Mise à jour de la métrique dans les méthodes de quasi-Newton réduites en optimisation avec contraintes d'égalité

Jean Charles Gilbert

RAIRO
MODÉLISATION MATHÉMATIQUE
ET ANALYSE NUMÉRIQUE

JEAN CHARLES GILBERT

Mise à jour de la métrique dans les méthodes de quasi-Newton réduites en optimisation avec contraintes d'égalité

RAIRO – Modélisation mathématique et analyse numérique, tome 22, n° 2 (1988), p. 251-288.

<http://www.numdam.org/item?id=M2AN_1988__22_2_251_0>

MATHEMATICAL MODELLING AND NUMERICAL ANALYSIS
MODÉLISATION MATHÉMATIQUE ET ANALYSE NUMÉRIQUE

(Vol. 22, n° 2, 1988, p. 251 à 288)

MISE A JOUR DE LA MÉTRIQUE DANS LES MÉTHODES DE QUASI-NEWTON RÉDUITES EN OPTIMISATION AVEC CONTRAINTES D'ÉGALITÉ (*)

par Jean Charles GILBERT (1)

Communiqué par P.-G. CIARLET

Résumé. — *En optimisation dans $\mathbb{R}^n$, en présence de m contraintes d'égalité non linéaires, on s'intéresse aux méthodes de quasi-Newton réduites pour lesquelles on étudie la mise à jour de la métrique On se trouve devant l'alternative suivante. Soit les contraintes sont linéarisées deux fois par itération et la convergence superlinéaire peut s'obtenir par la technique de Broyden, Dennis et Moré, soit les contraintes ne sont linéarisées qu'une fois par itération et l'introduction d'un critère de mise a jour s'avère nécessaire à l'obtention de la convergence superlinéaire.*

Abstract. — *In optimization in $\mathbb{R}^n$ with m nonlinear equality constraints, we are concerned with reduced quasi-Newton methods for which we analyse the update of the matrix. The following alternative occurs Either the constraints are linearized twice per iteration and the superlinear convergence can be obtained by Broyden, Dennis and Moré's technique, or the constraints are linearized only once per iteration and an update criterion has to be introduced in order to obtain the superlinear convergence*

1. INTRODUCTION

Les méthodes de quasi-Newton pour résoudre une équation non linéaire sur $\mathbb{R}^n$, soit $F(x) = 0$, sont des méthodes qui génèrent deux suites : une suite (x_k) d'approximations d'une solution x_* et une suite (J_k) de matrices d'ordre n, non singulières qui approximent $\nabla F(x_*)$, la jacobienne de F en x_*. Partant d'un point x_1 de $\mathbb{R}^n$ et d'une matrice d'ordre n non singulière J_1, les deux suites sont obtenues par les récurrences suivantes :

$$x_{k+1} = x_k - J_k^{-1} F(x_k) , \tag{1.1}$$

$$J_{k+1} = U(J_k, \gamma_k, \sigma_k) , \tag{1.2}$$

(*) Reçu en octobre 1985, révisé en novembre 1987.

(1) INRIA, Domaine de Voluceau, BP 105, Rocquencourt, 78153 Le Chesnay Cedex, France.

où,

$$\gamma_k := F(x_{k+1}) - F(x_k)\,, \tag{1.3}$$

$$\sigma_k := x_{k+1} - x_k \tag{1.4}$$

et U est une formule de mise à jour de J_k telle que J_{k+1} donnée par (1.2) satisfasse l'*équation de quasi-Newton* suivante :

$$\gamma_k = J_{k+1}\,\sigma_k\,. \tag{1.5}$$

Cette relation étant vérifiée au premier ordre en σ_k avec $J_{k+1} = \nabla F(x_k)$, on peut espérer que J_{k+1} vérifiant (1.5) sera une meilleure approximation de $\nabla F(x_k)$ que ne l'est J_k. La relation (1.1) montre que J_k joue le rôle de la jacobienne $\nabla F(x_k)$ dans la méthode de Newton, d'où le nom de méthode de quasi-Newton. La relation (1.5) montre que J_{k+1} est le « quotient » de γ_k par σ_k. Sur ces méthodes, on pourra consulter l'excellent article de revue de Dennis et Moré (1977).

Un avantage important des méthodes de quasi-Newton est de ne pas nécessiter le calcul des dérivées de F alors que l'on peut obtenir la *convergence superlinéaire* de la suite (x_k), c'est-à-dire :

$$\frac{\|x_{k+1} - x_*\|}{\|x_k - x_*\|} \to 0\,. \tag{1.6}$$

On s'intéressera dans ce travail à la résolution numérique du problème d'optimisation avec contraintes d'égalité qui s'écrit :

$$\min\,\{f(x) : c(x) = 0\}\,, \tag{1.7}$$

où f et c sont définies sur $\mathbb{R}^n$ à valeurs dans $\mathbb{R}$ et $\mathbb{R}^m$ respectivement. On suppose $m < n$. Si x_* est une solution locale de (1.7) et que la matrice jacobienne $A_* := \nabla c(x_*)$, de dimension $m \times n$, est surjective, il existe un unique multiplicateur de Lagrange λ_* dans $\mathbb{R}^m$ tel que les conditions nécessaires d'optimalité du premier ordre suivantes soient satisfaites :

$$\begin{cases} c(x_*) = 0\,, \\ \nabla f(x_*) + A_*^T \lambda_* = 0\,. \end{cases} \tag{1.8}$$

On a désigné par $\nabla f(x_*)$, $n \times 1$, le vecteur des dérivées partielles de f en x_* et par A_*^T la matrice transposée de A_*. La seconde équation de (1.8) exprime la nullité de la dérivée première par rapport à x du lagrangien

$$\ell(x, \lambda) := f(x) + c(x)^T \lambda\,. \tag{1.9}$$

De nombreuses méthodes de résolution du problème (1.7) se ramènent à la recherche de points stationnaires de f sur la variété $M_* := c^{-1}(0)$ *via* la

résolution du système (1.8). Par exemple, la *programmation quadratique successive* (PQS) consiste à résoudre le système (1.8) par des itérations de Newton (Wilson (1963)) ou de quasi-Newton (Han (1976)). Notons que dans cette dernière méthode, il y a mise à jour d'une matrice d'*ordre n* et non pas d'ordre $n + m$ comme pourrait le laisser penser la taille du système (1.8). En effet, la jacobienne du système d'optimalité (1.8) en (x, λ) s'écrit :

$$J(x, \lambda) = \begin{bmatrix} A(x) & 0 \\ L(x, \lambda) & A(x)^T \end{bmatrix} \tag{1.10}$$

où $A(x) := \nabla c(x)$ et $L(x, \lambda) := \nabla^2_{xx} \ell(x, \lambda)$ est le hessien par rapport à x du lagrangien (1.9). Dans la PQS, seule une approximation L_k de $L(x_k, \lambda_k)$ est mise à jour tandis que $A(x_k)$ est calculé. La méthode se présente sous la forme d'une succession de résolutions de problèmes quadratiques sous contraintes linéaires (d'où son nom). Étant donné x_k, on obtient x_{k+1} en résolvant le problème

$$\begin{aligned} \min \Big\{ & f'(x_k).(x - x_k) + \frac{1}{2} (x - x_k)^T L_k (x - x_k) : \\ & c(x_k) + c'(x_k).(x - x_k) = 0 \Big\} , \end{aligned} \tag{1.11}$$

dont la solution s'écrit $x_{k+1} = x_k + d_k^{\mathrm{PQS}}$ avec (*cf.* Gabay (1982b)) :

$$\begin{aligned} d_k^{\mathrm{PQS}} := & - A(x_k)^- \, c(x_k) \\ & - Z(x_k)^- \, (Z(x_k)^{-T} L_k \, Z(x_k)^-)^{-1} \\ & \times [g(x_k) - Z(x_k)^{-T} L_k \, A(x_k)^- \, c(x_k)] \end{aligned}$$

où $A(x_k)^-$ est un inverse à droite quelconque de $A(x_k)$ et $Z(x_k)^-$ est une *base* quelconque de $N(A(x_k))$, le noyau de $A(x_k)$, c'est-à-dire une matrice de dimension $n \times (n - m)$ dont les colonnes engendrent $N(A(x_k))$. On a noté $g(x)$ le *gradient réduit* de f en x sur la variété $M(x) := c^{-1}(c(x))$ qui est défini par

$$g(x) := Z(x)^{-T} \nabla f(x) . \tag{1.13}$$

Un autre point de vue est adopté par Gilbert (1986ab). Il consiste à mettre à profit la structure particulière de couplage d'équations que présente le système (1.8) et d'utiliser une méthode de résolution par « découplage ». L'algorithme se présente comme suit :

$$y_k := x_k - R_k \, c(x_k) , \tag{1.14}$$

$$x_{k+1} := y_k - Z(y_k)^- \, G_k^{-1} g(y_k) , \tag{1.15}$$

$$\begin{aligned} \lambda_{k+1} := & - A(y_k)^{-T} \nabla f(y_k) \\ & + A(y_k)^{-T} L_k \, Z(y_k)^- \, G_k^{-1} g(y_k) . \end{aligned} \tag{1.16}$$

Dans (1.14), R_k est un opérateur de dimension $n \times m$, asymptotiquement proche d'un inverse à droite quelconque A_*^- de A_*. En pratique, R_k sera pris égal à $A(x_k)^-$ ou $A(y_{k-1})^-$ selon que les contraintes sont linéarisées en x_k ou pas. On passe donc de x_k à y_k par un *pas de restauration* des contraintes qui tend à rendre y_k plus proche de M_* que ne l'est x_k. Les contraintes sont toujours linéarisées en y_k.

Dans (1.15), $Z(y_k)^-$ est une base (une matrice injective de dimension $n \times (n-m)$) de l'espace tangent à la variété $M(y_k)$ en y_k et G_k est une matrice d'ordre $n-m$ non singulière donnant une approximation de G_*, le hessien « projeté » du lagrangien dans le plan tangent $N(A_*)$ à M_* en x_*. En notant $L_* := L(x_*, \lambda_*)$ et en prenant une base quelconque Z_*^- de $N(A_*)$, le hessien projeté s'écrit :

$$G_* := Z_*^{-T} L_* Z_*^- . \tag{1.17}$$

Les métriques réduites G_k seront générées par une méthode de quasi-Newton les maintenant symétriques définies positives. Le déplacement tangent $t_k := x_{k+1} - y_k$ qui conduit de y_k à x_{k+1} par (1.15) est un *pas de minimisation* du critère f. En effet, $f'(y_k).t_k = -g(y_k)^T G_k^{-1} g(y_k)$ est négatif si G_k est définie positive. Ce déplacement t_k peut être obtenu en minimisant une approximation quadratique en y_k du lagrangien (1.9) dans le plan tangent à $M(y_k)$ en y_k. Si on note $t_k =: Z(y_k)^- h_k$, h_k est solution de

$$\min g(y_k)^T h + \frac{1}{2} h^T G_k h .$$

Dans (1.16), L_k est une matrice d'ordre n approximant L_*.

L'algorithme local (1.14)-(1.15) est une *méthode de quasi-Newton réduite* parce que la matrice à mettre à jour, G_k, est d'*ordre* $n-m$ alors qu'elle est d'ordre n dans les méthodes de quasi-Newton. Il est en effet remarquable que le calcul de la suite (x_k) ne demande pas le calcul de λ_{k+1} par (1.16) où l'approximation L_k du hessien complet intervient.

L'algorithme (1.14)-(1.15) a également été considéré par Coleman et Conn (1982ab) où la convergence de la suite (y_k) est étudiée. Contrairement à la suite (x_k), la suite (y_k) ne converge en général pas superlinéairement mais « seulement » superlinéairement en deux pas, c'est-à-dire :

$$\frac{\|y_{k+1} - x_*\|}{\|y_{k-1} - x_*\|} \to 0 .$$

Ce taux de convergence est moins bon que (1.6) et on ne peut pas obtenir mieux en général. Des contre-exemples ont en effet été donnés par Byrd (1985).

Un autre algorithme à métrique réduite a été étudié par divers auteurs

(*cf.* Powell (1978), Gabay (1982b) et Nocedal, Overton (1985)). Il s'obtient en négligeant dans le déplacement (1.12) de la PQS la partie du déplacement tangent qui fait intervenir l'approximation L_k du hessien complet du lagrangien. Il s'écrit :

$$x_{k+1} = x_k - A(x_k)^- \, c(x_k) - Z(x_k)^- \, G_k^{-1} \, g(x_k) \,, \tag{1.18}$$

où $A(x_k)^-$ est un inverse à droite quelconque de $A(x_k)$ (mise à part une hypothèse de régularité par rapport à x_k) et $Z(x_k)^-$ est une base quelconque de $N(A(x_k))$ (même restriction sur la régularité). La différence entre les algorithmes (1.14)-(1.15) et (1.18) se situe dans le pas de minimisation du critère qui dans le premier est tangent à la variété $M(y_k)$ tandis que dans le second il est tangent à la variété $M(x_k)$. Théoriquement, l'algorithme (1.18) est moins bon que l'algorithme (1.14)-(1.15) puisque la suite (x_k) générée par (1.8) ne peut en général converger « que » superlinéairement en deux pas (*cf.* Gabay (1982b)). Des contre-exemples ont également été donnés par Byrd (1985) et Yuan (1985).

L'objet de ce travail est d'étudier la mise à jour de la matrice G_k au cours des itérations dans l'algorithme (1.14)-(1.15). On utilisera la formule de BFGS. Deux variantes seront étudiées simultanément. Dans l'*algorithme QNR1,* les contraintes sont linéarisées deux fois par itération aux points x_k et y_k. L'opérateur de restauration R_k en (1.14) est pris égal à $A(x_k)^-$. C'est le cas le plus simple. La matrice est mise à jour à chaque itération et des conditions de convergence superlinéaire peuvent être obtenues en utilisant la technique de Broyden, Dennis et Moré (1973). Cet algorithme est dans son principe identique à celui de Coleman et Conn (1984). Toutefois, notre schéma de mise à jour est plus simple et nous montrons la convergence superlinéaire (en un pas) de la suite (x_k) alors que leur étude portait sur la suite (y_k) qui, répétons-le, ne converge en général pas superlinéairement (en un pas).

Dans l'*algorithme QNR2,* les contraintes ne sont plus linéarisées qu'au point y_k, ce qui peut représenter une diminution de coût appréciable pour certaines applications. L'opérateur R_k en (1.14) est alors $A(y_{k-1})^-$. Dans cette méthode, la matrice G_k n'est en général pas mise à jour à chaque itération et un critère décidant de l'opportunité de la mise à jour doit être introduit. Un choix judicieux de ce critère permet d'obtenir la convergence superlinéaire de la suite (x_k). L'idée d'utiliser un critère de mise à jour a également été utilisée par Nocedal et Overton (1985) pour l'étude de l'algorithme (1.18).

Cet article constitue une version révisée de la section 6 du rapport INRIA RR-482 dont la première partie révisée se trouve dans Gilbert (1986b). Une variante de l'algorithme étudié ici qui utilise un pas de restauration des contraintes supplémentaire est donnée dans Gilbert (1986c).

2. HYPOTHÈSES ET NOTATIONS

On supposera que les fonctions f et c sont de classe C_b^ν avec $\nu \geqq 3$, c'est-à-dire que f et c sont supposées être ν fois continûment différentiables sur $\mathbb{R}^n$ et avoir des dérivées partielles d'ordre k, $1 \leqq k \leqq \nu$, bornées sur $\mathbb{R}^n$. On supposera également que c est une *submersion* sur $\mathbb{R}^n$, c'est-à-dire que $A_x := \nabla c(x)$ est surjective quel que soit x dans $\mathbb{R}^n$. La surjectivité de $A_* := \nabla c(x_*)$ est une hypothèse couramment faite : elle implique la surjectivité de A_x dans un voisinage de x_*. Supposer que c soit une submersion sur $\mathbb{R}^n$ est une hypothèse assez restrictive. On sait alors qu'il existe un unique multiplicateur de Lagrange λ_* tel que le système d'optimalité (1.8) soit vérifié (*cf.* par exemple Fletcher (1981)). On supposera également que la condition suffisante d'optimalité du second ordre est satisfaite, à savoir que L_* est définie positive dans l'espace tangent $N(A_*)$ ce qui revient à supposer la définie positivité de $G_* := Z_*^{-T} L_* Z_*^-$ où Z_*^- est une base quelconque de $N(A_*)$.

La matrice surjective A_x étant donnée, on peut choisir arbitrairement un inverse à droite A_x^- de A_x et une base Z_x^- de $N(A_x)$. Ces choix étant faits, il est aisé de montrer l'existence d'une unique matrice surjective Z_x de dimension $(n-m) \times n$ telle que

$$Z_x A_x^- = 0 , \tag{2.1}$$

$$Z_x Z_x^- = I_{n-m} . \tag{2.2}$$

Alors $[A_x^- \; Z_x^-]$ est l'inverse de $[A_x^T \; Z_x^T]^T$ et on a la relation fort utile

$$I = A_x^- A_x + Z_x^- Z_x . \tag{2.3}$$

Si l'opérateur A_x est parfaitement défini, borné et de classe $C_b^{\nu-1}$ sur $\mathbb{R}^n$, il n'en va pas nécessairement de même pour l'inverse à droite A_x^- de A_x et pour la base Z_x^- de $N(A_x)$ pour lesquels un choix est à faire. On fera une hypothèse de régularité par rapport à x pour ces choix, à savoir, on supposera que l'application

$$x \to (A_x^-, Z_x^-) \text{ est bornée et de classe } C_b^{\nu-1} \text{ sur } \mathbb{R}^n . \tag{2.4}$$

Cette hypothèse est satisfaite pour A_x^- si on choisit

$$A_x^- = A_x^T (A_x A_x^T)^{-1} ,$$

mais d'autres choix sont possibles (*cf.* Gabay (1982a)). En ce qui concerne Z_x^-, la question est plus délicate bien que l'hypothèse précédente puisse

toujours être satisfaite localement par projection sur $N(A_x)$ d'une base Z_*^- de $N(A_*)$ (*cf.* Byrd et Schnabel (1986)).

On désignera par $\|.\|$ la norme l_2 de $\mathbb{R}^n$ ou $\mathbb{R}^m$. Les normes matricielles seront supposées subordonnées aux normes vectorielles :

$$\|A\| := \sup \{\|Av\| : \|v\| \leqq 1\} .$$

Si (v_k) est une suite d'un espace normé $(E, \|.\|)$ et (α_k) et (β_k) sont deux suites de nombres réels positifs, on dira que (v_k) est un grand O de (α_k) (on notera $v_k = O(\alpha_k)$) si la suite $(\|v_k\|/\alpha_k)$ est bornée et on dira que (v_k) est un petit o de (α_k) (on notera $v_k = o(\alpha_k)$) si la suite $(\|v_k\|/\alpha_k)$ converge vers zéro. On dira que (α_k) et (β_k) sont équivalentes (on notera $\alpha_k \sim \beta_k$) si $\alpha_k = O(\beta_k)$ et $\beta_k = O(\alpha_k)$. On notera v^i la i-ième composante d'un vecteur v de E (de dimension finie). On notera également $(u, v) = u^T v$ si u et v sont deux vecteurs de E. Pour r réel positif et v dans E, on notera $B(v, r)$ la boule de centre v et de rayon r. On notera $L(E)$ l'ensemble des opérateurs linéaires sur E.

3. UNE MÉTHODE DE QUASI-NEWTON RÉDUITE

3.1. Obtention de la méthode locale

L'algorithme (1.14)-(1.15) peut s'obtenir de la façon suivante (*cf.* Gilbert (1986b)). On construit d'abord un algorithme « idéal » dans le sens où on suppose connue l'évaluation au point cherché x_* des opérateurs intervenant dans la méthode. On obtient alors aisément des conditions sur ces opérateurs pour que localement, l'algorithme converge quadratiquement. On remplace ensuite ces opérateurs par leur évaluation en certains points ou par leur approximation quasi-Newtonienne.

Ces opérateurs sont construits d'une part à partir d'un inverse à droite *quelconque* de la jacobienne $X_* := [A_* \; 0]$ de la première équation du système d'optimalité (1.8) en x_*, à savoir :

$$X_*^- = \begin{bmatrix} A_*^- \\ U_* \end{bmatrix} ,$$

où A_*^- est un inverse à droite quelconque de $A_* := \nabla c(x_*)$, de dimension $n \times m$, et U_* est une matrice d'ordre m quelconque. D'autre part, on choisit une base *quelconque* de $N(X_*)$ dans $\mathbb{R}^{n+m}$ dont les vecteurs forment les n colonnes d'une matrice injective W_*^-. Alors nécessairement cette matrice a la forme suivante :

$$W_*^- = \begin{bmatrix} Z_*^- & 0 \\ 0 & I_m \end{bmatrix} R_* ,$$

où Z_*^- est une matrice de dimension $n \times (n-m)$ dont les colonnes forment une base quelconque de $N(A_*)$ et R_* est une matrice d'ordre n non singulière quelconque. Enfin, on construit l'inverse à droite *particulier* Y_*^- de la jacobienne $Y_* := [L_* \; A_*^T]$ de la seconde équation du système d'optimalité (1.8) en x_* qui vérifie $X_* \, Y_*^- = 0$. Cette condition détermine Y_*^- qui s'écrit $Y_*^- = W_*^- (Y_* \, W_*^-)^{-1}$ ou encore

$$Y_*^- = \begin{bmatrix} Z_* \, G_*^{-1} \, Z_*^T \\ A_*^{-T}(I - L_* \, Z_* \, G_*^{-1} \, Z_*^T) \end{bmatrix}.$$

L'algorithme « idéal » est le suivant :

$$\begin{bmatrix} y_k \\ \mu_k \end{bmatrix} = \begin{bmatrix} x_k \\ \lambda_k \end{bmatrix} - X_*^- \, c(x_k), \tag{3.1}$$

$$\begin{bmatrix} x_{k+1} \\ \lambda_{k+1} \end{bmatrix} = \begin{bmatrix} y_k \\ \mu_k \end{bmatrix} - Y_*^- (\nabla f(y_k) + A(y_k)^T \, \mu_k). \tag{3.2}$$

Partant du point courant x_k et du multiplicateur courant λ_k, l'étape (3.1) consiste à faire un pas de résolution de la première équation de (1.8) en utilisant un inverse à droite quelconque de sa jacobienne au point optimal. Ceci conduit au nouveau point y_k et au nouveau multiplicateur μ_k. L'étape (3.2) consiste à faire un pas de résolution de la seconde équation de (1.8) en utilisant l'inverse à droite particulier de sa jacobienne au point optimal qui rend ce déplacement tangent à la variété définie par la première équation de (1.8) : c'est la condition $X_* \, Y_*^- = 0$. Celle-ci suffit à rendre l'algorithme (3.1)-(3.2) localement (c'est-à-dire en supposant (x_1, λ_1) proche de (x_*, λ_*)) quadratiquement convergent.

Finalement, si dans (3.1), on remplace X_*^- par

$$\begin{bmatrix} R_k \\ U_k \end{bmatrix}$$

où R_k est un opérateur de restauration des contraintes qui approxime A_*^- et U_k est une matrice d'ordre m quelconque et si dans (3.2), on remplace Y_*^- par

$$\begin{bmatrix} Z(y_k)^- \, G_k^{-1} \, Z(y_k)^{-T} \\ A(y_k)^{-T} \, (I - L_k \, Z(y_k)^- \, G_k^{-1} \, Z(y_k)^{-T}) \end{bmatrix},$$

on obtient l'algorithme (1.14)-(1.16). On constate en effet que, $A(y_k) \, Z(y_k)^-$ étant nul ($Z(y_k)^-$ est une base de $N(A(y_k))$), μ_k n'intervient pas dans l'algorithme.

3.2. L'algorithme modèle

L'algorithme (1.14)-(1.15) peut être globalisé par la méthode utilisée par Han (1977) pour la PQS et il est alors plus naturel de considérer la suite (y_k). On introduit la fonctionnelle pénalisée exacte :

$$\theta_p(y) := f(y) + p\,\|c(y)\|_1 \tag{3.3}$$

où p est le paramètre de pénalisation et $\|.\|_1$ est la norme l_1 de $\mathbb{R}^m$. D'autres normes peuvent être utilisées et nous renvoyons à Bonnans et Gabay (1984) pour un cadre général. Si p est pris supérieur à $\|\lambda_*\|_\infty$ (où $\|.\|_\infty$ est la norme duale de $\|.\|_1$), x_* est un minimum local de θ_p : on dit que la pénalisation est exacte. Il est alors licite de chercher x_* en minimisant θ_p. Cette fonction n'est pas différentiable, mais on utilise les déplacements calculés par la méthode locale (1.14)-(1.15) pour obtenir un *arc de descente* de θ_p en y_k. On définit les déplacements

$$r_k := -R_k\,c(x_k)\,, \tag{3.4}$$

$$t_k := -Z(y_k)^-\,G_k^{-1}\,g(y_k)\,, \tag{3.5}$$

$$d_k := r_k + t_k\,, \tag{3.6}$$

$$e_k := t_k + r_{k+1}\,, \tag{3.7}$$

ainsi qu'un arc issu de y_k tangent à t_k que l'on paramétrise par ρ :

$$y_k(\rho) = y_k + \rho t_k + \rho^a\,r_{k+1}\,, \quad \text{avec} \quad a > 1\,. \tag{3.8}$$

On prend $y_{k+1} := y_k(\rho_k)$ avec ρ_k déterminé par une règle du type Armijo (1966) : étant donnés α dans $]0, 1/2[$ et β dans $]0, 1[$, on choisit ρ_k de la forme

$$\rho_k := \beta^{\ell_k}\,, \tag{3.9}$$

où ℓ_k est le plus petit entier naturel tel que

$$\theta_p(y_k(\beta^{\ell_k})) < \theta_p(y_k) + \beta^{\ell_k}\,\alpha f'(y_k)\cdot t_k - \beta^{a\ell_k}\,\alpha(p - \|\lambda(y_k)\|_\infty)\|c(y_k)\|_1\,. \tag{3.10}$$

Dans cette inégalité $\lambda(y_k)$ est défini comme suit :

$$\lambda(y) := -A(y)^{-T}\,\nabla f(y)\,. \tag{3.11}$$

C'est donc une estimation du multiplicateur λ_{k+1} donné par (1.16) lorsque l'on néglige le dernier terme qui converge vers 0 et qui a l'inconvénient de faire intervenir l'approximation L_k du hessien complet du lagrangien.

On montre (*cf.* Gilbert (1986b), lemme 6.1) que lorsque $0 < \underline{h} I \leqq G_k^{-1} \leqq \underline{h}^{-1} I$ (c'est-à-dire G_k^{-1} définie positive et bornée ainsi que G_k) et $\underline{p} + \|\lambda(y_k)\|_\infty \leqq p \leqq 1/\underline{p}$ ($\underline{h}$ et $\underline{p}$ sont des constantes positives), l'arc $y_k(\rho)$ est de descente pour θ_p en y_k et la règle (3.9)-(3.10) permet effectivement de déterminer un pas ρ_k positif. La borne inférieure pour p n'est pas connue a priori. Il est alors parfois nécessaire d'adapter p au cours des itérations afin que l'arc $y_k(\rho)$ soit de descente pour θ_p. On note alors p_k le paramètre de pénalisation à l'itération k. Si p_k varie, on ne minimise plus la même fonctionnelle θ_p à chaque itération. Néanmoins, on obtient un théorème de convergence globale si le choix de p_k satisfait aux conditions suivantes :

(3.12) $p_k \geqq \|\lambda(y_k)\|_\infty + \underline{p}$ pour tout k ,

(3.13) il existe un indice K tel que $k \geqq K$ et $p_{k-1} \geqq \|\lambda(y_k)\|_\infty + \underline{p}$ implique $p_k = p_{k-1}$,

(3.14) (p_k) est bornée si et seulement si p_k est modifié un nombre fini de fois .

Ces conditions sont satisfaites par exemple lorsque l'on prend la règle de Mayne et Polak (1982) : si $p_{k-1} \geqq \|\lambda(y_k)\|_\infty + \underline{p}$ alors $p_k := p_{k-1}$, sinon $p_k := \max((1+\delta)\, p_{k-1}, \|\lambda(y_k)\|_\infty + \underline{p})$, $\underline{p}$ et δ étant des réels positifs donnés au départ.

On peut à présent définir l'algorithme modèle global.

Algorithme QNR

1. Se donner α dans $]0, 1/2[$, β dans $]0, 1[$, a dans $]1, +\infty[$ et ε dans $]0, +\infty[$.
2. Choisir un point y_0 dans $\mathbb{R}^n$ et une matrice G_0 d'ordre $n - m$ symétrique définie positive.
3. Initialiser k à 0.
4. *Répéter :*
   4. 1. Linéariser les contraintes en y_k.
      Choisir un inverse à droite $A(y_k)^-$ de $\nabla c(y_k)$ et une base $Z(y_k)^-$ de $N(\nabla c(y_k))$ tel que l'on ait (2.4).
   4. 2. Calculer $\lambda(y_k) := -A(y_k)^{-T} \nabla f(y_k)$ et $g(y_k) := Z(y_k)^{-T} \nabla f(y_k)$.
   4. 3. *Si* $k \geqq 1$ *alors* Calculer G_k par mise à jour de G_{k-1}.
   4. 5. (* phase de minimisation *)
      Calculer $t_k := -Z(y_k)^- G_k^{-1} g(y_k)$ et $x_{k+1} := y_k + t_k$.
   4. 8. Évaluer R_{k+1} et $c(x_{k+1})$.

4. 9. (* phase de restauration *)
Calculer $r_{k+1} := -R_{k+1}\, c(x_{k+1})$.

4.10. (* test d'arrêt *)
Si $\|g(y_k)\| + \|c(x_{k+1})\|_1 \leqq \varepsilon$ *alors* Arrêter.

4.11. Adapter p_k selon les règles (3.12)-(3.14).

4.12. Déterminer un pas ρ_k sur l'arc (3.8) en utilisant la règle (3.9)-(3.10) et la fonctionnelle (3.3) (avec $p = p_k$).

4.13. Prendre $y_{k+1} := y_k(\rho_k)$.

4.14. $k := k + 1$.

Les étapes 4.4, 4.6 et 4.7 seront introduites plus loin tandis que les étapes 4.3 et 4.8 seront précisées plus loin.

L'analyse faite ci-après suppose implicitement que l'algorithme ne s'arrête pas après un nombre fini d'itérations, c'est-à-dire que y_k et x_k sont supposés différer de x_* pour tout indice k.

3.3. Rappel des résultats

On rappelle ici quelques résultats obtenus dans Gilbert (1986ab).

Sous les hypothèses de la Section 2, l'algorithme QNR est linéairement convergent dans un cadre local ((y_0, G_0) proche de (x_*, G_*) et $\rho_k = 1$ pour tout k) si G_k reste proche de G_*, ce qui est une condition assez restrictive. Ce résultat peut se déduire du lemme suivant qui est extrait de la preuve du théorème 4.2 dans Gilbert (1986b).

LEMME 3.1 : *Soit* κ *un réel dans* $]0, 1[$. *Il existe deux réels positifs* ε_1 *et* ε_2 *ne dépendant que de* κ *et de* f *et* c *dans un voisinage de* x_* *tels que si*

$$\begin{aligned} \|y_{k-1} - x_*\| &\leqq \varepsilon_1 , \\ \|x_k - x_*\| &\leqq \varepsilon_2 , \\ \|G_k - G_*\| &\leqq \varepsilon_2 , \end{aligned} \tag{3.15}$$

alors,

$$\|x_{k+1} - x_*\| \leqq \kappa \|x_k - x_*\| .$$

En fait la preuve a été donnée pour $R_k = A(y_{k-1})^-$ mais elle se généralise aisément au cas où $R_k = A(x_k)^-$, auquel cas (3.15) est superflu.

Lorsque (y_k) converge vers x_*, que ρ_k vaut 1 et que (G_k) et (G_k^{-1}) sont bornées, on a les équivalences suivantes :

$$\|d_k\| \sim \|x_k - x_*\| , \tag{3.16}$$

$$\|e_k\| \sim \|y_k - x_*\| . \tag{3.17}$$

On montre également une propriété de convergence globale, à savoir que la suite (y_k) générée par l'algorithme QNR vérifie

$$\|g(y_k)\| + \|c(y_k)\| \to 0 ,$$

dès que les suites (p_k), (G_k) et (G_k^{-1}) sont bornées. Il serait agréable de pouvoir se passer de ces conditions comme dans le cadre convexe sans contrainte (*cf.* Powell (1976)).

Lorsque (y_k) converge vers x_* et que ρ_k vaut 1 pour tout k, la convergence superlinéaire de (x_k) dépend de la qualité de l'approximation du hessien réduit du lagrangien G_* par G_k. On montre que ce taux de convergence est obtenu si et seulement si

$$(G_k - G_*) Z(y_k) t_k = o(\|x_k - x_*\|) . \tag{3.18}$$

Cette condition est à rapprocher de la condition nécessaire et suffisante de convergence superlinéaire de Dennis et Moré (1974) pour les méthodes de quasi-Newton sans contrainte. On voit donc que la convergence superlinéaire de (x_k) dépend de (3.18) mais également de l'admissibilité du pas unité après un nombre fini d'itérations.

La question de l'admissibilité du pas unité est traitée dans le théorème qui suit et que nous utiliserons plus loin. Soit $\mathbb{K}$ une sous-suite d'indices. On lui associe l'une des propriétés suivantes :

$$\|G_k - G_*\| \leqq M \quad \text{pour } k \text{ dans } \mathbb{K} , \tag{3.19}$$

$$(G_k - G_*) Z_* t_k = o(\|t_k\|) \quad \text{pour } k \text{ dans } \mathbb{K} , \tag{3.20}$$

$$t_k = O(\|r_{k+1}\|) \quad \text{pour } k \text{ dans } \mathbb{K} , \tag{3.21}$$

$$\rho_k < 1 \quad \text{et} \quad t_k = o(\|r_k\|) \quad \text{pour } k \text{ dans } \mathbb{K} . \tag{3.22}$$

On a alors le résultat suivant :

THÉORÈME 3.2 : *Soient (x_k), (y_k) et (G_k) les suites générées par l'algorithme* QNR. *Supposons que (y_k) converge vers x_* et qu'il existe un réel positif $\underline{h}$ tel que $\underline{h} I \leqq G_k^{-1} \leqq \underline{h}^{-1} I$ pour tout indice k. Soit $\mathbb{K}$ une sous-suite d'indices. Alors :*

(i) *il existe une constante positive $\bar{M}$ ne dépendant que de c, α et $\underline{h}$ telle que si la majoration* (3.19) *est vérifiée avec $M \leqq \bar{M}$ alors $\rho_k = 1$ pour $k \in \mathbb{K}$ après un nombre fini d'itérations,*

(ii) *si l'une des estimations* (3.20) *ou* (3.21) *est vérifiée alors $\rho_k = 1$ pour $k \in \mathbb{K}$ après un nombre fini d'itérations,*

(iii) *si la condition* (3.22) *est vérifiée alors $r_{k+1} = o(\|r_k\| \|t_k\|)$ pour $k \in \mathbb{K}$.*

L'admissibilité asymptotique du pas unité dépend donc soit de l'approximation de G_* par G_k *via* les conditions (3.19) ou (3.20), soit de l'importance relative du pas de minimisation t_k et du pas de restauration r_{k+1} *via* la condition (3.21).

La propriété (3.19) est très forte lorsque M est petit et est sans doute rarement satisfaite avec les méthodes de quasi-Newton (*cf.* Ge, Powell (1983)). Par contre, la propriété (3.20) sera en général vérifiée dans les algorithmes proposés. Remarquons que cette propriété d'approximation est plus forte que la propriété (3.18) (pour k dans $\mathbb{K}$). En effet, lorsque (G_k^{-1}) est bornée, on montre facilement que $t_k = O(\|y_k - x_*\|)$ et lorsque $\rho_k = 1$ pour tout k, on a $y_k - x_* = O(\|x_k - x_*\|)$. Dans ce cas, $t_k = O(\|x_k - x_*\|)$ et (3.20) implique (3.18) (pour k dans $\mathbb{K}$). Lorsque la propriété (3.20) n'est pas satisfaite, on montrera que (3.21) ou (3.22) doit avoirs lieu. Avec (3.21), on a également un pas unité asymptotiquement tandis que le résultat obtenu en (iii) permettra de montrer qu'il n'y a pas de sous-suites vérifiant (3.22).

Comme le montre les conditions (3.18) et (3.20), avoir G_k proche de G_* est important à deux titres. D'abord pour avoir la convergence superlinéaire (par (3.18)) ; mais pour cela il est nécessaire que ρ_k soit égal à 1, ce qui est réalisable (par (3.20)) lorsque G_k est proche de G_*. On voit donc qu'il est important de mettre G_k à jour même lorsque le pas ρ_k diffère de 1 et par conséquent d'étudier cette mise à jour dans un cadre global ($\rho_k \neq 1$) pour voir si la condition (3.20) peut être satisfaite. C'est ce que nous allons faire.

4. LA FORMULE DE BFGS

On utilisera pour la mise à jour de (G_k) une technique de quasi-Newton du type (1.2)-(1.4), en utilisant la formule de BFGS (*cf.* Broyden (1969), Fletcher (1970), Goldfarb (1970) et Shanno (1970)), c'est-à-dire avec G_{k+1} donnée par la formule :

$$G_{k+1} = G_k - \frac{G_k \sigma_k \sigma_k^T G_k}{(\sigma_k, G_k \sigma_k)} + \frac{\gamma_k \gamma_k^T}{(\gamma_k, \sigma_k)}, \tag{4.1}$$

où σ_k et γ_k sont deux vecteurs de $\mathbb{R}^{n-m}$. On notera schématiquement cette formule par $G_{k+1} = \text{BFGS}\ (G_k, \gamma_k, \sigma_k)$. Celle-ci est bien définie dès que (γ_k, σ_k) et $(\sigma_k, G_k \sigma_k)$ sont non nuls. De plus, si G_k est symétrique définie positive, G_{k+1} est définie positive si et seulement si (*cf.* Dennis, Moré (1977))

$$(\gamma_k, \sigma_k) > 0\ . \tag{4.2}$$

Comme on désire garder la définie positivité de G_k (*cf.* Section 3.3), la métrique ne sera pas mise à jour si la condition (4.2) n'est pas satisfaite. Lorsque (4.2) est vérifiée, on peut également générer les matrices inverses $H_k := G_k^{-1}$ par

$$H_{k+1} = \left[I - \frac{\sigma_k \gamma_k^T}{(\gamma_k, \sigma_k)} \right] H_k \left[I - \frac{\gamma_k \sigma_k^T}{(\gamma_k, \sigma_k)} \right] + \frac{\sigma_k \sigma_k^T}{(\gamma_k, \sigma_k)}, \tag{4.3}$$

formule que l'on note $H_{k+1} = \overline{\text{BFGS}}\ (H_k, \gamma_k, \sigma_k)$. Dans (4.1)-(4.3), le choix des vecteurs γ_k et σ_k de $\mathbb{R}^{n-m}$ dépendra de l'algorithme considéré.

On désignera par $\|\cdot\|_F$, la norme matricielle de Frobenius qui est définie par

$$\|H\|_F := (\text{tr}\ (H^T H))^{1/2} = \left(\sum_{i,j} H_{ij}^2 \right)^{1/2}.$$

Si M est une matrice carrée non singulière, on définit la norme matricielle $\|\cdot\|_{M\,F}$ par

$$\|H\|_{M,F} := \|MHM\|_F. \tag{4.4}$$

Le lemme suivant se déduit du lemme 5.2 de Broyden, Dennis et Moré (1973) en prenant $\beta = 1/3$ et en utilisant l'inégalité $(1 - 2\,\theta^2)^{1/2} \leqq 1 - \theta^2$.

LEMME 4.1 : *Soient M une matrice d'ordre $n - m$, symétrique, non singulière et γ et σ deux vecteurs de $\mathbb{R}^{n-m}$ vérifiant :*

$$(\gamma, \sigma) \neq 0 \quad \textit{et} \quad \|M\sigma - M^{-1}\gamma\| \leqq \frac{1}{3} \|M^{-1}\gamma\|. \tag{4.5}$$

Si H et H_* sont deux matrices d'ordre $n - m$, symétriques et si $\bar{H} := \overline{\text{BFGS}}\ (H, \gamma, \sigma)$, on a

$$\left\|\bar{H} - H_*\right\|_{M,F} \leqq \left(1 - \theta^2 + \frac{15}{4} \frac{\|M\sigma - M^{-1}\gamma\|}{\|M^{-1}\gamma\|}\right) \|H - H_*\|_{M,F} + 2(1 + 2(n-m)^{1/2}) \|M\|_F \frac{\|\sigma - H_*\gamma\|}{\|M^{-1}\gamma\|}, \tag{4.6}$$

où

$$\theta := \begin{cases} \dfrac{3^{1/2}}{4} \dfrac{\|M(H - H_*)\gamma\|}{\|H - H_*\|_{M,F} \|M^{-1}\gamma\|} & \textit{si} \quad H \neq H_*, \\ 0 \quad \textit{sinon}. \end{cases} \tag{4.7}$$

On utilisera ce lemme avec $M := H_*^{-1/2}$. Si H est une approximation de H_* et $\bar{H} := \overline{\text{BFGS}}\ (H, \gamma, \sigma)$, l'inégalité (4.6) montre que la distance entre $\bar{H}$ et H_* peut être contrôlée par la distance entre H et H_* et la manière dont γ et σ sont choisis par rapport à H_*. Le lemme suivant est extrait de la preuve du théorème 3 de Powell (1971). Il donne des conditions suffisantes sur les suites (γ_k) et (σ_k) pour que les suites de métriques (H_k) et (H_k^{-1}) soient bornées.

LEMME 4.2 : *Soient H_* une matrice d'ordre $n - m$, symétrique, définie positive et (γ_k) et (σ_k) deux suites de $\mathbb{R}^{n-m}$ telles que*

$$(\gamma_k, \sigma_k) > 0 \quad \textit{pour tout indice } k \textit{ et} \tag{4.8}$$

$$(\|\sigma_k - H_* \gamma_k\| / \|\gamma_k\|) \quad \textit{soit sommable .} \tag{4.9}$$

Si H_0 est une matrice d'ordre $n - m$, symétrique, définie positive, alors la formule de $\overline{\text{BFGS}}$ (4.3) génère à partir de H_0 une suite de matrices symétriques définies positives (H_k) telle que (H_k) et (H_k^{-1}) soient bornées.

Preuve : Soient $H_*^{1/2}$ la racine carrée symétrique définie positive de H_*, $G_*^{1/2} := (H_*^{1/2})^{-1}$ et

$$Q_k := I - \frac{H_*^{1/2} \gamma_k \gamma_k^T H_*^{1/2}}{(H_* \gamma_k, \gamma_k)} + \frac{G_*^{1/2} \sigma_k \sigma_k^T G_*^{1/2}}{(\gamma_k, \sigma_k)} .$$

Comme H_* est définie positive et (γ_k, σ_k) est positif, Q_k est bien définie et symétrique définie positive. On définit alors le vecteur z_k et les matrices J_k et R_k par :

$$z_k := Q_k^{1/2} H_*^{1/2} \gamma_k = Q_k^{-1/2} G_*^{1/2} \sigma_k ,$$
$$J_k := Q_k^{1/2} H_*^{1/2} G_k H_*^{1/2} Q_k^{1/2} ,$$
$$R_k := J_k - \frac{J_k z_k z_k^T J_k}{(J_k z_k, z_k)} .$$

R_k est bien définie car J_k est symétrique définie positive. Alors, grâce à (4.1), on a

$$Q_k^{1/2} H_*^{1/2} G_{k+1} H_*^{1/2} Q_k^{1/2} = R_k + \frac{z_k z_k^T}{\|z_k\|^2} .$$

Comme $R_k z_k = 0$ et $\|R_k\| \leqq \|J_k\|$, cette relation donne

$$\|Q_k^{1/2} H_*^{1/2} G_{k+1} H_*^{1/2} Q_k^{1/2}\| \leqq \max\ (1, \|J_k\|) ,$$

ou encore, en notant $m_k := \max(1, \|Q_k^{-1}\|) \max(1, \|Q_k\|)$,

$$\|H_*^{1/2} G_{k+1} H_*^{1/2}\| \leqq m_k \max(1, \|H_*^{1/2} G_k H_*^{1/2}\|).$$

De manière analogue, on obtient à partir de (4.3) :

$$\|G_*^{1/2} H_{k+1} G_*^{1/2}\| \leqq m_k \max(1, \|G_*^{1/2} H_k G_*^{1/2}\|).$$

Dès lors, il suffit de montrer que $\Pi(m_k) < +\infty$. Pour cela, on cherche une majoration de $\|Q_k\|$ et $\|Q_k^{-1}\|$. On écrit

$$Q_k - I = \frac{G_*^{1/2}(\sigma_k - H_* \gamma_k)\sigma_k^T G_*^{1/2}}{(\gamma_k, \sigma_k)} + \frac{H_*^{1/2}\gamma_k(\sigma_k - H_* \gamma_k)^T G_*^{1/2}}{(\gamma_k, \sigma_k)}$$
$$+ \frac{(H_* \gamma_k, \gamma_k) - (\gamma_k, \sigma_k)}{(\gamma_k, \sigma_k)(H_* \gamma_k, \gamma_k)} H_*^{1/2} \gamma_k \gamma_k^T H_*^{1/2}.$$

D'après (4.8) et (4.9), il existe des constantes positives C_1 et C_2 telles que $(\gamma_k, \sigma_k) \geqq C_1 \|\gamma_k\|^2$ et $\|\sigma_k\| \leqq C_2 \|\gamma_k\|$. En notant $\varepsilon_k := \|\sigma_k - H_* \gamma_k\| / \|\gamma_k\|$ et $C_3 := [C_2 \|G_*\| + \|H_*^{1/2}\| \, \|G_*^{1/2}\| + 1]/C_1$, la dernière relation donne

$$\|Q_k - I\| \leqq C_3 \varepsilon_k. \tag{4.10}$$

Soit $C_4 > C_3$. Pour k assez grand, disons $k \geqq K_1$, on a $C_3 \varepsilon_k \leqq 1 - C_3/C_4$. Alors, grâce à (4.10), on a :

$$\|Q_k\| \leqq 1 + C_3 \varepsilon_k,$$
$$\|Q_k^{-1}\| \leqq 1 + C_4 \varepsilon_k \quad \text{pour} \quad k \geqq K_1.$$

Ces deux inégalités et (4.9) montrent que $\Pi(m_k) < +\infty$. □

Le lemme suivant se déduit aisément du lemme 3.3 de Dennis et Moré (1974).

LEMME 4.3 : *Soient C_1, C_2 et C_3 trois constantes non négatives et (δ_k), (θ_k) et (χ_k) trois suites de réels non négatifs tels que*

$$\delta_{k+1} \leqq (1 - C_1 \theta_k + C_2 \chi_k) \delta_k + C_3 \chi_k \quad \textit{pour tout indice } k \textit{ et} \tag{4.11}$$

$$\sum_{k=0}^{\infty} \chi_k < +\infty. \tag{4.12}$$

Alors,

(i) (δ_k) *admet une limite* δ *et*

(ii) *si C_1 est positif, alors soit δ est nul, soit (θ_k) converge vers* 0.

5. DEUX ALGORITHMES

5.1. Choix des vecteurs γ_k et σ_k

La formule (4.3) montre qu'étant donnée H_k la qualité de H_{k+1} dépendra du choix des vecteurs γ_k et σ_k. Comme on désire que H_k soit proche de $H_* := G_*^{-1}$, un bon choix consisterait à prendre γ_k et σ_k tels que

$$\sigma_k = H_* \gamma_k . \tag{5.1}$$

En effet, comme H_{k+1} donné par (4.3) vérifie l'équation de quasi-Newton $\sigma_k = H_{k+1} \gamma_k$, le résultat de l'application de H_{k+1} et H_* à γ_k serait identique. H_* n'étant pas connue, on ne peut pas réaliser (5.1) exactement mais il est essentiel que cette relation soit vérifiée au premier ordre en γ_k, c'est-à-dire que l'on ait :

$$\sigma_k = H_* \gamma_k + o(\|\gamma_k\|) . \tag{5.2}$$

Cette condition apparaît en effet en (4.6) (avec $M = H_*^{-1/2}$) et en (4.9). H_* étant non singulière, cette relation montre que l'on a $\|\sigma_k\| \sim \|\gamma_k\|$ et par conséquent (5.2) est équivalente à

$$\gamma_k = G_* \sigma_k + o(\|\sigma_k\|) . \tag{5.3}$$

Ici, il est essentiel de remarquer que le gradient du gradient réduit g défini en (1.13) s'écrit (*cf.* par exemple Nocedal, Overton (1985)) :

$$\nabla g(x_*) = \nabla(Z_x^{-T}(\nabla f(x) + A_x^T \lambda_*))(x_*) = Z_*^{-T} L_* . \tag{5.4}$$

Comme $G_* = Z_*^{-T} L_* Z_*^{-}$, cette relation nous renseigne sur la manière de choisir γ_k et σ_k pour obtenir (5.3). γ_k doit être la différence de deux gradients réduits disons calculés aux points b_k et a_k, tels qu'asymptotiquement $(b_k - a_k)/\|b_k - a_k\|$ soit dans l'image de Z_*^{-}. Alors σ_k est pris de la forme $Z_*(b_k - a_k)$ où Z_* est fixé par les conditions (2.1) et (2.2) en $x = x_*$.

5.2. L'algorithme QNR1

Supposons que (y_k) converge vers x_*. Un premier choix de γ_k et σ_k est immédiat. Comme $t_k = x_{k+1} - y_k$ est dans l'image de $Z(y_k)^{-}$ (*cf.* formule (3.5)), on choisit :

$$\gamma_k := g(x_{k+1}) - g(y_k) , \tag{5.5}$$

$$\sigma_k := Z(y_k) t_k = - G_k^{-1} g(y_k) . \tag{5.6}$$

Dans (5.5), le calcul du gradient réduit en y_k et x_{k+1} nécessite la linéarisation des contraintes en ces deux points. Au lieu de (5.6), on pourrait également prendre $\sigma_k = Z(x_{k+1})\, t_k$ sans modifier les résultats (*cf.* Gilbert (1986a)). Ce choix de γ_k est plus simple que celui proposé par Coleman et Conn (1984). Montrons que l'on a bien (5.3). En utilisant (5.4), on obtient

$$\gamma_k = Z_*^{-T} L_*\, t_k + o(\|t_k\|)\,.$$

D'après (5.6), $t_k = Z(y_k)^-\, \sigma_k$ et comme $Z(y_k)^-$ est uniformément injective ($\|Z(y_k)^-\, u\| \geqq C\,\|u\|$, C constante positive indépendante de k), on a $\|t_k\| \sim \|\sigma_k\|$. Dès lors,

$$\gamma_k = Z_*^{-T} L_*\, Z_*^-\, \sigma_k + o(\|\sigma_k\|)\,,$$

c'est-à-dire (5.3).

On peut à présent définir l'algorithme QNR1.

Algorithme QNR1

Par rapport à l'algorithme QNR, on modifie/ajoute les étapes suivantes :

4.3. *Si* $k \geqq 1$
alors {Calculer $\gamma_{k-1} := g(x_k) - g(y_{k-1})$ et $\sigma_{k-1} := Z(y_{k-1})\, t_{k-1}$.
Si $(\gamma_{k-1}, \sigma_{k-1}) > 0$
alors $G_k :=$ BFGS $(G_{k-1}, \gamma_{k-1}, \sigma_{k-1})$
sinon $G_k := G_{k-1}$. }

4.6. Linéariser les contraintes en x_{k+1}.
Choisir un inverse à droite $A(x_{k+1})^-$ de $\nabla c(x_{k+1})$ et une base $Z(x_{k+1})^-$ de $N(\nabla c(x_{k+1}))$ tels que l'on ait (2.4).

4.7. Évaluer $g(x_{k+1}) := Z(x_{k+1})^{-T}\, \nabla f(x_{k+1})$.

4.8. $R_{k+1} := A(x_{k+1})^-$ et calculer $c(x_{k+1})$.

On voit que la métrique G_k n'est mise à jour que lorsque (γ_k, σ_k) est positif, ce qui permet de garantir la définie positivité de G_k. On définit

$$\mathbb{K}_1 := \{k \in \mathbb{N} : (\gamma_k, \sigma_k) > 0\}\,.$$

A l'étape 4.8, on a pu prendre $A(x_{k+1})^-$ comme opérateur de restauration des contraintes parce que celles-ci sont linéarisées en x_{k+1} (ainsi qu'un y_k). On pourrait également prendre $R_{k+1} = A(y_k)^-$.

5.3. L'algorithme QNR2

Le fait de devoir linéariser les contraintes aux deux points y_k et x_{k+1} peut dans certaines circonstances constituer un coût supplémentaire

important par rapport à l'algorithme modèle QNR : d'abord pour le calcul des éléments constituant $A(x_{k+1})$ et ensuite pour celui des opérateurs $A(x_{k+1})^-$ et $Z(x_{k+1})^-$. Par exemple, dans l'application traitées dans Blum, Gilbert et Thooris (1985) où il s'agit d'identifier une source non linéaire ($n - m \simeq 3$ paramètres) dans une équation aux dérivées partielles ($m \simeq 900$ contraintes), le temps de calcul pour linéariser les contraintes est environ 10 fois plus élevé que le temps de calcul de la phase de restauration des contraintes. Il est donc important de voir s'il est possible de se passer de cette linéarisation supplémentaire en prenant comme opérateur de restauration R_k la matrice $A(y_{k-1})^-$ et en choisissant :

$$\gamma_k := g(y_{k+1}) - g(y_k) , \tag{5.7}$$

$$\sigma_k := Z(y_k)\, v_k = \rho_k\, Z(y_k)\, t_k = -\rho_k\, G_k^{-1}\, g(y_k) , \tag{5.8}$$

où $v_k := y_{k+1} - y_k = \rho_k\, t_k + \rho_k^a\, r_{k+1}$. Si on suppose que (y_k) converge vers x_*, on a

$$\gamma_k = Z_*^{-T} L_*\, v_k + o(\|v_k\|) . \tag{5.9}$$

Pour avoir l'estimation (5.3), il faudrait que l'on ait $v_k = O(\|\sigma_k\|)$, ce qui n'est pas vrai en général (par exemple lorsque $t_k = 0$). Le lemme suivant permet de dégager des situations dans lesquelles l'estimation (5.3) a lieu.

LEMME 5.1 : *Soit (G_k) une suite de matrices d'ordre $n - m$ symétriques définies positives telles que (G_k) et (G_k^{-1}) soient bornées. Soient (x_k) et (y_k) les suites de points générées par l'algorithme modèle* QNR *avec les métriques (G_k) et $R_{k+1} = A(y_k)^-$ ou $A(x_{k+1})^-$. Supposons que (y_k) converge vers x_*. Alors, les relations suivantes sont équivalentes :*

$$\text{(i)} \qquad \rho_k^a\, c(y_k) = \rho_k\, o(\|t_k\|) ,$$

$$\text{(ii)} \qquad \rho_k^a\, r_{k+1} = \rho_k\, o(\|t_k\|) ,$$

$$\text{(iii)} \qquad v_k = Z_*^-\, Z_*\, v_k + \rho_k\, o(\|t_k\|) .$$

De plus, si l'une de ces relations est vérifiée, on a

$$\text{(iv)} \qquad \|v_k\| \sim \|Z_*\, v_k\| \sim \rho_k \|t_k\| .$$

Preuve : On utilise systématiquement l'équivalence (3.17) qui, en particulier, implique que $t_k = O(\|e_k\|)$ et $c(y_k) = O(\|e_k\|)$. En développant $c(x_{k+1})$ autour de y_k, on obtient :

$$c(x_{k+1}) = c(y_k) + o(\|t_k\|) . \tag{5.10}$$

On en déduit

$$r_{k+1} = -A_*^- c(y_k) + o(\|e_k\|)\,. \tag{5.11}$$

Dès lors, (i) implique que $\rho_k^q r_{k+1} = \rho_k o(\|t_k\|) + \rho_k^q o(\|e_k\|)$ et comme $\rho_k \leqq 1$ et $\|e_k\| \leqq \|t_k\| + \|r_{k+1}\|$, on en déduit (ii). Inversement, (ii), (5.11) et l'injectivité de A_*^- impliquent (i). Comme $t_k = Z_*^- Z_* t_k + o(\|t_k\|)$, on obtient avec (2.1) et (2.3) :

$$\begin{aligned} v_k &= Z_*^- Z_* v_k + A_*^- A_* v_k \\ &= Z_*^- Z_* v_k + \rho_k^q r_{k+1} + \rho_k o(\|t_k\|) + \rho_k^q o(\|r_{k+1}\|)\,, \end{aligned}$$

d'où on déduit l'équivalence entre (ii) et (iii). D'après (ii), $v_k = \rho_k t_k + \rho_k o(\|t_k\|)$ et donc $\|v_k\| \sim \rho_k \|t_k\|$. Cette équivalence, (iii) et l'injectivité de Z_*^- impliquent que $\|v_k\| \sim \|Z_* v_k\|$. □

Supposons à présent que les conditions (i)-(iv) du lemme 5.1 soient vérifiées. Alors (5.8) montre que $\|\sigma_k\| \sim \|v_k\|$ et

$$\sigma_k = Z_* v_k + o(\|\sigma_k\|)\,. \tag{5.12}$$

De (5.9) et (iii) on déduit que

$$\gamma_k = G_* Z_* v_k + o(\|\sigma_k\|)\,. \tag{5.13}$$

Dès lors, (5.12) et (5.13) montrent que l'on a l'estimation (5.3).

On l'a vu, il est important d'avoir (au moins) l'estimation (5.3) pour que la méthode de quasi-Newton génère des suites superlinéairement convergentes. D'après ce qui précède, cette estimation est vérifiée si on a (*cf.* la condition (ii) du lemme 5.1) :

$$\rho_k^{q-1} \|r_{k+1}\| \leqq \bar{\mu}_k \|t_k\|\,, \tag{5.14}$$

où $(\bar{\mu}_k)$ est une suite de réels positifs convergeant vers 0. Étant donnée une telle suite, on décide alors de ne mettre à jour G_k que si (5.14) est vérifiée, c'est-à-dire si le pas de restauration $\rho_k^q r_{k+1}$ est (asymptotiquement) petit devant le pas de minimisation $\rho_k t_k$. Si (5.14) n'est pas vérifiée, on prend G_{k+1} égal à G_k. Le test (5.14) est donc utilisé comme *critère de mise à jour*.

A présent, le problème se ramène à choisir judicieusement la suite $(\bar{\mu}_k)$ pour que lorsque (5.14) n'est pas vérifiée, on ait quand même la convergence superlinéaire. Il s'agit en fait de réaliser un compromis entre les deux situations suivantes :

— soit la métrique est rarement mise à jour ($(\bar{\mu}_k)$ décroît rapidement) et le vecteur γ_k est peu perturbé par rapport au vecteur $\bar{\gamma}_k := g(y_k + \rho_k t_k) - g(y_k)$ qui est la variation de gradient réduit normalement associée à σ_k,

— soit la métrique est souvent mise à jour ($(\bar{\mu}_k)$ décroît lentement) et le vecteur γ_k est fort perturbé par rapport à $\bar{\gamma}_k$.

On comprend qu'il est préférable que ce soit l'algorithme lui-même qui choisisse le taux de décroissance de la suite $(\bar{\mu}_k)$, en fonction de l'information dont il dispose à l'itération courante. Il me semble remarquable qu'un tel choix puisse être fait de façon à obtenir la convergence superlinéaire de toute la suite (x_k), et cela sous des hypothèses raisonnables. Pour cela, on désigne par $\mathbb{K}$ l'ensemble des indices k pour lesquels la matrice G_k est mise à jour. Étant donné un indice k, on désigne par (k^-) le plus grand indice précédent k pour lequel il y a eu mise à jour de G_k. De façon précise :

$$\begin{aligned} &k^- := 0 \quad si \quad i \in \mathbb{K} \Rightarrow i \geqq k\,, \\ &k^- := \max\,\{i \in \mathbb{K} : i < k\} \; sinon\,. \end{aligned} \tag{5.15}$$

On définit ensuite

$$k^{--} := (k^-)^-\,.$$

On prend alors la suite $(\bar{\mu}_k)$ définie par

$$\bar{\mu}_k := \mu_k \|e_{k^{--}}\| \tag{5.16}$$

où e_k est défini en (3.7) et (μ_k) est encore une suite décroissante de réels positifs qui, contrairement à $(\bar{\mu}_k)$, converge vers un nombre non nul. En fait, il suffirait de prendre une constante positive assez petite de telle sorte qu'asymptotiquement, on ait (σ_k, γ_k) positif lorsque le critère est satisfait (voir plus loin et section 7). Il se fait que l'algorithme peut lui-même adapter cette constante à une valeur convenable. Avec cette valeur de $\bar{\mu}_k$, le *critère de mise à jour* devient :

$$\rho_k^{a-1}\|r_{k+1}\| \leqq \mu_k \|e_{k^{--}}\|\,\|t_k\|\,. \tag{5.17}$$

Nocedal et Overton (1985) ont étudié la mise à jour de la matrice réduite dans l'algorithme (1.18). Ils utilisent également un critère de mise à jour semblable à (5.14) avec $\bar{\mu}_k$ donné *a priori* par

$$\bar{\mu}_k = \frac{\bar{\mu}_0}{(1+k)^{1+\nu}}\,,$$

où $\bar{\mu}_0$ et ν sont des constantes positives données a priori. On obtient alors la convergence superlinéaire (en deux pas, *cf.* les exemples de Byrd (1985) et de Yuan (1985)) dans un cadre local ($\rho_k = 1$ et (x_0, G_0) proche de

(x_*, G_*)) lorsque $\bar{\mu}_0$ est pris assez petit. Par rapport à ce choix, la valeur de $\bar{\mu}_k$ donnée en (5.16) me semble avoir deux avantages :

— elle n'est pas définie a priori mais gérée par l'algorithme en tenant compte de l'information courante,
— le critère peut être utilisé dans un cadre global.

On peut à présent définir l'algorithme QNR2.

Algorithme QNR2

Par rapport à l'algorithme QNR, on modifie/ajoute les étapes suivantes :

1. Se donner α dans $]0, 1/2[$, β dans $]0, 1[$, a dans $]1, +\infty[$, μ_0 dans $]0, +\infty[$ et ε dans $]0, +\infty[$.

4.3. *Si* $k \geqq 1$
alors {Calculer $\gamma_{k-1} := g(y_k) - g(y_{k-1})$ et $\sigma_{k-1} := \rho_{k-1} Z(y_{k-1})\, t_{k-1}$.
Si $(\gamma_{k-1}, \sigma_{k-1}) > 0$ *et* (5.17)
alors $G_k := \mathrm{BFGS}(G_{k-1}, \gamma_{k-1}, \sigma_{k-1})$
sinon $G_k := G_{k-1}$.}.

4.4. (* adaptation de μ_k^*)
Si $(\gamma_{k-1}, \sigma_{k-1}) \leqq 0$ *et* $\sigma_{k-1} \neq 0$ *et* (5.17)
alors $\mu_k := \mu_{k-1}/2$
sinon $\mu_k := \mu_{k-1}$.

4.8. $R_{k+1} := A(y_k)^-$ et calculer $c(x_{k+1})$.

Cet algorithme ne nécessite donc la linéarisation des contraintes qu'au seul point y_k. La restauration des contraintes se fait alors avec l'opérateur $R_{k+1} = A(y_k)^-$. La mise à jour de la matrice G_k ne se fait que si $k \in \mathbb{K}$ où

$$\mathbb{K} := \mathbb{K}_1 \cap \mathbb{K}_2 , \tag{5.18}$$

$$\mathbb{K}_1 := \{k \in \mathbb{N} : (\gamma_k, \sigma_k) > 0\} , \tag{5.19}$$

$$\mathbb{K}_2 := \{k \in \mathbb{N} : \rho_k^{a-1} \| r_{k+1} \| \leqq \mu_k \| e_{k--} \| \, \| t_k \| \} . \tag{5.20}$$

Pour QNR1, on a $\mathbb{K}_2 = \mathbb{N}$ puisqu'il n'y a pas de critère de mise à jour.

L'utilisation de la suite décroissante (μ_k) s'avère nécessaire pour éviter la situation où la matrice réduite n'est plus mise à jour du fait que (γ_k, σ_k) est non positif alors que le critère est satisfait. C'est-à-dire que l'on veut éviter la situation où

$$\mathbb{K} \text{ est borné et } \mathbb{K}_2 \backslash \mathbb{K}_1 \text{ est non borné} . \tag{5.21}$$

Cette situation est inconfortable puisque pour $k \in \mathbb{K}_2 \backslash \mathbb{K}_1$, on n'a ni $r_{k+1} = o(\| t_k \|)$ (car $\mathbb{K}$ est borné et donc $\| e_{k--} \|$ est constant), ni la

possibilité d'utiliser la négation de (5.17) (car $k \in \mathbb{K}_2$). On l'évite si μ_k devient suffisamment petit. En effet, pour $k \in \mathbb{K}_2$, on montrera que l'on a

$$\gamma_k = G_* \sigma_k + o(\|\sigma_k\|) + \mu_k o(\|e_{k^{--}}\| \|\sigma_k\|) . \tag{5.22}$$

Cette estimation affine la relation (5.3) que l'on retrouve lorsque $\mathbb{K}$ est non borné. Lorsque $k \in \mathbb{K}_2 \backslash \mathbb{K}_1$, on a en utilisant (5.22) :

$$0 \geqq (\gamma_k, \sigma_k) \geqq (\|H_*\|^{-1} - C\mu_k - \eta_k)\|\sigma_k\|^2 ,$$

où C est une constante positive et (η_k) converge vers 0. Dès lors, si μ_k devient assez petit, on a $\sigma_k = 0$, donc $t_k = 0$ et $r_{k+1} = 0$, grâce au critère (5.17). C'est-à-dire que l'algorithme s'arrête. La situation (5.21) est donc évitée si μ_k devient assez petit.

6. CONVERGENCE LOCALE

Le point de vue adopté dans cette section est local dans le sens où le résultat de convergence démontré (théorème 6.2) suppose que ρ_k est pris égal à 1 à chaque itération et que (y_0, G_0) est choisi proche de (x_*, G_*). On montre alors que (x_k) converge linéairement vers x_*.

Dans un premier temps, on montre grâce au lemme 4.1 que, quel que soit l'algorithme QNR1 ou QNR2 considéré, une inégalité du type (4.11) est vérifiée avec

$$\delta_k := |||H_k - H_*||| , \tag{6.1}$$

où $|||.|||$ est la norme matricielle définie par (*cf.* (4.4))

$$|||H||| = |||H|||_{G_*^{1/2}, F} .$$

Pour un indice k donné, on note, en fonction de l'algorithme considéré,

$$\pi(k) := \begin{cases} k & \text{pour QNR1 ,} \\ k^{--} & \text{pour QNR2 .} \end{cases}$$

On définit également les quantités suivantes :

$$\tau_k := \begin{cases} 1 & \text{si } k \in \mathbb{K} , \\ 0 & \text{sinon ,} \end{cases} \tag{6.2}$$

$$\theta_k := \begin{cases} \dfrac{3^{1/2}}{4} \quad \dfrac{\|G_*^{1/2}(H_k - H_*)\gamma_k\|}{|||H_k - H_*||| \, \|H_*^{1/2}\gamma_k\|} & \text{si } \gamma_k \neq 0 \quad \text{et} \quad H_k \neq H_* , \\ 0 \qquad \text{sinon .} \end{cases} \tag{6.3}$$

LEMME 6.1 : *Soit k un indice fixé. Soient $(y_i : 0 \leqq i \leqq k)$ et $(x_i : 1 \leqq i \leqq k+1)$ des points générés par l'un des algorithmes* QNR1 *ou* QNR2. *On suppose que ces points sont contenus dans la boule $B(x_*, \zeta)$. Alors, il existe des constantes positives ω_1, ω_2, ω_3, ω_4 et ε ne dépendent que de ζ, de μ_0 (pour* QNR2) *et de f et c sur $B(x_*, \zeta)$ telles que*

$$\|\gamma_k - G_* \sigma_k\| \leqq \omega_1 \chi_k \|\sigma_k\| \text{ pour } k \text{ dans } \mathbb{K}_2 , \tag{6.4}$$

où

$$\chi_k := \begin{cases} \|y_k - x_*\| + \|x_{k+1} - x_*\| \text{ pour QNR1} , \\ \mu_k \|e_k^{--}\| + \|y_k - x_*\| + \|x_{k+1} - x_*\| \text{ pour QNR2} . \end{cases}$$

Si de plus, $y_{\pi(k)}$, y_k, $x_{\pi(k)+1}$ et x_{k+1} sont dans $B(x_*, \varepsilon)$, alors

$$\|\sigma_k - H_* \gamma_k\| \leqq \omega_2 \chi_k \|\gamma_k\| \text{ pour } k \text{ dans } \mathbb{K}_2 , \tag{6.5}$$

$$\delta_{k+1} \leqq (1 - \tau_k \theta_k^2 + \omega_3 \tau_k \chi_k) \delta_k + \omega_4 \tau_k \chi_k . \tag{6.6}$$

Remarques : Dans (6.3), (6.4) et (6.5), γ_k et σ_k sont donnés par (5.5)-(5.6) ou (5.7)-(5.8) selon l'algorithme considéré. On ne suppose pas que le pas vaut un. On pourra donc utiliser ce lemme dans la section 7.

Preuve du lemme 6.1 : L'inégalité (6.4) s'obtient par les arguments utilisés dans les sections 5.2 et 5.3 que l'on reprend ici mais sans supposer la convergence de (y_k) vers x_*. La relation (6.5) se déduit alors de (6.4) lorsque χ_k est assez petit. Enfin, (6.6) provient de (6.5) et du lemme 4.1.

Soit C la constante de Lipschitz de l'application

$$x \to (A_x, A_x^-, Z_x^-, Z_x, \nabla g(x)) , \tag{6.7}$$

sur $B(x_*, \zeta)$. On désignera par C_1, C_2, ... des constantes positives ne dépendant que de ζ, de μ_0 (pour QNR2) et de f et c sur $B(x_*, \zeta)$.

1. Montrons d'abord les inégalités (6.4) et (6.5) pour l'algorithme QNR1. On a $t_k = Z(y_k)^- \sigma_k$ et donc

$$\|t_k - Z_*^- \sigma_k\| \leqq C \|y_k - x_*\| \|\sigma_k\| . \tag{6.8}$$

On en déduit

$$\|t_k\| \leqq C_1 \|\sigma_k\| . \tag{6.9}$$

D'autre part, en utilisant le caractère Lipschitzien de ∇g et (5.4), on obtient :

$$\|\gamma_k - Z_*^{-T} L_* t_k\| \leqq C \chi_k \|t_k\| .$$

Grâce à (6.8) et (6.9), on en déduit (6.4) avec $\omega_1 := C(C_1 + \|Z_*^{-T} L_*\|)$. Si y_k et x_{k+1} sont dans $B(x_*, \varepsilon)$ avec $\varepsilon \leqq \varepsilon_1 := (4 \omega_1 \|H_*\|)^{-1}$, on déduit de (6.4) que $\|\sigma_k\| \leqq 2 \|H_*\| \|\gamma_k\|$ et donc aussi (6.5) avec $\omega_2 := 2 \omega_1 \|H_*\|^2$.

2. Montrons à présent ces mêmes inégalités (6.4) et (6.5) pour l'algorithme QNR2. On suppose que $k \in \mathbb{K}_2$ et donc que le critère (5.17) est vérifié. Comme $\rho_k t_k = Z(y_k)^- \sigma_k$, on obtient :

$$\|\rho_k t_k - Z_*^- \sigma_k\| \leqq C \|y_k - x_*\| \|\sigma_k\| . \tag{6.10}$$

On en déduit :

$$\|\rho_k t_k\| \leqq C_2 \|\sigma_k\| . \tag{6.11}$$

En utilisant $v_k = \rho_k t_k + \rho_k^a r_{k+1}$, le critère (5.17), (6.10) et (6.11), on obtient :

$$\|v_k - Z_*^- \sigma_k\| \leqq C_3 \chi_k \|\sigma_k\| . \tag{6.12}$$

Comme $y_{k^{--}}$ et $y_{(k^{--})+1}$ sont dans $B(x_*, \zeta)$, $\chi_k \leqq C_4$ et l'inégalité (6.12) implique que

$$\|v_k\| \leqq C_5 \|\sigma_k\| . \tag{6.13}$$

Notons que, comme $y_{k+1} = y_k + \rho_k(x_{k+1} - y_k) - \rho_k^a A(y_k)^- c(x_{k+1})$ et $\rho_k \leqq 1$, on a :

$$\|y_{k+1} - x_*\| \leqq C_6(\|y_k - x_*\| + \|x_{k+1} - x_*\|) . \tag{6.14}$$

D'autre part, avec (5.7), le caractère Lipschitzien de ∇g et (5.4), puis (6.14) et (6.13), on a

$$\begin{aligned} \|\gamma_k - Z_*^{-T} L_* v_k\| &\leqq C(\|y_k - x_*\| + \|y_{k+1} - x_*\|)\|v_k\| \\ &\leqq C_7 \chi_k \|v_k\| \\ &\leqq C_8 \chi_k \|\sigma_k\| . \end{aligned}$$

Enfin, cette inégalité et (6.12) permettent d'obtenir (6.4) avec $\omega_1 := C_8 + C_3\|Z_*^{-T} L_*\|$. Remarquons qu'il existe une constante positive C_9 telle que si $y_{k^{--}}$, $x_{(k^{--})+1}$, y_k et x_{k+1} sont dans $B(x_*, \varepsilon)$, on a $\chi_k \leqq \varepsilon C_9$. Dès lors si $\varepsilon \leqq \varepsilon_2 := (2 C_9 \omega_1 \|H_*\|)^{-1}$, on déduit de (6.4) que $\|\sigma_k\| \leqq 2\|H_*\| \|\gamma_k\|$ et donc aussi (6.5) avec $\omega_2 := 2 \omega_1 \|H_*\|^2$.

3. On conclut en montrant l'inégalité (6.6). On peut supposer que k appartient à $\mathbb{K}$ sinon $\tau_k = 0$ et $\delta_{k+1} = \delta_k$ montrent que (6.6) est trivialement vérifiée. Si $k \in \mathbb{K}$, (γ_k, σ_k) est positif et on a (6.4) et (6.5). Dès lors si ε est pris assez petit pour que $\omega_2 \chi_k \|G_*\| \leqq 1/3$, les conditions (4.5) sont satisfaites avec $M = G_*^{1/2}$ et on peut appliquer le lemme 4.1. En utilisant la relation (6.5), l'inégalité (4.6) implique (6.6) avec $\omega_3 := 15 \omega_2 \|G_*\|/4$ et $\omega_4 := 2(1 + 2\sqrt{n-m})\, \omega_2 \|G_*^{1/2}\| \|G_*^{1/2}\|_F$. □

Le théorème suivant montre que la méthode de mise à jour de BFGS de H_k dans les algorithmes QNR1 et QNR2 avec le pas unité permet d'obtenir

la convergence linéaire de la suite (x_k) pour autant que (y_0, G_0) soit pris suffisamment proche de (x_*, G_*).

THÉORÈME 6.2 : *Soient* κ *un réel dans* $]0, 1[$ *et* s *une constante positive. Il existe une constante positive* ε *ne dépendant que de* κ, s, μ_0 *(pour* QNR2*) et de* f *et* c *dans un voisinage de* x_* *telle que si*

$$\|y_0 - x_*\| \leqq \varepsilon , \tag{6.15}$$

$$\|G_0 - G_*\| \leqq \varepsilon , \tag{6.16}$$

alors, l'algorithme QNR1 *ou* QNR2 *avec le pas unité génère des suites* (x_k) *et* (G_k) *telles que*

(i) (G_k) *et* (H_k) *sont bornées,*
(ii) $\|G_k - G_*\| \leqq s$ *pour tout indice* k,
(iii) $\|x_{k+1} - x_*\| \leqq \kappa \|x_k - x_*\|$ *pour tout indice* k.

Preuve : L'idée de la démonstration est tout à fait similaire à celle utilisée en optimisation sans contrainte. On montre que la distance entre H_k et H_* ne se détériore pas trop au cours des itérations (c'est le *lemme de détérioration bornée* de Broyden, Dennis, Moré (1973)) : on peut s'arranger pour que $|||H_k - H_*|||$ soit majoré par $2M$ si M est un majorant de $|||H_0 - H_*|||$.

1. On introduit d'abord un certain nombre de constantes positives permettant de déterminer ε. Soient $\kappa \in]0, 1[$ et $s > 0$ les constantes données dans l'énoncé. Soit $\eta > 0$ telle que pour tout H dans $L(\mathbb{R}^{n-m})$, on ait

$$\frac{1}{\eta} |||H||| \leq \|H\| \leq \eta |||H||| .$$

Soient $\varepsilon_1(\kappa)$ et $\varepsilon_2(\kappa)$ les constantes positives données par le lemme 3.1. Pour $\zeta > 0$, soient $\omega_1(\zeta)$, $\omega_2(\zeta)$, $\omega_3(\zeta)$, $\omega_4(\zeta)$ et $\varepsilon_3(\zeta)$ les constantes données par le lemme 6.1. On peut supposer que

$$\varepsilon_2 \leqq s \quad \text{et} \quad \varepsilon_3 \leqq \min(\varepsilon_1, \varepsilon_2, \zeta) .$$

On désignera par C_1, C_2, ... des constantes positives ne dépendant que de ζ, de μ_0 (pour QNR2) et de f et c sur $B(x_*, \zeta)$. Lorsque le pas ρ_k vaut 1 et lorsque les points $(y_i : 0 \leqq i \leqq k)$ et $(x_i : 1 \leqq i \leqq k+1)$ sont dans la boule $B(x_*, \varepsilon_3)$, on peut récrire l'inégalité (6.6). En négligeant le terme négatif du membre de droite, on obtient :

$$\delta_1 \leqq \delta_0 + C_1 \tau_0(\omega_3 \delta_0 + \omega_4)(\|y_0 - x_*\| + \|x_1 - x_*\|) , \tag{6.17}$$

$$\delta_{k+1} \leqq \delta_k + C_2 \tau_k(\omega_3 \delta_k + \omega_4)(\|x_{\pi(k)} - x_*\| + \|x_{\pi(k)+1} - x_*\| + \|x_k - x_*\| + \|x_{k+1} - x_*\|) \tag{6.18}$$

pour $k \geqq 1$. Soient C_3 et C_4 des constantes positives telles que lorsque y_{k-1}, x et z sont dans $B(x_*, \zeta)$, on ait

$$\|x_k - x_*\| \leqq (1 + C_3\|H_{k-1}\|)(\|y_{k-1} - x_*\|) , \tag{6.19}$$

$$\|x - A(z)^- c(x) - x_*\| \leqq C_4\|x - x_*\| . \tag{6.20}$$

On définit alors la constante δ par

$$\delta := \varepsilon_2/[2\eta^2\|H_*\|(\|G_*\| + \varepsilon_2)] . \tag{6.21}$$

Ensuite, on prend ε_4 positif tel que

$$\varepsilon_4 \leqq \varepsilon_3/(1 + C_4) \quad \text{et} \tag{6.22}$$

$$\varepsilon_4 \leqq \eta\delta(1 - \kappa)/[4\, C_2(2\,\eta\delta\omega_3 + \omega_4)(3 - 2\,\kappa)] . \tag{6.23}$$

Alors, il suffit de prendre ε positif tel que

$$\varepsilon \leqq \delta/[\|G_*\|(\|H_*\| + \delta)] , \tag{6.24}$$

$$\varepsilon \leqq \varepsilon_4/[1 + C_3(\|H_*\| + \delta)] , \tag{6.25}$$

$$\varepsilon \leqq \eta\delta/[C_1(\eta\delta\omega_3 + \omega_4)(2 + C_3\|H_*\| + C_3\,\delta)] . \tag{6.26}$$

2. On montre par récurrence que si (y_0, G_0) vérifie (6.15) et (6.16), alors

$$\|\!|H_k - H_*|\!\| \leqq 2\,\eta\delta \quad \text{pour} \quad k \geqq 0 , \tag{6.27}$$

$$\|x_{k+1} - x_*\| \leqq \kappa\|x_k - x_*\| \quad \text{pour} \quad k \geqq 1 , \tag{6.28}$$

$$y_k \in B(x_*, \varepsilon_3) \quad \text{et} \quad x_{k+1} \in B(x_*, \varepsilon_4) \quad \text{pour} \quad k \geqq 0 , \tag{6.29}$$

$$\delta_{k+1} \leqq \delta_k + 4\, C_2\,\tau_k(2\,\eta\delta\omega_3 + \omega_4)\|x_{\pi(k)} - x_*\| \quad \text{pour} \quad k \geqq 1 . \tag{6.30}$$

Montrons (6.27) et (6.29) pour $k = 0$. D'après (6.16) et (6.24), G_0 est non singulière et on a

$$\|H_0 - H_*\| \leqq \|H_0\|\,\|H_*\|\,\|G_0 - G_*\| \leqq \delta . \tag{6.31}$$

On en déduit $(6.27)_0$. Comme $\varepsilon \leqq \varepsilon_3$ (grâce à (6.25) et (6.22)), (6.15) montre que y_0 est dans $B(x_*, \varepsilon_3)$. Comme $\|H_0\| \leqq \|H_*\| + \delta$, on déduit de (6.19), (6.15) et (6.25) que x_1 est dans $B(x_*, \varepsilon_4)$.

Montrons à présent (6.27)-(6.30) pour $k = 1$. Grâce à $(6.29)_0$, on peut utiliser l'inégalité (6.17). Comme $\delta_0 \leqq \eta\delta$, cette inégalité (6.17), (6.19), (6.31) et (6.26) impliquent $(6.27)_1$. Cette estimation de δ_1 implique que $\|H_1 - H_*\| \leqq 2\,\eta^2\,\delta$ qui grâce à (6.21) montre que H_1 est non singulière et

que $\|G_1 - G_*\| \leqq \|G_1\| \|G_*\| \|H_1 - H_*\| \leqq \varepsilon_2$. Dès lors, on peut appliquer le lemme 3.1 qui montre $(6.28)_1$. Ensuite, $y_1 \in B(x_*, \varepsilon_3)$ grâce à (6.20) et (6.22). Tandis que $x_2 \in B(x_*, \varepsilon_4)$ grâce à $(6.28)_1$ et $(6.29)_0$. Finalement, $(6.30)_1$ s'obtient grâce à (6.18), $(6.27)_1$ et (6.28).

Supposons à présent que $(6.27)_i$-$(6.30)_i$ soient satisfaites pour $i = 1, \ldots, k-1$ et montrons qu'elles le sont encore pour $i = k$. En sommant les inégalités $(6.30)_i$ pour $i = 1, \ldots, k-1$ et en tenant compte de (6.31), puis de (6.28) et $\|x_1 - x_*\| \leqq \varepsilon_4$ et enfin de (6.23), on obtient :

$$\begin{aligned}
\delta_k &\leqq \delta_1 + 4\, C_2(2\,\eta\delta\omega_3 + \omega_4) \sum_{i=1}^{k-1} \tau_i \|x_{\pi(i)} - x_*\| \\
&\leqq \eta\delta + 4\, C_2(2\,\eta\delta\omega_3 + \omega_4)\left(\sum_{i=1}^{k-1} \|x_i - x_*\| + 2\,\varepsilon_4 \right) \\
&\leqq \eta\delta + 4\, C_2(2\,\eta\delta\omega_3 + \omega_4)\left(\frac{1}{1-\kappa} + 2 \right) \varepsilon_4 \\
&\leqq 2\,\eta\delta\,.
\end{aligned}$$

On a donc $(6.27)_k$. $(6.28)_k$ s'obtient alors grâce au lemme 3.1 car on a déjà vu que $(6.27)_k$ implique que $\|G_k - G_*\| \leqq \varepsilon_2$. $(6.29)_k$ et $(6.30)_k$ se déduisent alors comme précédemment.

3. Il reste à conclure. La condition (i) du théorème est claire puisqu'on a toujours $\|H_k - H_*\| \leqq 2\,\eta^2\,\delta$ et $\|G_k - G_*\| \leqq \varepsilon_2$. On a aussi (ii) puisqu'on a supposé $\varepsilon_2 \leqq s$. Enfin, on a (iii) grâce à (6.28). □

Sous les hypothèses du théorème précédent, (x_k) converge linéairement et (H_k) est bornée. Dès lors (y_k) converge également vers x_* et il existe deux constantes C_1 et C_2 telles que

$$\begin{aligned}
\|y_{k+2} - x_*\| &\leqq C_1 \|x_{k+2} - x_*\|\,, \\
\|x_{k+2} - x_*\| &\leqq \kappa \|x_{k+1} - x_*\|\,, \\
\|x_{k+1} - x_*\| &\leqq C_2 \|y_k - x_*\|\,.
\end{aligned}$$

On peut s'arranger pour que $C_1\, C_2\, \kappa < 1$ et donc obtenir la convergence linéaire *en deux pas* de (y_k) (*cf.* Gilbert (1986a), corollaire VI.4.3) : c'est un résultat de Coleman et Conn (1984).

7. CONVERGENCE SUPERLINÉAIRE

Dans cette section, nous allons montrer la convergence superlinéaire de la suite (x_k) générée par l'algorithme QNR1 (théorème 7.1) ou par l'algorithme QNR2 (théorème 7.2). Pour cela, nous supposerons que la suite (x_k) converge vers x_* et nous nous intéresserons à son taux de convergence.

THÉORÈME 7.1 : *Soient* (x_k), (y_k) *et* (G_k) *les suites générées par l'algorithme* QNR1. *On suppose que* (x_k) *et* (y_k) *convergent vers* x_* *avec*

$$\sum_{k=0}^{\infty} \|x_k - x_*\| < +\infty \quad \textit{et} \tag{7.1}$$

$$\sum_{k=0}^{\infty} \|y_k - x_*\| < +\infty \,. \tag{7.2}$$

Alors :
(i) (G_k) *et* (G_k^{-1}) *sont bornées,*
(ii) $(G_k - G_*)\,\sigma_k = o(\|\sigma_k\|)$ *pour* k *dans* $\mathbb{K}$,
(iii) $\rho_k = 1$ *pour* k *assez grand,*
(iv) (x_k) *converge superlinéairement,*
(v) (y_k) *converge superlinéairement en deux pas.*

Remarques : Les conditions (7.1) et (7.2) sont vérifiées sous les hypothèses du théorème 6.2 puisqu'alors (y_k) converge linéairement en deux pas. L'utilisation de la formule de BFGS et le fait que la positivité de (γ_k, σ_k) conditionne la mise à jour de G_k assurent la définie positivité des matrices réduites. Dans la conséquence (ii) du théorème, σ_k est donné en (5.6) et $\mathbb{K}$ est défini par

$$\mathbb{K} = \mathbb{K}_1 := \{k \in \mathbb{N} : (\gamma_k, \sigma_k) > 0\},$$

où γ_k est donné en (5.5). La conséquence (iii) veut dire que ρ_k vaut 1 après un nombre fini d'itérations.

Preuve du théorème 7.1 : Le schéma de la démonstration est le suivant. La propriété (i) provient du théorème de Powell (lemme 4.2). Pour (ii), on utilise la technique de Broyden, Dennis et Moré (1973) qui, pour l'essentiel, est condensée dans le lemme 6.1. L'estimation (ii) et le théorème 3.2 *via* la condition (3.20) permettent alors de montrer que ρ_k vaut 1 après un nombre fini d'itérations. Alors, asymptotiquement on se retrouve avec l'algorithme local (1.14)-(1.15) et grâce à (ii), on peut obtenir la condition (3.18) de convergence superlinéaire. Enfin, la propriété (v) est une conséquence immédiate de (iv).

1. *On prouve (i) et (ii).* Comme (x_k) et (y_k) convergent vers x_*, il existe un réel positif ζ tel que ces suites soient contenues dans la boule $B(x_*, \zeta)$. Soient alors $\omega_1(\zeta)$, $\omega_2(\zeta)$, $\omega_3(\zeta)$, $\omega_4(\zeta)$ et $\varepsilon(\zeta)$ les constantes données par le lemme 6.1. On peut supposer $\varepsilon \leqq \zeta$. Comme (x_k) et (y_k) convergent vers x_*, il existe un indice K_1 tel que x_k et y_k soient dans

$B(x_*, \varepsilon)$ pour $k \geqq K_1$. Ceci permet d'utiliser les inégalités (6.5) et (6.6) du lemme 6.1, dès que $k \geqq K_1$:

$$\|\sigma_k - H_* \gamma_k\| \leqq \omega_2 \chi_k \|\gamma_k\|, \tag{7.3}$$

$$\delta_{k+1} \leqq (1 - \tau_k \theta_k^2 + \omega_3 \tau_\kappa \chi_k) \delta_k + \omega_4 \tau_k \chi_k , \tag{7.4}$$

où, $\chi_k := \|y_k - x_*\| + \|x_{k+1} - x_*\|$ et δ_k, τ_k et θ_k sont définis en (6.1)-(6.3). D'après (7.1) et (7.2), (χ_k) est sommable et donc grâce à (7.3), l'hypothèse (4.9) du lemme 4.2 est satisfaite. Comme $H_{k+1} = H_k$ lorsque $k \notin \mathbb{K}$ et comme $(\gamma_k, \sigma_k) > 0$ lorsque $k \in \mathbb{K}$, on peut appliquer le lemme 4.2 qui conduit à (i). On peut également appliquer le lemme 4.3 puisque l'inégalité (7.4) a la forme (4.11) avec $(\tau_k \chi_k)$ sommable. Dès lors, (δ_k) converge, disons vers δ et soit $\delta = 0$, soit $(\tau_k \theta_k)$ converge vers 0. Dans le premier cas, (ii) est clairement vérifiée. Dans le second, on a $\tau_k(H_k - H_*) \gamma_k = o(\|\gamma_k\|)$ et grâce à (i), $\tau_k(G_k - G_*) H_* \gamma_k = o(\|\gamma_k\|)$. Cette estimation et $H_* \gamma_k = \sigma_k + o(\|\gamma_k\|)$ (*cf.* (7.3)) donnent (ii).

2. *On prouve (iii).* D'après (7.3), $\sigma_k = H_* \gamma_k + o(\|\gamma_k\|)$ et donc $\gamma_k = G_* \sigma_k + o(\|\sigma_k\|)$. G_* étant définie positive, il existe un indice $K_2 \geqq K_1$ et une constante positive C_1 tels que

$$(\gamma_k, \sigma_k) \geqq C_1 \|\sigma_k\|^2 \quad \text{pour} \quad k \geqq K_2 .$$

Dès lors, pour $k \geqq K_2$, soit $\sigma_k = 0$, soit $k \in \mathbb{K}$. Si $k \in \mathbb{K}$, la relation $\sigma_k = Z_* t_k + o(\|t_k\|)$ et (ii) montre que

$$(G_k - G_*) Z_* t_k = o(\|t_k\|) , \tag{7.5}$$

pour tout k dans $\mathbb{K}$; tandis que si $\sigma_k = 0$, on a $t_k = 0$. Dès lors, (7.5) et donc (3.20) est satisfaite pour tout indice k. D'après le théorème 3.2, ρ_k vaudra 1 pour k assez grand, disons, $k \geqq K_3$.

3. *On prouve (iv).* On suppose $k \geqq K_3$ et donc on a $\rho_k = 1$ et $y_k - x_* = O(\|x_k - x_*\|)$. On en déduit, avec (i), que $t_k = O(\|x_k - x_*\|)$. Cette dernière estimation et (7.5) montrent que (3.18) est vérifiée et par conséquent la suite (x_k) converge superlinéairement.

4. *On prouve (v).* Le pas ρ_k valant 1 pour $k \geqq K_3$ et (G_k^{-1}) étant bornée, on a $y_{k+2} - x_* = O(\|x_{k+2} - x_*\|)$ et $x_{k+1} - x_* = O(\|y_k - x_*\|)$. Comme $x_{k+2} - x_* = o(\|x_{k+1} - x_*\|)$, on voit que $y_{k+2} - x_* = o(\|y_k - x_*\|)$, c'est-à-dire que (y_k) converge superlinéairement en deux pas. □

THÉORÈME 7.2 : *Soient* (x_k), (y_k) *et* (G_k) *les suites générées par l'algorithme* QNR2. *On suppose que l'on a* (7.1), (7.2) *et*

$$C_1 := \sup_{k \geqq 0} \sup_{j \geqq 0} \frac{\|y_{k+j} - x_*\|}{\|y_k - x_*\|} < +\infty . \tag{7.6}$$

Alors :
(i) (G_k) *et* (G_k^{-1}) *sont bornées,*
(ii) $(G_k - G_*)\,\sigma_k = o(\|\sigma_k\|)$ *pour* k *dans* $\mathbb{K}$,
(iii) $y_{k+1} - x_* = o(\|y_k - x_*\|)$ *pour* k *dans* $\mathbb{K}$,
(iv) $\mathbb{K}_2 \backslash \mathbb{K}_1$ *est borné,*
(v) $\rho_k = 1$ *pour* k *assez grand,*
(vi) (x_k) *converge superlinéairement,*
(vii) (y_k) *converge superlinéairement en deux pas.*

Remarques : Par rapport aux hypothèses du théorème 7.1, on a rajouté (7.6). Cette hypothèse vient de ce qu'on a besoin de comparer $\|y_k - x_*\|$ et $\|y_{k^{--}} - x_*\|$ et que $((k) - (k^{--}))$ n'est pas nécessairement borné. Cette condition (7.6) est satisfaite sous les hypothèses du théorème 6.2. En effet, dans ce cas (H_k) est bornée et il existe donc une constante positive C telle que $\|y_{k+1} - x_*\| \leqq C\,\|y_k - x_*\|$. Alors, grâce à la convergence linéaire en deux pas de (y_k), on a certainement $C_1 \leqq C$. L'utilisation de la formule de BFGS sous la condition « (γ_k, σ_k) positif » assure la définie positivité des matrices G_k. Dans l'estimation (ii), σ_k est donné par (5.8) et $\mathbb{K}$ est défini en (5.18)-(5.20). La propriété (iv) a pour conséquence que la suite (μ_k) converge vers un nombre μ_∞ positif. Le « pour k assez grand » dans la propriété (v) et ci-après signifie « après un nombre fini d'itérations ».

Preuve du théorème 7.2 : Le schéma de la démonstration est semblable à celui de la preuve du théorème 7.1.

1. *On prouve (i) et (ii).* Soit $\zeta > 0$ tel que la boule $B(x_*, \zeta)$ contiennent les suites (x_k) et (y_k). Soient $\omega_1(\zeta)$, $\omega_2(\zeta)$, $\omega_3(\zeta)$, $\omega_4(\zeta)$ et $\varepsilon(\zeta)$ les constantes positives données par le lemme 6.1. On peut supposer $\varepsilon \leqq \zeta$. Comme (x_k) et (y_k) convergent vers x_*, il existe un indice K_1 tel que x_k et y_k soient dans $B(x_*, \varepsilon)$ pour $k \geqq K_1$. Par application du lemme 6.1, on obtient :

$$\|\gamma_k - G_*\,\sigma_k\| \leqq \omega_1\,\chi_k \|\sigma_k\| \quad \text{pour } k \text{ dans } \mathbb{K}_2 \text{ et } k \geqq K_1\,, \tag{7.7}$$

où, $\chi_k := \mu_k \|e_{k^{--}}\| + \|y_k - x_*\| + \|x_{k+1} - x_*\|$. Si $\mathbb{K}$ est borné, (i) est clairement vérifiée et pour (ii), il n'y a rien à démontrer. On suppose donc que $\mathbb{K}$ est non borné. Alors, il existe un indice $K_2 > K_1$ tel que $(k^{--}) \geqq K_1$ pour $k \geqq K_2$. En appliquant le lemme 6.1, on obtient :

$$\|\sigma_k - H_*\,\gamma_k\| \leqq \omega_2\,\chi_k \|\gamma_k\| \quad \text{pour } k \text{ dans } \mathbb{K}_2 \text{ et } k \geqq K_2 \tag{7.8}$$
($\mathbb{K}$ non borné),

$$\delta_{k+1} \leqq (1 - \tau_k\,\theta_k^2 + \omega_3\,\tau_k\,\chi_k)\,\delta_k + \omega_4\,\tau_k\,\chi_k \quad \text{pour } k \geqq K_2 \tag{7.9}$$
($\mathbb{K}$ non borné),

où δ_k, τ_k et θ_k sont définis en (6.1)-(6.3). Clairement, $r_{k+1} = O(\|x_{k+1} - x_*\|)$ et donc $e_k = O(\|y_k - x_*\| + \|x_{k+1} - x_*\|)$. Comme (μ_k) est bornée par μ_0, on voit qu'il existe une constante positive C_2 telle que

$$\chi_k \leqq C_2(\|y_{k^{--}} - x_*\| + \|x_{(k^{--})+1} - x_*\|) + \|y_k - x_*\| + \|x_{k+1} - x_*\|.$$

Dès lors, grâce à cette inégalité, (7.1) et (7.2), on voit que $(\tau_k \chi_k)$ est sommable. Alors les hypothèses du lemme 4.2 sont satisfaites pour la sous-suite $\mathbb{K}$, c'est-à-dire chaque fois que la matrice G_k est mise à jour par la formule de BFGS : (γ_k, σ_k) est positif pour $k \in \mathbb{K}$ et l'on a (7.8) avec $(\chi_k : k \in \mathbb{K})$ sommable. Ce lemme prouve (i). Grâce à (7.9), on peut également appliquer le lemme 4.3 (avec χ_k et θ_k dans (4.11) remplacé par $\tau_k \chi_k$ et $\tau_k \theta_k$). Comme dans la preuve du théorème 7.1, on en déduit (ii).

2. *On montre que* ρ_k *vaut 1 pour* $k \geqq K_3$ *dans* $\mathbb{K}$. Si $k \in \mathbb{K}$, on peut utiliser (ii). Comme $\sigma_k = \rho_k Z(y_k) t_k$, on obtient l'estimation

$$(G_k - G_*) Z(y_k) t_k = o(\|t_k\|) \quad \text{pour} \quad k \in \mathbb{K}. \tag{7.10}$$

C'est la condition (3.20) du théorème 3.2. Elle donne le résultat.

3. *On prouve (iii).* En développant $g(y_k)$ autour de x_*, en utilisant (5.4) et (i), on obtient :

$$t_k = - Z_*^- H_k Z_*^{-T} L_* (y_k - x_*) + o(\|y_k - x_*\|). \tag{7.11}$$

En développant $c(x_{k+1})$ autour de y_k et en utilisant $c'(y_k) \cdot t_k = 0$, puis en développant $c(y_k)$ autour de x_* et en utilisant (7.11), on obtient :

$$r_{k+1} = - A(y_k)^- c(x_{k+1}) = - A(y_k)^- c(y_k) + O(\|t_k\|^2)$$

$$r_{k+1} = - A_*^- A_* (y_k - x_*) + O(\|y_k - x_*\|^2). \tag{7.12}$$

On peut supposer $\mathbb{K}$ non borné. Si $k \in \mathbb{K}$, le critère est vérifié et comme $\rho_k = 1$, on a $r_{k+1} = o(\|t_k\|)$. Avec (7.11), cela donne $r_{k+1} = o(\|y_k - x_*\|)$. Dès lors (7.12) et l'injectivité de A_*^- montrent que

$$A_* (y_k - x_*) = o(\|y_k - x_*\|) \quad \text{pour} \quad k \in \mathbb{K}. \tag{7.13}$$

Dès lors, en utilisant (2.3), on peut récrire (7.11) comme suit

$$t_k = - Z_*^- H_k G_* Z_* (y_k - x_*) + o(\|y_k - x_*\|) \quad \text{pour} \quad k \in \mathbb{K}. \tag{7.14}$$

Comme $\rho_k = 1$, $y_{k+1} = y_k + t_k + r_{k+1}$ et avec (7.12), (7.13) et (7.14), on obtient pour k dans $\mathbb{K}$:

$$y_{k+1} - x_* = Z_*^- H_k (G_k - G_*) Z_* (y_k - x_*) + o(\|y_k - x_*\|). \tag{7.15}$$

En utilisant (7.10) et (7.14), on voit que le premier terme du membre de droite de (7.15) est un $o(\|y_k - x_*\|)$, ce qui prouve (iii).

4. *On prouve (iv).* Pour $k \in \mathbb{K}_2 \backslash \mathbb{K}_1$, $k \geqq K_1$ et $\sigma_k \neq 0$, on a grâce à (7.7) :

$$\begin{aligned} 0 \geqq (\gamma_k, \sigma_k) &= (G_* \, \sigma_k, \sigma_k) + (\gamma_k - G_* \, \sigma_k, \sigma_k) \\ &\geqq (\|H_*\|^{-1} - \omega_1 \chi_k)\|\sigma_k\|^2 . \end{aligned}$$

Comme il existe une constante positive C_3 et une suite de nombres positifs (η_k) convergeant vers 0 telles que $\chi_k \leqq C_3 \, \mu_k + \eta_k$, on obtient :

$$\omega_1 \, C_3 \, \mu_k \geqq \|H_*\|^{-1} - \omega_1 \, \eta_k \quad \text{pour } k \text{ dans } \mathbb{K}_2 \backslash \mathbb{K}_1 , \quad k \geqq K_1 , \quad \sigma_k \neq 0 .$$

Ce qui montre que $\mu_\infty > 0$ et que l'ensemble d'indices $\{k \in \mathbb{K}_2 \backslash \mathbb{K}_1 : \sigma_k \neq 0\}$ est borné. On en déduit (iv). En effet, dans le cas contraire, $\sigma_k = 0$ donc $t_k = 0$ pour k grand dans $\mathbb{K}_2 \backslash \mathbb{K}_1$. Alors le critère (5.17) impliquerait que $r_{k+1} = 0$ et donc que y_k est solution.

5. *On montre deux inégalités lorsque* ρ_{k-1} *égal* 1. Si $\rho_{k-1} = 1$, $y_k = x_k - A(y_{k-1})^- \, c(x_k)$. En développant $c(x_k)$ autour de x_* et en utilisant (2.3), on obtient

$$y_k - x_* = Z_*^- \, Z_* (x_k - x_*) - (A(y_{k-1})^- - A_*^-) \, A_* (x_k - x_*) + \\ + O(\|x_k - x_*\|^2) . \tag{7.16}$$

Alors, (7.12) donne

$$t_k = - Z_*^- \, H_k \, G_* \, Z_* (x_k - x_*) + o(\|x_k - x_*\|) . \tag{7.17}$$

(G_k^{-1}) étant bornée, on en déduit :

$$Z_* (x_k - x_*) = O(\|t_k\|) + o(\|x_k - x_*\|) . \tag{7.18}$$

Comme $x_{k+1} = y_k + t_k$, on obtient en utilisant (7.16) et (7.18) :

$$\|x_{k+1} - x_*\| \leqq C_4 \|t_k\| + o(\|x_k - x_*\|) . \tag{7.19}$$

C'est la première inégalité. De (7.12) et (7.16), on obtient :

$$r_{k+1} = A_*^- \, A_* \, (A(y_{k-1})^- - A_*^-) \, A_* (x_k - x_*) + O(\|x_k - x_*\|^2) .$$

Or $x_k - x_* = O(\|y_{k-1} - x_*\|)$, dès lors cette estimation s'écrit

$$\|r_{k+1}\| \leqq C_5 \|y_{k-1} - x_*\| \, \|x_k - x_*\| . \tag{7.20}$$

C'est la seconde inégalité.

6. *On prouve (v).* On a montré à l'étape 2 que si $k \in \mathbb{K} := \mathbb{K}_1 \cap \mathbb{K}_2$ et $k \geqq K_3$ alors $\rho_k = 1$. D'après l'étape 4, $\mathbb{K}_2 \backslash \mathbb{K}_1$ est borné. Dès lors, il reste à examiner la sous-suite d'indices $\mathbb{K}_2^c$ pour lesquels le critère (5.17) n'est pas satisfait. Par négation du critère et en notant que $\mu_k \geqq \mu_\infty > 0$ et que $\rho_k \leqq 1$, on obtient :

$$\|e_{k^{--}}\| \, \|t_k\| \leqq \frac{1}{\mu_\infty} \|r_{k+1}\|. \tag{7.21}$$

Si $\mathbb{K}$ est borné, k^{--} est un indice constant dès que k est assez grand et donc, d'après (7.21), $t_k = O(\|r_{k+1}\|)$. Alors d'après le théorème 3.2 (condition (3.21)), $\rho_k = 1$ pour k assez grand dans $\mathbb{K}_2^c$ (si $\mathbb{K}$ borné). Supposons à présent que $\mathbb{K}$ ne soit pas borné et que l'on n'ait pas $\rho_k = 1$ après un nombre fini d'itérations. Comme $\mathbb{K}$ est non borné et que $\rho_k = 1$ pour k grand dans $\mathbb{K}$, on peut trouver une sous-suite $\mathbb{I}$ d'indices i telle que :

$$\rho_{i-1} = 1 \,, \quad \rho_i \neq 1 \quad \text{pour} \quad i \in \mathbb{I} \subset \mathbb{K}_2^c \,. \tag{7.22}$$

On montre que cela conduit à une contradiction. Pour les indices i dans $\mathbb{I}$, on a en utilisant (7.21), puis (7.20) (ce qui est licite car $\rho_{i-1} = 1$), puis l'équivalence $\|e_i\| \sim \|y_i - x_*\|$ (d'où provient la constante C_6) et (7.6) (on a $(i^{--}) + 1 \leqq i - 1$) :

$$\begin{aligned} \|t_i\| &\leqq \frac{1}{\mu_\infty} \frac{\|r_{i+1}\|}{\|e_{i^{--}}\|} \\ &\leqq \frac{C_5}{\mu_\infty} \frac{\|y_{i-1} - x_*\|}{\|e_{i^{--}}\|} \|x_i - x_*\| \\ &\leqq \frac{C_1 C_5 C_6}{\mu_\infty} \frac{\|y_{(i^{--})+1} - x_*\|}{\|y_{i^{--}} - x_*\|} \|x_i - x_*\|. \end{aligned} \tag{7.23}$$

Comme $(i^{--}) \in \mathbb{K}$ qui est non borné, on voit, grâce à cette inégalité, à (iii) et à l'équivalence $\|d_k\| \sim \|x_k - x_*\|$ que

$$t_i = o(\|d_i\|) \,. \tag{7.24}$$

Comme $d_i = r_i + t_i$, on en déduit que $t_i = o(\|r_i\|)$ et d'après le théorème 3.2 (condition (3.22)) :

$$r_{i+1} = o(\|r_i\| \, \|t_i\|) \quad \text{pour} \quad i \in \mathbb{I} \,. \tag{7.25}$$

Or, grâce à (7.6), on a

$$\|r_i\| \leqq C_7 \|y_{i-1} - x_*\| \leqq C_1 C_7 \|y_{i^{--}} - x_*\|.$$

L'utilisation de cette inégalité et de (7.25) à la place de (7.20) en (7.23) conduit à $t_i = o(\|t_i\|)$ et donc $t_i = 0$ pour i assez grand dans $\mathbb{I}$. Or dans ce cas $\rho_i = 1$, ce qui contredit (7.22).

En conclusion, on a montré l'existence d'un indice $K_5 \geqq K_4$ tel que

$$\rho_k = 1 \quad \text{pour} \quad k \geqq K_5 .$$

7. *On prouve (vi).* On suppose $k > K_5$, de sorte que $\rho_k = \rho_{k-1} = 1$. Considérons en premier lieu les indices $k \in \mathbb{K} := \mathbb{K}_1 \cap \mathbb{K}_2$ (supposé non borné). En utilisant (7.10) et $t_k = O(\|x_k - x_*\|)$, on obtient (3.18) et donc

$$x_{k+1} - x_* = o(\|x_k - x_*\|) , \tag{7.26}$$

pour k dans $\mathbb{K}$. Comme $\mathbb{K}_2 \backslash \mathbb{K}_1$ est borné (étape 4), il reste à considérer les indices $k \in \mathbb{K}_2^c$. Lorsque $\mathbb{K}$ est borné, la négation du critère (5.17) donne $t_k = O(\|r_{k+1}\|)$ et les inégalités (7.20) et (7.19) fournissent l'estimation (7.26) pour $k \in \mathbb{K}_2^c$. Lorsque $\mathbb{K}$ est non borné, on peut utiliser (7.24) qui injecté dans (7.19) donne encore (7.26) pour $k \in \mathbb{K}_2^c$.

8. *On prouve (vii).* Voir la preuve du théorème 7.1. □

8. CONCLUSION

Nous avons développé dans ce travail deux techniques de mise à jour de la métrique dans les méthodes de quasi-Newton réduites. Toutes deux utilisent la formule de mise à jour de BFGS dont la propriété de maintien de la définie positivité des métriques a joué de façon essentielle dans l'obtention des résultats. Cette propriété apparaît d'ailleurs comme une condition naturelle puisqu'il s'agit d'approximer le hessien réduit du lagrangien que nous avons supposé défini positif.

Les deux méthodes se distinguent par le nombre de linéarisations des contraintes nécessaire à chaque itération. Dans l'algorithme QNR1, les contraintes sont linéarisées deux fois par itération. Dans l'algorithme QNR2, une seule linéarisation des contraintes suffit, mais la métrique ne peut plus être systématiquement mise à jour à chaque itération : un critère de mise à jour doit être utilisé. Les deux méthodes génèrent des suites superlinéairement convergentes (en un pas). Dans l'état actuel de la théorie, peu de choses peuvent être dites en plus. Il revient donc aux essais numériques la tâche de départager les deux méthodes. A ce sujet, il ne fait pas beaucoup de doute, me semble-t-il, que la palme revient à l'algorithme QNR1. En effet, dans les deux méthodes, il y a découplage des équations d'optimalité : on se rapproche de la variété par des itérations de Newton et on minimise la fonctionnelle par des itérations de quasi-Newton réduites.

Dans l'algorithme QNR1, une seule perturbation intervient dans l'approximation du hessien réduit : elle provient du mouvement du plan tangent au cours des itérations. Dans l'algorithme QNR2, une autre perturbation provient de l'inadéquation du mariage de γ_k et σ_k au sein de l'équation de quasi-Newton, le critère de mise à jour apparaissant alors comme l'outil permettant à l'algorithme QNR2 de contrôler cette perturbation.

Pour l'algorithme QNR2, un critère de mise à jour basé sur la comparaison de la grandeur des pas de minimisation et de restauration des contraintes s'est avéré nécessaire, sa forme pouvant varier suivant le choix de la suite $(\bar{\mu}_k)$ en (5.14). L'avantage du choix (5.16) réside essentiellement dans les résultats de convergence qu'il permet d'obtenir. On pourrait toutefois lui objecter la présence de l'indice (k^{--}) qui ne croît pas de façon strictement monotone et pourrait dès lors être à la base de phénomènes oscillatoires (comme par exemple une mise à jour qui ne se ferait qu'une fois sur deux). On pourrait également se demander si lorsqu'il n'y a pas de mise à jour de la métrique, il ne vaudrait pas mieux supprimer le pas de minimisation. Une réponse à ces questions est apportée dans Gilbert (–) où il est montré que l'on peut utiliser le critère de mise à jour suivant :

$$\|c(y_k)\| \leqq \mu \|e_{k-1}\| \, \|g(y_k)\|,$$

si l'on fait deux pas de restauration des contraintes par itération. Lorsque ce critère n'est pas satisfait, l'on peut, au choix, faire ou ne pas faire de pas de minimisation tout en conservant la convergence superlinéaire. Dans cette inégalité, μ est une constante quelconque : on évite donc également un problème de choix de constante. Cet algorithme peut alors s'interpréter comme une méthode GRG (Gradient Réduit Généralisé) dans laquelle le test d'arrêt de la phase de restauration des contraintes porte non pas sur $\|c(y_k)\|$ (*cf.* par exemple Mukai et Polak (1978)) mais sur le rapport $\|c(y_k)\| / \|g(y_k)\|$ entre le pas de restauration et le pas tangent, avec un seuil de tolérance $(\mu \|e_{k-1}\|)$ qui décroît au cours des itérations.

REMERCIEMENTS

Ce travail a été mené en partie durant l'année 1984-1985 au Centre d'Études Nucléaires de Fontenay-aux-Roses (France) grâce à une bourse de la Commission des Communautés Européennes et en partie à l'Institut National de Recherche en Informatique et en Automatique (INRIA), à Rocquencourt (France). Je remercie vivement ces organismes ainsi que J. F. Bonnans de l'INRIA pour les discussions fructueuses que nous avons eues et pour ses amicaux conseils.

RÉFÉRENCES

L. ARMIJO (1966). *Minimization of functions having Lipschitz continuous first partial derivatives.* Pacific Journal of Mathematics 16/1, 1-3.

J BLUM, J. Ch. GILBERT, B. THOORIS (1985). *Parametric identification of the plasma current density from the magnetic measurements and the pressure profile, code IDENTC* Report of JET contract number JT3/9008.

J. F. BONNANS, D. GABAY (1984). *Une extension de la programmation quadratique successive.* Lecture Notes in Control and Information Sciences 63, 16-31. A. Bensoussan, J. L Lions (eds). Springer-Verlag.

C. G. BROYDEN (1969). *A new double-rank minimization algorithm.* Notices of the American Mathematical Society 16, 670.

C G. BROYDEN, J. E. DENNIS, J. J MORE (1973). *On the local and superlinear convergence of quasi-Newton methods.* Journal of the Institute of Mathematics and its Applications 12, 223-245

R. H BYRD (1985). *An example of irregular convergence in some constrained optimization methods that use the projected hessian.* Mathematical Programming 32, 232-237.

R. H. BYRD, R. B. SCHNABEL (1986). *Continuity of the null space basis and constrained optimization.* Mathematical Programming 35, 32-41.

T F COLEMAN, A. R. CONN (1982a). *Nonlinear programming via an exact penalty function . asymptotic analysis* Mathematical Programming 24, 123-136.

T. F. COLEMAN, A. R. CONN (1982b). *Nonlinear programming via an exact penalty function : global analysis.* Mathematical Programming 24, 137-161.

T F. COLEMAN, A. R. CONN (1984). *On the local convergence of a quasi-Newton method for the nonlinear programming problem.* SIAM Journal on Numerical Analysis 21/4, 755-769.

J. E. DENNIS, J. J. MORE (1974) *A characterization of superlinear convergence and its application to quasi-Newton methods* Mathematics of Computation 28/126, 549-560.

J. E. DENNIS, J J MORE (1977). *Quasi-Newton methods, motivation and theory.* SIAM Review 19, 46-89.

R. FLETCHER (1970). *A new approach to variable metric algorithms.* The Computer Journal 13/3, 317-322.

R. FLETCHER (1981). *Practical Methods of Optimization.* Vol. 2 : Constrained Optimization. John Wiley & Sons.

D. GABAY (1982a). *Minimizing a differentiable function over a differential manifold.* Journal of Optimization Theory and Applications 37/2, 177-219.

D. GABAY (1982b). *Reduced quasi-Newton methods with feasibility improvement for nonlinearly constrained optimization.* Mathematical Programming Study 16, 18-44.

R. P. GE, M. J. D. POWELL (1983). *The convergence of variable metric matrices in unconstrained optimization* Mathematical Programming 27, 123-143.

J. Ch. GILBERT (1986a). *Une méthode à métrique variable réduite en optimisation avec contraintes d'égalité non linéaires* Rapport de recherche de l'INRIA RR-482, 78153 Le Chesnay Cedex, France.

J. Ch. GILBERT (1986b). *On the local and global convergence of a reduced quasi-Newton method* Rapport de recherche de l'INRIA RR-565, 78153 Le Chesnay Cedex, France (version révisée dans IIASA Working Paper WP-87-113).

J. Ch. GILBERT (1986c). *Une méthode de quasi-Newton réduite en optimisation sous contraintes avec priorité à la restauration.* Lecture Notes in Control and Information Sciences 83, 40-53. A. Bensoussan, J. L. Lions (eds), Springer-Verlag.

J. Ch. GILBERT (—) (en préparation).

D. GOLDFARB (1970). *A family of variable metric methods derived by variational means.* Mathematics of Computation 24, 23-26.

S. P. HAN (1976). *Superlinearly convergent variable metric algorithms for general nonlinear programming problems.* Mathematical Programming 11, 263-282.

S. P. HAN (1977). *A globally convergent method for nonlinear programming.* Journal of Optimization Theory and Applications 22/3, 297-309.

D. Q. MAYNE, E. POLAK (1982). *A superlinearly convergent algorithm for constrained optimization problems.* Mathematical Programming Study 16, 45-61.

H. MUKAI, E. POLAK (1978). *On the use of approximations in algorithms for optimization problems with equality and inequality constraints.* SIAM Journal on Numerical Analysis 15/4, 674-693.

J. NOCEDAL, M. L. OVERTON (1985). *Projected Hessian updating algorithms for nonlinearly constrained optimization.* SIAM Journal on Numerical Analysis 22/5, 821-850.

M. J. D. POWELL (1971). *On the convergence of the variable metric algorithm.* Journal of the Institute of Mathematics and its Applications 7, 21-36.

M. J. D. POWELL (1976). *Some global convergence properties of a variable metric algorithm for minimization without exact line searches.* Nonlinear Programming, SIAM-AMS Proceedings, Vol. 9, American Mathematical Society, Providence, R.I.

M. J. D. POWELL (1978). *The convergence of variable metric methods for nonlinearly constrained optimization calculations.* Nonlinear Programming 3, 27-63. O. L. Mangasarian, R. R. Meyer, S. M. Robinson (eds), Academic Press, New York.

D. F. SHANNO (1970). *Conditioning of quasi-Newton methods for function minimization.* Mathematics of Computation 24, 647-656.

R. B. WILSON (1963). *A simplicial algorithm for concave programming.* Ph. D. Thesis. Graduate School of Business Administration, Havard Univ., Cambridge, MA.

Y. YUAN (1985). *An only 2-step Q-superlinear convergence example for some algorithms that use reduced Hessian approximations* Mathematical Programming 32, 224-231.