

Chapitre 3

Les méthodes numériques pour l'optimisation sans contraintes

3.1 Introduction

Comme la stationnarité de f est une condition nécessaire d'optimalité, pratiquement toutes les méthodes de minimisation sans contraintes dans \mathbb{R}^n consistent à rechercher un point x^* stationnaire ($\nabla f(x^*) = 0$). Ce problème est équivalent à la résolution d'un système d'équations non linéaires :

$$g_i(x) = \frac{\partial f}{\partial x_i}(x) = 0, \quad i = 1, \dots, n.$$

On peut chercher à résoudre directement ce système, ce qui conduit à la méthode de Newton. Cependant, il faut se rappeler que cette méthode présente quelques inconvénients :

- i) la divergence de la méthode, si le point de départ est trop éloigné de x^* ,
- ii) elle exige que la fonction f soit deux fois continûment différentiable et exige le calcul des dérivées secondes.

C'est pourquoi, les méthodes les plus couramment utilisées sont basées sur un principe général assez simple, mais dont l'analyse des conditions de convergence sont complexes : on les appelle méthodes de descente. Il s'agit de procédures itératives de construction d'une suite de points $\{x_k\}$ telle que la fonction f décroît d'une itération à une autre, i.e.,

$$f(x^{k+1}) < f(x^k), \quad k = 0, 1, \dots$$

Le processus itératif général de construction de la suite $\{x_k\}$ a la forme suivante :

$$x^{k+1} = x^k + \theta_k d_k, \quad k = 0, 1, \dots \quad (3.1)$$

où d_k est une direction d'amélioration ou de descente et θ_k le pas le long de cette direction.

Le problème restant posé et qui différencie les différentes méthodes numériques est le choix du pas θ_k et de la direction d_k à chaque itération.

3.2 La méthode du gradient

3.2.1 Interprétation physique du gradient

Pour une fonction différentiable $f : \mathbb{R}^n \rightarrow \mathbb{R}$, le gradient $\nabla f(x^0)$ de f au point x^0 se définit comme suit :

$$f(x^0 + \Delta x) - f(x^0) = (\nabla f(x^0))^T \Delta x + o(\|\Delta x\|), \quad (3.2)$$

où

$$\lim_{\|\Delta x\| \rightarrow 0} \frac{o(\|\Delta x\|)}{\|\Delta x\|} = 0.$$

Une autre interprétation du gradient peut se faire de la manière suivante : soit $x^0 \in \mathbb{R}^n$. De x^0 , on choisit une direction de déplacement $d \in \mathbb{R}^n$, $\|d\| = 1$, et on pose :

$$x(\theta) = x^0 + \theta d, \quad \theta \geq 0.$$

Le point $x(\theta)$ est le point atteint après un déplacement le long de la direction d avec un pas θ . Estimons la variation de la valeur de la fonction f en passant du point x^0 au point $x^0 + \theta d$. La relation (3.2) prend la forme suivante :

$$f(x^0 + \theta d) - f(x^0) = \theta \left[(\nabla f(x^0))^T d + \frac{o(\theta)}{\theta} \right].$$

Pour des valeurs de $\theta \geq 0$ voisines de zéro, le signe de l'expression $f(x^0 + \theta d) - f(x^0)$ dépendra du signe de $(\nabla f(x^0))^T d$.

Pour avoir la relation $f(x^0 + \theta d) < f(x^0)$, ie., une diminution de la valeur de $f(x)$ lors du passage du point x^0 au point $x^0 + \theta d$, on devrait choisir $\theta \geq 0$ assez petit et la direction $d \in \mathbb{R}^n$ de façon à avoir $(\nabla f(x^0))^T d < 0$.

Définition 3.2.1. La direction $d \in \mathbb{R}^n$ est appelée direction de descente, si

$$(\nabla f(x^0))^T d < 0 \quad (3.3)$$

Le processus itératif (3.1) de construction de la suite $\{x_k\}$ prend la forme

$$x^{k+1} = x^k + \theta_k d_k, \quad k = 0, 1, \dots \quad (3.4)$$

avec, si $\nabla f(x^k) \neq 0$, la direction d_k est choisie de façon que

$$[\nabla f(x^k)]^T d_k < 0, \quad (3.5)$$

et le pas θ_k est positif. Si $\nabla f(x^k) = 0$, alors le processus itératif est arrêté et $x^{k+1} = x^k$ (ce qui est équivalent à choisir $d_k = 0$).

En vertu de la relation (3.5) de d_k et du gradient $\nabla f(x^k)$, on appelle les algorithmes de ce type "méthodes du gradient". Certains auteurs réservent le terme "méthode du gradient" pour le cas particulier où $d_k = -\nabla f(x^k)$.

Il existe une large variété de possibilités pour choisir la direction d_k et le pas θ_k dans les méthodes du gradient.

3.2.2 Sélection des directions de descente

Plusieurs méthodes du gradient s'écrivent sous la forme :

$$x^{k+1} = x^k - \theta_k D_k \nabla f(x^k), \quad (3.6)$$

où D_k est une matrice symétrique définie positive. Dans ce cas, la direction d_k s'écrit

$$d_k = -D_k \nabla f(x^k)$$

et la condition de descente

$$[\nabla f(x^k)]^T d_k < 0$$

peut s'écrire

$$- [\nabla f(x^k)]^T D_k \nabla f(x^k) < 0$$

et qui est vérifiée puisque $D_k > 0$.

On donne quelques exemples de choix de la matrice D_k .

Méthode de la descente la plus rapide (méthode de la plus forte pente, steepest descent)

On pose $D_k = I$, $k = 0, 1, 2, \dots$ où I est la matrice identité $n \times n$. Dans ce cas, le processus itératif (3.6) prend la forme

$$x^{k+1} = x^k - \theta_k \nabla f(x^k), \quad k = 0, 1, 2, \dots \quad (3.7)$$

Nous allons justifier ci-après l'appellation de méthode de la plus forte pente. En partant du point x^k et en se mouvant le long de la direction d , la fonction $f(x)$ va varier avec une vitesse donnée par la dérivée directionnelle de f au point x^k dans la direction d , soit :

$$f'(x^k; d) = \lim_{\theta \searrow 0} \frac{f(x^k + \theta d) - f(x^k)}{\theta} = [\nabla f(x^k)]^T d.$$

Pour des directions d telles que $\|d\| = 1$, la vitesse de variation de la fonction f sera maximale pour

$$d_k = \frac{\nabla f(x^k)}{\|\nabla f(x^k)\|} \quad (3.8)$$

En effet, on a d'après l'inégalité de Cauchy-Schwartz

$$[\nabla f(x^k)]^T d \leq \|\nabla f(x^k)\| \|d\| = \|\nabla f(x^k)\|$$

et pour d_k vérifiant (3.8), on a

$$[\nabla f(x^k)]^T d_k = \frac{[\nabla f(x^k)]^T [\nabla f(x^k)]}{\|\nabla f(x^k)\|}.$$

Ainsi, le gradient $\nabla f(x^k)$ constitue la meilleure direction de croissance à partir du point x^k . La direction opposée à celle du gradient ($-\nabla f(x^k)$) est la meilleure direction de descente à partir de x^k .

L'inconvénient de la méthode de descente rapide est sa lente convergence.

3.2.3 Sélection du pas

Il existe un certain nombre de règles de choix du pas θ_k dans la méthode du gradient. Nous allons présenter les plus couramment utilisées.

a) Règle du minimum

Elle consiste à choisir, à l'itération k , le pas θ_k minimisant la fonction f dans la direction d_k , c-à-d

$$f(x^k + \theta_k d_k) = \min_{\theta \geq 0} f(x^k + \theta d_k). \quad (3.9)$$

b) Règle du minimum réduit

C'est une version de la règle du minimum qui est la plus utilisée. On fixe un scalaire $s > 0$ et θ_k est défini par

$$f(x^k + \theta_k d_k) = \min_{\theta \in [0, s]} f(x^k + \theta d_k). \quad (3.10)$$

Remarque 3.2.1. La résolution des problèmes d'optimisation (3.9) et (3.10) se fait par des méthodes d'optimisation unidimensionnelle.

c) Règle du pas constant

Ici, on fixe pour toutes les itérations

$$\theta_k = s, \quad k = 0, 1, \dots \quad (3.11)$$

La règle du pas constant est la plus simple. Toutefois, si le pas est grand, il y-a risque de divergence et quand le pas est trop petit, la vitesse de convergence serait très lente. Ainsi la règle du pas constant n'est utile que pour les problèmes où la valeur appropriée du pas est connue ou peut être facilement calculée.

Exemple 3.2.1. On considère la fonctionnelle quadratique suivante :

$$f(x_1, x_2) = 8x_1^2 + 4x_1x_2 + 5x_2^2.$$

On applique la méthode du gradient à pas optimal, en partant du point initial

$$x^{(0)} = \begin{pmatrix} x_1^{(0)} \\ x_2^{(0)} \end{pmatrix} = \begin{pmatrix} 10 \\ 10 \end{pmatrix},$$

pour trouver le minimum de f .

- On calcule le gradient de f :

$$\nabla f(x) = \begin{pmatrix} 16x_1 + 4x_2 \\ 4x_1 + 10x_2 \end{pmatrix}.$$

- On calcule la valeur du gradient au point $x^{(0)}$:

$$\nabla f(x^{(0)}) = \begin{pmatrix} 200 \\ 140 \end{pmatrix}.$$

- On calcule le point $x^{(1)}$:

$$x^{(1)} = x^{(0)} - \alpha^{(0)} \nabla f(x^{(0)}) = \begin{pmatrix} 10 \\ 10 \end{pmatrix} - \alpha^{(0)} \begin{pmatrix} 200 \\ 140 \end{pmatrix} = \begin{pmatrix} 10 - 200\alpha^{(0)} \\ 10 - 140\alpha^{(0)} \end{pmatrix}.$$

- On cherche la valeur optimale de $\alpha^{(0)}$. Pour ce, on calcule $f(x^{(1)})$ puis on détermine $\alpha^{(0)}$ permettant de minimiser $f(x^{(1)})$. On a

$$\begin{aligned} f(x^{(1)}) &= 8(10 - 200\alpha^{(0)})^2 + 4(10 - 200\alpha^{(0)})(10 - 140\alpha^{(0)}) + 5(10 - 140\alpha^{(0)})^2 \\ &= -59600 + 1060000\alpha^{(0)} \\ &= \varphi(\alpha^{(0)}). \end{aligned}$$

On calcule $\alpha^{(0)}$ tel que $\varphi'(\alpha^{(0)}) = 0$, d'où $\alpha^{(0)} = 0.056$.

- On trouve ainsi

$$x^{(1)} = \begin{pmatrix} -1.20 \\ 2.16 \end{pmatrix}.$$

- On cherche de la même façon tous les points $x^{(k)}$ jusqu'à ce que le critère d'arrêt soit satisfait.

Exemple 3.2.2. Utiliser la méthode de la plus forte pente puis de la plus forte pente accélérée (p=2) pour trouver le minimum de la fonction f définie de \mathbb{R}^2 dans \mathbb{R} comme suit :

$$f(x_1, x_2) = x_1^4 - 2x_1^3 + \frac{1}{6}x_2^2 - 3x_1 - x_1x_2.$$

On a

$$\nabla f(x_1, x_2) = \begin{pmatrix} 4x_1^3 - 6x_1^2 - 3 - x_2 \\ \frac{1}{3}x_2 - x_1 \end{pmatrix}.$$

Soit $x^0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$. On trouve la direction de la plus forte pente

$$-\nabla f(x^0) = \begin{pmatrix} 3 \\ 0 \end{pmatrix}.$$

On minimise donc

$$g_0(\lambda) = f(x^0 - \lambda \nabla f(x^0)) = 81\lambda^4 - 54\lambda^3 - 6\lambda.$$

On trouve

$$\lambda_0 \approx 0.582.$$

Notre premier pas nous conduit donc au point

$$x^1 = x^0 - \lambda_0 \nabla f(x^0) = \begin{pmatrix} 1.746 \\ 0 \end{pmatrix}.$$

On réitère pour obtenir les valeurs suivantes :

i	x^i	$-\nabla f(x^i)$	$g_i(\lambda)$	λ_i	x^i
0	$\begin{pmatrix} 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 3 \\ 0 \end{pmatrix}$	$81\lambda^4 - 54\lambda^3 - 6\lambda$	0.582	$\begin{pmatrix} 1.746 \\ 0 \end{pmatrix}$
1	$\begin{pmatrix} 1.746 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 1.746 \end{pmatrix}$	$0.508\lambda^2 - 3.049\lambda - 6.590$	3	$\begin{pmatrix} 1.746 \\ 5.238 \end{pmatrix}$
2	$\begin{pmatrix} 1.746 \\ 5.238 \end{pmatrix}$	$\begin{pmatrix} 5.238 \\ 0 \end{pmatrix}$	$752\lambda^4 + 716\lambda^3 + 214\lambda^2 - 27\lambda - 11$	0.05	$\begin{pmatrix} 2.01 \\ 5.238 \end{pmatrix}$
3	$\begin{pmatrix} 2.01 \\ 5.238 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 0.264 \end{pmatrix}$	$0.012\lambda^2 - 0.07\lambda - 11.904$	3	$\begin{pmatrix} 2.01 \\ 6.03 \end{pmatrix}$

Pour la plus forte pente accélérée, après avoir fait 2 pas de la plus forte pente standard, nous devons faire un pas accéléré, en utilisant la direction $x^2 - x^0$. On minimise donc la fonction

$$g_2(\lambda) = f(x^0 + \lambda(x^2 - x^0)) = 9.29\lambda^4 - 10.65\lambda^3 - 4.57\lambda^2 - 5.238\lambda.$$

On trouve

$$\lambda_2 = 1.172$$

et on atteint le point

$$x^3 = x^0 + \lambda_2(x^2 - x^0) = \begin{pmatrix} 2.05 \\ 6.137 \end{pmatrix},$$

qui est près du minimum global.

3.3 La méthode de Newton

La méthode de Newton est obtenue en posant dans (3.6)

$$D_k = (\nabla^2 f(x^k))^{-1}, \quad k = 0, 1, \dots \quad (3.12)$$

ce qui donne

$$x^{k+1} = x^k - \theta_k (\nabla^2 f(x^k))^{-1} \nabla f(x^k), \quad k = 0, 1, \dots$$

en supposant que $\nabla^2 f(x^k)$ est une matrice définie positive. Dans le cas où $\nabla^2 f(x^k)$ n'est pas définie positive, une certaine modification est nécessaire.

L'idée de la méthode de Newton peut s'expliquer de deux manières :

- Lors de la minimisation d'une fonction sur \mathbb{R}^n , le problème consistait à chercher un point $x^* \in \mathbb{R}^n$ tel que

$$\nabla f(x^*) = 0.$$

La méthode numérique de Newton appliquée à l'équation

$$g(x^*) = \nabla f(x^*) = 0$$

construit les approximations successives de la manière suivante :

$$x^{k+1} = x^k - \left[\frac{\partial g(x^k)}{\partial x} \right]^{-1} g(x^k), \quad k = 0, 1, \dots \quad (3.13)$$

où

$$g = \begin{pmatrix} g_1 \\ \vdots \\ g_n \end{pmatrix}, \quad \frac{\partial g(x^k)}{\partial x} = \left(\frac{\partial g_i(x^k)}{\partial x_j} \right)_{1 \leq i, j \leq n}.$$

En tenant compte du fait que

$$\frac{\partial g(x^k)}{\partial x} = \nabla^2 f(x^k) = H(x^k),$$

la formule itérative (3.13) s'écrit alors

$$x^{k+1} = x^k - [\nabla^2 f(x^k)]^{-1} \nabla f(x^k), \quad k = 0, 1, \dots \quad (3.14)$$

Le point x^{k+1} peut être interprété comme la racine de l'approximation linéaire de la fonction

$$g(x) = \nabla f(x)$$

au voisinage de x^k . On a

$$g(x) \simeq g(x^k) + \frac{\partial g(x^k)}{\partial x}(x - x^k) = 0,$$

d'où

$$x - x^k = - \left[\frac{\partial g(x^k)}{\partial x} \right]^{-1} g(x^k)$$

et

$$x = x^{k+1} = x^k - \left[\frac{\partial g(x^k)}{\partial x} \right]^{-1} g(x^k)$$

ou

$$x^{k+1} = x^k - [\nabla^2 f(x^k)]^{-1} \nabla f(x^k).$$

- Une autre interprétation de la méthode de Newton est la suivante : elle consiste à minimiser l'approximation quadratique $q(x)$ de $f(x)$ au voisinage de x^k .

$$f(x) \simeq q(x) = f(x^k) + [\nabla f(x^k)]^T (x - x^k) + \frac{1}{2} (x - x^k)^T \nabla^2 f(x^k) (x - x^k).$$

Si le hessien $H(x^k) = \nabla^2 f(x^k)$ est définie positif, le point x^{k+1} de la méthode de Newton est alors pris comme étant le minimum de la fonction $q(x)$:

$$0 = \nabla q(x^{k+1}) = \nabla f(x^k) + \nabla^2 f(x^k)(x^{k+1} - x^k) = 0.$$

D'où la formule itérative

$$x^{k+1} = x^k - [\nabla^2 f(x^k)]^{-1} \nabla f(x^k).$$

Cette formule itérative est une itération de la méthode de Newton. Elle correspond à l'itération plus générale

$$x^{k+1} = x^k - \theta_k D_k \nabla f(x^k), \quad k = 0, 1, \dots \quad (3.15)$$

où $\theta_k = 1$, $D_k = [\nabla^2 f(x^k)]^{-1}$, ou encore

$$x^{k+1} = x^k + \theta_k d_k$$

avec $\theta_k = 1$, $d_k = - [\nabla^2 f(x^k)]^{-1} \nabla f(x^k)$.

Remarque 3.3.1. Lorsque le hessien $\nabla^2 f(x^k)$ n'est pas défini positif, la direction de déplacement d_k dans la méthode de Newton peut ne pas être une direction de descente. En effet, l'expression

$$[\nabla f(x^k)]^T d_k = -[\nabla f(x^k)]^T [\nabla^2 f(x^k)]^{-1} \nabla f(x^k) = -d_k^T \nabla^2 f(x^k) d_k$$

peut être éventuellement positive et dans ce cas, la direction d_k ne serait plus une direction de descente.

Exemple 3.3.1. On cherche à obtenir le zéro de la fonction

$$F : (x, y, z) \mapsto \begin{pmatrix} x^2 + y^2 + z^2 - 3 \\ x^2 + y^2 - z - 1 \\ x + y + z - 3 \end{pmatrix}.$$

La méthode de Newton s'écrit comme suit :

$$\begin{cases} x^{(0)} \text{ est donné} \\ x^{(k+1)} = x^{(k)} - (\nabla F(x^{(k)}))^{-1} F(x^{(k)}). \end{cases}$$

Dans notre cas, on a

$$\nabla F(x) = \begin{pmatrix} 2x & 2y & 2z \\ 2x & 2y & -1 \\ 1 & 1 & 1 \end{pmatrix}.$$

Premier cas : On choisit $x^{(0)} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$.

Itération 1 : On a

$$F(x^{(0)}) = \begin{pmatrix} -1 \\ -1 \\ -1 \end{pmatrix}$$

et

$$\nabla F(x^{(0)}) = \begin{pmatrix} 2 & 0 & 2 \\ 2 & 0 & -1 \\ 1 & 1 & 1 \end{pmatrix}$$

d'où

$$x^{(1)} = \begin{pmatrix} \frac{3}{2} \\ \frac{1}{2} \\ 1 \end{pmatrix}.$$

Itération 2 : On a

$$F(x^{(1)}) = \begin{pmatrix} \frac{-1}{2} \\ \frac{3}{2} \\ -1 \end{pmatrix}$$

et

$$\nabla F(x^{(1)}) = \begin{pmatrix} 3 & 1 & 2 \\ 3 & 1 & -1 \\ 1 & 1 & 1 \end{pmatrix}$$

d'où

$$x^{(2)} = \begin{pmatrix} \frac{5}{4} \\ \frac{3}{4} \\ 1 \end{pmatrix}.$$

Itération 3 : On a

$$F(x^{(2)}) = \begin{pmatrix} \frac{1}{8} \\ \frac{1}{8} \\ 0 \end{pmatrix}$$

et

$$\nabla F(x^{(2)}) = \begin{pmatrix} 5/2 & 3/2 & 2 \\ 5/2 & 3/2 & -1 \\ 1 & 1 & 1 \end{pmatrix}$$

d'où

$$x^{(3)} = \begin{pmatrix} \frac{9}{8} \\ \frac{1}{8} \\ 1 \end{pmatrix}.$$

On a ainsi

$$F(x^{(3)}) = \begin{pmatrix} \frac{1}{32} \\ \frac{1}{32} \\ 0 \end{pmatrix}.$$

Deuxième cas : On choisit $x^{(0)} = (0 \ 0 \ 0)$.

Itération 1 : On a

$$F(x^{(0)}) = \begin{pmatrix} 3 \\ -1 \\ -3 \end{pmatrix}$$

et

$$\nabla F(x^{(0)}) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 1 & 1 & 1 \end{pmatrix}.$$

On tombe ainsi sur l'un des cas pathologiques de la méthode de Newton, lorsque la matrice ∇F est singulière.

3.4 La méthode du gradient conjugué

Dans cette méthode, on utilise un modèle quadratique de f , nous allons donc d'abord présenter l'algorithme appliqué à une fonction fortement convexe quadratique.

3.4.1 Cas linéaire

Soit $f(x) = \frac{1}{2}x^T Ax - bx$, avec A symétrique définie positive. On sait alors qu'il existe un minimum unique sur \mathbb{R}^n donné par $x^* = A^{-1}b$. Dans la suite, on note $r(x) = Ax - b = \nabla f(x)$ le résidu.

Définition 3.4.1. Les vecteurs non nuls $\{d_1, \dots, d_p\}$ sont dits conjugués par rapport à la matrice A si

$$\text{pour tout } i \neq j \in \{1, \dots, p\} : d_i^T A d_j = 0.$$

Lemme 3.4.1. *Un ensemble de vecteurs conjugués par rapport à A est un ensemble de vecteurs linéairement indépendants.*

Posons $\phi(\rho) = f(x + \rho d)$, alors

$$\phi'(\rho) = \rho d^T A d + A x d - b d$$

et

$$f(x + \rho d) = \frac{1}{2}(x + \rho d)^T A(x + \rho d) - b(x + \rho d).$$

D'où

$$\phi'(\rho) = 0 \Rightarrow \rho = -\frac{r(x)d}{d^T A d}, \quad (d \neq 0). \quad (3.16)$$

Théorème 3.4.1. *Soit $x^{(0)} \in \mathbb{R}^m$ et $\{d_0, \dots, d_{m-1}\}$ un ensemble de vecteurs conjugués par rapport à A . On considère la suite définie par*

$$x^{(k+1)} = x^{(k)} + \rho_k d_k, \quad \text{où} \quad \rho_k = -\frac{r(x^{(k)})d_k}{d_k^T A d_k}.$$

Alors $r(x^{(k)})d_i = 0$ pour $i = 0, \dots, k-1$ et $x^{(k)}$ minimise f et la suite $(x^{(k)})$ converge en au plus n itérations.

Pour pouvoir utiliser la méthode itérative du théorème précédent, il faut construire un ensemble de vecteurs conjugués par rapport à A . Or, afin de réduire le nombre d'opérations, on se propose de déduire $d^{(k)}$ à partir de $d^{(k-1)}$ uniquement.

Dans l'algorithme du gradient conjugué linéaire, on prend

$$d^{(k)} = -\nabla f(x^{(k)}) + \beta_k d^{(k-1)}$$

avec β_k tel que $d^{(k)T} Ad^{(k-1)} = 0$, on en déduit

$$\begin{aligned} d^{(k)T} Ad^{(k-1)} &= (-\nabla f(x^{(k)}) + \beta_k d^{(k-1)})^T Ad^{(k-1)} \\ &= -r(x^{(k)T}) Ad^{(k-1)} + \beta_k d^{(k-1)T} Ad^{(k-1)} \end{aligned}$$

d'où

$$\beta_k = \frac{r(x^{(k)T}) Ad^{(k-1)}}{d^{(k-1)T} Ad^{(k-1)}}. \quad (3.17)$$

Comme $d^{(k)T} \nabla f(x^{(k)}) = -\|r(x^{(k)})\|_2^2$, on déduit bien une direction de descente.

Remarque 3.4.1. On rappelle que s'il existe un vecteur d tel que $d^T \nabla f(x^*) < 0$, alors d est une direction de descente de f en x^* .

Algorithme : Gradient conjugué linéaire

1. Initialisation Choisir $x^{(0)}$ et poser $k = 0$, $r^{(0)} = Ax^{(0)} - b$, $d^{(0)} = -r^{(0)}$.
2. Itération k Tant que $\|r(x^{(k)})\| > \epsilon$, faire

$$\begin{aligned} \rho_k &= \frac{r^{(k)T} r^{(k)}}{d^{(k)T} Ad^{(k)}} \\ x^{(k+1)} &= x^{(k)} + \rho_k d^{(k)} \\ r^{(k+1)} &= r^{(k)} + \rho_k Ad^{(k)} \\ \beta_{k+1} &= \frac{r^{(k+1)T} r^{(k+1)}}{r^{(k)T} r^{(k)}} \\ d^{(k+1)} &= -r^{(k+1)} + \beta_{k+1} d^{(k)} \\ k &= k + 1 \end{aligned}$$

Fin tant que

Notons que les formules pour $r^{(k)}$ et β_{k+1} ont changé, ceci est possible grâce au théorème suivant qui affirme en particulier que les $d^{(k)}$ sont conjugués.

Théorème 3.4.2. Soit $x^{(k)} \neq x^*$, obtenu après la $k^{\text{ème}}$ itération de l'algorithme du gradient conjugué linéaire avec $r^{(k)}$ et β_{k+1} définis par les équations (8.17) et (8.18) respectivement. Alors, on a

- (i) $r^{(k)}r^{(i)} = 0$, pour $i = 0, \dots, k-1$;
- (ii) $\text{Vect}\{r^{(0)}, \dots, r^{(k)}\} = \text{Vect}\{r^{(0)}, Ar^{(0)} \dots, A^k r^{(k)}\}$;
- (iii) $\text{Vect}\{d^{(0)}, \dots, d^{(k)}\} = \text{Vect}\{d^{(0)}, Ad^{(0)} \dots, A^k d^{(k)}\}$;
- (iv) $d^{(k)T} Ad^{(i)} = 0$ pour $i = 0, \dots, k-1$;

Commentaires. • On a

$$\rho_k = -\frac{r^{(k)T} d^{(k)}}{d^{(k)T} Ad^{(k)}}, \quad (3.18)$$

et dans l'algorithme, on a

$$\rho_k = \frac{r^{(k)T} r^{(k)}}{d^{(k)T} Ad^{(k)}} \quad (3.19)$$

Montrons l'équivalence de (3.18) et (3.19). On a $d^{(k)} = -r^{(k)} + \beta_k d^{(k-1)}$, donc (3.18) s'écrit comme suit :

$$\begin{aligned} \rho_k &= -\frac{r^{(k)T}}{d^{(k)T} Ad^{(k)}} (-r^{(k)} + \beta_k d^{(k-1)}) \\ &= \frac{r^{(k)T} r^{(k)}}{d^{(k)T} Ad^{(k)}} - \beta_k \frac{r^{(k)T} d^{(k-1)}}{d^{(k)T} Ad^{(k)}} \\ &= \frac{r^{(k)T} r^{(k)}}{d^{(k)T} Ad^{(k)}} \end{aligned}$$

car $r^{(k)T} d^{(k-1)} = 0$.

- On a $x^{(k+1)} = x^{(k)} + \rho_k d^{(k)}$ et $r^{(k)} = Ax^{(k)} - b$ d'où

$$\begin{aligned} r^{(k+1)} - r^{(k)} &= (Ax^{(k+1)} - b) - (Ax^{(k)} - b) \\ &= (Ax^{(k+1)} - x^{(k)}) \\ &= \rho_k Ad^{(k)} \end{aligned}$$

d'où

$$r^{(k+1)} = r^{(k)} + \rho_k Ad^{(k)}.$$

- Montrons que

$$\beta_k = \frac{r^{(k)T} Ad^{(k-1)}}{d^{(k-1)T} Ad^{(k-1)}}$$

et

$$\beta_k = \frac{r^{(k)T} r^{(k)}}{r^{(k-1)T} r^{(k-1)}}$$

sont équivalentes.

On a

$$r^{(k)} - r^{(k-1)} = \rho_{k-1} Ad^{(k-1)}$$

donc

$$r^{(k)T} Ad^{(k-1)} = r^{(k)T} \frac{r^{(k)} - r^{(k-1)}}{\rho_{k-1}}$$

d'où

$$\beta_k = r^{(k)T} \frac{r^{(k)} - r^{(k-1)}}{\rho_{k-1} d^{(k-1)T} Ad^{(k-1)}}$$

or

$$\rho_k = \frac{r^{(k)T} r^{(k)}}{d^{(k)T} Ad^{(k)}} \Rightarrow \rho_{k-1} d^{(k-1)T} Ad^{(k-1)} = r^{(k-1)T} r^{(k-1)}$$

d'où

$$\beta_k = r^{(k)T} \frac{r^{(k)} - r^{(k-1)}}{r^{(k-1)T} r^{(k-1)}} = \frac{r^{(k)T} r^{(k)}}{r^{(k-1)T} r^{(k-1)}}$$

car

$$r^{(k)T} r^{(k-1)} = 0, \quad r^{(k)} = d^{(k)} - \beta_{k-1} d^{(k-1)}$$

et $r^{(k)}$ est orthogonal à toutes les directions $d^{(0)}, d^{(1)}, \dots, d^{(k)}$.

□

- Remarque 3.4.2.* 1. Dans la démonstration du théorème, il est essentiel de choisir $d^{(0)} = -\nabla f(x^{(0)})$. C'est à partir de $d^{(0)}$ que l'on construit des gradients orthogonaux et des directions conjuguées.
2. L'algorithme du gradient conjugué linéaire est surtout utile pour résoudre des grands systèmes creux.
3. La convergence peut être assez rapide, si A admet seulement p valeurs propres distinctes, la convergence a lieu en au plus p itérations.

3.4.2 Cas non linéaire

Pour pouvoir appliquer l'algorithme du gradient conjugué linéaire au cas d'une fonction f quelconque, il faut :

- remplacer $r(x^{(k)})$ par $\nabla f(x^{(k)})$;
- déterminer ρ_k par une minimisation unidimensionnelle.

Algorithme du Gradient conjugué non linéaire (Fletcher-Reeves)

1. Initialisation Choisir $x^{(0)}$, poser $k = 0$ et calculer $f^{(0)} = f(x^{(0)})$ et $\nabla f^{(0)} = \nabla f(x^{(0)})$, $d^{(0)} = -\nabla f^{(0)}$.

2. Itération k Tant que $\|\nabla f^{(k)}\| > \epsilon$, faire
- Déterminer ρ_k
 - Calculer $x^{(k+1)} = x^{(k)} + \rho_k d^{(k)}$
 - Calculer $\nabla f^{(k+1)} = \nabla f(x^{(k+1)})$
 - Calculer $\beta_{k+1} = \frac{\nabla f^{(k+1)T} \nabla f^{(k+1)}}{\nabla f^{(k)T} \nabla f^{(k)}}$
 - Calculer $d^{(k+1)} = -\nabla f^{(k+1)} + \beta_{k+1} d^{(k)}$
 - Poser $k = k + 1$.
- Fin tant que

3.5 La méthode de relaxation

La dernière méthode que nous présentons permet de ramener un problème de minimisation dans \mathbb{R}^n à la résolution successive de n problèmes de minimisation dans \mathbb{R} , à chaque itération.

On cherche à minimiser $f : \mathbb{R}^n \rightarrow \mathbb{R}$. On pose $X = (x_1, \dots, x_n)$.

Le principe de la méthode est le suivant : étant donné un itéré X^k de coordonnées (x_1^k, \dots, x_n^k) , on fixe toutes les composantes sauf la première et on minimise sur la première :

$$\min f(x, x_2^k, \dots, x_n^k), x \in \mathbb{R}.$$

On obtient ainsi la première coordonnée de l'itéré suivant X^{k+1} que l'on note x_1^{k+1} ; on peut, pour effectuer cette minimisation dans \mathbb{R} , utiliser par exemple la méthode de Newton dans \mathbb{R} . On recommence ensuite en fixant la première coordonnée à x_1^{k+1} et les $n - 2$ dernières comme précédemment. On minimise sur la deuxième et ainsi de suite.

L'algorithme obtenu est :

Algorithme de relaxation

1. Initialisation Poser $k = 0$ et choisir $X^0 \in \mathbb{R}^n$.
2. Itération k Pour i variant de 1 à n , on calcule la solution x_i^{k+1} de

$$\min f(x_1^{k+1}, x_2^{k+1}, \dots, x_i^{k+1}, x, x_{i+1}^{k+1}, \dots, x_n^{k+1}), x \in \mathbb{R}.$$

3. Critère d'arrêt – Si $\|x_{k+1} - x_k\| < \epsilon$, stop.
- Sinon, poser $k = k + 1$ et aller à 2.