l'égalité

$$\|I_n\|_n = \max_{x \neq 0} \frac{\|I_n x\|_n}{\|x\|_n} = 1,$$

donc la norme de Schur-Frobenius n'est pas une norme induite. ◇

Proposition B.10 *Pour toute matrice $A \in \mathbb{C}^{n\times n}$ et toute matrice untaire $U \in \mathbb{C}^{n\times n}$*

$$\|UA\|_2 = \|AU\|_2 = \|A\|_2.$$

Démonstration : On écrit

$$\|UA\|_2 = \max_{x\neq 0} \frac{\|UAx\|_2}{\|x\|_2} = \max_{x\neq 0} \frac{(UAx, UAx)}{\|x\|_2} = \max_{x\neq 0} \frac{(Ax, U^*UAx)}{\|x\|_2} \overset{U^*U=I_n}{=} \max_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \|A\|_2;$$

par ailleurs, on a

$$\|AU\|_2 = \max_{x\neq 0} \frac{\|AUx\|_2}{\|x\|_2} = \max_{x\neq 0} \frac{\|AQx\|_2}{\|Qx\|_2} = \|A\|_2.$$

◇

B.4 Normes des matrices et valeurs propres

Les valeurs propres d'une matrice A de $\mathbb{C}^{n\times n}$, notées λ_i ou $\lambda_i(A)$, appartiennent *a priori* à $\mathbb{C}$.

Définition B.11 *le* **rayon spectral** *de la matrice $A \in \mathbb{C}^{n\times n}$ est le* réel positif

$$\rho(A) = \max_i |\lambda_i(A)|.$$

Proposition B.12 *Pour toute matrice $A \in \mathbb{C}^{n\times n}$ et toute norme matricielle $\|\cdot\|$*

$$\rho(A) \leq \|A\|.$$

Démonstration : Soit λ une valeur propre de A telle que $|\lambda| = \rho(A)$, et $p \in \mathbb{C}^n$ un vecteur propre associé : $Ap = \lambda p$. Puisque $p \neq 0$, il existe $q \in \mathbb{C}^n$ tel que la matrice pq^T appartient à $\mathbb{C}^{n\times n} \setminus \{0\}$. Comme $\|\cdot\|$ est une norme matricielle, on a par définition $\|A(pq^T)\| \leq \|A\| \, \|pq^T\|$. Par ailleurs :

$$\rho(A)\|pq^T\| = |\lambda| \, \|pq^T\| = \|\lambda \, pq^T\| = \|Apq^T\|.$$

Ainsi, $\rho(A)\|pq^T\| \leq \|A\| \, \|pq^T\|$ et comme $\|pq^T\| \neq 0$, on a démontré le résultat. ◇

Proposition B.13 *Pour toute matrice $A \in \mathbb{C}^{n\times n}$ et tout $\varepsilon > 0$, il existe une norme matricielle $\|\cdot\|$ telle que*

$$\|A\| - \varepsilon \leq \rho(A).$$

Démonstration : Pour obtenir ce résultat, on utilise une propriété importante des matrices carrées (voir la Proposition A.21) : toute matrice $A \in \mathbb{C}^{n\times n}$ peut s'écrire sous la forme $A = QTQ^*$, où Q est une matrice unitaire et T une matrice triangulaire supérieure dont la diagonale est formée des valeurs propres de la matrice A (ces valeurs propres peuvent

être complexes ou réelles, nulles ou non, distinctes ou non et ne sont pas rangées suivant leur module) :

$$T = \begin{pmatrix} \lambda_1 & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & \ddots & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & \ddots & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & \ddots & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & 0 & \ddots & \mathrm{x} \\ 0 & 0 & 0 & 0 & 0 & \lambda_n \end{pmatrix}.$$

Soit δ un nombre réel strictement positif, on construit la matrice diagonale $D \in \mathbb{C}^{n\times n}$

$$D = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \delta & 0 & 0 & 0 & 0 \\ 0 & 0 & \delta^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & \delta^{n-1} \end{pmatrix};$$

alors

$$(QD)^{-1}A(QD) = D^{-1}TD = \begin{pmatrix} \lambda_1 & \delta T_{1,2} & \delta^2 T_{1,3} & \dots & \dots & \delta^{n-1}T_{1,n} \\ 0 & \lambda_2 & \delta T_{2,3} & \ddots & \ddots & \delta^{n-2}T_{2,n} \\ 0 & 0 & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & \ddots & \delta^2 T_{n-2,n} \\ 0 & 0 & 0 & 0 & \ddots & \delta T_{n-1,n} \\ 0 & 0 & 0 & 0 & 0 & \lambda_n \end{pmatrix}.$$

Pour une matrice $A \in \mathbb{C}^{n\times n}$ et un réel ε donnés, on peut trouver $\delta > 0$ tel que

$$\text{pour tout } i = 1, \cdots, n-1, \quad \max_{i+1\le j\le n} |\delta^{j-i}T_{i,j}| < \varepsilon/n;$$

alors la norme matricielle (dépendant de A et ε) définie pour toute matrice $B \in \mathbb{C}^{n\times n}$ par

$$\|B\|_{A,\varepsilon} = \|(QD)^{-1}B(QD)\|_\infty$$

vérifie l'inégalité

$$\|A\|_{A,\varepsilon} \le \max_i |\lambda_i| + \varepsilon$$

Cette norme est la norme matricielle subordonnée à la norme vectorielle

$$v \mapsto \|(QD)^{-1}v\|_\infty$$

car

$$\|(QD)^{-1}B(QD)\|_\infty = \max_{y\neq 0} \frac{\|(QD)^{-1}B(QD)y\|_\infty}{\|y\|_\infty} = \max_{x\neq 0} \frac{\|(QD)^{-1}Bx\|_\infty}{\|(QD)^{-1}x\|_\infty}.$$

◇

Proposition B.14 *Pour toute matrice* $A \in \mathbb{C}^{n\times n}$

$$\|A\|_2 = \sqrt{\rho(A^*A)}.$$

Pour toute matrice $A \in \mathbb{C}^{n\times n}$ hermitienne

$$\rho(A) = \|A\|_2.$$

Démonstration : Par définition

$$\|A\|_2^2 = \max_{x\neq 0} \frac{\|Ax\|_2^2}{\|x\|_2^2} = \max_{x\neq 0} \frac{(Ax, Ax)}{(x, x)} = \max_{x\neq 0} \frac{(x, A^*Ax)}{(x, x)}.$$

Evaluons le dernier maximum : $A^*A \in \mathbb{C}^{n\times n}$ est une matrice hermitienne, elle admet donc une base de vecteurs propres orthogonaux $\{v_1, v_2, \ldots, v_n\}$ (voir la Proposition A.20). Ses valeurs propres $\lambda_i(A^*A)$ sont réelles et positives. On obtient pour tout vecteur $x \in \mathbb{C}^n$

$$x = \sum_{i=1}^n \alpha_i v_i, \quad \text{et} \quad A^*Ax = \sum_{i=1}^n \lambda_i(A^*A)\alpha_i v_i.$$

Ainsi

$$\frac{(x, A^*Ax)}{(x, x)} = \frac{\sum_{i=1}^n \lambda_i(A^*A)|\alpha_i|^2(v_i, v_i)}{\sum_{i=1}^n |\alpha_i|^2(v_i, v_i)} \leq \max_{i=1,n}(\lambda_i(A^*A)) = \rho(A^*A).$$

Par ailleurs, si on choisit $x_0 = v_{i_0}$ pour i_0 tel que $\lambda_{i_0}(A^*A) = \rho(A^*A)$, on trouve

$$\frac{(x_0, A^*Ax_0)}{(x_0, x_0)} = \rho(A^*A).$$

Dans le cas d'une matrice A hermitienne, on choisit cette fois une base $\{v_1, v_2, \ldots, v_n\}$ de $\mathbb{C}^n$ de vecteurs propres orthogonaux de A. On trouve :

$$\frac{(Ax, Ax)}{(x, x)} = \frac{\sum_{i=1}^n |\alpha_i|^2|\lambda_i(A)|^2(v_i, v_i)}{\sum_{i=1}^n |\alpha_i|^2(v_i, v_i)} \leq (\max_{i=1,n}(|\lambda_i(A)|)^2 = (\rho(A))^2.$$

Et si on choisit $x_0 = v_{i_0}$ pour i_0 tel que $|\lambda_{i_0}(A)| = \rho(A)$, on trouve

$$\frac{(Ax_0, Ax_0)}{(x_0, x_0)} = (\rho(A))^2.$$

◇

Corollaire B.15 *Pour toute matrice* $A \in \mathbb{C}^{n\times n}$

$$\|A\|_2 \leq \|A\|_F \leq \sqrt{n}\|A\|_2.$$

Démonstration : On a par définition $\|A\|_F = tr(A^*A)$ et en outre

$$\|A\|_2^2 = \rho(A^*A) \leq tr(A^*A) \leq n\,\rho(A^*A) = n\,\|A\|_2^2.$$

◇

Définition B.16 *Soit A une matrice de $\mathbb{C}^{n\times n}$. A toute valeur propre de A^*A on associe la* **valeur singulière** $\sigma(A)$ *via la relation*

$$\sigma(A) = \sqrt{\lambda(A^*A)}.$$

De manière classique, on range les valeurs singulières de la matrice A par valeur décroissante :

$$\sigma_n(A) \leq \sigma_{n-1}(A) \leq \ldots \leq \sigma_2(A) \leq \sigma_1(A).$$

La Proposition B.14 se résume à $\|A\|_2 = \sigma_1(A)$.
Enfin, on vérifie facilement le résultat suivant

Proposition B.17 *Pour toute matrice $A \in \mathbb{C}^{n\times n}$ la norme de Schur-Frobenius vérifie*

$$\|A\|_F^2 = \sum_{i=1,n} \sigma_i^2(A).$$

Démonstration : Par définition, $\|A\|_F^2 = tr(A^*A) = \sum_{i=1,n} \lambda_i(A^*A) = \sum_{i=1,n} \sigma_i^2(A)$. ◇

B.5 Suites de vecteurs. Suites de matrices

L'analyse théorique des algorithmes de calcul numérique utilise la notion de suite de vecteurs et étudie leur convergence éventuelle. On notera $\{x_k\}_{k\in\mathbb{N}}$ une suite d'éléments d'un espace vectoriel `E` sur $\mathbb{C}$, et on dira que la suite $\{x_k\}_{k\in\mathbb{N}}$ converge vers l'élément x de `E` si

$$\lim_{k\to\infty} \|x_k - x\| = 0,$$

que l'on écrit de manière classique $\lim_{k\to\infty} x_k = x$. Dans le cas particulier d'un espace vectoriel `E` de dimension finie n, cette définition est indépendante de la norme choisie, et la convergence de la suite $\{x_k\}_{k\in\mathbb{N}}$ vers x est équivalente à la convergence de chacune des suites de composantes $\{(x_k)_i\}_{k\in\mathbb{N}}$ vers x_i pour $i = 1, 2, \ldots, n$, par rapport à une base quelconque.

On définit de manière analogue une suite de matrices $\{A_k\}_{k\in\mathbb{N}} \in \mathbb{C}^{m\times n}$ et on peut utiliser la définition de la convergence d'une suite de vecteurs dans l'espace vectoriel $\mathbb{C}^{m\times n}$ de dimension finie $n \times m$. On dira donc que $\{A_k\}_{k\in\mathbb{N}}$ converge vers la matrice A de $\mathbb{C}^{m\times n}$ si

$$\lim_{k\to\infty} \|A_k - A\| = 0,$$

et on écrira $\lim_{k\to\infty} A_k = A$. Cependant dans de nombreux algorithmes les éléments de la suite de matrices considérée sont de la forme $A_k = A^k$, puissances successives d'une matrice donnée $A \in \mathbb{C}^{n\times n}$. Dans ce cas particulier, on relie la convergence de cette suite au rayon spectral de la matrice A.

Théorème B.18 *Pour toute matrice $A \in \mathbb{C}^{n\times n}$ les conditions suivantes sont équivalentes :*

(i) $\lim_{k\to\infty} A^k = 0$ *;*

(ii) $\forall x \in \mathbb{C}^n$, $\lim_{k\to\infty} A^k x = 0$ *;*

(iii) $\rho(A) < 1$ *;*

(iv) il existe une norme induite telle que $\|A\| < 1$.

Démonstration : On montre l'équivalence par implication circulaire.

(i) $\Longrightarrow$ (ii) : pour toute norme vectorielle $\|\cdot\|$ et sa norme matricielle induite, on a la majoration

$$\forall x \in \mathbb{C}^n \qquad \|A^k x\| \leq \|A^k\| \times \|x\|.$$

d'où le résultat.

(ii) $\Longrightarrow$ (iii) : soit λ une valeur propre de A telle que $\rho(A) = |\lambda|$ et soit $u \neq 0$ un vecteur propre associé, alors

$$\|A^k u\| = \|\lambda^k u\| = |\lambda|^k \|u\| = \rho(A)^k \|u\|.$$

Si on suppose $\rho(A) \geq 1$ alors $\|A^k u\| \geq \|u\|$, ce qui contredit (ii).

(iii) $\Longrightarrow$ (iv) : est une conséquence de la Proposition B.13 (prendre $\varepsilon = (1 - \rho(A))/2$).

(iv) $\Longrightarrow$ (i) : puisque $\|A^k\| \leq \|A\|^k$, on a bien $\lim_{k\to\infty} \|A^k\| = 0$ pour la norme matricielle telle que $\|A\| < 1$. ◇

Remarque B.19 *Il faut souligner que l'utilisation d'une norme matricielle arbitraire pour évaluer la convergence d'une suite peut amener à une mauvaise conclusion. Par contre, la valeur du rayon spectral de la matrice est toujours pertinente.*

Corollaire B.20 *Soit* $A \in \mathbb{C}^{n\times n}$ *et* $\|\cdot\|$ *une norme matricielle :*

$$\lim_{k\to\infty} \|A^k\|^{1/k} = \rho(A).$$

Démonstration : D'après la Proposition B.12, on a pour tout k, $\rho(A^k) \leq \|A^k\|$. Par ailleurs,

$$\rho(A)^k = (\max_i |\lambda_i(A)|)^k = \max_i |\lambda_i(A)^k| \leq \max_i |\lambda_i(A^k)| = \rho(A^k),$$

puisque si $\lambda_i(A)$ est valeur propre de A, alors $\lambda_i(A)^k$ est valeur propre de A^k. D'où l'on déduit que $\rho(A) \leq \|A^k\|^{1/k}$ pour tout k.
Il faut démontrer que

$$\forall \varepsilon > 0,\ \exists k_\varepsilon,\ \forall k \geq k_\varepsilon, \quad \|A^k\|^{1/k} \leq \rho(A) + \varepsilon.$$

Or, si on introduit $A_\varepsilon = (\rho(A) + \varepsilon)^{-1} A$, on a $\rho(A_\varepsilon) < 1$ d'où $\lim_{k\to\infty} \|(A_\varepsilon)^k\| = 0$ d'après le Théorème B.18. En particulier, il existe k_ε tel que pour tout $k \geq k_\varepsilon$, on a $\|(A_\varepsilon)^k\| < 1$, ou, en d'autres termes, $\|A^k\| < (\rho(A) + \varepsilon)^k$, ce qui est le résultat cherché.

◇

Annexe C

En dimension infinie

Nous commençons par quelques rappels sur les espaces de Hilbert. Ensuite, nous énumérons les principaux résultats mathématiques permettant de résoudre les problèmes associés aux équations aux dérivées partielles elliptiques, complétées de conditions aux limites, lorsqu'ils sont écrits sous forme variationnelle [11, 2]. Pour finir, nous rappelons quelques résultats fondamentaux pour l'approximation des formulations variationnelles.

C.1 Espaces de Hilbert

Les espaces vectoriels sont définis sur $\mathbb{C}$. Ceci étant dit, les notions ci-dessous sont aisément transposables à des espaces vectoriels définis sur $\mathbb{R}$.
Par définition, un espace topologique est *séparable* s'il contient un sous-ensemble dénombrable dense ; un espace de *Banach* est un espace vectoriel complet muni d'une norme ; un espace de *Hilbert* est un espace vectoriel muni d'un produit scalaire, complet par rapport à la norme induite par le produit scalaire. Pour rappel, dans un espace vectoriel, un produit scalaire $(\cdot,\cdot)$ possède les propriétés suivantes :

- Il est *linéaire* par rapport à la première variable :
  $\forall a_1, a_2 \in \mathbb{C},\ \forall v_1, v_2, w \in V,\ (a_1 v_1 + a_2 v_2, w) = a_1(v_1, w) + a_2(v_2, w).$
- Il est *antilinéaire* par rapport à la deuxième variable :
  $\forall a_1, a_2 \in \mathbb{C},\ \forall v, w_1, w_2 \in V,\ (v, a_1 w_1 + a_2 w_2) = \overline{a_1}(v, w_1) + \overline{a_2}(v, w_2).$
- Il est *hermitien* :
  $\forall v, w \in V,\ (v, w) = \overline{(w, v)}.$
- Il est *défini-positif* :
  $\forall v \in V \setminus \{0\},\ (v, v) > 0.$

Alors, $\|v\| : V \to \mathbb{R}$ définie par $\|v\| = (v, v)^{1/2}$ est une norme sur l'espace vectoriel. De plus, on a l'inégalité de Cauchy-Schwarz : $\forall v, w \in V,\ |(v, w)| \leq \|v\|\,\|w\|$. Dans la suite, on note $(\cdot,\cdot)_V$ le produit scalaire, et $\|\cdot\|_V$ la norme induite sur l'espace vectoriel V.

Soit V un espace de Hilbert. Son espace dual[54], noté V', est l'espace vectoriel des formes *antilinéaires* et continues sur V, muni de la norme

$$\|f\|_{V'} = \sup_{v \in V \setminus \{0\}} \frac{|\langle f, v\rangle_V|}{\|v\|_V}.$$

54. V' peut être appelé l'espace antidual. Nous choisissons la dénomination espace dual, car elle s'applique également pour les espaces vectoriels définis sur $\mathbb{R}$, et les formes linéaires et continues.

De façon générique, $\langle f, v\rangle_V$ dénote l'action de la forme f sur l'élément v. Lorsque l'appartenance aux espaces V et V' est claire, nous omettrons l'indice $_V$ pour écrire simplement $\langle f, v\rangle$.
Pour $v \in V$ donné, $f_v : w \mapsto (v, w)_V$ définit un élément de V'. D'après le théorème de Riesz C.3 ci-après, $v \mapsto f_v$ est une isométrie bijective de V dans V'. De plus, on peut transporter la structure d'espace de Hilbert à V' en définissant son produit scalaire via $(f_v, f_w)_{V'} = (v, w)_V$, pour tout $f_v, f_w \in V'$.
Soit W un deuxième espace de Hilbert. Nous utilisons des formes continues et *sesquilinéaires* [55] dans $V \times W$. La forme sesquilinéaire $a : V \times W \to \mathbb{C}$, $(v, w) \mapsto a(v, w)$ est *continue* si la quantité

$$|||a||| = \sup_{v \in V\setminus\{0\}, w \in W\setminus\{0\}} \frac{|a(v, w)|}{\|v\|_V \|w\|_W}$$

est bornée. Lorsque $a(\cdot, \cdot)$ est sesquilinéaire et continue dans $V \times W$, elle définit un unique opérateur (borné) A de V dans W' (on écrit $A \in \mathcal{L}(V, W')$) :

$$\forall (v, w) \in V \times W,\ \langle Av, w\rangle_W = a(v, w). \tag{C.1}$$

Classiquement, la norme $|||A||| = \sup_{v \in V\setminus\{0\}} \frac{\|Av\|_{W'}}{\|v\|_V}$ est égale à $|||a|||$.
On peut également définir sa *transposée conjuguée* $A^\dagger$ de W dans V' :

$$\forall (v, w) \in V \times W,\ \langle A^\dagger w, v\rangle_V = \overline{a(v, w)}.$$

Pour une forme continue et *bilinéaire* [56] a sur des espaces de Hilbert V et W définis sur $\mathbb{R}$, on introduit A de V dans W' comme ci-dessus, respectivement sa *transposée* A^t de W dans V' – sans conjugaison –.
Evidemment, à partir d'un opérateur (borné) A de V dans W', on pourrait définir une forme continue et sesquilinéaire (ou bilinéaire) sur $V \times W$ via (C.1).

Finissons par la notion d'espace pivot, qui est une autre conséquence du théorème de Riesz C.3 ci-dessous. Soit H un espace de Hilbert. Pour $g \in H'$, soit $h \in H$ la solution de

$$\begin{cases} \textit{Trouver } h \in H \textit{ tel que} \\ \forall h' \in H,\ (h, h')_H = \langle g, h'\rangle_H. \end{cases}$$

L'application $g \mapsto h$ est une isométrie bijective de H' dans H. A partir de là, on peut choisir d'*identifier* H' à H.

Définition C.1 `(espace pivot)` *Soit H un espace de Hilbert. Dès lors que H' est identifié à H – par l'intermédiaire de l'application $g \mapsto h$ introduite ci-dessus – H est appelé l'*espace pivot.

Il suit la

Proposition C.2 *Soient H et V deux espaces de Hilbert, tels que V soit un sous-espace vectoriel dense de H, et tel que l'injection canonique $i_{V \to H}$ soit continue. Lorsque H est choisi comme espace pivot, on peut identifier H à un sous-espace vectoriel de V'.*

55. Une forme sesquilinéaire est linéaire par rapport à la première variable et antilinéaire par rapport à la deuxième variable.

56. Une forme bilinéaire est linéaire par rapport aux première et seconde variables.

Qui plus est, lorsque les hypothèses de la proposition sont vraies, on vérifie que l'application $i_{H\to V'}$ est continue (et injective), et que $i_{H\to V'}H$ est dense dans V'. Par conséquent, on peut écrire

$$V \subset H \overset{\text{(pivot)}}{=} H' \subset V',$$

avec des injections continues de sous-espaces vectoriels denses les uns dans les autres.

C.2 Résultats fondamentaux

Soit V un espace de Hilbert. Soit f un élément de V', on introduit le problème

$$\begin{cases} \textit{Trouver } u \in V \textit{ tel que} \\ \forall v \in V,\ (u,v)_V = \langle f, v\rangle. \end{cases} \tag{C.2}$$

On appelle (C.2) une *formulation variationnelle.*
Le premier résultat est le théorème de Riesz.

Théorème C.3 `(Riesz)` *Pour tout $f \in V'$, le problème (C.2) possède une solution et une seule u dans V. De plus, on a l'égalité $\|u\|_V = \|f\|_{V'}$.*

On a un deuxième résultat qui généralise le théorème de Riesz.

Définition C.4 *Soit $a(\cdot,\cdot)$ une forme continue et sesquilinéaire sur $V \times V$. La forme a est* coercive *si*

$$\exists \alpha > 0,\ \forall v \in V,\ |a(v,v)| \geq \alpha \|v\|_V^2.$$

Soit $a(\cdot,\cdot)$ une forme continue et sesquilinéaire sur $V \times V$. Soit f un élément de V', nous introduisons une seconde formulation variationnelle

$$\begin{cases} \textit{Trouver } u \in V \textit{ tel que} \\ \forall v \in V,\ a(u,v) = \langle f, v\rangle. \end{cases} \tag{C.3}$$

Définition C.5 *Lorsqu'un problème (C.3) possède une solution et une seule dans V qui dépend continûment de la donnée f, c'est-à-dire*

$$\exists C > 0,\ \forall f \in V',\ \textit{(C.3) a une solution et une seule } u, \textit{ avec } \|u\|_V \leq C\|f\|_{V'},$$

on dit qu'il est bien posé *(au sens d'Hadamard).*

Notons qu'il est possible de reformuler le problème (C.3) comme suit.

$$\begin{cases} \textit{Trouver } u \in V \textit{ tel que} \\ Au = f \text{ dans } V'. \end{cases} \tag{C.4}$$

Il est important de noter que l'opérateur (borné) A^{-1} est bien défini (et continu) de V' dans V si, et seulement si, le problème (C.3) est bien posé. Dans ce cas, les opérateurs A et A^{-1} sont des isomorphismes.

Le deuxième résultat, appelé le théorème de Lax-Milgram, propose une condition suffisante pour garantir le caractère bien posé du problème (C.3).

Théorème C.6 (Lax-Milgram) *Supposons que la forme continue et sesquilinéaire a soit coercive. Alors, le problème (C.3) est bien posé.*

Remarque C.7 *On pourrait également définir la coercivité des formes sesquilinéaires selon*

$$\exists \alpha > 0,\ \exists \theta \in [0, 2\pi[,\quad \forall v \in V,\ Re[\exp(\imath\theta)\, a(v,v)] \geq \alpha \, \|v\|_V^2.$$

Pour une forme continue et sesquilinéaire, cette définition est équivalente *à la définition C.4, cf. [6]. Nous utilisons la définition C.4 dans la suite.*
De plus, pour des formes a sur un espace de Hilbert V défini sur $\mathbb{R}$*, les deux définitions reviennent à*

$$\exists s \in \{-1, +1\},\ \exists \alpha > 0,\ \forall v \in V,\ s\, a(v,v) \geq \alpha \, \|v\|_V^2.$$

Concernant les problèmes (C.2-C.3), on constate que la solution $u \in V$ possède deux caractéristiques principales : tout d'abord, qu'elle est *mesurée selon une norme donnée* – $\|\cdot\|_V$ – ; et ensuite qu'elle est *définie par son action* sur tous les champs de V, ou par une équation posée dans V'.

Plutôt que d'imposer la coercivité (condition suffisante) de la forme sesquilinéaire, on peut considérer une *condition de stabilité*, également appelée une *condition inf-sup*. Ce type de condition est aussi très utile lorsque les arguments de la forme sesquilinéaire n'appartiennent pas au même espace fonctionnel [20, 4, 2]. Soient donc W un deuxième espace de Hilbert, et $a(\cdot,\cdot)$ une forme continue et sesquilinéaire sur $V \times W$. Soit $f \in W'$, nous introduisons une troisième formulation variationnelle

$$\left\{ \begin{array}{l} \textit{Trouver } u \in V \textit{ tel que} \\ \forall w \in W,\ a(u,w) = \langle f, w \rangle. \end{array} \right. \tag{C.5}$$

Ces problèmes généralisent les problèmes (C.3), pour lesquels $W = V$.
Le caractère *bien posé* (au sens d'Hadamard) s'exprime cette fois par

$$\exists C > 0,\ \forall f \in W',\ \text{(C.5) a une solution et une seule } u, \text{ avec } \|u\|_V \leq C \|f\|_{W'}.$$

Il est possible de reformuler le problème (C.5) à l'aide de l'opérateur $A \in \mathcal{L}(V, W')$ associé à la forme a par (C.1).

$$\left\{ \begin{array}{l} \textit{Trouver } u \in V \textit{ tel que} \\ Au = f \text{ dans } W'. \end{array} \right. \tag{C.6}$$

Comme précédemment, l'opérateur (borné) A^{-1} est bien défini (et continu) de W' dans V si, et seulement si, le problème (C.5) est bien posé.

Définition C.8 *Soit a une forme continue et sesquilinéaire sur $V \times W$.*
Elle vérifie une condition de stabilité *si*

$$\exists \alpha' > 0,\ \forall v \in V,\ \sup_{w \in W \setminus \{0\}} \frac{|a(v,w)|}{\|w\|_W} \geq \alpha' \, \|v\|_V. \tag{C.7}$$

Elle vérifie une condition de solvabilité *si*

$$\{w \in W\ :\ \forall v \in V,\ a(v,w) = 0\} = \{0\}. \tag{C.8}$$

Remarque C.9 *Lorsque $W = V$, la coercivité d'une forme sesquilinéaire implique condition de stabilité (avec $\alpha' = \alpha$) et condition de solvabilité pour la même forme.*

On a alors le résultat ci-dessous.

Proposition C.10 *Supposons que la forme continue et sesquilinéaire a vérifie une condition de stabilité (C.7) avec un $\alpha' > 0$. Alors, $Ker(A) = \{0\}$, $Im(A)$ est fermée dans W', et A est une bijection de V dans $Im(A)$. Par conséquent, pour tout $f \in Im(A)$, le problème (C.5) possède une solution u et une seule dans V, et de plus $\alpha' \|u\|_V \leq \|f\|_{W'}$. Enfin, si la forme a vérifie la condition de solvabilité (C.8), alors $Im(A) = W'$ et le problème (C.5) est bien posé.*

Théorème C.11 (Banach-Necas-Babuska) *Soit a une forme continue et sesquilinéaire sur $V \times W$. Le problème (C.5) est bien posé si, et seulement si, la forme a vérifie une condition de stabilité (C.7) et une condition de solvabilité (C.8).*

Corollaire C.12 (Banach-Necas-Babuska) *Supposons que $W = V$, et que la forme a soit continue, sesquilinéaire et hermitienne sur $V \times V$. Le problème (C.5) est bien posé si, et seulement si, la forme a vérifie une condition de stabilité (C.7).*

Introduisons maintenant une condition *a priori* intermediaire (cf. [5]).

Définition C.13 *Soit $a(\cdot, \cdot)$ une forme continue et sesquilinéaire sur $V \times W$. Elle est* $\mathbb{T}$-coercive *si*

$$\exists \mathbb{T} \in \mathcal{L}(V, W),\ \textit{bijective},\ \exists \underline{\alpha} > 0,\ \forall v \in V,\ |a(v, \mathbb{T}v)| \geq \underline{\alpha} \|v\|_V^2.$$

Soit $a(\cdot, \cdot)$ une forme continue, sesquilinéaire et hermitienne sur $V \times V$. Elle est $\mathbb{T}$-coercive *si*

$$\exists \mathbb{T} \in \mathcal{L}(V),\ \exists \underline{\alpha} > 0,\ \forall v \in V,\ |a(v, \mathbb{T}v)| \geq \underline{\alpha} \|v\|_V^2.$$

Lorsque la forme est hermitienne (cas $W = V$) le caractère bijectif de $\mathbb{T}$ n'est donc plus requis.

Théorème C.14 *Soit $a(\cdot, \cdot)$ une forme continue et sesquilinéaire sur $V \times W$. La forme a est $\mathbb{T}$-coercive si, et seulement si, elle vérifie une condition de stabilité et la condition de solvabilité.*

Dans le cadre de la théorie inf-sup, un operateur $\mathbb{T}$ réalisant la $\mathbb{T}$-coercivité est parfois appelé un opérateur inf-sup.

Pour résumer, dans le cas où $W = V$, pour assurer que les problèmes (C.3-C.4) sont bien posés, et que l'opérateur correspondant A défini par (C.1) est un isomorphisme :

- une *condition suffisante* est que la forme a soit coercive (voir le théorème de Lax-Milgram C.6) ;
- une *condition nécessaire et suffisante* est que la forme a vérifie une condition de stabilité et la condition de solvabilité, ou de façon équivalente que la forme a soit $\mathbb{T}$-coercive (voir les théorèmes C.11 et C.14).

C.3 Problèmes mixtes

On introduit les problèmes mixtes, qui sont des problèmes à plusieurs inconnues, sous la forme abstraite suivante, appelée *formulation variationnelle mixte* [20, 4] :

$$\left\{\begin{array}{l} \textit{Trouver } (u,p) \in V \times Q \textit{ tel que} \\ \forall v \in V, \quad a(u,v) + b(v,p) = \langle f, v\rangle \quad \text{(i)} \\ \forall q \in Q, \qquad\quad b(u,q) = 0. \qquad \text{(ii)} \end{array}\right. \tag{C.9}$$

Les espaces V et Q sont des espaces de Hilbert, la forme a est sesquilinéaire et continue sur $V \times V$, la forme b est sesquilinéaire et continue sur $V \times Q$, et $f \in V'$ est la donnée. L'équation (C.9)-(ii) est souvent appelée *équation de contrainte*.
Comment garantir le caractère bien posé du problème (C.9) ?

Soit $K := \{v \in V \ : \ \forall q \in Q,\ b(v,q) = 0\}$ le noyau de la forme b dans V. Comme b est continue, K est un sous-espace vectoriel fermé de V : muni de $\|\cdot\|_V$ et $(\cdot,\cdot)_V$, K est un espace de Hilbert. On a donc la décomposition orthogonale $V = K \overset{\perp}{\oplus} K^\perp$. D'après (C.9)-(ii), on a $u \in K$, et on peut récrire le problème (C.9) sous la forme équivalente :

$$\left\{\begin{array}{l} \textit{Trouver } (u,p) \in K \times Q \textit{ tel que} \\ \forall v \in K, \qquad\quad a(u,v) = \langle f, v\rangle \quad \text{(i)} \\ \forall v' \in K^\perp, \quad a(u,v') + b(v',p) = \langle f, v'\rangle. \quad \text{(ii)} \end{array}\right. \tag{C.10}$$

Considèrons la résolution du problème (C.10) en deux étapes :

1. On résout d'abord le problème en u, qui s'écrit sous forme variationnelle :

$$\left\{\begin{array}{l} \textit{Trouver } u \in K \textit{ tel que} \\ \forall v \in K,\ a(u,v) = \langle f, v\rangle. \end{array}\right.$$

Ce problème admet une unique solution, dépendant continûment de la donnée, si *par exemple* (voir les remarques C.16) la forme a est coercive sur $K \times K$, c'est-à-dire s'il existe une constante $\alpha > 0$ telle que pour tout $v \in K$, $|a(v,v)| \geq \alpha \|v\|_V^2$ (cette condition est suffisante d'après le théorème de Lax-Milgram C.6).
On fait cette hypothèse de coercivité par la suite. L'inconnue u est caractérisée par (C.10)-(i) et, si on considère que $f \in K'$, on a :

$$\|u\|_V \leq \frac{1}{\alpha}\|f\|_{K'}. \tag{C.11}$$

2. On résout ensuite le problème en p avec comme données $u \in V$ et $f \in (K')^\perp$, qui s'écrit :

$$\left\{\begin{array}{l} \textit{Trouver } p \in Q \textit{ tel que} \\ \forall v' \in K^\perp,\ b(v',p) = \langle f, v'\rangle - a(u,v'). \end{array}\right. \tag{C.12}$$

C'est une formulation variationnelle généralisée. Elle est bien posée si la forme b satisfait les conditions de Banach-Necas-Babuska (voir le théorème C.11), c'est-à-dire la *condition de stabilité* (C.7) :

$$\exists \beta > 0,\ \forall q \in Q, \quad \sup_{v' \in K^\perp \setminus \{0\}} \frac{|b(v',q)|}{\|v'\|_V} \geq \beta \, \|q\|_Q, \tag{C.13}$$

et la *condition de solvabilité* (C.8) :

$$\text{Soit } v' \in K^\perp \ : \ \forall q \in Q,\ b(v', q) = 0 \ \Rightarrow \ v' = 0. \tag{C.14}$$

Examinons ces deux conditions.
Pour la condition (C.14), par définition du noyau K on trouve que $v' \in K^\perp \cap K = \{0\}$. En d'autres termes, la condition (C.14) est toujours vérifiée.
Pour la condition (C.13), par définition, si $v \in K$, on a automatiquement $b(v, q) = 0$, d'où la condition finale pour assurer l'existence et l'unicité de p :

$$\exists \beta > 0,\ \forall q \in Q,\ \sup_{v \in V \setminus \{0\}} \frac{|b(v, q)|}{\|v\|_V} \geq \beta \, \|q\|_Q. \tag{C.15}$$

Dans ce cas, on a : $\|p\|_Q \leq \dfrac{1}{\beta} \dfrac{|b(v', p)|}{\|v'\|_V}$ pour un $v' \in K^\perp$ réalisant (C.13) avec $q = p$. D'où :

$$\|p\|_Q \leq \frac{1}{\beta} \left(\|f\|_{(K^\perp)'} + |||a|||\|u\|_V \right) \leq \frac{1}{\beta} \left(\|f\|_{(K^\perp)'} + \frac{|||a|||}{\alpha} \|f\|_{K'} \right). \tag{C.16}$$

La seconde inégalité est obtenue d'après la résolution du problème en u (voir l'étape **1.**). D'après ce qui précède, on peut établir le théorème suivant :

Théorème C.15 *Le problème (C.9) admet une unique solution si*
- *La forme sesquilinéaire et continue a est coercive sur $K \times K$;*
- *La forme sesquilinéaire et continue b vérifie la condition de stabilité* (C.15).

De plus, la solution (u, p) dépend continûment de la donnée $f \in V'$, cf. (C.11) et (C.16).

Remarque C.16 *Pour l'étape* **1.**, *et donc pour déterminer u, il faut et il suffit que la forme a vérifie une condition de stabilité et une condition de solvabilité sur $K \times K$.*
On peut aussi résoudre des problèmes mixtes où (C.9)-(ii) *est remplacé par*

$$\forall q \in Q,\ b(u, q) = \langle g, q \rangle,$$

avec $g \in Q'$ une seconde donnée. On obtient le caractère bien posé de ce problème mixte (avec contrainte non-nulle) sous les mêmes *hypothèses que celles du théorème C.15, cf. [4].*

C.4 Eléments de théorie de l'approximation

Nous nous intéressons à l'*approximation* de la formulation variationnelle (C.5), posée dans des espaces de Hilbert V et W de *dimensions infinies*, avec $f \in W'$ et $a(\cdot, \cdot)$ une forme continue et sesquilinéaire sur $V \times W$. On note A l'opérateur de $\mathcal{L}(V, W')$ associé à la forme a via (C.1). Nous supposons que (C.5) est bien posé.

C.4.1 Approximation des formes

Pour approcher le problème (C.5), restons pour l'instant dans $V \times W$, et considérons une suite infinie $(a^\eta)_{\eta > 0}$ de formes continues et sesquilinéaires sur $V \times W$, ainsi que $(f^\eta)_{\eta > 0}$, une suite infinie d'éléments de W'. Par convention, le paramètre η tend vers 0 et, lorsque

η tend vers 0, les formes a^η (resp. f^η) "approchent" la forme de départ a (resp. f) en un sens à préciser. On définit les problèmes approchés, un pour chaque $\eta > 0$, par

$$\left\{ \begin{array}{l} \textit{Trouver } u^\eta \in V \textit{ tel que} \\ \forall w \in W,\ a^\eta(u^\eta, w) = \langle f^\eta, w \rangle. \end{array} \right. \tag{C.17}$$

A l'aide d'un opérateur, ces problèmes approchés s'écrivent

$$\left\{ \begin{array}{l} \textit{Trouver } u^\eta \in V \textit{ tel que} \\ A^\eta u^\eta = f^\eta \text{ dans } W', \end{array} \right. \tag{C.18}$$

avec $A^\eta \in \mathcal{L}(V, W')$ défini par

$$\forall (v, w) \in V \times W,\ \langle A^\eta v, w \rangle = a^\eta(v, w).$$

L'hypothèse de base est que le problème approché (C.17) est bien posé. D'après le théorème C.11, on sait que la forme a^η est stable (avec une constante de stabilité notée $\alpha^\eta > 0$ ci-dessous), et de plus $Im(A^\eta) = W'$.
Pour $\eta > 0$, on introduit le *terme de consistance*

$$Cons_{a,\eta}(v) = \|(A - A^\eta)v\|_{W'} = \sup_{w \in W \setminus \{0\}} \frac{|(a - a^\eta)(v, w)|}{\|w\|_W}, \tag{C.19}$$

qui exprime les écarts entre la forme ou l'opérateur exact et leurs approximations.

Théorème C.17 *Soit $\eta > 0$ donné. Si que le problème (C.17) est bien posé, alors*

$$\|u - u^\eta\|_V \leq |||A^{-1}|||(Cons_{a,\eta}(u^\eta) + \|f - f^\eta\|_{W'}). \tag{C.20}$$

Remarquons que $|||A - A^\eta||| = \sup_{v \in V \setminus \{0\}} \{Cons_{a,\eta}(v)/\|v\|_V\}$. Par conséquent, si la suite de solutions approchées $(u^\eta)_{\eta > 0}$ est bornée, on en déduit sa convergence vers la soution exacte u sous les conditions

$$\lim_{\eta \to 0} \|f - f^\eta\|_{W'} = 0 \text{ et } \lim_{\eta \to 0} |||A - A^\eta||| = 0. \tag{C.21}$$

Par ailleurs, pour démontrer que la suite $(u^\eta)_{\eta > 0}$ est bornée, on peut utiliser le résultat ci-dessous.

Théorème C.18 *Soit $\eta > 0$ donné. Si que le problème (C.17) est bien posé, alors*

$$\|u - u^\eta\|_V \leq \frac{1}{\alpha^\eta}(Cons_{a,\eta}(u) + \|f - f^\eta\|_{W'}). \tag{C.22}$$

C.4.2 Approximations de Galerkin

Considérons d'abord le cas $W = V$, ce qui permet d'introduire les méthodes d'approximation dites de *Galerkin*. Ensuite nous examinerons le cas général des méthodes d'approximation dites de *Petrov-Galerkin*, pour lesquelles $W \neq V$ est possible (voir §C.4.3). Pour approcher le problème (C.3), soit $(V_\delta)_{\delta > 0}$ une suite de sous-espaces vectoriels de dimension finie. On note $n(\delta)$ la dimension de V_δ. Par convention, le paramètre δ tend vers 0 et on suppose que $\lim_{\delta \to 0} n(\delta) = +\infty$, de sorte que V_δ peut "approcher" V en un certain sens

(voir ci-dessous (C.26), (C.30) et (C.40)).
Lorsque $(V_\delta)_{\delta>0}$ sont des sous-espaces vectoriels de V, on parle d'*approximation conforme*.
On peut alors définir les problèmes approchés, ou discrets (un pour chaque valeur de δ), par

$$\left\{ \begin{array}{l} \textit{Trouver } u_\delta \in V_\delta \textit{ tel que} \\ \forall v_\delta \in V_\delta,\ a(u_\delta, v_\delta) = \langle f, v_\delta \rangle. \end{array} \right. \tag{C.23}$$

A l'aide d'un opérateur, ces problèmes discrets s'écrivent

$$\left\{ \begin{array}{l} \textit{Trouver } u_\delta \in V_\delta \textit{ tel que} \\ A_\delta u_\delta = f \text{ dans } (V_\delta)', \end{array} \right. \tag{C.24}$$

où $A_\delta \in \mathcal{L}(V_\delta, (V_\delta)')$, et f est considéré comme un élément de $(V_\delta)'$:

$$\forall v_\delta, w_\delta \in V_\delta,\ \langle A_\delta v_\delta, w_\delta \rangle_{V_\delta} = a(v_\delta, w_\delta),\ \langle f, v_\delta \rangle_{V_\delta} = \langle f, v_\delta \rangle_V.$$

On a un résultat suffisant sur le caractère bien posé des problèmes discrets.

Proposition C.19 *Si la forme a est coercive sur $V \times V$, alors le problème discret est bien posé pour tout $\delta > 0$.*

Lorsque la forme a est coercive, on note $\alpha > 0$ une constante de coercivité.
Ci-dessous, nous proposons des outils pour estimer l'*erreur* $u - u_\delta$ entre la solution exacte u et la solution discrète ou approchée u_δ : on parle d'*estimation d'erreur*.

Théorème C.20 `(lemme de Céa)` *Si la forme a est coercive sur $V \times V$, alors*

$$\|u - u_\delta\|_V \leq \frac{|||a|||}{\alpha} \inf_{v_\delta \in V_\delta} \|u - v_\delta\|_V. \tag{C.25}$$

Remarque C.21 *La donnée f qui n'apparaît pas explicitement dans (C.25) est présente via la solution u, qui dépend linéairement de f (voir la définition C.5 : $\|u\|_V \leq C\|f\|_{V'}$ avec $C > 0$ indépendant de f).*
Par ailleurs, on a $\inf_{v_\delta \in V_\delta} \|u - v_\delta\|_V \leq \|u - u_\delta\|_V$, et il résulte de l'estimation d'erreur (C.25) que les deux quantités $\inf_{v_\delta \in V_\delta} \|u - v_\delta\|_V$ et $\|u - u_\delta\|_V$ sont du même ordre.

Ci-dessus, le *terme d'approximabilité* $\inf_{v_\delta \in V_\delta} \|u - v_\delta\|_V$ correspond à la capacité d'approximation des éléments de V par ceux de V_δ. Puisque la dimension de V_δ tend vers l'infini lorsque δ tend vers 0, on s'attend à ce que le terme d'approximabilité tende vers 0, ce qui entraîne que l'erreur tend elle aussi vers 0. Pour ce faire, on a besoin d'une hypothèse du type :

$$\begin{array}{c} \exists V_+ \subset V,\ V_+ \text{ dense dans } V \text{ et } \forall \delta,\ \exists r_\delta \in \mathcal{L}(V_+, V_\delta) \text{ tels que} \\ \forall v \in V_+,\ \lim_{\delta \to 0} \|v - r_\delta v\|_V = 0. \end{array} \tag{C.26}$$

L'hypothèse (C.26) peut être vue comme une *propriété d'approximabilité minimale*. En effet, on peut facilement établir le résultat ci-dessous.

Proposition C.22 *Si la propriété d'approximabilité minimale (C.26) est vraie, alors*

$$\forall v \in V,\ \lim_{\delta \to 0} \left(\inf_{v_\delta \in V_\delta} \|v - v_\delta\|_V \right) = 0. \tag{C.27}$$

Lorsque la propriété d'approximabilité minimale est vraie, on déduit donc du lemme de Céa que l'error $u - u_\delta$ tend vers 0 dans V quand δ tend vers 0.

Théorème C.23 *Si la forme a est coercive sur $V \times V$ et si la propriété d'approximabilité minimale (C.26) est vraie, alors l'erreur tend vers 0 quand δ tend vers 0 :*

$$\lim_{\delta \to 0} \|u - u_\delta\|_V = 0. \tag{C.28}$$

On peut proposer des estimations d'erreurs plus "précises". Celles-ci reposent sur la définition de méthodes d'approximation *ad hoc*. Pour cela, on peut supposer que les données f sont plus régulières. Plus précisément, pour H un sous-espace vectoriel normé de V', introduisons :

$$\tilde{V} := \{\tilde{v} \in V \ : \ \exists f \in H \text{ tel que } \tilde{v} \text{ solution de (C.3) avec la donnée } f\}$$

muni de la norme $\|\tilde{v}\|_{\tilde{V}} = \|f\|_H$. L'idée est de définir des opérateurs discrets allant de $\tilde{V}$ dans V_δ. Par exemple, il peut s'agir d'opérateurs d'interpolation ou de projection $(\pi_\delta)_{\delta>0}$, qui dépendent de la méthode d'approximation choisie. On fait une hypothèse du type :

$$\begin{gathered}\forall \delta, \ \exists \pi_\delta \in \mathcal{L}(\tilde{V}, V_\delta) \text{ et } \exists \epsilon \ : \ \mathbb{R}^+ \to \mathbb{R}^+, \lim_{\delta \to 0} \epsilon(\delta) = 0 \text{ tels que} \\ \forall \tilde{v} \in \tilde{V}, \ \|\tilde{v} - \pi_\delta \tilde{v}\|_V \leq \epsilon(\delta) \, \|\tilde{v}\|_{\tilde{V}}.\end{gathered} \tag{C.29}$$

Par construction, $\pi_\delta \tilde{v}$ peut servir à borner $\inf_{v_\delta \in V_\delta} \|\tilde{v} - v_\delta\|_V$. Ceci conduit à la *propriété d'approximabilité uniforme*, à savoir

$$\sup_{\tilde{v} \in \tilde{V} \setminus \{0\}} \left(\frac{1}{\|\tilde{v}\|_{\tilde{V}}} \inf_{v_\delta \in V_\delta} \|\tilde{v} - v_\delta\|_V \right) \leq \epsilon(\delta), \tag{C.30}$$

et à des estimations d'erreur améliorées pour les solutions appartenant $\tilde{V}$, où cette fois la mesure de la donnée f apparaît explicitement, sous la forme $\|f\|_H$.

Théorème C.24 *Si la forme a est coercive sur $V \times V$ et si la propriété d'approximabilité uniforme (C.30) est vraie, alors pour $f \in H$, l'erreur est bornée par :*

$$\|u - u_\delta\|_V \leq \frac{|||a|||}{\alpha} \, \epsilon(\delta) \, \|f\|_H. \tag{C.31}$$

Considérons maintenant des problèmes discrets qui comportent des formes différentes des formes a et f. Cette situation se produit lorsque l'action des formes sur les éléments de V_δ est calculée de manière approchée (par exemple, l'utilisation de formules de quadrature pour calculer des intégrales). Dans ce cas, on écrit les problèmes discrets sous la forme

$$\begin{cases} \textit{Trouver } u_\delta \in V_\delta \textit{ tel que} \\ \forall v_\delta \in V_\delta, \ a_\delta(u_\delta, v_\delta) = \langle f_\delta, v_\delta \rangle. \end{cases} \tag{C.32}$$

Pour obtenir le caractère bien posé et pour en déduire des estimations d'erreur, on introduit la notion "discrète" de coercivité uniforme.

Définition C.25 *La famille de formes sesquilinéaires $(a_\delta)_{\delta>0}$ est* uniformément V_δ-coercive *si*

$$\exists \alpha^\star > 0, \ \forall \delta, \ \forall v_\delta \in V_\delta, \ |a_\delta(v_\delta, v_\delta)| \geq \alpha^\star \, \|v_\delta\|_V^2.$$

On en déduit le caractère bien posé des problèmes discrets comme précédemment.

Proposition C.26 *Si la famille de formes sesquilinéaires $(a_\delta)_{\delta>0}$ est uniformément V_δ-coercive, alors le problème discret (C.32) est bien posé.*

Pour $\delta > 0$ et $v_\delta \in V_\delta$, introduisons

$$Cons_{f,\delta} = \sup_{w_\delta \in V_\delta \setminus \{0\}} \frac{|\langle f - f_\delta, w_\delta \rangle|}{\|w_\delta\|_V}, \tag{C.33}$$

$$Cons_{a,\delta}(v_\delta) = \sup_{w_\delta \in V_\delta \setminus \{0\}} \frac{|(a - a_\delta)(v_\delta, w_\delta)|}{\|w_\delta\|_V}, \tag{C.34}$$

$$Cons_{a,\delta} = \sup_{\substack{v_\delta \in V_\delta \setminus \{0\} \\ w_\delta \in V_\delta \setminus \{0\}}} \frac{|(a - a_\delta)(v_\delta, w_\delta)|}{\|v_\delta\|_V \, \|w_\delta\|_V}. \tag{C.35}$$

Ce sont des termes de consistance, au sens où ils expriment les écarts entre les formes exactes (a et f) et les formes approchées (a_δ et f_δ).

Théorème C.27 `(Premier lemme de Strang)` *Si la famille de formes sesquilinéaires $(a_\delta)_{\delta>0}$ est uniformément V_δ-coercive, alors l'erreur $\|u - u_\delta\|_V$ est bornée par*

$$\|u - u_\delta\|_V \leq C \left[\inf_{v_\delta \in V_\delta} \left(\|u - v_\delta\|_V + Cons_{a,\delta}(v_\delta) \right) + Cons_{f,\delta} \right], \tag{C.36}$$

avec $C := \max\left(\frac{1}{\alpha^\star}, \frac{|||a|||}{\alpha^\star} + 1\right) > 0$ indépendant de δ.

Si les termes de consistance tendent vers 0, à savoir :

$$\lim_{\delta \to 0} Cons_{f,\delta} = 0 \quad \text{et} \quad \lim_{\delta \to 0} Cons_{a,\delta} = 0, \tag{C.37}$$

on en déduit la convergence sous réserve que la propriété d'approximabilité minimale soit vraie (cf. proposition C.22).

Remarque C.28 *Dans le cadre précédent, c'est-à-dire avec des formes discrètes différentes des formes exactes, la notion "discrète" de famille $(a_\delta)_{\delta>0}$ uniformément V_δ-coercive a remplacé celle de forme exacte a coercive.*

Passons à l'*approximation non-conforme*, c'est-à-dire lorsque $V_\delta \not\subset V$. Dans ce cas, on doit d'abord définir une norme sur $V + V_\delta$, notée $\| \cdot \|_{V,\delta}$, telle que $\|v\|_{V,\delta} = \|v\|_V$ pour tout $v \in V$. De la même façon, les formes doivent être prolongées à $V + V_\delta$: typiquement, on définit a_δ^{NC} et f_δ^{NC} sur $V + V_\delta$ vérifiant

$$\forall v, w \in V, \ a_\delta^{NC}(v, w) = a(v, w) \text{ et } \langle f_\delta^{NC}, v \rangle = \langle f, v \rangle.$$

Dans le cadre non-conforme, les problèmes discrets sont alors écrits sous la forme

$$\left\{ \begin{array}{l} \textit{Trouver } u_\delta \in V_\delta \textit{ tel que} \\ \forall v_\delta \in V_\delta, \ a_\delta^{NC}(u_\delta, v_\delta) = \langle f_\delta^{NC}, v_\delta \rangle. \end{array} \right. \tag{C.38}$$

Définition C.29 *La famille de formes sesquilinéaires* $(a_\delta^{NC})_{\delta>0}$ *est* uniformément $(V + V_\delta) \times V_\delta$-continue *si*

$$\exists A^{NC} > 0,\ \forall \delta,\ \forall (v, w_\delta) \in (V + V_\delta) \times V_\delta,\ |a_\delta^{NC}(v, w_\delta)| \leq A^{NC} \|v\|_{V,\delta} \|w_\delta\|_{V,\delta}.$$

Définition C.30 *La famille de formes sesquilinéaires* $(a_\delta^{NC})_{\delta>0}$ *est* uniformément V_δ-coercive *si*

$$\exists \alpha^{NC} > 0,\ \forall \delta,\ \forall v_\delta \in V_\delta,\ |a_\delta^{NC}(v_\delta, v_\delta)| \geq \alpha^{NC} \|v_\delta\|_{V,\delta}^2.$$

Proposition C.31 *Si la famille de formes sesquilinéaires* $(a_\delta^{NC})_{\delta>0}$ *est uniformément* V_δ*-coercive, alors le problème discret (C.38) est bien posé.*

On introduit ensuite, pour $\delta > 0$, un nouveau terme de consistance qui exprime les écarts entre les formes prolongées :

$$Cons_\delta^{NC} = \sup_{w_\delta \in V_\delta \setminus \{0\}} \frac{|a_\delta^{NC}(u, w_\delta) - \langle f_\delta^{NC}, w_\delta \rangle|}{\|w_\delta\|_{V,\delta}}.$$

Théorème C.32 (`Deuxième lemme de Strang`) *Si la famille de formes sesquilinéaires* $(a_\delta^{NC})_{\delta>0}$ *est uniformément* $(V + V_\delta) \times V_\delta$*-continue et uniformément* V_δ*-coercive, alors l'erreur* $\|u - u_\delta\|_{V,\delta}$ *est majorée par*

$$\|u - u_\delta\|_{V,\delta} \leq C \left[\inf_{v_\delta \in V_\delta} \|u - v_\delta\|_{V,\delta} + Cons_\delta^{NC} \right], \tag{C.39}$$

avec $C := \max\left(\frac{1}{\alpha^{NC}}, \frac{A^{NC}}{\alpha^{NC}} + 1\right) > 0$ *indépendant de* δ.

Dans le cadre non-conforme, on aura donc convergence si :

$$\begin{array}{c} \exists V_+ \subset V,\ V_+ \text{ dense dans } V \text{ et } \forall \delta,\ \exists r_\delta^{NC} \in \mathcal{L}(V_+, V_\delta) \text{ tels que} \\ \forall v \in V_+,\ \lim_{\delta \to 0} \|v - r_\delta^{NC} v\|_{V,\delta} = 0\,, \end{array} \tag{C.40}$$

et si le terme de consistance tend vers 0 :

$$\lim_{\delta \to 0} \left[\sup_{f \in V',\ \|f\|_{V'}=1} Cons_\delta^{NC} \right] = 0. \tag{C.41}$$

Théorème C.33 *Si la famille de formes sesquilinéaires* $(a_\delta^{NC})_{\delta>0}$ *est uniformément* $(V + V_\delta) \times V_\delta$*-continue et uniformément* V_δ*-coercive, si le terme de consistance* $Cons_\delta^{NC}$ *tend vers* 0 *et si la propriété d'approximabilité minimale (C.40) est vraie, alors l'erreur tend vers* 0 *quand* δ *tend vers* 0 *:*

$$\lim_{\delta \to 0} \|u - u_\delta\|_{V,\delta} = 0. \tag{C.42}$$

C.4.3 Approximations de Petrov-Galerkin

Dans ce paragraphe on étudie un cas plus général, à savoir que $V = W$ ou $V \neq W$ sont possibles. On étudie maintenant l'approximation du problème (C.5) : soient $(V_\delta)_{\delta>0}$ et $(W_\delta)_{\delta>0}$ deux suites de sous-espaces vectoriels de dimension finie. Comme précédemment (cf. introduction C.4.2), on veut que V_δ (resp. W_δ) "approche" V (resp. W) en un certain sens.
Dans le cas de l'*approximation conforme*, on a $V_\delta \subset V$ et $W_\delta \subset W$ pour tout $\delta > 0$. Les problèmes approchés, ou discrets, sont

$$\left\{ \begin{array}{l} \textit{Trouver } u_\delta \in V_\delta \textit{ tel que} \\ \forall w_\delta \in W_\delta,\ a_\delta(u_\delta, w_\delta) = \langle f_\delta, w_\delta \rangle, \end{array} \right. \tag{C.43}$$

avec des formes discrètes a_δ et f_δ qui peuvent être différentes respectivement de a et f. A l'aide d'un opérateur, les problèmes discrets s'écrivent (C.24), l'équation étant posée dans $(W_\delta)'$.

Définition C.34 *La famille de formes sesquilinéaires* $(a_\delta)_{\delta>0}$ *est* uniformément $V_\delta \times W_\delta$-stable *si*

$$\exists \alpha_\dagger > 0,\ \forall \delta,\ \forall v_\delta \in V_\delta,\quad \sup_{w_\delta \in W_\delta \setminus \{0\}} \frac{|a_\delta(v_\delta, w_\delta)|}{\|w_\delta\|_W} \geq \alpha_\dagger \, \|v_\delta\|_V. \tag{C.44}$$

On peut également passer par la T-coercivité discrète uniforme.

Définition C.35 *La famille des formes* $(a_\delta)_\delta$ *est* uniformément $\mathtt{T}_\delta$-coercive *si, et seulement si*

$$\begin{array}{c} \exists \alpha^\star, \beta^\star > 0,\ \forall \delta > 0,\ \exists \mathtt{T}_\delta \in \mathcal{L}(V_\delta, W_\delta), \\ |||\mathtt{T}_\delta||| \leq \beta^\star \textit{ et } \forall v_\delta \in V_\delta,\ |a_\delta(v_\delta, \mathtt{T}_\delta v_\delta)| \geq \alpha^\star \|v_\delta\|_V^2. \end{array} \tag{C.45}$$

Remarque C.36 *En règle générale, la* $\mathbb{T}$*-coercivité discrète uniforme se déduit simplement de la* $\mathbb{T}$*-coercivité exacte. Il suffit d'ajouter des* $_\delta$*, et de choisir les espaces d'approximation compatibles avec les conditions requises pour la "bonne" définition de l'opérateur.*

On a le résultat suivant, à rapprocher du théorème C.14.

Théorème C.37 *Soit une famille* $(a_\delta)_\delta$ *de formes sesquilinéaires, continues et uniformément bornées. Les deux assertions ci-dessous sont équivalentes :*

- *la famille des formes* $(a_\delta)_\delta$ *est uniformément* $V_\delta \times W_\delta$*-stable ;*
- *la famille des formes* $(a_\delta)_\delta$ *est uniformément* $\mathtt{T}_\delta$*-coercive.*

Le caractère bien posé est une conséquence des conditions (C.44) ou (C.45).

Proposition C.38 *Si la famille de formes sesquilinéaires* $(a_\delta)_{\delta>0}$ *est uniformément* $V_\delta \times W_\delta$*-stable (cf. (C.44)) et si* $dim V_\delta = dim W_\delta$ *pour tout* δ*, alors le problème discret (C.43) est bien posé pour tout* δ*.*

On peut également obtenir une estimation d'erreur avec des termes de consistence du type (C.33), (C.34) et (C.35) pour tout δ et $v_\delta \in V_\delta$, le supremum étant ici pris sur $w_\delta \in W_\delta \setminus \{0\}$.

Théorème C.39 (Premier lemme de Strang) *Si la famille de formes sesquilinéaires $(a_\delta)_{\delta>0}$ est uniformément $V_\delta \times W_\delta$-stable et si $\dim V_\delta = \dim W_\delta$ pour tout δ, alors l'erreur $\|u-u_\delta\|_V$ est bornée par*

$$\|u-u_\delta\|_V \leq C\left[\inf_{v_\delta\in V_\delta}\left(\|u-v_\delta\|_V + Cons_{a,\delta}(v_\delta)\right) + Cons_{f,\delta}\right], \tag{C.46}$$

avec $C := \max\left(\frac{1}{\alpha_\dagger}, \frac{|||a|||}{\alpha_\dagger}+1\right) > 0$ *indépendant de* δ.

Dans le cas d'une *approximation non-conforme* ($V_\delta \not\subset V$, avec la norme $\|\cdot\|_{V,\delta}$ sur $V+V_\delta$, resp. $W_\delta \not\subset W$, avec la norme $\|\cdot\|_{W,\delta}$ sur $W+W_\delta$), on peut garantir le caractère bien posé, sous les hypothèses $\dim V_\delta = \dim W_\delta$ pour tout $\delta > 0$, si la famille des formes $(a_\delta^{NC})_{\delta>0}$ est uniformément $V_\delta \times W_\delta$-stable. Si on ajoute l'hypothèse de $(V+V_\delta)\times W_\delta$-continuité uniforme, on aboutit à une estimation d'erreur.
Dans le cadre non-conforme, le problème discret s'écrit

$$\begin{cases} \textit{Trouver } u_\delta \in V_\delta \textit{ tel que} \\ \forall w_\delta \in W_\delta,\ a_\delta^{NC}(u_\delta, w_\delta) = \langle f_\delta^{NC}, w_\delta\rangle, \end{cases} \tag{C.47}$$

et les définitions sont les suivantes.

Définition C.40 *La famille de formes sesquilinéaires $(a_\delta^{NC})_{\delta>0}$ est* uniformément $V_\delta\times W_\delta$-stable *si*

$$\exists \alpha_\dagger^{NC} > 0,\ \forall\delta,\ \forall v_\delta \in V_\delta,\ \sup_{w_\delta\in W_\delta\setminus\{0\}} \frac{|a_\delta^{NC}(v_\delta, w_\delta)|}{\|w_\delta\|_{W,\delta}} \geq \alpha_\dagger^{NC}\,\|v_\delta\|_{V,\delta}. \tag{C.48}$$

Elle est uniformément $(V+V_\delta)\times W_\delta$-continue *si*

$$\exists A_\dagger^{NC} > 0,\ \forall\delta,\ \forall (v,w_\delta)\in (V+V_\delta)\times W_\delta,\ |a_\delta^{NC}(v,w_\delta)| \leq A_\dagger^{NC}\,\|v\|_{V,\delta}\,\|w_\delta\|_{W,\delta}.$$

Proposition C.41 *Si la famille de formes sesquilinéaires $(a_\delta^{NC})_{\delta>0}$ est uniformément $V_\delta\times W_\delta$-stable (cf. (C.48)) et si $\dim V_\delta = \dim W_\delta$ pour tout δ, alors le problème discret (C.47) est bien posé pour tout δ.*

Introduisons, pour $\delta > 0$, le terme de consistence exprimant les écarts entre les formes prolongées :

$$Cons_{\delta,\dagger}^{NC} = \sup_{w_\delta\in W_\delta\setminus\{0\}} \frac{|a_\delta^{NC}(u,w_\delta) - \langle f_\delta^{NC}, w_\delta\rangle|}{\|w_\delta\|_{W,\delta}}.$$

Théorème C.42 (Deuxième lemme de Strang) *Si la famille de formes sesquilinéaires $(a_\delta^{NC})_{\delta>0}$ est uniformément $(V+V_\delta)\times W_\delta$-continue et uniformément $V_\delta\times W_\delta$-stable et si $\dim V_\delta = \dim W_\delta$ pour tout δ, alors l'erreur $\|u-u_\delta\|_{V,\delta}$ est majorée par*

$$\|u-u_\delta\|_{V,\delta} \leq C\left[\inf_{v_\delta\in V_\delta}\|u-v_\delta\|_{V,\delta} + Cons_{\delta,\dagger}^{NC}\right], \tag{C.49}$$

avec $C := \max\left(\frac{1}{\alpha_\dagger^{NC}}, \frac{A_\dagger^{NC}}{\alpha_\dagger^{NC}}+1\right) > 0$ *indépendant de* δ.

C.4.4 Approximation des formes et des espaces

Il est possible de mélanger les techniques d'approximation de §C.4.1 d'une part, et de §C.4.2 ou de §C.4.3 d'autre part. Partant du problème (C.17) pour un certain paramètre $\eta > 0$, on peut ensuite résoudre un problème approché de celui-ci, caractérisé par un paramètre $\delta > 0$. Pour optimiser les résultats sur l'erreur $u - u_\delta^\eta$, il faudra relier les paramètres η et δ entre eux.

Annexe D

Distributions et espaces fonctionnels

Nous renvoyons à [3, §2] pour plus de détails. Ci-dessous, on se place dans $\mathbb{R}^d$, avec $d \in \mathbb{N} \setminus \{0\}$.
On appelle multi-indice un $\alpha = (\alpha_1, \cdots, \alpha_d) \in \mathbb{N}^d$, et on note $|\alpha| = \sum_{j=1,d} \alpha_j$. La dérivée partielle d'ordre α est notée

$$\partial_\alpha f = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}}.$$

$d\boldsymbol{x} = dx_1 dx_2 \cdots dx_d$ est la mesure de Lebesgue dans $\mathbb{R}^d$.
On dit que $\mathcal{O}$ est un **domaine** de $\mathbb{R}^d$ si, et seulement si, $\mathcal{O}$ est un sous-ensemble de $\mathbb{R}^d$, ouvert, borné, connexe, dont la frontière $\partial\mathcal{O}$ est "suffisamment régulière" [57]. Le vecteur $\boldsymbol{n}$ désigne le vecteur normal unitaire sortant de $\partial\mathcal{O}$. On considèrera de façon générique $\Gamma \subset \partial\mathcal{O}$ une partie de $\partial\mathcal{O}$ de mesure non-nulle, elle-même de frontière $\partial\Gamma$ "suffisamment régulière".
Enfin, $C^\infty(\overline{\mathcal{O}})$ est composé des restrictions à $\overline{\mathcal{O}}$ des fonctions de $C^\infty(\mathbb{R}^d)$.

D.1 Distributions

On rappelle que $\mathcal{D}(\mathcal{O})$ est l'espace des fonctions $\mathcal{C}^\infty$ à support [58] compact dans $\mathcal{O}$. On utilise la **convergence des suites** pour définir la topologie sur $\mathcal{D}(\mathcal{O})$. Soit $(f_k)_k$ une suite d'éléments de $\mathcal{D}(\mathcal{O})$: on dit qu'elle *converge dans* $\mathcal{D}(\mathcal{O})$ *vers* f si, et seulement si,

(i) il existe un sous-ensemble compact K de $\mathcal{O}$ tel que $\mathrm{supp}(f_k) \subset K$, pour k suffisamment grand ;
(ii) pour tous les multi-indices $\alpha \in \mathbb{N}^d$, $(\partial_\alpha f_k)_k$ converge uniformément vers $\partial_\alpha f$ dans K.

Une forme linéaire et continue T définie sur $\mathcal{D}(\mathcal{O})$ est appelée une **distribution**. L'ensemble des distributions est un espace vectoriel, noté $\mathcal{D}'(\mathcal{O})$. Soit $T \in \mathcal{D}'(\mathcal{O})$ et $f \in \mathcal{D}(\mathcal{O})$: l'action de T sur f est écrit à l'aide de **crochets de dualité**, c'est-à-dire

$$\langle T, f \rangle.$$

57. On associe deux propriétés à cette définition : 1. l'ouvert $\mathcal{O}$ est localement d'un seul côté de sa frontière en tout point de celle-ci ; 2. (par exemple) la frontière est régulière par morceaux.

58. Par définition, le **support d'une fonction** $\boldsymbol{x} \mapsto f(\boldsymbol{x})$ est l'adhérence de l'ensemble des points $\boldsymbol{x}$ pour lesquels $f(\boldsymbol{x}) \neq 0$.

D'après la définition de la topologie de $\mathcal{D}(\mathcal{O})$, une forme linéaire T est *continue* si, et seulement si,

$$\forall (f_k)_k, f \in \mathcal{D}(\mathcal{O}) \text{ tels que } f_k \to f \text{ dans } \mathcal{D}(\mathcal{O}), \quad \langle T, f_k \rangle \to \langle T, f \rangle.$$

On peut enfin introduire la **dérivation au sens des distributions**. Soit $T \in \mathcal{D}'(\mathcal{O})$, sa $j^{\text{ème}}$ dérivée partielle ($1 \leq j \leq d$) est une distribution définie par

$$\forall f \in \mathcal{D}(\mathcal{O}), \quad \langle \frac{\partial T}{\partial x_j}, f \rangle = -\langle T, \frac{\partial f}{\partial x_j} \rangle.$$

Pour $1 \leq i \leq d$, l'opérateur ∂_i est continu de $\mathcal{D}'(\mathcal{O})$ dans lui-même. Dans la suite, les opérateurs de dérivation sont considérés au sens des distributions.

D.2 Espaces fonctionnels

On rappelle que l'espace fonctionnel $L^p(\mathcal{O})$ est composé des fonctions f mesurables au sens de Lebesgue sur $\mathcal{O}$, et telles que

$$\begin{cases} \text{si } 1 \leq p < \infty : & \|f\|_{L^p(\mathcal{O})} := \left\{ \displaystyle\int_{\mathcal{O}} |f|^p \, d\boldsymbol{x} \right\}^{1/p} < \infty \\ \text{si } p = \infty : & \|f\|_{L^\infty(\mathcal{O})} := \text{esssup}_{\boldsymbol{x} \in \mathcal{O}} |f(\boldsymbol{x})| < \infty \end{cases} .$$

On suppose que ces fonctions sont à valeurs complexes [59]. Pour $p \in [1, \infty]$, on écrit que $f_1 = f_2$ dans $L^p(\mathcal{O})$ pour dire que $f_1, f_2 \in L^p(\mathcal{O})$ et que $f_1(\boldsymbol{x}) = f_2(\boldsymbol{x})$ presque pour tout $\boldsymbol{x}$ dans $\mathcal{O}$. Pour $p \in [1, \infty]$, $L^p(\mathcal{O})$ est un espace de Banach muni de la norme $\| \cdot \|_{L^p(\mathcal{O})}$. Et, pour $1 \leq p < \infty$, $L^p(\mathcal{O})$ est séparable.
Enfin, pour $p \in [1, \infty]$, $W^{1,p}(\mathcal{O}) := \{ v \in L^p(\mathcal{O}) \ : \ \partial_i v \in L^p(\mathcal{O}), \ 1 \leq i \leq d \}$ est un espace de Banach muni de la norme produit.

On définit les espaces fonctionnels suivants sur $\mathcal{O}$, qui sont des espaces de Hilbert munis du produit scalaire associé :

— $L^2(\mathcal{O}), \quad (u, v)_{L^2(\mathcal{O})} = \displaystyle\int_{\mathcal{O}} u \, \overline{v} \, d\boldsymbol{x}$;
$\mathbf{L}^2(\mathcal{O}) := L^2(\mathcal{O})^d, \quad (\boldsymbol{p}, \boldsymbol{q})_{\mathbf{L}^2(\mathcal{O})} = \displaystyle\int_{\mathcal{O}} \boldsymbol{p} \cdot \overline{\boldsymbol{q}} \, d\boldsymbol{x}.$

— $H^1(\mathcal{O}) := \left\{ v \in L^2(\mathcal{O}) \ : \ \partial_i v \in L^2(\mathcal{O}), \ 1 \leq i \leq d \right\} = \left\{ v \in L^2(\mathcal{O}) \ : \ \mathbf{grad}\, v \in \mathbf{L}^2(\mathcal{O}) \right\}$,
$(u, v)_{H^1(\mathcal{O})} = (u, v)_{L^2(\mathcal{O})} + \sum_{i=1,d} (\partial_i u, \partial_i v)_{L^2(\mathcal{O})} = (u, v)_{L^2(\mathcal{O})} + (\mathbf{grad}\, u, \mathbf{grad}\, v)_{\mathbf{L}^2(\mathcal{O})}$,

pour $m \geq 2$: $H^m(\mathcal{O}) := \left\{ v \in L^2(\mathcal{O}) \ : \ \partial_i v \in H^{m-1}(\mathcal{O}), \ 1 \leq i \leq d \right\}$,
$(u, v)_{H^m(\mathcal{O})} = (u, v)_{L^2(\mathcal{O})} + \sum_{i=1,d} (\partial_i u, \partial_i v)_{H^{m-1}(\mathcal{O})}$,

pour $s \in \mathbb{R}^+ \setminus \mathbb{N}$: on écrit $s = m + \sigma$, $m \in \mathbb{N}$ et $\sigma \in]0, 1[$ et on définit $H^s(\mathcal{O})$ par interpolation entre $H^m(\mathcal{O})$ et $H^{m+1}(\mathcal{O})$, avec la convention $H^0(\mathcal{O}) = L^2(\mathcal{O})$;

59. Bien sûr, on peut les considérer à valeurs réelles ! Dans ce cas, on enlève la conjugaison plus bas...

pour $s \geq 0$: $H_0^s(\mathcal{O})$ est défini comme l'adhérence de $\mathcal{D}(\mathcal{O})$ dans $H^s(\mathcal{O})$;

pour $s \geq 0$, $C^\infty(\overline{\mathcal{O}})$ est dense dans $H^s(\mathcal{O})$.

— $\mathbf{H}(\mathrm{div}, \mathcal{O}) := \left\{ \boldsymbol{q} \in \mathbf{L}^2(\mathcal{O}) \ : \ \mathrm{div}\, \boldsymbol{q} \in L^2(\mathcal{O}) \right\}$,
$(\boldsymbol{p}, \boldsymbol{q})_{\mathbf{H}(\mathrm{div},\mathcal{O})} = (\boldsymbol{p}, \boldsymbol{q})_{\mathbf{L}^2(\mathcal{O})} + (\mathrm{div}\, \boldsymbol{p}, \mathrm{div}\, \boldsymbol{q})_{L^2(\mathcal{O})}$;

$\mathbf{H}_0(\mathrm{div}, \mathcal{O})$ est défini comme l'adhérence de $\mathcal{D}(\mathcal{O})$ dans $\mathbf{H}(\mathrm{div}, \mathcal{O})$;

$C^\infty(\overline{\mathcal{O}})^d$ est dense dans $\mathbf{H}(\mathrm{div}, \mathcal{O})$.

Comme $\mathcal{O}$ est un domaine de $\mathbb{R}^d$ ($d = 1, 2, 3$), on peut en particulier définir la normale unitaire sortante $\boldsymbol{n}$ presque partout sur la frontière $\partial\mathcal{O}$, et on a $\boldsymbol{n} \in \mathbf{L}^\infty(\partial\mathcal{O})$. Pour une partie $\Gamma \subset \partial\mathcal{O}$, on définit les espaces fonctionnels suivants sur Γ, qui sont des espaces de Hilbert munis du produit scalaire associé :

— $L^2(\Gamma) := \left\{ \mu \text{ mesurable sur } \Gamma \ : \ \displaystyle\int_\Gamma |\mu|^2 \, d\Gamma < +\infty \right\}, \quad (\mu, \mu')_{L^2(\Gamma)} = \displaystyle\int_\Gamma \mu \, \overline{\mu'} \, d\Gamma.$

— $H^{1/2}(\Gamma) = \left\{ \mu \in L^2(\Gamma) \ : \ \displaystyle\int_\Gamma \int_\Gamma \frac{|\mu(\boldsymbol{x}) - \mu(\boldsymbol{y})|^2}{|\boldsymbol{x} - \boldsymbol{y}|^d} d\Gamma_{\boldsymbol{x}} d\Gamma_{\boldsymbol{y}} < \infty \right\},$
$(\mu, \mu')_{H^{1/2}(\Gamma)} = (\mu, \mu')_{L^2(\Gamma)} + \displaystyle\int_\Gamma \int_\Gamma \frac{(\mu(\boldsymbol{x}) - \mu(\boldsymbol{y}))\,(\mu'(\boldsymbol{x}) - \mu'(\boldsymbol{y}))}{|\boldsymbol{x} - \boldsymbol{y}|^d} d\Gamma_{\boldsymbol{x}} d\Gamma_{\boldsymbol{y}}.$

— $\widetilde{H}^{1/2}(\Gamma) = \left\{ u \in H^{1/2}(\Gamma) \ : \ \tilde{u} \in H^{1/2}(\partial\mathcal{O}) \right\}$, où $\tilde{u} \in L^2(\partial\mathcal{O})$ est le prolongement de u à tout $\partial\mathcal{O}$, égal à 0 sur $\partial\mathcal{O} \setminus \bar{\Gamma}$.

On a les inclusions strictes

$$H^{1/2}(\Gamma) \subsetneq L^2(\Gamma) \subsetneq (H^{1/2}(\Gamma))' \text{ et } \widetilde{H}^{1/2}(\Gamma) \subsetneq L^2(\Gamma) \subsetneq (\widetilde{H}^{1/2}(\Gamma))'.$$

On note que si $\lambda \in L^2(\Gamma)$ et $\mu \in H^{1/2}(\Gamma)$, on a

$$\langle \lambda, \mu \rangle_{H^{1/2}(\Gamma)} = \int_\Gamma \lambda \, \mu \, d\Gamma.$$

D.3 Théorèmes de trace

Pour $u \in C^\infty(\overline{\mathcal{O}})$ on note $u_{|\Gamma}$ sa trace sur une partie $\Gamma \subset \partial\mathcal{O}$.

D.3.1 Trace des fonctions de $H^1(\mathcal{O})$

Théorème D.1 *L'application* **trace** $\gamma_0 \ : \ u \mapsto u_{|\partial\mathcal{O}}$ *de* $C^\infty(\overline{\mathcal{O}})$ *sur* $\partial\mathcal{O}$ *se prolonge par continuité en une application continue de* $H^1(\mathcal{O})$ *dans* $L^2(\partial\mathcal{O})$ *:*

$$\exists c > 0, \ \forall u \in H^1(\mathcal{O}), \quad \|\gamma_0(u)\|_{L^2(\partial\mathcal{O})} \leq c \|u\|_{H^1(\mathcal{O})}.$$

L'application trace est surjective et continue de $H^1(\mathcal{O})$ *sur* $H^{1/2}(\partial\mathcal{O})$ *:*

$\exists c > 0, \ \forall u \in H^1(\mathcal{O}), \quad \|\gamma_0(u)\|_{H^{1/2}(\partial\mathcal{O})} \leq c \|u\|_{H^1(\mathcal{O})}$;
$\exists c > 0, \ \forall \mu \in H^{1/2}(\partial\mathcal{O}), \quad \exists u \in H^1(\mathcal{O})$ *tel que* $\gamma_0(u) = \mu$ *et* $\|u\|_{H^1(\mathcal{O})} \leq c \|\mu\|_{H^{1/2}(\partial\mathcal{O})}$.

Pour le dernier point, l'élément $u \in H^1(\mathcal{O})$ est appelé un **relèvement** de $g \in H^{1/2}(\partial\mathcal{O})$.

On a l'identification

$$H^1_0(\mathcal{O}) = \left\{ v \in H^1(\mathcal{O}) \ : \ v_{|\partial\mathcal{O}} = 0 \right\}.$$

On peut définir de même une application trace sur Γ, de $H^1(\mathcal{O})$ dans $L^2(\Gamma)$, où la trace est notée $v_{|\Gamma}$. Cette application est surjective de $H^1(\mathcal{O})$ dans $H^{1/2}(\Gamma)$.

Soit

$$C^\infty_\Gamma(\overline{\mathcal{O}}) := \{ v \in C^\infty(\overline{\mathcal{O}}) \ : \ v = 0 \text{ dans un voisinage de } \Gamma \}.$$

Si on note $H^1_{0,\Gamma}(\mathcal{O})$ l'adhérence dans $H^1(\mathcal{O})$ de $C^\infty_\Gamma(\overline{\mathcal{O}})$, on peut démontrer que :

$$H^1_{0,\Gamma}(\mathcal{O}) = \left\{ v \in H^1(\mathcal{O}) \ : \ v_{|\Gamma} = 0 \right\}.$$

En outre, l'ensemble des traces sur Γ de $C^\infty_{\partial\mathcal{O}\setminus\bar{\Gamma}}(\overline{\mathcal{O}})$ est dense dans $L^2(\Gamma)$.

Enfin on note que si $v \in H^1(\mathcal{O})$ s'annule dans un voisinage de $\partial\Gamma$, alors $v_{|\Gamma} \in \widetilde{H}^{1/2}(\Gamma)$.

Théorème D.2 `(inégalité de Poincaré)` *Soit $\mathcal{O}$ un domaine de $\mathbb{R}^d$ et $\Gamma \subset \partial\mathcal{O}$ une partie de mesure non-nulle. Alors*

$$\exists C_P > 0, \ \forall v \in H^1_{0,\Gamma}(\mathcal{O}), \quad \|v\|_{L^2(\mathcal{O})} \leq C_P \, \|\, \mathbf{grad}\, v\|_{\mathbf{L}^2(\mathcal{O})}.$$

D.3.2 Trace normale des fonctions de $\mathbf{H}(\mathrm{div}, \mathcal{O})$

Théorème D.3 *L'application* **trace normale** $\gamma_1 : \boldsymbol{q} \mapsto \gamma_1(\boldsymbol{q}) = \boldsymbol{q} \cdot \boldsymbol{n}_{|\partial\mathcal{O}}$ *de $\mathcal{C}^\infty(\overline{\mathcal{O}})^d$ sur $\partial\mathcal{O}$ se prolonge par continuité en une application continue et surjective de $\mathbf{H}(\mathrm{div}, \mathcal{O})$ dans l'espace dual $(H^{1/2}(\partial\mathcal{O}))'$:*

$\exists c > 0, \ \forall \boldsymbol{q} \in \mathbf{H}(\mathrm{div}, \mathcal{O}), \quad \|\gamma_1(\boldsymbol{q})\|_{(H^{1/2}(\partial\mathcal{O}))'} \leq c \|\boldsymbol{q}\|_{\mathbf{H}(\mathrm{div},\mathcal{O})}$;
$\exists c > 0, \ \forall \mu \in (H^{1/2}(\partial\mathcal{O}))', \ \exists \boldsymbol{q} \in \mathbf{H}(\mathrm{div}, \mathcal{O})$ *tel que* $\gamma_1(\boldsymbol{q}) = \mu$ *et* $\|\boldsymbol{q}\|_{\mathbf{H}(\mathrm{div},\mathcal{O})} \leq c \|\mu\|_{(H^{1/2}(\partial\mathcal{O}))'}$.

Pour le dernier point, l'élément $\boldsymbol{q} \in \mathbf{H}(\mathrm{div}, \mathcal{O})$ est appelé un **relèvement** de $g \in (H^{1/2}(\partial\mathcal{O}))'$.
On a l'identification

$$\mathbf{H}_0(\mathrm{div}, \mathcal{O}) = \left\{ \boldsymbol{q} \in \mathbf{H}(\mathrm{div}, \mathcal{O}) \ : \ \boldsymbol{q} \cdot \boldsymbol{n}_{|\partial\mathcal{O}} = 0 \right\}.$$

Soit Γ une partie stricte de $\partial\mathcal{O}$. Pour $\boldsymbol{q} \in \mathbf{H}(\mathrm{div}, \mathcal{O})$, on sait d'après le théorème D.3 que $\boldsymbol{q} \cdot \boldsymbol{n}_{|\partial\mathcal{O}} \in (H^{1/2}(\partial\mathcal{O}))'$. Par contre en général $\boldsymbol{q} \cdot \boldsymbol{n}_{|\Gamma} \notin (H^{1/2}(\Gamma))'$, alors qu'on a automatiquement $\boldsymbol{q} \cdot \boldsymbol{n}_{|\Gamma} \in (\widetilde{H}^{1/2}(\Gamma))'$.

D.3.3 Formules d'intégration par parties

On a les formules d'intégration par parties suivantes :

$$\forall v \in H^1_0(\mathcal{O}), \ \forall \boldsymbol{q} \in \mathbf{H}(\mathrm{div}, \mathcal{O}), \ \int_{\mathcal{O}} (\boldsymbol{q} \cdot \overline{\mathbf{grad}\, v} + \mathrm{div}\, \boldsymbol{q}\, \overline{v})\, d\boldsymbol{x} = 0, \tag{D.1}$$

$$\forall v \in H^1(\mathcal{O}), \ \forall \boldsymbol{q} \in \mathbf{H}(\mathrm{div}, \mathcal{O}), \ \int_{\mathcal{O}} (\boldsymbol{q} \cdot \overline{\mathbf{grad}\, v} + \mathrm{div}\, \boldsymbol{q}\, \overline{v})\, d\boldsymbol{x} = \langle \boldsymbol{q} \cdot \boldsymbol{n}\, ,\, v \rangle_{H^{1/2}(\partial\mathcal{O})}. \tag{D.2}$$

Si on veut "découper" les crochets de dualité dans (D.2) sur Γ et $\Gamma' = \partial\mathcal{O} \setminus \bar{\Gamma}$ deux parties (strictes) de $\partial\mathcal{O}$, on a deux possibilités.
(D.2-i) Soit il faut supposer que la trace normale $\boldsymbol{q}\cdot\boldsymbol{n}_{|\partial\mathcal{O}}$ est plus régulière que $(H^{1/2}(\partial\mathcal{O}))'$, par exemple que $\boldsymbol{q}\cdot\boldsymbol{n}_{|\partial\mathcal{O}} \in L^2(\partial\mathcal{O})$. Dans ce cas, on peut écrire pour tout $v \in H^1(\mathcal{O})$ que :

$$\langle \boldsymbol{q}\cdot\boldsymbol{n}\,,\, v\rangle_{H^{1/2}(\partial\mathcal{O})} = \int_{\partial\mathcal{O}} \boldsymbol{q}\cdot\boldsymbol{n}\, v\, d\Gamma = \int_{\Gamma} \boldsymbol{q}\cdot\boldsymbol{n}\, v\, d\Gamma + \int_{\Gamma'} \boldsymbol{q}\cdot\boldsymbol{n}\, v\, d\Gamma.$$

(D.2-ii) Soit il faut choisir $v \in H^1(\mathcal{O})$ qui s'annule au voisinage de $\partial\Gamma = \partial\Gamma'$. Dans ce cas, $v_{|\Gamma} \in \widetilde{H}^{1/2}(\Gamma)$ et $v_{|\Gamma'} \in \widetilde{H}^{1/2}(\Gamma')$, et on peut écrire pour tout $\boldsymbol{q} \in \mathbf{H}(\mathrm{div},\mathcal{O})$ que :

$$\langle \boldsymbol{q}\cdot\boldsymbol{n}\,,\, v\rangle_{H^{1/2}(\partial\mathcal{O})} = \langle \boldsymbol{q}\cdot\boldsymbol{n}_{|\Gamma}\,,\, v_{|\Gamma}\rangle_{\widetilde{H}^{1/2}(\Gamma)} + \langle \boldsymbol{q}\cdot\boldsymbol{n}_{|\Gamma'}\,,\, v_{|\Gamma'}\rangle_{\widetilde{H}^{1/2}(\Gamma')}.$$

Bibliographie

[1] **L. M. Adams, H. F. Jordan**, Is SOR color-blind ?, *SIAM Journal on Scientific and Statistical Computing*, **7** (1986).

[2] **E. Bécache, P. Ciarlet, C. Hazard, E. Lunéville**, *La méthode des éléments finis. De la théorie à la pratique. II. Compléments*, Les Presses de l'ENSTA, Coll. Les Cours (2010).

[3] **F. Assous, P. Ciarlet, S. Labrunie**, *Mathematical foundations of computational electromagnetism*, Springer-Verlag, Berlin (2018).

[4] **D. Boffi, F. Brezzi, M. Fortin**, *Mixed and hybrid finite element methods and applications*, Springer-Verlag, Berlin (2013).

[5] **A.-S. Bonnet-Ben Dhia, P. Ciarlet, C. M. Zwölf**, Time harmonic wave diffraction problems in materials with sign-shifting coefficients, *J. Comput. Appl. Math.*, **234**, 1912–1919 (2010). (Corrigendum *J. Comput. Appl. Math.*, **234**, 2616 (2010))

[6] **H. Brezis**, *Analyse fonctionnelle. Théorie et applications*, Masson, Paris (1983). *Functional analysis, Sobolev spaces and Partial Differential Equations*, Universitext, Springer (2011).

[7] **F. Chatelin**, *Valeurs propres de matrices*, Masson, Paris (1988).

[8] **J. Céa**, *Optimisation : Théorie et algorithmes*, Dunod, Paris (1971)

[9] **P. Ciarlet, E. Jamelot, F. Kpadonou**, Domain decomposition methods for the diffusion equation with low-regularity solution, *Computers and Mathematics with Applications*, **74**, p. 2369–2384 (2017).

[10] **P. Ciarlet, L. Giret, E. Jamelot, F. Kpadonou**, Numerical analysis of the mixed finite element method for the neutron diffusion eigenproblem with heterogeneous coefficients, *Math. Mod. Num. Anal.*, **52**, p. 2003-2035 (2018).

[11] **P. Ciarlet, E. Lunéville**, *La méthode des éléments finis. De la théorie à la pratique. I. Concepts généraux*, Les Presses de l'ENSTA, Coll. Les Cours (2009).

[12] **P. Ciarlet, H. Zidani**, *Optimisation quadratique*, Cours AO 101, ENSTA.

[13] **P. G. Ciarlet**, *Introduction to numerical linear algebra and optimisation*, Cambridge Texts in Applied Mathematics, Cambridge (1989).

[14] **M. Costabel, M. Dauge, S. Nicaise**, *Singularities of electromagnetic fields in polyhedral domains*, M2AN Math. Model. Numer. Anal., **33**, 3, pp. 627-–649, (1999).

[15] **R. Dautray, J.-L. Lions**, *Analyse mathématique et calcul numérique pour les sciences et les techniques*, Volume 1 : Modèles physiques, Masson, Paris (1987).

[16] **R. Dautray, J.-L. Lions**, *Analyse mathématique et calcul numérique pour les sciences et les techniques*, Volume 6 : Méthodes intégrales et numériques, Masson, Paris (1988).

[17] **V. Dolean, P. Jolivet, F. Nataf**, *An introduction to domain decomposition methods*, Algorithms, theory, and parallel implementation, SIAM, Philadelphia (2015).

[18] **A. Ern, J.-L. Guermond**, *Theory and Practice of Finite Elements*, Springer Series in Applied Mathematical, **159**, Springer, New York (2004).

[19] **M. Gander**, Optimized Schwarz methods, *SIAM J. Numer. Anal.*, **44**, pp. 699-–731 (2006).

[20] **V. Girault, P.-A. Raviart**, *Finite element methods for Navier-Stokes equations*, Springer Series in Computational Mathematics, **5**, Springer Verlag, Berlin (1986).

[21] **A. S. Householder**, *The theory of matrices in numerical analysis*, Blaisdell Publishing Company (1970).

[22] **F. Jean**, *Stabilité et commande des systèmes dynamiques. Cours et exercices corrigés*, Les Presses de l'ENSTA, Coll. Les Cours (2011).

[23] **P.-L. Lions**, *On the Schwarz alternating method I*, First International Symposium on Domain Decomposition Methods for Partial Differential equations (1988).

[24] **P.-L. Lions**, *On the Schwarz alternating method III : a variant for nonoverlapping subdomains*, Third International Symposium Domain Decomposition Methods for Partial Differential equations (1990).

[25] **F. Magoulès, F.-X. Roux**, *Calcul scientifique parallèle*, Dunod, Paris (2013).

[26] **G. Meurant**, *Computer solution of large linear systems*, Elsevier, New York (1999).

[27] **A. Modave**, *Calcul scientifique parallèle*, Master Analyse, Modélisation et Simulation de l'Université Paris-Saclay et de l'Institut Polytechnique de Paris, `https://ams301.pages.math.cnrs.fr` (2019).

[28] **F. Nataf, F. Nier**, *Convergence rate of some domain decomposition methods for overlapping and nonoverlapping subdomains*, Numer. Math., **75**, pp. 357-377 (1997).

[29] **M. A. Olshanskii, E. E. Tyrtyshnikov**, *Iterative methods for linear systems*, SIAM, Philadelphia (2014).

[30] **J. M. Ortega, R. J. Plemmons**, Extensions of the Ostrowski-Reich theorem for SOR iterations, *Linear Algebra and its Applications*, **28** (1979).

[31] **B. N. Parlett**, *The symmetric eigenvalue problem*, Prentice Hall, Englewood Cliffs (1980).

[32] **A. Quarteroni, A. Valli**, *Domain decomposition methods for partial differential equations*, Oxford Science Publications (1999).

[33] **Y. Saad**, *Numerical methods for large eigenvalue problems*, 2nd Edition, SIAM (2011).

[34] **H. A. Schwarz**, Uber einige Abbildungsaufgaben, *Ges. Math Abh.*, **11**, pp. 65–83 (1869).

[35] **W. Shakespeare**, *Much ado about nothing* (ca. 1598).

[36] **B. I. Wohlmuth**, *Hierarchical a posteriori error estimators for Mortar finite element methods with Lagrange multipliers*, SIAM J. Numer. Anal., **36**, p. 1636-1658 (1999).

[37] **D. M. Young**, *Iterative solution of large linear systems*, Academic Press, New York (1971).

Index