المدرسة الوطنية للمهندسين بتونس

école nationale d'ingénieurs de Tunis

Analyse Numérique

Radhia Bessi & Maher Moakher

2015–2016

Table des matières

Chapitre 1

Méthodes directes pour la résolution des systèmes linéaires

1.1 Introduction

La résolution des systèmes linéaires de grandes tailles est l'un des plus importants problèmes en analyse numérique. Le but de ce chapitre est d'étudier des méthodes de résolution numérique d'un linéaire $Ax = b$, où A est une matrice carrée inversible.

Pour motivation, commençons par citer le problème mécanique classique suivant qui conduit à la résolution d'un système linéaire.

La déformation x d'une corde élastique, fixée aux bords et soumise à un champ de force f, peut se traduire par l'équation différentielle suivante

$$\begin{cases} -x''(t) = f(t), & t \in [0,1] \\ x(0) = x(1) = 0, \end{cases} \tag{1.1}$$

pour f une fonction continue sur $[0,1]$. En général, il n'est pas possible d'expliciter la solution exacte de ce problème. L'idée donc est de chercher une solution approchée x^h de x en prenant une subdivision $0 = t_0 \leq t_1 \leq \cdots \leq t_{n+1} = 1$, avec $t_i = ih$, $i = 0, \ldots, n+1$ et $h = \dfrac{1}{n+1}$. On se limitera à calculer $x^h(t_i) = x_i \simeq x(t_i)$, pour $i = 0, \ldots, n+1$ et par interpolation par exemple, on peut avoir x^h sur tout l'intervalle $[0,1]$.

Si on suppose que notre solution x est de classe C^2, alors on a les deux développement de Taylor suivants :

$$x(t_{i+1}) = x(t_i + h) = x(t_i) + x'(t_i)h + x''(t_i)\frac{h^2}{2} + O(h^3),$$

et

$$x(t_{i-1}) = x(t_i - h) = x(t_i) - x'(t_i)h + x''(t_i)\frac{h^2}{2} + O(h^3).$$

Si on fait la somme de deux égalités on obtient

$$x''(t_i) = \frac{x(t_i + h) - 2x(t_i) + x(t_i - h)}{h^2} + O(h).$$

Si on néglige les termes d'ordre $O(h)$ dans l'expression de la dérivée seconde, le problème discrétisé pour résoudre notre équation devient

$$\begin{cases} -\dfrac{x_{i+1} - 2x_i + x_{i-1}}{h^2} = f(x_i), \quad i = 1, \ldots, n \\ x_0 = x_{n+1} = 0 \end{cases} \tag{1.2}$$

qui est équivalent au système linéaire $A_h x = b_h$, où

$$A_h = \frac{1}{h^2} \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & -1 & 2 & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{pmatrix} \quad x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \text{ et } b_h = \begin{pmatrix} f(t_1) \\ \vdots \\ f(t_n) \end{pmatrix}.$$

La solution approchée est d'autant plus proche de la solution exacte x lorsque n est grand.

On rappele que la solution unique $x = (x_1, \ldots, x_n)$ d'un système $Ax = b$, pour $A = (a_{ij})_{1 \leq i,j \leq n}$ une matrice inversible et $b = (b_1, \ldots, b_n)^T$ est donnée par les formules de **Cramer** suivantes :

$$x_i = \frac{\det A_i}{\det A}, \quad i = 1, \ldots, n,$$

où det désigne le déterminant et A_i est la matrice d'ordre n obtenue à partir de A en remplaçant sa colonne i par le vecteur de second membre b. Donc la résolution d'un système linéaire d'ordre n, par les formules de Cramer fait intervenir le calcul de $n + 1$ déterminants dont chaque déterminant nécessite de l'ordre de $n!$ multiplications. La méthode de Cramer devient trop coûteuse même pour des matrices de tailles assez petites. D'où l'idée de concevoir des méthodes qui en général donnent la solution à une valeur près, mais avec un nombre raisonnable d'opérations.

Pour simplifier, on se limitera à résoudre numériquement un système linéaire réel sachant qu'un système linéaire complexe est équivalent à deux systèmes linéaires réels représentants ses parties réelles et imaginaire.

On distinguera deux méthodes numériques pour résoudre un système linéaire. Les méthodes directes, dont certaines font l'objet de ce chapitre et les méthodes itératives qui seront développées dans les chapitres suivants.

1.2 Principe des méthodes directes

On appelle méthode directe pour résoudre un système de type $Ax = b$, une méthode qui donne x après un nombre fini d'opérations élémentaires. Le principe de ces méthodes est de se ramener à un système linéaire équivalent, mais qui est plus simple à résoudre. C'est le cas par exemple où la matrice du système est diagonale, triangulaire ou orthogonale.

1.2.1 Rappels et notations sur les vecteurs et les matrices

Tout au long de ce cours on utilisera les notations suivantes :

Pour n un entier naturel non nul fixé, on note par $\mathbb{R}^n$ l'espace vectoriel des vecteurs colonnes à coefficients réels $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$. Le vecteur ligne $x^T = (x_1, \ldots, x_n)$ désigne la transposée du vecteur $x \in \mathbb{R}^n$.

Le produit scalaire dans $\mathbb{R}^n$ sera noté $(\cdot, \cdot)$ et est défini par

$$(x, y) = \sum_{i=1}^{n} x_i y_i,$$

et la norme associée est la norme euclidienne donnée par

$$\|x\|_2 = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{\frac{1}{2}} = (x, x)^{\frac{1}{2}}.$$

On notera l'espace vectoriel des matrices à n lignes et p colonnes et à coefficients dans $\mathbb{R}$ par $\mathcal{M}_{n,p}(\mathbb{R})$ et par $\mathcal{M}_n(\mathbb{R})$ l'ensemble des matrices carrées d'ordre n à coefficients dans $\mathbb{R}$.

La transposée d'une matrice $A = (a_{ij})_{1 \leq i \leq n, 1 \leq j \leq p} \in \mathcal{M}_{n,p}(\mathbb{R})$ par

$$A^T = (a_{ji})_{1 \leq i \leq p, 1 \leq j \leq n} \in \mathcal{M}_{p,n}(\mathbb{R}).$$

La matrice transposée vérifie

$$(Ax, y) = (Ax)^T y = x^T A^T y = (x, A^T y) \text{ pour tout } x \in \mathbb{R}^p \text{ et } y \in \mathbb{R}^n.$$

Définition 1.2.1 *Soit $A = (a_{ij})_{1 \leq i,j \leq n} \in \mathcal{M}_n(\mathbb{R})$. A est dite :*

- *diagonale si $a_{ij} = 0$ pour tout $i \neq j$.*
- *triangulaire inférieure si $a_{ij} = 0$ pour tout $i < j$.*
- *triangulaire supérieure si $a_{ij} = 0$ pour tout $i > j$.*
- *inversible si son déterminant est non nul ou s'il existe une matrice $B \in \mathcal{M}_n(\mathbb{R})$ telle que*

$$AB = BA = I_n$$

où I_n est la matrice unité d'ordre n donnée par

$$I_n = \begin{pmatrix} 1 & & \bigcirc \\ & \ddots & \\ \bigcirc & & 1 \end{pmatrix}.$$

La matrice B est appelée inverse de A et sera notée A^{-1}.

- *semblable à une matrice $D \in \mathcal{M}_n(\mathbb{R})$ s'il existe une matrice inversible P telle que $D = P^{-1}AP$.*
- *diagonalisable si elle est semblable à une matrice diagonale.*

On peut vérifier facilement que

Proposition 1.2.1

1. *Le produit de deux matrices triangulaires supérieures (repectivement inférieures) est une matrice triangulaire supérieure (respectivement inférieure).*
2. *Une matrice $A = (a_{ij})_{1\leq,i,j\leq n} \in \mathcal{M}_n(\mathbb{R})$ triangulaire est inversible si et seulement si $a_{ii} \neq 0$, pour tout $i = 1, \ldots, n$. De plus A^{-1} est de même type que A et $(A^{-1})_{ii} = \dfrac{1}{a_{ii}}$, pour tout $i = 1, \ldots, n$.*

1.2.2 Résolution d'un système triangulaire

Soit le système trianglaire supérieur

$$\left\{\begin{array}{lllll} u_{11}x_1+ & u_{12}x_2+ & \cdots & u_{1n}x_n & = b_1 \\ & u_{22}x_2+ & \cdots & u_{2n}x_n & = b_2 \\ & & \ddots & \vdots & \\ & & & u_{nn}x_n & = b_n \end{array}\right. . \tag{1.3}$$

Si U est la matrice triangulaire supérieure suivante :

$$U = \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ & u_{22} & \cdots & u_{2n} \\ & \bigcirc & \ddots & \vdots \\ & & & u_{nn} \end{pmatrix},$$

alors, le système (1.3) est équivalent à $Ux = b$ où $b = (b_i)_{1\leq i\leq n}$. On utilise la méthode de remontée pour résoudre ce système dont l'algorithme est le suivant :

$x_n = \frac{b_n}{u_{nn}}$.
Pour $k = (n-1), \ldots, 1$
$$x_k = \left(b_k - \sum_{i=k+1}^{n} u_{ki}x_i\right)\frac{1}{u_{kk}}$$
Fin de la boucle sur k.

A chaque étape k on a une division, $(n-k)$ multiplication pour $1 \leq k \leq n$ et $(n-k)$ additions pour $k \leq n-1$. Au total le nombre d'opérations est

$$\sum_{k=1}^{n}(n-k+1) + \sum_{k=1}^{n-1}(n-k) = \frac{n(n+1)}{2} + \frac{n(n-1)}{2} = n^2.$$

Pour la résolution d'un système triangulaire inférieur $Lx = b$, on utilise la méthode de descente dont l'algorithme est le suivant :

$$\left|\begin{array}{l} x_1 = \dfrac{b_1}{l_{11}}. \\ \text{Pour } k = 2, \ldots, n \\ \quad x_k = \left(b_k - \displaystyle\sum_{i=1}^{k-1} l_{ki} x_i\right) \dfrac{1}{l_{kk}}. \\ \text{Fin de la boucle sur } k. \end{array}\right.$$

Cette méthode nécessite aussi n^2 opérations.

1.3 Factorisation LU

1.3.1 Rappel sur la méthode de Gauss

Parmi les méthodes directes classiques pour résoudre un système linéaire $Ax = b$, pour A une matrice carrée inversible d'ordre n, on cite la méthode d'élimination de **Gauss** dont le principe est d'effectuer un nombre fini d'opérations algébriques linéaires sur les lignes de la matrice A et sur b, pour se ramener à un système triangulaire supérieur équivalent.

Exemple 1.3.1 *Soit à résoudre le système linéaire d'ordre 4, suivant :*

$$\left\{\begin{array}{rcrcrcrcl} x_1 & - & 3x_2 & - & x_3 & & & = & 2 \\ -x_1 & + & x_2 & & & + & 2x_4 & = & 3 \\ & & x_2 & - & x_3 & & & = & 1 \\ 2x_1 & + & x_2 & & & - & x_4 & = & 0 \end{array}\right.$$

Sous forme matricielle ce système s'écrit sous la forme $Ax = b$, pour

$$A = \begin{pmatrix} \boxed{1} & -3 & -1 & 0 \\ -1 & 1 & 0 & 2 \\ 0 & 1 & -1 & 0 \\ 2 & 1 & 0 & -1 \end{pmatrix} \quad et \quad b = \begin{pmatrix} 2 \\ 3 \\ 1 \\ 0 \end{pmatrix}.$$

Dans la première étape on élimine la première inconnue x_1 des équations 2,3 et 4 en combinant chacune avec la première équation. Afin d'éliminer il faut d'abord vérifier que x_1 apparait dans la première équation. Si ce n'est pas le cas, il faut permuter l'équation avec une autre dont le coefficient de x_1 est non nul. On choisit comme premier pivot α_1 le coefficient de x_1 dans la nouvelle première équation qui est appelée ligne de pivot. Dans notre exemple $\alpha_1 = 1$. Eliminer x_1 des autres équations revient à annuler les coefficients de la première colonne de A en dessous de la diagonale. Ceci revient

dans la méthode de Gauss à remplacer chaque ligne L_i de A et de b par $L_i - \frac{a_{i1}}{\alpha_1}L_1$, pour $i = 2, 3$ et 4. Dans notre cas

$$L_2 \longleftarrow L_2 + L_1, \quad L_3 \longleftarrow L_3, \quad L_4 \longleftarrow L_4 - 2L_1.$$

Après cette première étape, le système équivalent a comme nouvelle matrice et comme second membre

$$A^{(1)} = \begin{pmatrix} 1 & -3 & -1 & 0 \\ 0 & \boxed{-2} & -1 & 2 \\ 0 & 1 & -1 & 0 \\ 0 & 7 & 2 & -1 \end{pmatrix} \text{ et } b^{(1)} = \begin{pmatrix} 2 \\ 5 \\ 1 \\ -4 \end{pmatrix}.$$

Dans la deuxième étape c'est la deuxième ligne qui joue le role de ligne de pivot si x_2 est présent (Sinon, on permute la deuxième équation avec une autre sans toucher la première). Le coefficient de x_2 devient le nouveau pivot α_2 qui vaut -2 dans cet exemple. Pour annuler les coefficients de la deuxième colonne en dessous de la diagonale, on fait les opérations

$$L_3 \longleftarrow L_3 + \tfrac{1}{2}L_2, \quad L_4 \longleftarrow L_4 + \frac{7}{2}L_2.$$

Le système équivalent a pour matrice et second membre

$$A^{(2)} = \begin{pmatrix} 1 & -3 & -1 & 0 \\ 0 & -2 & -1 & 2 \\ 0 & 0 & \boxed{-\frac{3}{2}} & 1 \\ 0 & 0 & -\frac{3}{2} & 6 \end{pmatrix} \text{ et } b^{(2)} = \begin{pmatrix} 2 \\ 5 \\ \frac{7}{2} \\ \frac{27}{2} \end{pmatrix}.$$

Enfin, pour éliminer x_3 de la quatrième équation, on utilise le pivot $\alpha_3 = -\frac{3}{2}$ et on fait $L_4 \longleftarrow L_4 - L_3$. La matrice et le second membre de système équivalent sont

$$A^{(3)} = \begin{pmatrix} 1 & -3 & -1 & 0 \\ 0 & -2 & -1 & 2 \\ 0 & 0 & -\frac{3}{2} & 1 \\ 0 & 0 & 0 & 5 \end{pmatrix} \text{ et } b^{(3)} = \begin{pmatrix} 2 \\ 5 \\ \frac{7}{2} \\ 10 \end{pmatrix}.$$

Le dernier système est triangulaire supérieur. On résout par la méthode de remontée ce système triangulaire supérieur pour obtenir $x = \begin{pmatrix} 1 \\ 0 \\ -1 \\ 2 \end{pmatrix}$ qui est aussi la solution de (S).

Remarque 1.3.1 *Si on note $U = A^{(3)}$, alors on peut vérifier qu'on a*

$$A = LU, \quad \text{avec } L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -\frac{1}{2} & 1 & 0 \\ 2 & -\frac{7}{2} & 1 & 1 \end{pmatrix}$$

est la matrice triangulaire inférieure dont les termes diagonaux valent 1 et dont le coefficient l_{ik} de la matrice L sous la diagonale est exactement le terme $\dfrac{a_{ik}^{(k)}}{\alpha_k}$, où $a_{ik}^{(k)}$ désigne les coefficients de $A^{(k)}$, matrice d'élimination de Gauss à l'étape k.

Algorithme de Gauss et nombre d'opérations

Si pour $i = 1, \ldots, n$ et $p = 1, \ldots, n$, on désigne par $A[i, p : n]$ les termes de la ligne i d'une matrice A situés dans l'ordre entre les colonnes p et n, alors l'algorithme de la méthode d'élimination de Gauss pour résoudre un système $Ax = b$, s'écrit :

Pour $k = 1, \ldots, n-1$,
  Recherche du pivot et permutation des lignes si nécessaire
  Pour $i = k+1, \ldots, n$
    $\ell_{ik} = \frac{a_{ik}}{a_{kk}}$
    $A[i, k : n] = A[i, k : n] - \ell_{ik} A[k, k : n]$
    $b_i = b_i - \ell_{ik} b_k$
  Fin de la boucle sur i
Fin de la boucle sur k.

Puis, on résout le système triangulaire obtenu.

A chaque étape k de l'élimination de x_k il faut effectuer :

- $(n-k)$ divisions,
- $(n-k+1)(n-k)$ multiplications,
- $(n-k+1)(n-k)$ additions.

Le nombre total d'opérations est donc :

$$3\sum_{k=1}^{n-1}(n-k) + 2\sum_{k=1}^{n-1}(n-k)^2 = 2\sum_{p=1}^{n-1} p + 2\sum_{p=1}^{n-1} p^2.$$

Si on utilise le fait que $\displaystyle\sum_{p=1}^{n} p = \frac{n(n+1)}{2}$ et que $\displaystyle\sum_{p=1}^{n} p^2 = \frac{n(n+1)(2n+1)}{6}$ on tire que le nombre d'opérations dans la méthode d'élimination de Gauss est $\dfrac{3}{2}n(n-1) + 2\dfrac{n(n-1)(2n-1)}{6}$ qui est de l'ordre de $\dfrac{2}{3}n^3$ lorsque n est très grand.

Remarque 1.3.2 *Le choix du pivot peut avoir une grande importance dans le calcul comme le montrera l'exemple suivant.*

Soit à résoudre le système

$$\begin{pmatrix} 10^{-p} & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

dont la solution exacte est $\begin{pmatrix} -\dfrac{1}{1-10^{-p}} \\ \dfrac{1}{1-10^{-p}} \end{pmatrix}$.

Après la première étape d'élimination de Gauss, on obtient

$$\begin{pmatrix} 10^{-p} & 1 \\ 0 & 1-10^{p} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ -10^{p} \end{pmatrix}.$$

Pour p tel que sur une machine $1-10^{p} \simeq -10^{p}$ on obtient $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \simeq \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ qui n'est pas proche de la solution exacte.

Cet exemple montre l'effet des erreurs d'arrondi qui proviennent de la division par des pivots trop petits. La stratégie de pivot partiel consiste à choisir comme pivot α_k vérifiant $|\alpha_k| = \max\limits_{k\leq i\leq n} |a_{ik}^{(k)}|$ le coefficient de module maximum de la partie de la colonne k en dessous de la diagonale.

Dans l'exemple (1.3.1)*, l'élimination de Gauss était faite sans stratégie de pivot. Avec la stratégie de pivot partiel, et dans la première étape, on doit choisir le premier pivot $a_{31} = 2$ au lieu de $a_{11} = 1$.*

1.3.2 Factorisation LU

Dans cette partie on développe la technique de la factorisation LU qui permet aussi de résoudre un système $Ax = b$ en se ramenant à la résolution d'un système triangulaire inférieur puis d'un système triangulaire supérieur.

Proposition 1.3.1 *Soit $A = (a_{ij})_{1\leq i,j\leq n} \in \mathcal{M}_n(\mathbb{R})$ une matrice carrée d'ordre n telle que toutes les n sous-matrices $A^{(k)} = (a_{ij})_{1\leq i,j\leq k}$, pour $k = 1, \ldots, n$, sont inversibles.*

Alors, il existe une matrice triangulaire inférieure L dont les coefficients diagonaux sont égaux à 1, et une matrice triangulaire supérieure U telle que $A = LU$. De plus cette décomposition, dite décomposition ou factorisation LU de A, est unique.

***Preuve*. Unicité :** On suppose que $A = L_1U_1 = L_2U_2$ avec $L_i, i = 1, 2$ triangulaire inférieure vérifiant $(L_i)_{kk} = 1$ pour tout $k = 1, \ldots, n$, $U_i, i = 1, 2$ triangulaire supérieure. Alors les matrices U_1 et U_2 sont inversibles, puisque $0 \neq \det A = \det U_1 = \det U_2$. On a alors $L_2^{-1}L_1 = U_2U_1^{-1}$, ce qui implique que $L_2^{-1}L_1$ est à la fois une matrice triangulaire inférieure et triangulaire supérieure, donc $L_2^{-1}L_1$ est une matrice diagonale. Mais comme $(L_1)_{kk} = (L_2^{-1})_{kk} = 1$ on en déduit que $L_2^{-1}L_1 = I_n$ et par suite $L_2 = L_1$ et $U_2 = U_1$.
Existence : Notons d'abord que si une telle factorisation existe pour $L = (\ell_{ij})_{1\leq i,j\leq n}$ triangulaire inférieure $U = (u_{ij})_{1\leq i,j\leq n}$ est triangulaire supérieure, alors

$$a_{ij} = \sum_{k=1}^{\min(i,j)} l_{ik}u_{kj}, \text{ pour tout } 1 \leq i, j \leq n \tag{1.4}$$

On montre l'existence par récurrence sur n, l'ordre de la matrice A à factoriser.

Si $n = 2$, $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ vérifiant $a \neq 0$ et $det(A) = ad - bc \neq 0$, alors A admet la factorisation

$$A = \begin{pmatrix} 1 & 0 \\ \dfrac{c}{a} & 1 \end{pmatrix} \begin{pmatrix} a & b \\ 0 & \dfrac{ad-bc}{a} \end{pmatrix}.$$

Supposons que le résultat est vrai jusqu'à l'ordre $n-1$. Soit $A = (a_{ij})$ une matrice d'ordre n telle que les sous-matrices $A^{(k)}$, pour $k = 1, \ldots, n$ sont inversibles. D'après l'hypothèse de récurrence la matrice $A^{(n-1)}$ qui est d'ordre $n-1$ admet une factorisation $\tilde{L}\tilde{U}$, où $\tilde{L} = (l_{ij})_{1 \leq i,j \leq n-1}$ est une matrice triangulaire inférieure d'ordre $n-1$ à diagonale unité et $\tilde{U} = (u_{ij})_{1 \leq i,j \leq n-1}$ est une matrice triangulaire supérieure d'ordre $n-1$ dont les termes diagonaux sont tous non nuls ($u_{ii} \neq 0 \; \forall \, i = 1, \ldots, n-1$). On détermine les coefficients représentant la n-ième colonne de U puis la n-ième ligne de L de la façon suivante :

$$\begin{cases} u_{1n} = a_{1n}, \\ u_{2n} = a_{2n} - l_{21}u_{2n}, \\ \quad \vdots \\ u_{pn} = a_{pn} - \displaystyle\sum_{k=1}^{p-1} l_{pk}u_{kn}, \\ \quad \vdots \\ u_{nn} = a_{nn} - \displaystyle\sum_{k=1}^{n-1} l_{nk}u_{kn}, \end{cases}$$

puis les coefficients

$$\begin{cases} l_{n1} = \dfrac{a_{n1}}{u_{11}}, \\ l_{n2} = \dfrac{a_{n2} - l_{n1}u_{12}}{u_{22}}, \\ \quad \vdots \\ l_{np} = \dfrac{a_{np} - \displaystyle\sum_{k=1}^{p-1} l_{nk}u_{kp}}{u_{pp}} \\ \quad \vdots \\ l_{n,n-1} = \dfrac{a_{n,n-1} - \displaystyle\sum_{k=1}^{n-2} l_{nk}u_{kn}}{u_{n-1,n-1}}. \end{cases}$$

Ensuite, on définit les deux matrices d'ordre n :

$$L = \begin{pmatrix} & & & 0 \\ & \tilde{L} & & \vdots \\ l_{n1} & \cdots & l_{n,n-1} & 1 \end{pmatrix} \text{ et } U = \begin{pmatrix} & & & u_{1n} \\ & \tilde{U} & & \vdots \\ 0 & \cdots & 0 & u_{nn} \end{pmatrix}.$$

On note que L est triangulaire inférieure à diagonale unité, U est triangulaire supérieure avec coefficients diagonaux non nuls et on peut vérifier facilement que $A = LU$. □

Exemple 1.3.2 *Reprenons la matrice A de système donné dans l'exemple* (1.3.1)*, où*

$$A = \begin{pmatrix} 1 & -3 & -1 & 0 \\ -1 & 1 & 0 & 2 \\ 0 & 1 & -1 & 0 \\ 2 & 1 & 0 & -1 \end{pmatrix}.$$

On peut vérifier que $\det(A^{(k)}) \neq 0$*; $k = 1, ..., 4$, donc A admet une factorisation LU, avec $L = (\ell_{ij})_{1\leq,i,j\leq 4}$ triangulaire inférieure vérifiant $\ell_{ii} = 1,,\ i = 1, ..., 4$ et $U = (u_{ij})_{1\leq i,j\leq 4}$ est une matrice triangulaire supérieure. Pour calculer L et U on utilise l'égalité* (1.4) *où on commence par calculer les termes $u_{1j}, j = 1, ..., 4$, de la première ligne de U en utilisant le coefficient a_{1j}. Comme* $\inf(1, j) = 1$ *et $\ell_{11} = 1$, on obtient alors*

$$u_{1j} = a_{1j},\ j = 1, ..., 4.$$

Soit donc

$$u_{11} = 1,\ u_{12} = -3,\ u_{13} = -1,\ \text{et } u_{14} = 0.$$

Connaissant la première ligne de U et avant de passer à sa deuxième ligne, on calcule les coefficients $\ell_{i1},\ i = 2, 3$ et 4 car pour $i = 1$, par hypothèse $\ell_{11} = 1$. Or $\inf(i, 1) = 1$ *et de* (1.4) *on obtient $a_{i1} = u_{11}\ell_{i1},\ i = 2, 3$ et 4. Comme on connait u_{11}, on déduit*

$$\ell_{i1} = \frac{a_{i1}}{u_{11}},\ i = 2, 3 \text{ et } 4.$$

Soit donc

$$\ell_{21} = -1,\ \ell_{31} = 0,\ \ell_{41} = 2$$

Ayant la premiere ligne U et la première colonne de L, on peut calculer les coefficients u_{2j} de la deuxième ligne de U, pour $j \geq 2$, puisque U est triangulaire supérieure. Par identification de terme a_{2j} dans (1.4)*, on obtient alors*

$$a_{2j} = \ell_{21}u_{1j} + \ell_{22}u_{2j},$$

Puisque on connait le terme u_{1j} de la première ligne de U et ℓ_{21} de la première colonne de U, et comme $\ell_{22} = 1$, on déduit

$$u_{2j} = a_{2j} - \ell_{21}u_{1j}, j = 2, 3 \text{ et } 4.$$

Dans notre cas

$$u_{22} = -2,\ u_{23} = -1 \text{ et } u_{24} = 2.$$

La connaissance des termes de deux premières lignes de U et de la première colonne de L permet de déduire la deuxième colonne de L et d'avoir,

$$\ell_{32} = (a_{32} - \ell_{31}u_{12})/u_{22} = -\frac{1}{2} \text{ et } \ell_{42} = (a_{42} - \ell_{41}u_{12})/u_{22} = -\frac{7}{2}.$$

De même, on calcule dans l'ordre les termes u_{33} et u_{34} de la troisième ligne de U puis le terme ℓ_{34} de la troisième colonne de L par les formules suivantes :

$$u_{3j} = a_{3j} - (\ell_{31}u_{1j} + \ell_{32}u_{2j}),\ j = 3 \text{ et } 4,\ \ell_{43} = (a_{43} - \ell_{41}u_{13} - \ell_{42}u_{23})/u_{33}.$$

On trouve alors

$$u_{33} = -\frac{3}{2},\ u_{34} = 1 \text{ puis } \ell_{34} = 1.$$

Enfin, reste u_{44} le seul terme à calculer de la colonne 4 de la matrice U qui est donné par

$$u_{44} = a_{44} - (\ell_{41}u_{14} + \ell_{42}u_{24} + \ell_{43}u_{34}) = 5.$$

On obtient alors les matrices L et U données dans la remarque (1.3.1).

Algorithme de la factorisation LU

De (1.4) et pour $p = 1, \ldots, n$, en lisant dans l'ordre les termes de la p-ième ligne eu dessus de la diagonale de la matrice à factoriser A, on obtient la p-ième ligne de U, puis au fur et on mesure qu'on écrit les termes de la p-ième colonnes de A en dessus de la diagonale, on déduit la p-ième colonne de L. Ceci peut se traduire par l'algorithme suivant :

Pour $p = 1, \ldots, n$,
  Pour $j = p, \ldots, n$,
    $u_{pj} := a_{pj} - \sum_{k=1}^{p-1} l_{pk}u_{kj}$.
  Fin de la boucle sur j

  Pour $i = p+1, \ldots, n$,
    $l_{ip} := \left(a_{ip} - \sum_{k=1}^{p-1} l_{ik}u_{kp}\right)/u_{pp}$.
  Fin de la boucle sur i.
Fin de la boucle sur p.

Remarque 1.3.3 *De la preuve ou de l'algorithme de la factorisation LU de la matrice A on tire que toutes les sous matrices $A^{(k)}$ de A admettent aussi une factorisation $L^{(k)}U^{(k)}$ pour $L^{(k)} = (l_{ij})_{1\leq i,j\leq k}$ et $U^{(k)} = (u_{ij})_{1\leq i,j\leq k}$, pour $k = 1, \ldots, n$. Ainsi la matrice A admet une factorisation LU si et seulement si toutes les matrices $A^{(k)}$ sont inversibles.*

Nombre d'opérations

A chaque étape p, pour calculer $u_{pj}, j = p, \ldots, n$ il faut effectuer

- $p - 1$ multiplications,
- $p - 1$ additions,

donc $2(p - 1)$ opérations.

Le nombre total d'opérations pour déterminer la p-ième ligne de U est

$$\sum_{j=p}^{n} 2(p-1) = 2(n-p+1)(p-1).$$

Ensuite pour calculer le terme l_{ip}, $i = p+1, \ldots, n$ il faut

- 1 division,
- $p-1$ multiplications,
- $p-1$ additions.

Donc le calcul de la p-ième colonne de L nécessite

$$\sum_{i=p+1}^{n} (2(p-1)+1) = (n-p)(2p-1).$$

Le nombre total d'opérations pour effectuer la factorisation LU de A est

$$\sum_{p=1}^{n}[2(n-p+1)(p-1) + (n-p)(2p-1)] = (4n+1)\sum_{p=1}^{n} p - 4\sum_{p=1}^{n} p^2 + 2n.$$

Si on utilise le fait que $\displaystyle\sum_{p=1}^{n} p^2 = \frac{n(n+1)(2n+1)}{6}$ et que $\displaystyle\sum_{p=1}^{n} p = \frac{n(n+1)}{2}$ on tire que le nombre d'opérations pour effectuer la factorisation LU d'une matrice d'ordre n est de l'ordre de $\frac{2}{3}n^3$.

Résolution de $Ax = b$ pour $A = LU$

Le système $Ax = b$ donne $LUx = b$. Si on pose $y = Ux$, il suffit de résoudre le système triangulaire inférieur $Ly = b$ puis le système triangulaire supérieur $Ux = y$, i.e.,

$$\text{Résoudre } Ax = b \Leftrightarrow \begin{cases} \text{Résoudre } Ly = b, \\ \text{puis résoudre } Ux = y. \end{cases}$$

Puisque la résolution d'un système triangulaire, nécessite n^2 opérations, le nombre d'opérations pour résoudre un système linéaire $Ax = b$, par la factorisation $A = LU$ est de l'ordre $\frac{2}{3}n^3$.

Remarques 1.3.1

1. *On utilise la factorisation LU d'une matrice A surtout pour résoudre plusieurs systèmes linéaires avec la même matrice A mais avec des seconds membres différents. Si on a par exemple à résoudre p systèmes linéaires avec la même matrice A, le coût est de l'ordre de $\frac{2}{3}n^3 + 2pn^2$ au lieu de $p\frac{2}{3}n^3$.*

2. Puisque

$$\det A = \det L \det U = \prod_{i=1}^{n} u_{ii},$$

la factorisation LU permet aussi de calculer $\det A$ *en seulement* $\frac{2}{3}n^3$ *opérations au lieu de* $(n!-1)(n-1)$ *opérations si on utilise la formule de Leibniz*

$$\det A = \sum_{\sigma \in S_n} \epsilon(\sigma) \prod_{i=1}^{n} a_{\sigma(i),i},$$

où S_n *est l'ensemble des permutations des nombres* $\{1, \ldots, n\}$ *et* $\epsilon(\sigma)$ *est la signature de la permutation* σ*.*

Matrices bandes

Définition 1.3.1 *Une matrice* $A = (a_{ij})_{1 \leq i,j \leq n} \in \mathcal{M}_n(\mathbb{R})$ *est dite matrice bande s'il existe* $n_b \in \mathbb{N}$*, tel que* $a_{ij} = 0$ *pour* $|i-j| > n_b$*. Autrement dit les coefficients non nuls de la matrice* A *sont situés dans une bande de longueur* $2n_b + 1$*, centrée sur la diagonale principale. Pour* $n_b = 1$*, la matrice est dite* **tridiagonale**.

Proposition 1.3.2 *La factorisation* $A = LU$ *conserve la structure bande.*

Preuve. On montre cette propriété par récurrence sur les étapes de la factorisation de la matrice A.

La matrice A étant bande, la propriété est vraie à l'étape 0. Supposons que la propriété est vraie jusqu'à l'étape $p-1$ où on a donc $l_{ij} = u_{ij} = 0$ pour $1 \leq i, j \leq p-1$, $|i-j| > n_b$. A l'étape p on a

$$u_{pj} := a_{pj} - \sum_{k=1}^{p-1} l_{pk} u_{kj}, \text{ pour } j = p, \ldots, n$$

et

$$l_{ip} := \frac{a_{ip} - \sum_{k=1}^{p-1} l_{ik} u_{kp}}{u_{pp}}, \text{ pour } i = p+1, \ldots, n.$$

Soit p tel que $|p-j| > n_b$. Alors $a_{pj} = 0$ puisque A est une matrice bande.
Si $p - j > n_b > 0$ alors $u_{pj} = 0$, puisque U est triangulaire supérieure.
Si $p - j < -n_b < 0$, donc pour $1 \leq k \leq p-1$, on a $k - j \leq p - 1 - j < p - j < -n_b$. D'après l'hypothèse de récurrence $u_{kj} = 0$. Ainsi $u_{pj} = 0$ et la propriété est vérifiée pour la matrice U à l'étape p. De même, si p est tel que $|i-p| > n_b$, alors $a_{pj} = 0$.
Si $i - p < -n_b < 0$ alors $l_{ip} = 0$, puisque L est triangulaire inférieure.
Si $i - p > n_b < 0$, donc pour $1 \leq k \leq p-1 < p$, on a $i - k \leq i - p < -n_b$. D'après l'hypothèse de récurrence $l_{ik} = 0$. Ainsi $l_{ip} = 0$ et la propriété est vérifiée aussi pour la matrice L à l'étape p.
□

1.4 Factorisation de Cholesky

1.4.1 Rappels sur les matrices symétriques

Définition 1.4.1 *Soit $A = (a_{ij})_{1\leq i,j\leq n} \in \mathcal{M}_n(\mathbb{R})$. La matrice transposée de A, notée A^T est définie par*

$$A^T = (a_{ji})_{1\leq i,j\leq n}.$$

L'opérateur transposé vérifie, pour toutes marices A et B de $\mathcal{M}_n(\mathbb{R})$,

1. $(A+B)^T = A^T + B^T$
2. $(AB)^T = B^T A^T$
3. Si A est inversible, alors A^T est inversible et $(A^T)^{-1} = (A^{-1})^T$.
4. $(Ax, y) = (x, A^T y)$ pout tout $x, y \in \mathbb{R}^n$.

Définition 1.4.2 *Soit $A \in \mathcal{M}_n(\mathbb{R})$. La matrice A est dite :*

1. *symétrique si $A^T = A$.*
2. *orthogonale si $AA^T = A^T A = I_n$.*

Remarque 1.4.1 *Si $A \in \mathcal{M}_n(\mathbb{R})$ est une matrice orthogonale, alors A est inversible et $A^{-1} = A^T$ puisqu'on a $AA^T = A^T A = I_n$. De plus les vecteurs lignes ainsi que les vecteurs colonnes forment une base orthonormée de $\mathbb{R}^n$.*

Définition 1.4.3 *:* **Vecteurs et valeurs propres**

Soit $A \in \mathcal{M}_n(\mathbb{R})$. On rappelle que $\lambda \in \mathbb{C}$ est une valeur propre de A si $det(A - \lambda I_n) = 0$. Un vecteur $x \in \mathbb{C}^n$ non nul est dit vecteur propre de A associé à la valeur propre λ si $Ax = \lambda x$.

Un résultat connu concernant la réduction de matrices symétriques est le suivant :

Proposition 1.4.1 *Soit $A \in \mathcal{M}_n(R)$ une matrice symétrique. Alors, il existe une matrice orthogonale $O \in \mathcal{M}_n(\mathbb{R})$ telle que la matrice $D = O^T AO$ soit diagonale réelle.*

Remarques 1.4.1 *Si $A \in \mathcal{M}_n(\mathbb{R})$ symétrique et O une matrice orthogonale telle que*

$$O^T AO = D = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix},$$

alors $\lambda_1, \lambda_2, \ldots, \lambda_n$ sont les valeurs propres de A. De plus si $f_1, f_2, \ldots, f_n$ sont les vecteurs colonnes de O, alors f_i est un vecteur propre de A associé à la valeur λ_i pour $i = 1, \ldots, n$ et la famille $(f_1, \ldots, f_n)$ forme une base orthonormée des vecteurs propres de A.

Conséquences 1.4.1

1. *Toute matrice symétrique est diagonalisable.*

2. *Les valeurs propres d'une matrice symétrique sont toutes réelles.*

Définition 1.4.4 *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique. La matrice A est dite semi-définie positive si*

$$(Ax, x) \geq 0 \quad \textit{pour tout } x \in \mathbb{R}^n.$$

Si de plus A vérifie, $(Ax, x) = 0$ si et seulement si $x = 0$, alors A est dite définie positive.

Proposition 1.4.2 *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique.*

1. *La matrice A est semi-définie positive si et seulement si toutes ses valeurs propres sont positives.*
2. *La matrice A est définie positive si et seulement toutes ses valeurs propres sont strictement positives.*

Preuve. Soit $A \in \mathcal{M}_n(\mathbb{R})$. On suppose que A est semi-définie positive (resp. définie positive). Soit λ une valeur propre de A et $x \in \mathbb{R}^n$ (non nul) un vecteur propre associé à la valeur propre λ. On a $Ax = \lambda x$, donc

$$(Ax, x) = \lambda(x, x) = \lambda \|x\|_2^2.$$

Comme A est semi-définie positive (resp. définie positive), donc $(Ax, x) \geq 0$ (resp. > 0) et par conséquent

$$\lambda \|x\|_2^2 \geq 0 \text{ (resp. } > 0)$$

ou $\lambda \geq 0$ (resp. > 0).

Supposons que toutes les valeurs propres de A sont positives (resp. strictement positives) et montrons que A est semi-définie positive (resp. strictement positives). On sait qu'il existe une base orthonormée $(f_1, \ldots, f_n)$ de vecteurs propres de A. Soit $x \in \mathbb{R}^n$, (resp. $x \in \mathbb{R}^n \setminus \{0\}$,

$$x = \sum_{i=1}^{n} x_i f_i.$$

Donc

$$Ax = \sum_{i=1}^{n} x_i \lambda_i f_i,$$

et

$$(Ax, x) = \sum_{i=1}^{n} \lambda_i x_i^2 \geq 0 \text{ (resp. } > 0). \tag{1.5}$$

□

Remarque 1.4.2 *De* (1.5) *on tire facilement que si A est une matrice symétrique d'ordre n et si λ_1, λ_n sont respectivement la plus petite et la plus grande valeur propre de A, alors pour tout $x \in \mathbb{R}^n$, on a*

$$\lambda_1 \|x\|_2^2 \leq (Ax, x) \leq \lambda_n \|x\|_2^2. \tag{1.6}$$

L'inégalité (1.6) sera fort utile pour la suite.

1.4.2 Factorisation des matrices symétriques

Proposition 1.4.3 *Soit $A = (a_{ij})_{1\leq i,j\leq n} \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique définie positive. Alors la matrice A admet une factorisation LU avec les éléments diagonaux de U sont strictement positifs.*

Preuve. Soit pour $k = 1, \ldots, n$, $A^{(k)} = (a_{ij})_{1\leq i,j\leq k} \in \mathcal{M}_k(\mathbb{R})$. Montrons que la matrice symétrique $A^{(k)}$ est inversible. En effet, pour tout $x = (x_1, \ldots, x_k)^T \in \mathbb{R}^k \backslash \{0\}$, on pose $\tilde{x} = (x_1, \ldots, x_k, 0, \ldots, 0)^T \in \mathbb{R}^n$. Alors $\tilde{x} \neq 0$ et on a $(A^{(k)}x, x) = (A\tilde{x}, \tilde{x}) > 0$. La matrice $A^{(k)}$ est par conséquent définie positive, donc inversible. La matrice A admet donc une factorisation LU.

Montrons que $u_{ii} > 0$, $i = 1, \ldots, n$. Soit la matrice diagonale

$$D = \begin{pmatrix} u_{11} & & \\ & \ddots & \\ & & u_{nn} \end{pmatrix}.$$

Puisque

$$\det A = \det L \det U = \prod_{i=1}^{n} u_{ii} \neq 0,$$

donc la matrice diagonale D est inversible. Soit

$$S = D^{-1}U = \left(\frac{u_{ij}}{u_{ii}}\right)_{1\leq i,j\leq n},$$

alors $U = DS$ et

$$A = LDS = S^T DL^T.$$

La matrice S^T est triangulaire inférieure à diagonale unité ($(S^T)_{ii} = 1, i = 1, \ldots, n$) et la matrice DL^T est une matrice triangulaire supérieure. D'après l'unicité de factorisation LU, on a $L = S^T$ et $U = DL^T$. Ainsi $A = LDL^T$.

Soit $x \in \mathbb{R}^n \setminus \{0\}$ et y tel que $x = L^T y$, alors y est aussi non nul et on a

$$(Dx, x) = (DL^T y, L^T y) = (LDL^T y, y) = (Ay, y) > 0,$$

donc la matrice diagonale symétrique D est définie positive, ses valeurs propres $u_{ii}, i = 1, \ldots, n$ sont donc strictement positives. □

Corollaire 1.4.1 *:* **Factorisation de Crout**

De la preuve de la proposition précédente on déduit que si A est une matrice symétrique inversible et admettant une factorisation LU, alors A admet la factorisation $A = LDL^T$ où D est une matrice diagonale. La factorisation LDL^T est appelée factorisation de **Crout**.

1.4.3 Factorisation de Cholesky

Proposition 1.4.4 *Soit A une matrice symétrique définie positive. Alors il existe une matrice triangulaire inférieure B à* **éléments diagonaux strictement positifs** *telle que $A = BB^T$. De plus cette factroisation est unique. Une telle factorisation est dite factorisation de* **Cholesky** *de A.*

Preuve. On sait que A admet une factoriasion $LDL^T = LU$ avec $(L)_{ii} = 1$ et $(D)_{ii} = u_{ii} > 0$ pour $i = 1, \ldots, n$. Si on pose

$$D' = \begin{pmatrix} \sqrt{u}_{11} & & \\ & \ddots & \\ & & \sqrt{u}_{nn} \end{pmatrix},$$

alors

$$A = (LD')(D'L^T) = BB^T.$$

L'unicité découle de celle de la factorisation LU. □

Remarque 1.4.3 *La matrice B de la factorisation de* **Cholesky** *de la matrice symétrique $A = LU$ vérifie $B = L\,\mathrm{diag}(\sqrt{u_{ii}})$ où*

$$\mathrm{diag}(\sqrt{u_{ii}}) = \begin{pmatrix} \sqrt{u}_{11} & & \\ & \ddots & \\ & & \sqrt{u}_{nn} \end{pmatrix}.$$

Algorithme de la factorisation de Cholesky

Il suffit d'identifier le produit A et BB^T pour $B = (b_{ij})$ une matrice triangulaire inférieure. On en déduit

$$a_{ij} = \sum_{k=1}^{\min(i,j)} b_{ik} b_{jk}, \quad \forall\ 1 \leq i, j \leq n.$$

On commence par calculer terme à terme la première colonne de la matrice triangulaire B, puis on passe à la deuxième colonne et ensuite connaissant les $(j-1)$ premières colonnes de B, on peut déduire, d'abord le terme de la diagonale b_{jj}, puis les termes de la j-ème colonne se trouvant en dessous de la diagonale jusqu'on on obtient toute la matrice triangulaire inférieure B. Ceci se traduit aussi par l'algorithme suivant :

Pour $j = 1, \ldots, n$,

$$b_{jj} := \sqrt{a_{jj} - \sum_{k=1}^{j-1} b_{jk}^2}.$$

  Pour $i = j+1, \ldots, n$,

$$b_{ij} := \left(a_{ij} - \sum_{k=1}^{j-1} b_{ik} b_{jk} \right) / b_{jj}.$$

  Fin de la boucle sur i.

Fin de la boucle sur j.

Remarque 1.4.4 *Clairement, le coût de la méthode de Cholesky est de l'ordre de* $\frac{n^3}{3}$ *qui est la moitié de celui de la factorisation* LU*, puisque ici le calcul de la matrice* B *donne immédiatement sa matrice transposée* B^T*.*

Chapitre 2

Méthodes itératives pour la résolution des systèmes linéaires

2.1 Introduction

Pour des systèmes linéaires de grande taille, les méthodes directes de factorisations (de type LU ou de Cholesky) deviennent couteuses en temps de calcul ou en place mémoire. L'idée alors de ne plus chercher à résoudre exactement le système linéaire $Ax = b$, mais d'approcher sa solution x par une suite de vecteurs $(x^{(k)})$ vérifiant

$$\lim_{k\to+\infty} \|x^{(k)} - x\| = 0.$$

Dans ce chapitre, la suite $(x^{(k)})$ est construite à l'aide d'une relation de récurrence simple de type $x^{(k+1)} = F(x^{(k)})$, pour une application affine $F(x) = Bx + c$ où B dépend de A et c dépend de A et b.

2.2 Rappels sur les normes matricielles

On rappelle que l'espace vectoriel complexe $\mathbb{C}^n$ est muni du produit hermitien défini pour $x = (x_i)_{1\leq i\leq n}$ et $y = (y_i)_{1\leq i\leq n} \in \mathbb{C}^n$, par

$$(x, y) = \sum_{i=1}^{n} x_i \bar{y}_i,$$

et la norme euclidienne associé à ce produit scalaire est

$$\|x\|_2 = (\sum_{i=1}^{n} |x_i|^2)^{\frac{1}{2}}.$$

Dans ce chapitre, on note par $\mathcal{M}_n(\mathbb{K})$ l'espace vectoriel des matrices carrées d'ordre n à coefficients dans $\mathbb{K}$ où $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$. On rappelle que pour $x = (x_i)_{1\leq i\leq n} \in \mathbb{K}^n$ on a les normes vectorielles suivantes :

- $\|x\|_p = (\sum_{i=1}^{n} |x_i|^p)^{\frac{1}{p}}$, pour $p \in [1, +\infty[$; norme de Hölder d'indice p.
- $\|x\|_\infty = \max_{1\leq i\leq n} |x_i|$, norme infinie.

2.2.1 Normes matricielles subordonnées

Définition 2.2.1 *On appelle norme matricielle sur $\mathbb{K}^n$ toute application $\|\cdot\|$ définie sur $\mathcal{M}_n(\mathbb{K})$ et à valeurs dans $\mathbb{R}_+$ vérifiant pour tout $A, B \in \mathcal{M}_n(\mathbb{K})$ et pour tout $\alpha \in \mathbb{K}$:*
- *$\|A\| = 0 \Leftrightarrow A = O_n$, où O_n est la matrice nulle d'ordre n,*
- *$\|\alpha A\| = |\alpha| \, \|A\|$,*
- *$\|A + B\| \leq \|A\| + \|B\|$,*
- *$\|AB\| \leq \|A\| \, \|B\|$.*

Définition 2.2.2 *:***Normes matricielles subordonnées**

Toute norme vectorielle $\|\cdot\|$ de $\mathbb{K}^n$ définit une norme matricielle de la façon suivante

$$\forall A \in \mathcal{M}_n(\mathbb{K}),\ \|A\| = \sup_{x\in\mathbb{K}^n,\ x\neq 0} \frac{\|Ax\|}{\|x\|} = \sup_{\|x\|\leq 1} \|Ax\| = \sup_{\|x\|=1} \|Ax\|,$$

dite norme matricielle subordonnée ou induite (à cette norme vectorielle).

Toute norme matricielle subordonnée vérifie :

1. $\|Ax\| \leq \|A\|\|x\|$, pour toute matrice $A \in \mathcal{M}_n(\mathbb{K})$ et pour tout vecteur $x \in \mathbb{K}^n$.
2. $\|I_n\| = 1$.

On notera $\|\cdot\|_p$ la norme matricielle subordonnée assoiée à la norme vectorielle d'indice p.

Exemple 2.2.1 *La norme (dite de* **Frobenius***) définie, pour $A = (a_{ij})_{1\leq i,j\leq n} \in \mathcal{M}_n(\mathbb{K})$, par*

$$\|A\| = (\sum_{i,j=1}^{n} |a_{ij}|^2)^{\frac{1}{2}},$$

n'est pas subordonnée puisque $\|I_n\| = \sqrt{n} \neq 1$.

Définition 2.2.3 *:***Rayon spectral**

Soient $A \in \mathcal{M}_n(\mathbb{K})$ et $\lambda_i \in \mathbb{C}$, $1 \leq i \leq n$ les valeurs propres de A. On rappelle que le spectre de A, qu'on note $Sp(A)$, est l'ensemble des valeurs propres de A.

On appelle rayon spectral de A, le réel positif, noté $\rho(A) = \max_{1\leq i\leq n} |\lambda_i|$ qui est le maximum des modules des valeurs propres de A.

On a alors :

Proposition 2.2.1 *Les normes matricielles subordonnées aux normes vectorielles $\|\cdot\|_1$, $\|\cdot\|_2$ et $\|\cdot\|_\infty$ sont données, pour une matrice $A = (a_{ij})_{1\leq i,j\leq n}$, par :*

- $\|A\|_2 = \begin{cases} \sqrt{\rho(A^TA)} = \sqrt{\rho(AA^T)} = \|A^T\|_2 & \textit{si } A \textit{ est une matrice réelle,} \\ \sqrt{\rho(A^*A)} = \sqrt{\rho(AA^*)} = \|A^*\|_2 & \textit{si } A \textit{ est une matrice complexe,} \end{cases}$
  où pour $A = (a_{ij})_{1\leq i,j\leq n} \in \mathcal{M}_n(\mathbb{C})$, $A^* = \bar{A}^T = (\bar{a}_{ji})_{1\leq i,j\leq n}$ *la matrice adjointe de* A.
- $\|A\|_1 = \max\limits_{1\leq j\leq n} \sum\limits_{i=1}^{n} |a_{ij}|$,
- $\|A\|_\infty = \max\limits_{1\leq i\leq n} \sum\limits_{j=1}^{n} |a_{ij}|$.

Preuve.

1. Soit $A = (a_{ij})_{1\leq i,j\leq n}$ une matrice non nulle de $\mathcal{M}_n(\mathbb{R})$.

$$\|A\|_2^2 = \sup_{x\in\mathbb{R}^n, x\neq 0} \frac{\|Ax\|_2^2}{\|x\|_2^2} = \sup_{x\in\mathbb{R}^n, x\neq 0} \frac{(Ax, Ax)}{\|x\|_2^2} = \sup_{x\in\mathbb{R}^n, x\neq 0} \frac{(A^TAx, x)}{\|x\|_2^2}.$$

   La matrice $B = A^TA$ étant symétrique semi définie positive puisque $B^T = B$ et $(Bx, x) = (Ax, Ax) = \|Ax\|_2^2 \geq 0$, pour tout $x \in \mathbb{R}^n$. D'après (1.6), on a

$$\|A\|_2^2 \leq \rho(A^TA).$$

   D'autre part, si x est un vecteur propre de $B = A^TA$ associé à la valeur propre $\rho(A^TA)$, alors

$$\|Ax\|_2^2 = (Bx, x) = \rho(A^TA)\|x\|^2,$$

   donc

$$\frac{\|Ax\|_2}{\|x\|_2} = \sqrt{\rho(A^TA)} \leq \|A\|_2,$$

   et finalement

$$\|A\|_2 = \sqrt{\rho(A^TA)}.$$

   Le cas d'une matrice complexe se fait de la même manière quitte à remplacer $\mathbb{R}$ par $\mathbb{C}$ et A^T par A^*.
2. A faire en exercice.
3. Soit $x = (x_i)_{1\leq i\leq n} \in \mathbb{C}^n$. On a alors,

$$\|Ax\|_\infty = \max_{1\leq i\leq n} |(Ax)_i| = \max_{1\leq i\leq n} |\sum_{j=1}^{n} a_{ij}x_j| \leq \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|\ |x_j| \leq (\max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|)\|x\|_\infty,$$

   et par conséquent $\|A\|_\infty \leq \max\limits_{1\leq i\leq n} \sum\limits_{j=1}^{n} |a_{ij}|$. D'autre part, si i_0 est tel que

$$\sum_{j=1}^{n} |a_{i_0j}| = \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|.$$

Comme $A \neq 0$, alors sa i_0-ème colonne est non nulle et il existe au moins j tel que $a_{i_0 j} \neq 0$. Posons, pour $j = 1, \ldots, n$,

$$x_j = \begin{cases} \dfrac{\bar{a}_{i_0 j}}{|a_{i_0 j}|} & \text{si } a_{i_0 j} \neq 0 \\ 0 & \text{sinon} \end{cases}.$$

Alors $\|x\|_\infty = 1$ et

$$|(Ax)_{i_0}| = |\sum_{j=1}^{n} a_{i_0 j} \frac{\bar{a}_{i_0 j}}{|a_{i_0 j}|}| = \sum_{j=1}^{n} |a_{i_0 j}| = \max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}|.$$

Par suite $\|A\|_\infty = \max\limits_{1 \leq i \leq n} \sum\limits_{j=1}^{n} |a_{ij}|$.

□

Remarque 2.2.1 *Si $A \in \mathcal{M}_n(\mathbb{R})$ est une matrice symétrique, alors*

$$\|A\|_2 = \rho(A). \tag{2.1}$$

Car les valeurs propres de A^2 sont les carrées des valeurs propres de A lorsque cette dernière est diagonalisable et donc

$$\rho(A^2) = \rho(A)^2.$$

La relation (2.1) devient une inégalité pour une matrice quelconque et pour les autres normes. Plus précisément :

Proposition 2.2.2 *On a :*

1. *Pour toute matrice A et pour toute norme matricielle $\|\cdot\|$, subordonnée ou non, on a :*

$$\rho(A) \leq \|A\|.$$

2. *Pour tout $\epsilon > 0$ et pour toute matrice $A \in \mathcal{M}_n(\mathbb{C})$, il existe une norme matricielle subordonnée notée $\|\cdot\|_{A,\epsilon}$ telle que*

$$\|A\|_{A,\epsilon} \leq \rho(A) + \epsilon.$$

Preuve.

1. Soient $A \in \mathcal{M}_n(\mathbb{C})$, λ une valeur propore de A telle que $|\lambda| = \rho(A)$ et $x \in \mathbb{C}^n, x \neq 0$ un vecteur propore associé à la valeur propre λ. Alors $Ax = \lambda x$.

   La matrice de taille n définie par $xx^* = (x_i \bar{x}_j)_{1 \leq i,j \leq n}$ est une matrice non nulle puisque x est un vecteur non nul. Alors $Axx^* = \lambda xx^*$ et par suite

$$\|Axx^*\| = \rho(A)\|xx^*\| \leq \|A\|\|xx^*\|.$$

   Ce qui donne $\rho(A) \leq \|A\|$.

2. Soit $A \in \mathcal{M}_n(\mathbb{C})$, on sait que A est semblable à une matrice triangulaire supérieure complexe, il existe donc $(f_1, \ldots, f_n)$ une base de $\mathbb{C}^n$ et une matrice triangulaire supérieure $T = (\lambda_{ij})_{1 \leq i,j \leq n}$ telles que
$$Af_i = \sum_{j \geq i} \lambda_{ij} f_j,$$
avec, si $P = (f_1, \ldots, f_n)$, alors $P^{-1}AP = T$.
Pour un réel $\eta \in]0, 1[$, on définit $B = (\tilde{e}_1, \tilde{e}_2, \ldots, \tilde{e}_n)$ la base de l'espace vectoriel $\mathbb{C}^n$ par,
$$\tilde{e}_1 = f_1, \tilde{e}_2 = \eta f_2, ..., \tilde{e}_n = \eta^{n-1} f_n.$$
Si $x = \sum_{j=1}^{n} \alpha_i \tilde{e}_i \in \mathbb{C}^n$, on définit la norme vectorielle de x par
$$\|x\| := \max_{1 \leq j \leq n} |\alpha_j|.$$
Il est clair que cette norme dépend de A et de η.
Soit $\epsilon > 0$, montrons que si η est bien choisi, la norme subordonnée associée vérifie
$$\|A\| \leq \rho(A) + \epsilon.$$
Pour tout $i \in \{1, \ldots, n\}$,
$$A\tilde{e}_i = \eta^{i-1} A f_i = \eta^{i-1} \sum_{1 \leq i \leq j} \lambda_{ij} f_j = \sum_{1 \leq i \leq j} \eta^{i-1} \eta^{1-j} \lambda_{ij} \tilde{e}_j.$$
Donc
$$A\tilde{e}_i = \sum_{1 \leq i \leq j} \eta^{i-j} \lambda_{ij} \tilde{e}_j.$$
Si $x = \sum_{i=1}^{n} \alpha_i \tilde{e}_i$, alors
$$Ax = \sum_{i=1}^{n} \alpha_i \sum_{1 \leq i \leq j} \eta^{i-j} \lambda_{ij} \tilde{e}_j = \sum_{j=1}^{n} (\sum_{i=j}^{n} \eta^{i-j} \lambda_{ij} \alpha_i) \tilde{e}_j.$$
$$\|Ax\| = \max_{1 \leq j \leq n} |\sum_{i=j}^{n} \eta^{i-j} \lambda_{ij} \alpha_i| \leq \max_{1 \leq j \leq n} \left[|\lambda_{jj}\|\alpha_j| + \eta \sum_{i=1}^{n} |\lambda_{ij}\|\alpha_i| \right]$$
$$\leq \rho(A)\|x\| + \eta \max_{1 \leq j \leq n} \sum_{i=1}^{n} |\lambda_{ij}|\|x\|.$$
Par conséquent
$$\frac{\|Ax\|}{\|x\|} \leq \rho(A) + \eta \|T_A\|_1,$$

donc

$$\|A\| \leq \rho(A) + \eta\|T_A\|_1.$$

Pour $\eta \in]0,1[$ tel que $\eta\|T_A\|_1 \leq \epsilon$, on a bien

$$\|\cdot\|_{A,\varepsilon} = \|A\| \leq \rho(A) + \epsilon.$$

□

On utilse ce dernier résultat pour montrer la proposition suivante :

Proposition 2.2.3 *Soit B une matrice carrée de $\mathcal{M}_n(\mathbb{C})$. Les conditions suivantes sont équivalentes :*

i) $\lim_{k\to+\infty} B^k x = 0$ *pour tout vecteur x de $\mathbb{C}^n$,*

ii) $\rho(B) < 1$*,*

iii) $\|B\| < 1$ *pour au moins une norme matricielle subordonnée.*

Preuve.

i) ⇒ ii) Soit λ une valeur propre de B telle que $\rho(B) = |\lambda|$ et soit x un vecteur propre associé à la valeur propre λ de B. On a $Bx = \lambda x$, donc $B^k x = \lambda^k x$ et par conséquent $\|B^k x\| = \rho(B)^k\|x\| \to 0$ quand $k \to +\infty$ si et seulement si $\rho(B) < 1$.

ii) ⇒ iii) Pour $\epsilon = \dfrac{1-\rho(B)}{2}$, il existe une norme matricielle subordonnée $\|\cdot\|$ telle que

$$\|B\| \leq \rho(B) + \epsilon = \rho(B) + \frac{1-\rho(B)}{2} = \frac{\rho(B)+1}{2} < 1.$$

iii) ⇒ i) On a $\|B^k x\| \leq \|B\|^k\|x\| \to 0$ quand $k \to +\infty$.

□

2.3 Méthodes itératives

Pour résoudre un système linéaire $Ax = b$ pour $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible et $b \in \mathbb{R}^n$, on utilise dans ce chapitre des méthodes itératives dont le principe est d'écrire A comme la différence de deux matrices $A = M - N$, où M est une matrice inversible (M est en général diagonale, triangulaire ou facile à inverser). Le système $Ax = b$ est équivalent à $x = M^{-1}(Nx + b)$. On approche donc la solution du système $Ax = b$ par la suite $(x^{(k)})$ définie par

$$x^{(k+1)} = M^{-1}(Nx^{(k)} + b), \quad k > 0,$$

en partant d'un $x^{(0)}$ donné.

Remarques 2.3.1

1. *On n'a pas toujours besoin de calculer M^{-1}, mais il faut savoir calculer la solution de $Mx^{(k+1)} = Nx^{(k)} + b$, pour $x^{(0)}$ donnée.*

2. *Il est clair que si la suite $(x^{(k)})$ converge, elle converge vers la solution unique x de $Ax = b$. On dit dans ce cas que la méthode itérative correspondante est* **convergente** *pour la résolution de système $Ax = b$.*
3. *Si on considère $e^{(k)}$ l'erreur à l'étape k, $e^{(k)} = x^{(k)} - x$, alors $Mx^{(k+1)} = b + Nx^{(k)}$ et $Mx = b + Nx$ et par conséquent $e^{(k+1)} = M^{-1}Ne^{(k)} = \cdots = (M^{-1}N)^k e^{(0)}$. La matrice $M^{-1}N = B$ est appelée* **matrice d'itération** *de la méthode. La suite $(x^{(k)})$ vérifie donc $x^{(k+1)} = Bx^{(k)} + c$, pour $c = M^{-1}b$.*

En général on a :

Proposition 2.3.1 *Si $B = M^{-1}N$ et $c = M^{-1}b$, alors la suite $(x^{(k)})$ donnée par*

$$x^{(k+1)} = Bx^{(k)} + c, \text{ pour } x^{(0)} \text{ donné}, \tag{2.2}$$

converge vers x la solution unique de $x = Bx + c$, pour tout choix de $x^{(0)}$, si et seulement si la matrice d'itération B vérifie $\rho(B) < 1$.

Preuve. La suite donnée par $x^{(0)}$ et $x^{(k+1)} = Bx^{(k)} + c$ converge vers x pour tout choix de $x^{(0)}$ si et seulement si $(B^k e^{(0)}) \to 0$ quand $k \to +\infty$ pour tout $e^{(0)} = x^{(0)} - x$. Ceci étant vrai si et seulement si $\rho(B) < 1$. □

Remarque 2.3.1 *Ainsi, pour montrer que la suite générée par la méthode itérative (2.2) est convergente, il suffit de vérifier que le rayon spectral de sa matrice d'itération B est plus petit que 1 ou encore que pour au moins une norme quelconque $\|\cdot\|$, (subordonnée ou non) $\|B\| < 1$.*

Un premier résultat de convergence d'une méthode itérative concerne les matrices symétriques définies positives.

Proposition 2.3.2 *:***Cas d'une matrice symétrique définie positive**
Soit A une matrice symétrique définie positive et M, N deux matrices telles que $A = M - N$, avec M inversible. Si la matrice symétrique $M^T + N$ est définie positive, alors $\rho(M^{-1}N) < 1$ et la suite définie par $x^{(k+1)} = M^{-1}Nx^{(k)} + M^{-1}b$ converge vers x solution de $Ax = b$ pour tout choix de $x^{(0)}$.

Preuve. Si A est symétrique, alors la matrice $M^T + N$ est toujours symétrique. En effet

$$(M^T + N)^T = M + N^T = (A + N) + N^T = A^T + N^T + N = (A + N)^T + N = M^T + N.$$

Montrons que si A est définie positive et $M^T + N$ est définie positive, alors

$$\rho(M^{-1}N) < 1.$$

Soit $\lambda \in \mathrm{Sp}(M^{-1}N)$ et soit $x \in \mathbb{C}^n$, $x = y + iz \neq 0$, $y, z \in \mathbb{R}^n$ tel que $M^{-1}Nx = \lambda x$. On a $M^{-1}Nx = \lambda x$, donc

$$Nx = \lambda Mx. \tag{2.3}$$

Par suite

$$(Nx, x) = \lambda(Mx, x).$$

Comme $A = M - N$, donc $A = M(I - M^{-1}N)$ et par conséquent

$$Ax = M(x - M^{-1}Nx) = M(x - \lambda x) = (1 - \lambda)Mx.$$

D'où

$$(Ax, x) = (Ay, y) + (Az, z) = (1 - \lambda)(Mx, x) > 0,$$

et donc $\lambda \neq 1$. De plus on a

$$(Nx, x) = \lambda(Mx, x) \text{ et } (M^T y, y) = (y, My) = (My, y)$$

impliquant

$$\begin{aligned}((M^T + N)x, x) &= ((M^T + N)y, y) + ((M^T + N)z, z) = (1 + \lambda)(Mx, x) \\ &= \frac{1 + \lambda}{1 - \lambda}(Ax, x) = \frac{1 + \lambda}{1 - \lambda}[(Ay, y) + (Az, z)].\end{aligned}$$

Comme les matrices A et $M^T + N$ sont définies positives, alors nécessairement le nombre complexe λ est tel que $\dfrac{1 - \lambda}{1 + \lambda} > 0$. Ceci est équivalent à $\lambda \in \mathbb{R}$ et $-1 < \lambda < 1$. □

Dans les trois méthodes itératives classiques qu'on va étudier, on utilise la décomposition suivante de $A = (a_{ij})_{1 \leq i,j \leq n} \in \mathcal{M}_n(\mathbb{R})$:

$$A = D - E - F,$$

avec

– La matrice diagonale

$$D = \begin{pmatrix} a_{11} & & \bigcirc \\ & \ddots & \\ \bigcirc & & a_{nn} \end{pmatrix}.$$

– La matrice triangulaire inférieure $E = \begin{pmatrix} 0 & 0 & \dots & \dots & 0 \\ -a_{21} & 0 & \dots & \cdots & 0 \\ -a_{31} & -a_{32} & 0 & & \vdots \\ \vdots & & \ddots & \ddots & \\ -a_{n1} & -a_{n2} & \dots & -a_{nn-1} & 0 \end{pmatrix}$ qui représente l'opposé de la partie en dessous de la diagonale.

– La matrice triangulaire supérieure $F = \begin{pmatrix} 0 & -a_{12} & -a_{13} \cdots & & -a_{1n} \\ \vdots & 0 & -a_{23} & \dots & -a_{2n} \\ & & \ddots & \ddots & \vdots \\ 0 & & & 0 & -a_{n-1n} \\ 0 & \dots & \dots & \dots & 0 \end{pmatrix}$ qui représente l'opposé de la partie en dessus de la diagonale.

2.3.1 Méthode de *Jacobi*

Dans la méthode de *Jacobi* on choisit $M = D$ et donc $N = E + F = D - A$ où on suppose que la matrice diagonale D est inversible ($a_{ii} \neq 0$ pour tout $i = 1, \ldots, n$). La matrice d'itération est $J = D^{-1}(E + F)$ et la suite $(x^{(k)})$ dans la méthode de *Jacobi* est donnée par :

1. Initialisation : $x^{(0)} \in \mathbb{R}^n$ donnée.
2. Pour $k \geq 0$, on calcule $x^{(k+1)}$ solution de $Dx^{(k+1)} = (E + F)x^{(k)} + b$

Les composantes de $x^{(k+1)}$ sont données en fonction de celles de $x^{(k)}$ par

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left[b_i - \sum_{i \neq j} a_{ij} x_j^{(k)} \right], \ i = 1, \ldots, n.$$

Il est bien de noter que les composantes $x_i^{(k+1)}$ de $x^{(k+1)}$ sont calculées les uns indépendamment des autres et dépendent seulement de l'itéré précédent $x^{(k)}$.

Exemple 2.3.1 *Etudier la convergence de la méthode de Jacobi pour résoudre le système $Ax = b$, pour*

$$A = \begin{pmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{pmatrix} \text{ et } b \in \mathbb{R}^3.$$

Dans ce cas

$$D = I_3, \ E = \begin{pmatrix} 0 & 0 & 0 \\ -1 & 0 & 0 \\ -2 & -2 & 0 \end{pmatrix} \text{ et } F = \begin{pmatrix} 0 & -2 & 2 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix}.$$

La matrice d'itération est $J = D^{-1}(E + F) = D - A = \begin{pmatrix} 0 & -2 & 2 \\ -1 & 0 & -1 \\ -2 & -2 & 0 \end{pmatrix}$. *Le polynôme caractéristique de J est*

$$P_J(\lambda) = \det(J - \lambda I_3) = -\lambda^3.$$

Donc $\rho(J) = 0 < 1$ et par conséquent la méthode de Jacobi est convergente pour résoudre le système $Ax = b$. Par exemple, pour $b = \begin{pmatrix} 1 \\ 3 \\ 5 \end{pmatrix}$, la suite $(x^{(k)}) \in \mathbb{R}^3$ construite par la méthode de Jacobi converge vers le vecteur $\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$ solution unique système $Ax = b$ pour tout choix de $x^{(0)} \in \mathbb{R}^3$.

On montrera la convergence de la méthode de Jacobi pour les matrices à diagonale strictement dominante :

Définition 2.3.1 *Une matrice $A = (a_{ij})_{1\leq i,j\leq n} \in \mathcal{M}_n(\mathbb{C})$ est dite à diagonale strictement dominante si elle vérifie*

$$|a_{ii}| > \sum_{j\neq i} |a_{ij}| \quad \forall\, i \in \{1,\ldots,n\}.$$

Proposition 2.3.3 *Soit A une matrice à diagonale strictement dominante. Alors A est inversible et la méthode de Jacobi converge pour résoudre le système $Ax = b$, pour tout $b \in \mathbb{R}^n$.*

Preuve. Montrons que A est inversible. Soit $x = (x_i)_{1\leq i\leq n} \in \mathbb{C}^n$ tel que $Ax = 0$. Alors, pour tout $i = 1,\ldots,n$,

$$\sum_{j=1}^{n} a_{ij}x_j = 0.$$

Ou encore

$$a_{ii}x_i = \sum_{j\neq i}^{n} a_{ij}x_j.$$

Par conséquent

$$|a_{ii}\|x_i| \leq \sum_{j\neq i}^{n} |a_{ij}\|x_j|.$$

Si $x \neq 0$, alors $\|x\|_\infty \neq 0$ et il existe $i_0 \in \{1,\ldots,n\}$ tel que

$$|x_{i_0}| = \max_{1\leq i\leq n} |x_i| = \|x\|_\infty.$$

Ainsi

$$|a_{i_0i_0}| \leq \sum_{j\neq i_0} |a_{i_0j}| \frac{|x_j|}{|x_{i_0}|} \leq \sum_{j\neq i_0} |a_{i_0j}| < 1.$$

Ce qui est une contradiction, donc $x = 0$ et A ne peut être que inversible.

Pour la méthode de *Jacobi* on a

$$J = \begin{pmatrix} 0 & -\frac{a_{12}}{a_{11}} & \cdots & -\frac{a_{1n}}{a_{11}} \\ -\frac{a_{21}}{a_2} & 0 & \cdots & -\frac{a_{2n}}{a_{22}} \\ \vdots & & \ddots & \\ -\frac{a_{n1}}{a_{nn}} & \cdots & -\frac{a_{nn-1}}{a_{nn}} & 0 \end{pmatrix}$$

qui vérifie

$$\|J\|_\infty = \max_{1\leq i\leq n} \sum_{j\neq i} \frac{|a_{ij}|}{|a_{ii}|} < 1.$$

Comme

$$\rho(J) < \|J\|_\infty < 1,$$

et donc la méthode de Jacobi converge. □

2.3.2 Méthode de relaxation

Dans la méthode de relaxation, on introduit un paramètre réel non nul ω et on prend

$$M = \frac{1}{\omega}D - E,\ N = M - A = \frac{1-\omega}{\omega}D + F.$$

La matrice d'itération est donc

$$\begin{aligned}\mathcal{L}_\omega &= (\tfrac{1}{\omega}D - E)^{-1}(\tfrac{1-\omega}{\omega}D + F)\\ &= \omega(I - \omega D^{-1}E)^{-1}D^{-1}D\tfrac{1}{\omega}((1-\omega)I_n + \omega D^{-1}F)\\ &= (I - \omega D^{-1}E)^{-1}((1-\omega)I_n + \omega D^{-1}F).\end{aligned}$$

- Si $\omega < 1$, on parle de sous relaxation.
- Si $\omega > 1$, on parle de sur relaxation.
- Si $\omega = 1$ la méthode est dite de *Gauss-Seidel*.

Indépendamment de la matrice A, on a le résultat de divergence suivant :

Proposition 2.3.4 *La méthode de relaxation diverge pour tout $\omega \in \mathbb{R}\backslash]0,2[$.*

Preuve. On a

$$\det(\mathcal{L}_\omega) = \det(I_n - \omega D^{-1}E)^{-1}\det((1-\omega)I_n + \omega D^{-1}F).$$

La matrice $(I_n - \omega D^{-1}E)$ est triangulaire inférieure dont les coefficients diagonaux sont égaux à 1, donc $(I_n - \omega D^{-1}E)^{-1}$ est aussi triangulaire inférieure à coefficients diagonaux sont aussi 1. La matrice $(1-\omega)I_n + \omega D^{-1}F$ est triangulaire supérieure ayant $1-\omega$ comme termes diagonaux. Ainsi $\det(\mathcal{L}_\omega) = (1-\omega)^n$. Comme $|1-\omega|^n = |\det(\mathcal{L}_\omega)| \leq \rho(\mathcal{L}_\omega)^n$, donc $|1-\omega| \leq \rho(\mathcal{L}_\omega)$ et si $\omega \in \mathbb{R} \setminus [0,2]$, alors $|1-\omega| > 1$. Par conséquent $\rho(\mathcal{L}_\omega) > 1$ et la méthode de relaxation diverge. □

Ainsi, la méthode de relaxation n'est intéressante que pour $0 < \omega < 2$.

Proposition 2.3.5 *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique définie positive. Alors la méthode de relaxation converge si et seulement $\omega \in]0,2[$.*

Preuve. Il suffit de montrer que $M^T + N$ est définie positive.

$$M^T + N = (\frac{1}{\omega}D - E)^T + \frac{1-\omega}{\omega}D + F = \frac{1}{\omega}D - F + \frac{1-\omega}{\omega}D + F = \frac{2-\omega}{\omega}D.$$

Comme A est définie positive, les coefficients diagonaux de A qui sont les éléments de la diagonale de la matrice D, sont strictement positifs, puisque chaque terme $a_{ii} = (Ae_i, e_i)$, pour $i = 1, \ldots, n$, pour e_i est le $i-$ième vecteur de la base canonique de $\mathbb{R}^n$. La matrice $M^T + N$ est donc définie positive si et seulement si $\dfrac{2-\omega}{\omega} > 0$. Ce qui est équivalent à $\omega \in]0,2[$. □

2.3.3 Méthode de *Gauss-Seidel*

C'est la méthode de relaxation pour $\omega = 1$. On choisit donc pour cette méthode la matrice triangulaire inférieure $M = D - E$ représentant la partie inférieure avec la diagonale de la matrice

A. Donc $N = F$ qui est une matrice triangulaire supérieure à diagonale nulle. On suppose que M est inversible ($a_{ii} \neq 0$ pour tout $i = 1, \ldots, n$). Dans ce cas la matrice d'itération est

$$\mathcal{L}_1 = (D - E)^{-1}F,$$

et la méthode de *Gauss-Seidel* s'écrit :

1. Initialisation : $x^{(0)} \in \mathbb{R}^n$ donnée,
2. Pour $k \geq 0$, on calcule$x^{(k+1)}$ solution de $(D - E)x^{(k+1)} = Fx^{(k)} + b$

La i-ème composante de $x^{(k+1)}$ est donnée en fonction des composantes $x_j^{(k+1)}$, $j < i$ de $x^{(k+1)}$ et des composantes $x_j^{(k)}$, $j > i$ de $x^{(k)}$ par

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left[b_i - \sum_{j<i} a_{ij} x_j^{(k+1)} - \sum_{j>i} a_{ij} x_j^{(k)} \right], \ i = 1, \ldots, n.$$

Contrairement à la méthode de Jacobi, le calcul des composantes $x_i^{(k+1)}$ de $x^{(k+1)}$ doit s'effectuer dans l'ordre $i = 1, \ldots, n$.

Exemple 2.3.2 *Reprenons le système de l'exemple* (2.3.1) *où la matrice* $A = \begin{pmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{pmatrix}$. *La matrice d'itération de la méthode de Gauss-Seidel est*

$$\mathcal{L}_1 = (D - E)^{-1}F = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 2 & 2 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 0 & -2 & 2 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -2 & 2 \\ 0 & 2 & -3 \\ 0 & 0 & 2 \end{pmatrix}.$$

Les valeurs propres de la matrice tridiagonale $\mathcal{L}_1$ *sont donc 0 et 2 et par conséquent* $\rho(\mathcal{L}_1) = 2 > 1$. *La méthode de Gauss-Seidel est divergente pour résoudre le système* $Ax = b$.

On sait déja que si A est une matrice symétrique définie positive, alors la méthode de Gauss-Seidel est convergente ($\omega = 1 \in]0, 2[$). On montre aussi que

Proposition 2.3.6 *Si* $A \in \mathcal{M}_n(\mathbb{R})$ *une matrice à diagonale strictement dominante, alors, la méthode de* Gauss-Seidel *pour la résolution de système* $Ax = b$ *est convergente.*

Preuve. Montrons que si A est à diagonale strictement dominante, alors $\rho(\mathcal{L}_1) < 1$.

Soit $x = (x_i)_{1 \leq i \leq n}$ un vecteur propre de $\mathcal{L}_1$ pour une valeur propre λ de $\mathcal{L}_1 = (D - E)^{-1}F$. Alors $\mathcal{L}_1 x = \lambda x$ et donc $\lambda(D - E)x = Fx$.

Par conséquent

$$\lambda(a_{ii}x_i + \sum_{j<i} a_{ij}x_j) = \sum_{j>i} a_{ij}x_j \ \forall i \in \{1,\dots,n\}.$$

Pour $i \in \{1,\dots,n\}$, si $x_i \neq 0$ on a alors

$$\lambda(a_{ii} + \sum_{j<i} a_{ij}\frac{x_j}{x_i}) = \sum_{j>i} a_{ij}\frac{x_j}{x_i},$$

ou encore

$$|\lambda| \leq \frac{\sum_{j>i} |a_{ij}|\frac{|x_j|}{|x_i|}}{|a_{ii} + \sum_{j<i} a_{ij}\frac{x_j}{x_i}|} \leq \frac{\sum_{j>i} |a_{ij}|\frac{|x_j|}{|x_i|}}{|a_{ii}| - \sum_{j<i} |a_{ij}|\frac{|x_j|}{|x_i|}}.$$

En particulier, si i_0 est tel que $|x_{i_0}| = \max_{1\leq i\leq n} |x_i| = \|x\|_\infty$, alors

$$|\lambda| \leq \frac{\sum_{j>i_0} |a_{i_0 j}|\frac{|x_j|}{|x_{i_0}|}}{|a_{i_0 i_0}| - \sum_{j<i_0} |a_{i_0 j}|\frac{|x_j|}{|x_{i_0}|}} < 1.$$

Par suite $\rho(\mathcal{L}_1) < 1$. □

2.3.4 Vitesse de convergence

En pratique, il faut tenir compte de la vitesse de convergence. Autrement dit, entre deux méthodes itératives convergentes, on choisit celle qui donne la solution plus rapidement que l'autre. Un critère de mesure de cette vitesse est l'évolution de l'erreur $\|x^{(k)} - x\|$ qui dépend du rayon spectral de la matrice d'itération. En effet, si on a une suite convergente $(x^{(k)})$ définie par $x^{(k+1)} = Bx^{(k)} + c$, pour $x^{(0)}$ donné, sa limite est x vérifiant $x = Bx + c$, et on a

$$x^{(k)} - x = B(x^{(k-1)} - x) = B^k(x^{(0)} - x)$$

et par conséquent

$$\|x^{(k)} - x\|_2 = \|B^k(x^{(0)} - x)\|_2 \leq \|B\|_2^k \|x^{(0)} - x\|_2.$$

En particulier, si B est symétrique, alors, $\|B\|_2 = \rho(B)$ et il vient

$$\|x^{(k)} - x\|_2 \leq \rho(B)^k \|x^{(0)} - x\|_2.$$

Ainsi, la suite $(x^{(k)})$ converge plus vite vers x d'autant que $\rho(B)$ est plus petit. Ceci reste encore vrai pour une norme matricielle quelconque et pour une matrice B quelconque d'après le lemme suivant

Lemme 2.3.1 *Si $B \in \mathcal{M}_n(\mathbb{C})$ et $\|.\|$ est une norme matricielle subordonnée, alors*

$$\rho(B) = \lim_{k\to+\infty} \|B^k\|^{\frac{1}{k}}.$$

La démonstration de ce lemme est laissée en exercice. On compare alors la vitesse de convergence de deux méthodes itératives de type (2.2) de la façon suivante :

Définition 2.3.2 *Pour la recherche d'une solution x d'un système $Ax = b$, une méthode itérative de matrice d'itération B est dite plus rapide qu'une autre de matrice d'itération $\tilde{B}$, si $\rho(B) < \rho(\tilde{B})$.*

En général, on ne peut rien dire sur la convergence d'une de deux méthodes de *Jacobi* et de *Gauss-seidel* connaissant la nature de convergence de l'autre. Cependant, pour une matrice tridiagonale on a :

Proposition 2.3.7 *:* **Cas d'une matrice tridiagonale**

Soit A une matrice tridiagonale. Alors les rayons spectraux de Jacobi *et* Gauss-Seidel *sont liés par la relation*

$$\rho(\mathcal{L}_1) = (\rho(J))^2.$$

Donc les deux méthodes convergent ou divergent simultanèment. Lorsqu'elles convergent, la méthode de Gauss-Seidel *est plus rapide.*

Preuve. $A = D - E - F$, avec

$$A = \begin{pmatrix} a_{11} & a_{12} & & \bigcirc \\ a_{21} & \ddots & \ddots & \\ & \ddots & \ddots & a_{n-1n} \\ \bigcirc & & a_{nn-1} & a_{nn} \end{pmatrix}, \; D = \begin{pmatrix} a_{11} & & & \bigcirc \\ & \ddots & & \\ & & \ddots & \\ \bigcirc & & & a_{nn} \end{pmatrix},$$

$$E = \begin{pmatrix} 0 & & & \bigcirc \\ -a_{21} & \ddots & & \\ & \ddots & \ddots & \\ \bigcirc & & -a_{nn-1} & 0 \end{pmatrix} \text{ et } F = \begin{pmatrix} 0 & -a_{12} & & \bigcirc \\ 0 & \ddots & \ddots & \\ & \ddots & \ddots & -a_{n-1n} \\ \bigcirc & & & 0 \end{pmatrix}.$$

Montrons que pour tout $\alpha, \mu \in \mathbb{C}^*$, $\det(\mu D - E - F) = \det(\mu D - \alpha E - \alpha^{-1} F)$. On pose

$$A_\mu := \mu D - E - F = \begin{pmatrix} \mu a_{11} & a_{12} & & \bigcirc \\ a_{21} & \ddots & \ddots & \\ & \ddots & \ddots & a_{n-1n} \\ \bigcirc & & a_{nn-1} & \mu a_{nn} \end{pmatrix}$$

et

$$A_{\alpha,\mu} := \mu D - \alpha E - \alpha^{-1} F = \begin{pmatrix} \mu a_{11} & \alpha^{-1} a_{12} & & \bigcirc \\ \alpha a_{21} & \ddots & \ddots & \\ & \ddots & \ddots & \alpha^{-1} a_{n-1n} \\ \bigcirc & & \alpha a_{nn-1} & \mu a_{nn} \end{pmatrix}.$$

Soit la matrice diagonale

$$P = \begin{pmatrix} \alpha & & & \bigcirc \\ & \alpha^2 & & \\ & & \ddots & \\ \bigcirc & & & \alpha^n \end{pmatrix}.$$

Alors P est inversible et

$$P^{-1}A_{\alpha,\mu} = \begin{pmatrix} \alpha^{-1}\mu a_{11} & \alpha^{-2}a_{12} & & \bigcirc \\ \alpha^{-1}a_{21} & \ddots & \ddots & \\ & \ddots & \ddots & \alpha^{-n}a_{n-1n} \\ \bigcirc & & \alpha^{1-n}a_{nn-1} & \alpha^{-n}\mu a_{nn} \end{pmatrix},$$

$$P^{-1}A_{\alpha,\mu}P = \begin{pmatrix} \mu a_{11} & a_{12} & & \bigcirc \\ a_{21} & \ddots & \ddots & \\ & \ddots & \ddots & a_{n-1n} \\ \bigcirc & & a_{nn-1} & \mu a_{nn} \end{pmatrix} = A_\mu.$$

Donc

$$\det(P^{-1}A_{\alpha,\mu}P) = \det A_\mu.$$

Montrons que si P_1 le polynôme caractéristique de $\mathcal{L}_1$ et P_J le polynôme caractéristique de J, alors

$$P_1(\lambda) = \lambda^{\frac{n}{2}} P_J(\lambda^{\frac{1}{2}}).$$

En effet, sachant que $D-E$ est triangulaire supérieure de diagonale celle de la matrice inversible D, donc elle est aussi inversible et on a, pour $\lambda \in \mathbb{C}^*$,

$$\begin{aligned} P_1(\lambda) &= \det((D-E)^{-1}F - \lambda I_n) \\ &= \det(D-E)^{-1}\det(F - \lambda D - \lambda E) \\ &= \det D^{-1}\det(-\lambda E + F - \lambda D) \\ &= \lambda^{\frac{n}{2}}\det D^{-1}\det(-\lambda^{\frac{1}{2}}D + \lambda^{\frac{1}{2}}E + \lambda^{-\frac{1}{2}}F) \\ &= \lambda^{\frac{n}{2}}\det D^{-1}\det(-\lambda^{-\frac{1}{2}}D + E + F) \\ &= \lambda^{\frac{n}{2}}\det(D^{-1}(E+F) - \lambda^{\frac{1}{2}}I_n) \\ &= \lambda^{\frac{n}{2}}P_J(\lambda^{1/2}). \end{aligned}$$

En conclusion, $\lambda \in \mathrm{Sp}(J)$ si et seulement si $\lambda^2 \in \mathrm{Sp}(\mathcal{L}_1)$. Ainsi $\rho(\mathcal{L}_1) = (\rho(J))^2$. □

2.3.5 Critère ou test d'arrêt

En général, dans les méthodes itératives pour la détermination de $\bar{x}$ solution d'un problème quelconque, on construit une suite $(x^{(k)})$ qui converge vers $\bar{x}$. En pratique, et si on n'exige pas un critère d'arret, le nombre d'itérations pour calculer les termes de cette suite pourrait être infini. Le

calcul donc devrait être interrompu dès qu'on obtient une solution approchée à ε prés, vérifiant par exemple $\|Ax^{(k)} - b\| \leq \varepsilon$ si $\bar{x}$ est aussi solution de $Ax = b$, ou $\|x^{(k+1)} - x^{(k)}\| \leq \varepsilon$ dans le cas général, pour une tolérence ε donnée et une norme vectorielle $\|\cdot\|$ choisie. Il existe aussi d'autres critères d'arrêt comme

$$\frac{\|Ax^{(k)} - b\|}{\|b\|} \leq \varepsilon$$

ou

$$\frac{\|x^{(k)} - x^{(k-1)}\|}{\|x^{(k)}\|} \leq \varepsilon$$

pour les algorithmes de résolution des systèmes linéaires.

2.4 Conditionnement

La notion du conditionnent d'une matrice peut servir à établir des majorations des erreurs d'arrondi dues au erreurs sur les données. En pratique, par exemple pour un système linéaire $Ax = b$, les données A et b ne sont données qu'à une erreur près. Cette perturbation des données peut engendrer une grande erreur de la solution. C'est le cas de l'exemple suivant :

Soit le système linéaire $Ax = b$, pour

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 4 + 10^{-6} \end{pmatrix} \quad \text{et} \quad b = \begin{pmatrix} 3 \\ 6 + 10^{-6} \end{pmatrix}.$$

Ce système admet une solution unique $x = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$.

Si on change légèrement b par $\tilde{b}$, avec $\tilde{b} = \begin{pmatrix} 3 \\ 6 - 10^{-6} \end{pmatrix}$, la solution du système $Ax = \tilde{b}$ est $\tilde{x} = \begin{pmatrix} 5 \\ -1 \end{pmatrix}$. Ainsi une petite perturbation de la donnée b implique une grande modification de la solution.

En général, si x est tel que $Ax = b$ et $x + \delta x$ vérifiant $A(x + \delta x) = b + \delta b$, alors $A\delta x = \delta b$ ou encore $\delta x = A^{-1}b$. Par conséquent

$$\|\delta x\| \leq \|A^{-1}\|\|\delta b\|, \tag{2.4}$$

pour $\|\cdot\|$ une norme matricielle subordonnée.

D'autre part $\|b\| = \|Ax\| \leq \|A\|\|x\|$ et donc

$$\frac{1}{\|x\|} \leq \frac{\|A\|}{\|b\|}. \tag{2.5}$$

De (2.4) et (2.5), on déduit la majoration d'erreur suivante :

$$\frac{\|\delta x\|}{\|x\|} \leq \|A\|\|A^{-1}\|\frac{\|\delta b\|}{\|b\|}. \tag{2.6}$$

On remarque que pour $\|\delta x\|$ soit assez petit pour une petite perturbation de b, il suffit que le terme $\|A\|\|A^{-1}\|$ soit aussi assez petit,

Définition 2.4.1 *On appelle conditionnement d'une matrice inversible A relatif à une norme matricielle subordonnée $\|\cdot\|$, le réel positif, qu'on note* $\mathrm{Cond}(A)$

$$\mathrm{Cond}(A) = \|A\|\|A^{-1}\|.$$

Proposition 2.4.1 *Pour toute matrice inversible $A \in \mathcal{M}_n(\mathbb{K})$ on a :*

1. $\mathrm{Cond}(A) = \mathrm{Cond}(A^{-1})$,
2. $\mathrm{Cond}(\alpha A) = \mathrm{Cond}(A)$, *pour tout* $\alpha \in \mathbb{K}^*$,
3. $\mathrm{Cond}(A) \geq 1$,
4. $\mathrm{Cond}_2(A) = \dfrac{\max\limits_{1\leq i\leq n} |\lambda_i|}{\min\limits_{1\leq i\leq n} |\lambda_i|}$, *si A est une matrice réelle symétrique (ou complexe Hermitienne) inversible de valeurs propres $(\lambda_i)_{1\leq i\leq n}$. Ici* Cond_2 *désigne le conditionnement relatif à la norme* $\|\cdot\|_2$.

En plus de la majoration (2.6), on a en général :

Proposition 2.4.2 *Soient A et δA deux matrices de $\mathcal{M}_n(\mathbb{K})$, avec A inversible et soient b et δb deux vecteurs de $\mathbb{K}^n$ tel que $b \neq 0$.*

1. *Si $Ax = b$ et $A(x + \delta x) = b + \delta b$, alors on a*
$$\frac{\|\delta x\|}{\|x\|} \leq \mathrm{Cond}(A)\frac{\|\delta b\|}{\|b\|}.$$
2. *Si $Ax = b$ et $(A + \delta A)(x + \delta x) = b$, alors*
$$\frac{\|\delta x\|}{\|x + \delta x\|} \leq \mathrm{Cond}(A)\frac{\|\delta A\|}{\|A\|}.$$

Preuve.

1. Déja montré
2. Si $Ax = b$ et $(A + \delta A)(x + \delta x) = b$, alors $\delta A(x + \delta x) = -A\delta x$ et donc $\delta x = -A^{-1}\delta A(x + \delta x)$ et on obtient ainsi
$$\|\delta x\| \leq \|A^{-1}\|\|\delta A\|\|x + \delta x\| = \mathrm{Cond}(A)\frac{\|\delta A\|}{\|A\|}\|x + \delta x\|.$$
Il suffit de multiplier par $\dfrac{1}{\|x + \delta x\|}$ pour conclure le résultat désiré.

□

Préconditionnement d'un système linéaire

Au lieu de résoudre un système $Ax = b$, on multiplie à gauche par une matrice inversible P et résoudre le système $PAx = Pb$, où la matrice P est choisie pour que PA soit bien conditionnée. Ce dernier système s'appelle système préconditionné.

Chapitre 3

Optimisation sans contraintes

On rappelle que tout minimum ou maximum x d'une fonction réelle dérivable $f : \mathbb{R} \to \mathbb{R}$ satisfait $f'(x) = 0$. Ce résultat s'applique aussi pour les fonctions $J : \mathbb{R}^n \to \mathbb{R}$ dérivables.

Ce chapitre vise dans sa première partie à rappeler quelques notions élémentaires concernant les problèmes d'optimisation non contraints sur $\mathbb{R}^n$. Ensuite, et dans la dernière partie de ce chapitre, des méthodes de descente de type gradient seront développées pour calculer numériquement un minimum $\bar{x}$ d'une fonction J sur $\mathbb{R}^n$.

Comme application on considère le problème de minimisation d'une fonction quadratique dont la résolution est équivalente à la résolution d'un système linéaire $Ax = b$, où A est une matrice symétrique définie positive.

3.1 Optimisation sur $\mathbb{R}^n$

Dans de nombreux domaines d'applications d'analyse numérique, on est amené à minimiser une fonction J et chercher $\bar{x}$ vérifiant

$$J(\bar{x}) \leq J(x),\ \forall\ x \in \mathbb{R}^n, \tag{3.1}$$

où la fonction J, dite objectif, est une fonction numérique à variable un vecteur x de $\mathbb{R}^n$. C'est un problème de minimisation qu'on peut noter aussi

$$\min_{x \in \mathbb{R}^n} J(x). \tag{P}$$

Pour maximiser une fonction J il suffit de minimiser $(-J)$, un problème d'optimisation sur $\mathbb{R}^n$ peut donc s'écrire toujours sous la forme de (P).

Tout vecteur $\bar{x}$ solution de (3.1) est appelé solution optimale de (P). On dit aussi que $\bar{x}$ présente (ou tout simplement) un minimum de J sur $\mathbb{R}^n$.

Du théorique au numérique, pour la résolution du problème (P) ou de (3.1) on s'intéresse aux questions suivantes :

1. Existe-il une solution et est elle unique ?
2. Comment caractériser cette solution ?
3. Comment approcher d'une manière efficace cette solution ?

Commençons par développer la première étape concernant l'existence puis l'unicité d'une solution d'un problème d'optimisation sans contraintes.

3.1.1 Existence et unicité d'un minimum

On rappelle tout d'abord qu'une suite de vecteurs $(x^{(k)})$ est dite bornée de $\mathbb{R}^n$ s'il existe $R > 0$ tel que

$$\|x^{(k)}\| \leq R, \text{ pour tout entier } k,$$

i.e., la suite $(x^{(k)})$ est contenue dans une boule de centre 0 et de rayon $R > 0$.

On rappelle aussi que de toute suite bornée de $\mathbb{R}^n$, on peut en extraire une sous suite convergente.

Proposition 3.1.1 *On suppose que J est continue et qu'elle vérifie*

$$\lim_{\|x\| \to +\infty} J(x) = +\infty. \tag{3.2}$$

Alors le problème de minimisation (P) admet au moins une solution.

Preuve. Soit

$$\alpha = \inf_{x \in \mathbb{R}^n} J(x).$$

Montrons que α est atteint, i.e., il existe $\bar{x} \in \mathbb{R}^n$ tel que $J(\bar{x}) = \alpha$.

D'après la propriété de la borne inférieure, il existe une suite $(x^{(k)})_k$ de $\mathbb{R}^n$

$$\lim_{k \to \infty} J(x^{(k)}) = \alpha.$$

Alors la suite $(x^{(k)})$ est bornée. En effet, si la suite $(x^{(k)})$ est non bornée, elle vérifie, pour une sous suite, notée aussi $(x^{(k)})$, $\lim_{k \to +\infty} ||x^{(k)}|| = +\infty$ et comme $\lim_{\|x\| \to +\infty} J(x^{(k)}) = +\infty$, donc $\alpha = +\infty$, ce qui est une contradiction.

Comme $(x^{(k)})$ est bornée, on peut en extraire une sous suite $(x^{(k)})$ convergente vers $\bar{x}$. Comme J est continue, alors

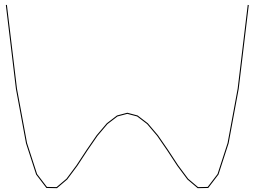
$$J(\overline{x}) = \lim_{k \to +\infty} J(x^{(k)}) = \inf_{x \in \mathbb{R}^n} J(x).$$

D'où l'existence d'au moins d'une solution pour le problème (P). □

Remarques 3.1.1

1. *Si J vérifie* (3.2)*, on dit qu'elle est infinie à l'infinie ou qu'elle est coercive.*
2. *Une fonction J continue et minorée sur $\mathbb{R}^n$ admet une borne inférieure, mais pas forcément de minimum si elle n'est pas infinie à l'infini, voici un contre exemple : La fonction $x \mapsto \exp(x)$ est positive, mais elle n'a pas de minimum sur $\mathbb{R}$.*

Remarque 3.1.1 *Ce résultat donne l'existence d'un minimum mais il ne donne pas l'unicité. La fonction $x \mapsto x^4 - 2x^2$ admet deux minima atteints en $x = -1$ et $x = 1$.*



La fonction$x \mapsto x^4 - 2x$

On peut avoir des résultats d'unicité si on utilise la convexité.

Définition 3.1.1 *Soit J une fonction définie sur $\mathbb{R}^n$ à valeurs dans $\mathbb{R}$.*

1. *On dit que J est convexe sur $\mathbb{R}^n$ si :*

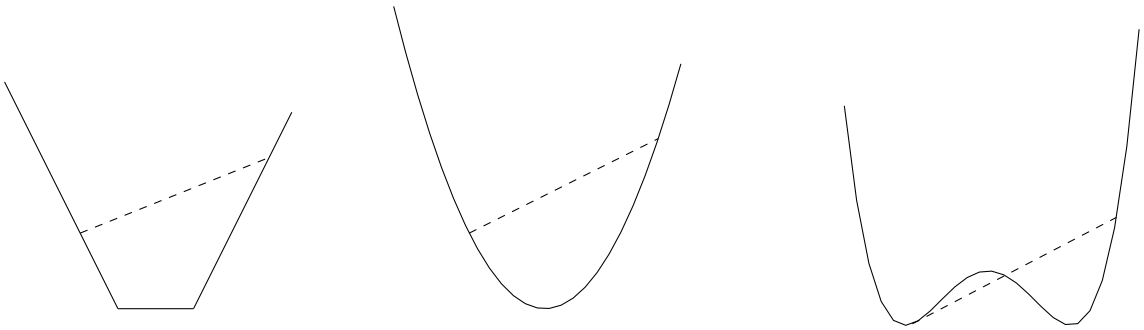
$$\forall\ (x,y)\ \in\ \mathbb{R}^n \times \mathbb{R}^n,\ \forall\ t \in [0,1],\ J((1-t)x) + ty) \leq (1-t)J(x) + tJ(y).$$

2. *On dit que J est strictement convexe sur $\mathbb{R}^n$ si :*

$$\forall\ (x,y)\ \in\ \mathbb{R}^n \times \mathbb{R}^n,\ \textit{tel que } x \neq y,\ \forall\ t \in]0,1[,\ J((1-t)x + ty) < (1-t)J(x) + tJ(y).$$

Remarques 3.1.2

1. *Graphiquement, une fonction est convexe si sa courbe se trouve en dessous de ses cordes.*



Fonction convexe *Fonction strictement convexe* *Fonction non convexe*

2. *Une fonction J est dite concave si $(-J)$ est convexe.*

Exemples 3.1.1

1. *L'application $x \mapsto \|x\|_2^2$ est strictement convexe sur $\mathbb{R}^n$.*
2. *Toute forme linéaire $J(x) = (b,x) = b^T x$ où $b \in \mathbb{R}^n$ est convexe mais non strictement convexe.*

Proposition 3.1.2 *Si la fonction $J : \mathbb{R}^n \longrightarrow \mathbb{R}$ est convexe, alors l'ensemble des solutions du problème (P) est convexe. Si de plus J est strictement convexe et elle admet un minimum sur $\mathbb{R}^n$, il est unique.*

Preuve. Soit S l'ensemble des solutions optimales. Montrons que S est convexe. Soit $(x_1, x_2) \in S \times S$, si $t \in [0,1]$, et comme J est convexe on a donc :

$$\begin{aligned} J((1-t)x_1 + tx_2) &\leq (1-t)J(x_1) + tJ(x_2) \\ &\leq (1-t)J(x) + tJ(x) = J(x), \quad \forall\, x \ \in \mathbb{R}^n \end{aligned}$$

On en déduit que $(1-t)x_1 + tx_2 \in S$, pour tout $t \in [0,1]$. D'où la convexité de S.

Si de plus J est strictement convexe, on suppose qu'il existe deux solutions x_1 et x_2 de (P) tel que $x_1 \neq x_2$. Alors $\dfrac{x_1 + x_2}{2} \in S$ et en vertu de la stricte convexité de J il vient que,

$$J\left(\frac{x_1 + x_2}{2}\right) < \tfrac{1}{2}J(x_1) + \tfrac{1}{2}J(x_2) = J(x_1).$$

Contradiction, ainsi $x_1 = x_2$. □

3.1.2 Conditions d'optimalité

On se limite dans ce cours à donner les conditions d'optimalité du premier ordre qui font intervenir seulement les dérivées d'ordre 1 de la fonction à minimiser.

Fonction différentiable

Définition 3.1.2 *On dit que $J : \mathbb{R}^n \to \mathbb{R}$ est différentiable au point $x \in \mathbb{R}^n$, s'il existe $p \in \mathbb{R}^n$ et une fonction $\epsilon : \mathbb{R}^n \to \mathbb{R}$ tels que*

$$J(x+h) = J(x) + (p,h) + \|h\|\epsilon(h) \quad \textit{avec} \quad \lim_{h \to 0} \epsilon(h) = 0.$$

S'il existe, le vecteur p est appelé la différentielle (ou la dérivée) de J en x et il est noté $J'(x)$. On dit que J est différentiable sur $\mathbb{R}^n$ si elle est différentiable en tout point de $\mathbb{R}^n$.

Remarque 3.1.2 *Si on note $(e_1, \ldots, e_n)$ la base canonique de $\mathbb{R}^n$, on peut vérifier que $J : \mathbb{R}^n \to \mathbb{R}$ est différentiable en x. De plus, le gradient de J en x est*

$$J'(x) = \nabla J(x) = \begin{pmatrix} \frac{\partial J}{\partial x_1}(x) \\ \vdots \\ \frac{\partial J}{\partial x_n}(x) \end{pmatrix}, \tag{3.3}$$

où $\dfrac{\partial J}{\partial x_i}(x) = \lim\limits_{t \to 0} \dfrac{J(x + te_i) - J(x)}{t}$, $i = 1, \ldots, n$ sont les dérivées partielles de J.

Exercice 3.1.1 *: On pourra montrer en exercice que :*

1. *Toute forme linéaire $J(x) = (b,x) = b^T.x$ où $b \in \mathbb{R}^n$, est différentiable et on a*

$$J'(x) \cdot h = (b,h) = b^T h,$$

donc

$$J'(x) = \nabla J(x) = b.$$

2. *La fonctionnelle* $J : x \mapsto J(x) = \frac{1}{2}a(x,x)$, *où* a *est une forme bilinéaire symétrique continue sur* $\mathbb{R}^n$, *alors*

$$J'(x) \cdot h = a(x,h).$$

Donc

$$J'(x) = a(x,\cdot).$$

En particulier, si A *est une matrice symétrique, alors* $J : x \to \dfrac{1}{2}(Ax,x)$ *est différentiable et on a :*

$$J'(x) = \nabla J(x) = Ax. \tag{3.4}$$

Fonction différentiable convexe

On peut caractériser la convexité et la convexité stricte des fonctions différentiables. C'est l'objet de deux propositions suivantes :

Proposition 3.1.3 *Soit* $J : \mathbb{R}^n \to \mathbb{R}$ *une fonction différentiable. Alors les propriétés suivantes sont équivalentes :*

i) J *est convexe sur* $\mathbb{R}^n$;

ii) *pour tout* $(x,y) \in \mathbb{R}^n \times \mathbb{R}^n$ *on a :*

$$J(y) \geq \ J(x) + (\nabla J(x), y - x);$$

iii) ∇J *est un opérateur monotone, i.e.,* $\forall\ (x,y) \in \mathbb{R}^n \times \mathbb{R}^n$ *on a :*

$$(\nabla J(x) - \nabla J(y), x - y) \geq 0.$$

Preuve.

i) $\Rightarrow$ **ii)** Supposons que J est convexe, alors pour tout $t \in]0,1]$ et pour tout $(x,y) \in \mathbb{R}^n \times \mathbb{R}^n$ on a :

$$J((1-t)x + ty) \leq (1-t)J(x) + tJ(y).$$

Ou encore

$$J(y) \geq J(x) + \frac{J(x + t(y-x)) - J(x)}{t}.$$

En faisant tendre t vers 0 on obtient

$$J(y) \geq J(x) + (\nabla J(x), y - x).$$

ii) $\Rightarrow$ **iii)** Si $x, y \in \mathbb{R}^n$, alors

$$J(y) \geq J(x) + (\nabla J(x), y - x) \geq J(y) + (\nabla J(y), x - y) + (\nabla J(x), y - x).$$

Par conséquent

$$(\nabla J(y) - \nabla J(x), y - x) \geq 0.$$

iii) $\Rightarrow$ **i)** Pour montrer que J est convexe il suffit de montrer que, pour tout x, y dans $\mathbb{R}^n$, la fonction g définie sur $[0,1]$ par

$$g(t) = (1-t)J(x) + tJ(y) - J((1-t)x + ty)$$

est positive. Il est clair que g est dérivable sur $[0,1]$ et que

$$g'(t) = -J(x) + J(y) - (\nabla J(x + t(y-x)), y - x).$$

Pour $t_1 \neq t_2$ on a :

$$(g'(t_1) - g'(t_2))(t_1 - t_2) = (\nabla J(x + t_2(y-x)) - \nabla J(x + t_1(y-x)), y - x)(t_2 - t_1) \leq 0.$$

Donc g' est décroissante. Comme g est continue sur $[0,1]$ et dérivable sur $]0,1[$ vérifiant $g(0) = g(1) = 0$, d'après le théorème de Rolle, il existe $c \in]0,1[$ tel que $g'(c) = 0$, ce qui entraîne que g est positive pour tout $t \in [0,1]$.

□

On peut montrer d'une façon analogue que :

Proposition 3.1.4 *Si $J : \mathbb{R}^n \to \mathbb{R}$ une fonction différentiable, alors les propriétés suivantes sont équivalentes :*

i) *J est strictement convexe sur $\mathbb{R}^n$;*

ii) *pour tout $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$ tels que $x \neq y$, on a :*

$$J(y) > \ J(x) + (\nabla J(x), y - x); \tag{3.5}$$

iii) *∇J est un opérateur strictement monotone i.e., $\forall\ (x, y) \in \mathbb{R}^n \times \mathbb{R}^n$ tels que $x \neq y$, on a :*

$$(\nabla J(x) - \nabla J(y), x - y) > 0.$$

Remarque 3.1.3 *D'après (iii) de la proposition* (3.1.3)*, une fonction dérivable $J : \mathbb{R} \to \mathbb{R}$ est convexe sur $\mathbb{R}$ si et seulement si sa fonction dérivée J' est croissante. Si de plus J est deux fois dérivable, alors J est convexe sur $\mathbb{R}$ si et seulement si J'' est positive.*

Exemple 3.1.1 *Si A est une matrice symétrique, la fonction $J : \mathbb{R}^n \to R; x \mapsto \frac{1}{2}(Ax, x) - (b, x)$ est strictement convexe sur $\mathbb{R}^n$ si et seulement si*

$$(\nabla J(x) - \nabla J(y), x - y) = (A(x-y), x-y) > 0 \text{ pour tout } x, y \in \mathbb{R}^n,\ x \neq y.$$

On rappelle que cette dernière propriété sur A est vérifiée si et seulement si A est définie positive.

Condition nécessaire et suffisante d'optimalité

En général, pour déterminer un minimum d'une fonction J on cherche parmi les points critiques qui vérifient la condition nécessaire d'optimalité (qui devient aussi suffisante lorsque J est convexe) suivante :

Proposition 3.1.5 *Soit $J : \mathbb{R}^n \to \mathbb{R}$ une fonction différentiable en $\bar{x}$*

1. *Si $\bar{x}$ réalise un minimum de J sur $\mathbb{R}^n$, alors*

$$\nabla J(\bar{x}) = 0. \tag{3.6}$$

2. *Si la fonction J est convexe et si $\nabla J(\bar{x}) = 0$, alors $\bar{x}$ réalise un minimum de J sur $\mathbb{R}^n$.*

Remarque 3.1.4 *La condition* (3.6) *est connue sous le nom de l'équation d'Euler qui est sans la convexité de J est en général une condition nécessaire mais non suffisante d'optimalité.*

La fonction $x \mapsto x^4 - 2x^2$ tracée dans la remarque (3.1.1) *est un contre exemple puisque sa dérivée s'annule en $x = 0$ et $x = \pm 1$, mais $x = 0$ ne réalise pas un minimum pour cette fonction.*

La fonction $J : x \mapsto x^3$ vérifie $J'(0) = 0$ mais 0 n'est ni un minimum ni un maximum de J sur $\mathbb{R}$.

3.1.3 Problème d'optimisation quadratique

Si la fonction objectif est de type quadratique, c.-à.-d, si elle est de la forme

$$J(x) = \tfrac{1}{2}(Ax, x) - (b, x),$$

pour $A \in \mathcal{M}_n(\mathbb{R})$ symétrique et $b \in \mathbb{R}^n$, alors le problème de minimisation (P) est dit quadratique et on a :

Proposition 3.1.6 *Si A est symétrique définie positive, alors le problème quadratique (P) admet une solution unique $\bar{x}$ vérifiant :*

$$A\bar{x} = b.$$

Preuve. Si on note par λ_1 la plus petite valeur propre de A, on a (voir (1.6) dans chapitre 1) :

$$\lambda_1 \|x\|_2^2 \leq (Ax, x) \ \forall x \in \mathbb{R}^n. \tag{3.7}$$

D'après la remarque (**??**) et l'inégalité de Cauchy-Schawrz, on déduit que J vérifie

$$J(x) \geq \frac{\lambda_1}{2}\|x\|_2^2 - (b, x) \geq \frac{\lambda_1}{2}\|x\|_2^2 - \|b\|_2\|x\|_2 \underset{\|x\|_2 \to +\infty}{\to} 0 .$$

De plus, J est strictement convexe, donc J admet un minimum unique $\bar{x}$ sur $\mathbb{R}^n$ vérifiant

$$\nabla J(\bar{x}) = A\bar{x} - b = 0.$$

□

3.1.4 Problème aux moindres carrés

Pour $A \in \mathcal{M}_{pn}(\mathbb{R})$ et $b \in \mathbb{R}^p$, avec $n < p$, le système $Ax = b$ ne possède pas, en général, de solution. On peut se contenter alors à chercher x qui minimise le carré de la norme de $Ax - b$ et donc ainsi formuler le problème de moindres carrés suivant :

$$\min_{x \in \mathbb{R}^n} \tfrac{1}{2}\|Ax - b\|_2^2. \tag{3.8}$$

Remarque 3.1.5 *Pour chercher la solution de* (3.8) *et afin de pouvoir appliquer l'équation d'Euler, on minimise la norme au carré de $Ax - b$ au lieu de $\|Ax - b\|_2$ puisque l'application norme n'est jamais différentiable en* 0. *Clairement, les deux problèmes sont équivalents.*

Proposition 3.1.7 *Le problème* (3.8) *admet au moins une solution. De plus toute, solution $\bar{x}$ de* (3.8) *est aussi solution du système*

$$A^T Ax = A^T b.$$

Le problème (3.8) *admet une solution unique si et seulement si la matrice A est injective.*

Avant de montrer cette proposition, rappelons le résultat classique d'algèbre linéaire

Lemme 3.1.1 *Pour toute matrice $A \in \mathcal{M}_{pn}(\mathbb{R})$, on a $Im(A^T A) = Im(A^T)$.*

Preuve. Il est clair que $\mathrm{Im}(A^T A) \subset \mathrm{Im}(A^T)$. De plus, d'après le théorème de rang, on sait que

$$\dim \mathrm{Im}(A) + \dim(\ker(A)) = dim(\mathbb{R}^n) = n = \dim \mathrm{Im}(A^T) + \dim(\ker(A)).$$

Or on $\ker(A) = \ker(A^T A)$, puisque $\ker(A) \subset \ker(A^T A)$ et si $x \in \ker(A^T A)$, $(A^T Ax, x) = (Ax, Ax) = \|Ax\|_2^2 = 0$, donc $Ax = 0$ et par suite $x \in \ker(A^T A)$. Appliquons encore une fois le théorème de rang à la matrice $A^T A$, on obtient

$$\dim \mathrm{Im}(A^T A) + \dim(\ker(A^T A)) = dim(\mathbb{R}^n) = n.$$

Par conséquent, $\dim \mathrm{Im}(A^T A) = \dim \mathrm{Im}(A^T)$ et les deux espaces sont donc égaux.
□

Preuve. de la proposition (3.1.7)

Ici, la fonction objectif est

$$J(x) = \tfrac{1}{2}(A^T Ax, x) - (A^T b, x) + \tfrac{1}{2}\|b\|_2^2.$$

La matrice $A^T A$ est toujours symétrique semi-définie positive puisque $(A^T Ax, x) = \|Ax\|_2^2 \geq 0$ pour tout vecteur x de $\mathbb{R}^n$. La fonction quadratique J est donc convexe. Par conséquent, $\bar{x}$ est solution de (3.8) si et seulement

$$\nabla J(\bar{x}) = A^T A\bar{x} - A^T b = 0.$$

Comme $\mathrm{Im}(A^T) = \mathrm{Im}(A^T A)$on déduit que le système $A^T Ax = A^T b$ admet au moins une solution, d'où l'existence d'au moins un minimum de J sur $\mathbb{R}^n$. L'unicité de minimum est assurée si et seulement si le système $A^T Ax = A^T b$ admet une solution unique, pour tout vecteur b, donc si et seulement si la matrice $A^T A$ est bijective. Ceci étant vérifié si et seulement A est injective. □

3.2 Algorithmes de descente et méthodes du gradients

3.2.1 Méthodes de descente

Les méthodes numériques de descente en général sont des méthodes itératives dont le but est de déterminer $\bar{x}$ réalisant le minimum d'une fonction J sur $\mathbb{R}^n$, en utilisant des directions de déplacement permettant de se rapprocher le plus possible de $\bar{x}$. Dans ce chapitre on va utiliser des méthodes numériques de descente du gradient pour un problème de minimisation sans contrainte quelconque dont la convergence sera étudiée seulement pour le problème quadratique :

$$J(x) := \min_{x\in\mathbb{R}^n} \frac{1}{2}(Ax, x) - (b, x), \tag{3.9}$$

où A est une matrice symétrique définie positive et b un vecteur de $\mathbb{R}^n$.

Ces méthodes s'appliquent alors aussi pour résoudre numériquement un système linéaire $Ax = b$, pour A une matrice symétrique définie positive.

Définition 3.2.1 *Soit J une fonction de $\mathbb{R}^n$ à valeurs dans $\mathbb{R}$. Soit $x \in \mathbb{R}^n$. On dit que $d \in \mathbb{R}^n$, avec $d \neq 0$, est une direction de descente de J en x s'il existe $\alpha_0 > 0$ tel que*

$$J(x + \alpha d) \leq J(x),\ \forall\ \alpha \in [0, \alpha_0].$$

Ainsi, une méthode de descente pour la recherche de $\bar{x}$ solution de

$$\min_{x\in\mathbb{R}^n} J(x),$$

consiste à construire une suite $(x^{(k)})_{k\in\mathbb{N}}$ de la manière suivante :

- Initialisation : $x^{(0)} \in \mathbb{R}^n$,
- Pour $k \geq 0$:
  on cherche la direction de descente $d^{(k)}$ au point $x^{(k)}$ et on détermine le pas α_k,
  Ensuite, on calcule $x^{(k+1)} = x^{(k)} + \alpha_k d^{(k)}$.

Proposition 3.2.1 *Soient $J : \mathbb{R}^n \to \mathbb{R}$ une fonction différentiable, $x, d \in \mathbb{R}^n$ avec $d \neq 0$. Si d est une direction de descente en x, alors :*

$$(\nabla J(x), d) \leq 0.$$

Preuve. Soit la fonction φ de $\mathbb{R}$ dans $\mathbb{R}$ définie par : $\varphi(\alpha) = J(x + \alpha d)$. Alors φ est dérivable

$$\varphi'(\alpha) = (\nabla J(x + \alpha d), d).$$

On considère $d \in \mathbb{R}^n$ une direction de descente au point x alors par définition, il existe $\alpha_0 > 0$ tel que :

$$J(x + \alpha d) \leq J(x), \forall\ \alpha \in [0, \alpha_0].$$

Comme d est une direction de descente, on peut écrire, pour tout $\alpha \in [0, \alpha_0]$,

$$\varphi(\alpha) \leq \varphi(0),$$

et donc pour tout $\alpha \in]0, \alpha_0]$

$$\frac{\varphi(\alpha) - \varphi(0)}{\alpha - 0} \leq 0.$$

En passant à la limite lorsque α tend vers 0, on déduit que $\varphi'(0) \leq 0$, ce qui signifie que

$$(\nabla J(x), d) \leq 0.$$

3.2.2 Méthodes du gradient

Dans ces méthodes les directions de descente s'expriment en fonction du gradient de la fonction à minimiser.

Remarque 3.2.1 *Ces méthodes itératives de descente ne sont pas aussi finies. Il faut donc un critère d'arrêt qui permet d'arréter les itérations dès que l'itéré $x^{(k)}$ s'approche de la solution $\bar{x}$ du problème à minimiser. Comme test d'arrêt classique pour ce type d'algorithmes consiste à s'arrêter à l'itération k si $\|\nabla J(x^{(k)})\| \leq \varepsilon$ ou si $\|x^{(k+1)} - x^{(k)}\| \leq \varepsilon$, pour ε une précision donnée.*

3.2.3 Méthode du gradient à pas fixe

La méthode du gradient à pas fixe $\alpha > 0$ consiste à choisir comme direction de descente $d^{(k+1)}$ à l'étape $k+1$, $d^{(k+1)} = -\nabla J(x^{(k)})$. Pour un problème quadratique l'algorithme s'écrit comme suit :

1. On choisit $x^{(0)} \in \mathbb{R}^n$ et un pas $\alpha > 0$.
2. Pour $k \geq 0$, on calcule : $\left| \begin{array}{l} d^{(k)} = -\nabla J(x^{(k)}) = b - Ax^{(k)}, \\ x^{(k+1)} = x^{(k)} + \alpha d^{(k)} \end{array} \right.$

Proposition 3.2.2 *:* **Méthode du gradient à pas fixe pour un problème quadratique**
Si $J(x) = \frac{1}{2}(Ax, x) - (b, x)$, où A est une matrice symétrique définie positive, alors, la méthode du gradient à pas fixe converge si et seulement si $\alpha \in]0, \dfrac{2}{\rho(A)}[$.

Preuve. On a, pour $k \geq 0$,

$$x^{(k+1)} = x^{(k)} - \alpha(Ax^{(k)} - b) = (I - \alpha A)x^{(k)} + \alpha b.$$

La matrice d'itération est $B_\alpha = I - \alpha A$ dont le spectre est

$$\mathrm{Sp}(B_\alpha) = \{1 - \alpha\lambda;\ \lambda \in \mathrm{Sp}(A)\}\,.$$

Si on note par λ_1 et λ_n respectivement la plus petite et la plus grande valeur propre de A, alors

$$1 - \alpha\lambda_n \leq 1 - \alpha\lambda \leq 1 - \alpha\lambda_1,\ \forall\lambda \in \mathrm{Sp}(A).$$

Ainsi,

$$\rho(B_\alpha) = \max(|1 - \alpha\lambda_1|, |1 - \alpha\lambda_n|).$$

Par conséquent, la méthode du gradient à pas fixe converge si et seulement si $\rho(B_\alpha) < 1$. Donc si et seulement si $0 < \alpha < \frac{2}{\lambda_1}$ et $0 < \alpha < \frac{2}{\lambda_n}$. Cette dernière condition est bien équivalente à

$$0 < \alpha < \frac{2}{\rho(A)}.$$

□

3.2.4 Méthodes du gradient à pas optimal

Méthodes de descente à pas optimal

Une méthode de descente générale est du type $x^{(k+1)} = x^{(k)} + \alpha_k d^{(k)}$, pour $x^{(0)}$ donné et pour une direction $d^{(k)}$ connue, est dite à pas optimal si l'on choisit le pas α_k de manière à minimiser la fonction $\alpha \mapsto J(x^{(k)} + \alpha d^{(k)})$.

Calculons le pas α_k pour un problème quadratique. Ceci revient à chercher $\alpha \in \mathbb{R}$ vérifie, pour $x, d \in \mathbb{R}^n$ tel que $d \neq 0$:

$$\forall\ r \in \mathbb{R}, \quad J(x + \alpha d) \leq J(x + rd). \tag{Q}$$

Lemme 3.2.1 *Si la matrice A est symétrique définie positive, alors (Q) admet une unique solution donnée par*

$$\alpha = -\frac{(Ax - b, d)}{(Ad, d)}.$$

Preuve. On considère la fonction $f : \mathbb{R} \to \mathbb{R}$ définie par :

$$f(r) = J(x + rd).$$

On a

$$\begin{aligned} f(r) &= J(x + rd) \\ &= \frac{1}{2}(A(x + rd), \bar{x} + rd) - (b, x + rd) \\ &= \frac{1}{2}((Ax, x) + 2r(Ax, d) + r^2(Ad, d)) - (b, x) - r(b, x) \\ &= \frac{1}{2}r^2(Ad, d) + r(Ax - b, d) + \frac{1}{2}(Ax, x) - (b, x). \end{aligned}$$

Puisque $(Ad, d) > 0$ (car A est définie positive et $d \neq 0$), on en déduit que la fonction f (polynôme de degré 2) admet un minimum global au point $r = \alpha$ donné par :

$$\alpha = -\frac{(Ax - b, d)}{(Ad, d)}.$$

□

Remarque 3.2.2 *Pour un problème quadratique du type* (3.9)*, une méthode de descente à pas optimal* $\alpha_k = -\dfrac{(Ax^{(k)} - b, d^{(k)})}{(Ad^{(k)}, d^{(k)})}$ *donne une relation d'orthogonalité entre la direction de descente* $d^{(k)}$ *et le gradient* $g^{(k+1)}$ *pour*

$$g^{(k)} = \nabla J(x^{(k)}) = Ax^{(k)} - b = -d^{(k)}.$$

En effet

$$(g^{(k+1)}, d^{(k)}) = (Ax^{(k+1)} - b, d^{(k)}) = (Ax^{(k)} - b, d^{(k)}) + \alpha_k (Ad^{(k)}, d^{(k)}) = 0. \tag{3.10}$$

Algorithme du gradient à pas optimal

La méthode du gradient à pas optimal est une méthode de descente à pas optimal dont la direction de descente est donnée par le gradient.

Le principe de cette méthode s'écrit :

- Initialisation : $x^{(0)} \in \mathbb{R}^n$ donné.
- Pour $k \geq 0$,
  On calcule $d^{(k)} = -\nabla J(x^{(k)})$,
  On choisit $\alpha_k \geq 0$ tel que : $J(x^{(k)} + \alpha_k d^{(k)}) \leq J(x^{(k)} + \alpha d^{(k)}),$ pour tout $\alpha \geq 0$.
  On pose $x^{(k+1)} = x^{(k)} + \alpha_k d^{(k)}$.

Ainsi, l'algorithme du gradient à pas optimal pour un problème quadratique s'écrit :

1. Initialisation : $x^{(0)} \in \mathbb{R}^n$ donné,
2. Pour $k \geq 0$, on calcule :

$$\left| \begin{array}{l} d^{(k)} = b - Ax^{(k)}, \\ \alpha_k = \dfrac{\| d^{(k)} \|_2^2}{(Ad^{(k)}, d^{(k)})}, \\ x^{(k+1)} = x^{(k)} + \alpha_k d^{(k)}. \end{array} \right.$$

Proposition 3.2.3 *Si* A *est une matrice symétrique définie positive, alors la méthode du gradient à pas optimal converge.*

Preuve. Par construction, la suite $(J(x^{(k)}))$ est décroissante et on sait qu'elle est minorée par $J(\bar{x})$, donc elle est convergente. Par conséquent,

$$\lim_{k\to+\infty} J(x^{(k+1)}) - J(x^{(k)}) = 0.$$

Or

$$J(x^{(k+1)}) - J(x^{(k)}) = \frac{\alpha_k^2}{2}(Ad^{(k)}, d^{(k)}) + \alpha_k(Ax^{(k)} - b, d^{(k)}).$$

Sachant que

$$\alpha_k = \frac{\| d^{(k)} \|_2^2}{(Ad^{(k)}, d^{(k)})} \text{ et } d^{(k)} = -g^{(k)} = -(Ax^{(k)} - b),$$

donc

$$J(x^{(k+1)}) - J(x^{(k)}) = -\tfrac{1}{2}\frac{\|g^{(k)}\|_2^4}{(Ag^{(k)}, g^{(k)})}.$$

De (1.6) on déduit

$$\frac{1}{2\lambda_n}\|g^{(k)}\|_2^2 \leq J(x^{(k)}) - J(x^{(k+1)}),$$

où λ_n est la plus grande valeur propre de A. Donc

$$\|g^{(k)}\|_2^2 \underset{k\to+\infty}{\to} 0\,,$$

et par conséquent,

$$\|x^{(k)} - \bar{x}\|_2 = \|A^{-1}(Ax^{(k)} - b)\|_2 \leq \|A^{-1}\|_2\|g^{(k)}\|_2 \underset{k\to+\infty}{\to} 0\,.$$

□

3.2.5 Méthodes du gradient conjugué

Cette méthode s'applique directement pour chercher la solution unique $\bar{x}$ du problème

$$\min_{x\in\mathbb{R}^n} J(x),$$

pour

$$J(x) = \tfrac{1}{2}(Ax, x) - (b, x), \tag{P}$$

où A est une matrice symétrique définie positive.

Dans les algorithmes des deux méthodes de descente précédentes, les termes de la suite $(x^{(k))})$ s'approchent de plus en plus vers la solution cherchée $\bar{x}$ sans généralement jamais l'atteindre. D'où l'idée de la méthode du gradient conjugué où, comme on le verra, la suite $(x^{(k)})$ générée par l'algorithme de cette méthode devient stationnaire à partir d'un certain nombre d'itérations (au plus n), et elle vaut $\bar{x}$.

Définition 3.2.2 *Soit A est une matrice symétrique définie positive de $\mathcal{M}_n(\mathbb{R})$.*

1. *Deux vecteurs non nuls $d^{(1)}, d^{(2)}$ de $\mathbb{R}^n$ sont dits $A-$conjugués si $(Ad^{(1)}, d^{(2)}) = 0$.*
2. *Une famille $(d^{(1)}, d^{(2)}, \ldots, d^{(p)})$ de p vecteurs non nuls de $\mathbb{R}^n$ est dite $A-$conjuguée si*

$$(Ad^{(i)}, d^{(j)}) = 0 \ \ \forall \ \ i, j \in \{1, \ldots, p\}, \ i \neq j.$$

Remarque 3.2.3 *On peut facilement vérifier que*

1. *Toute famille $A-$conjuguée de p vecteurs non nuls de $\mathbb{R}^n$ est libre.*
2. *Une famille $A-$conjuguée de n vecteurs non nuls de $\mathbb{R}^n$ est une base de $\mathbb{R}^n$.*

L'idée de la méthode du gradient conjugué, qui est aussi une méthode de descente à pas optimal, est de construire la direction de descente en fonction du gradient et de la direction précédente de descente.

On part de $x^{(0)}$ donné et $d^{(0)} = -\nabla J(x^{(0)}) = -(Ax^{(0)} - b) = -g^{(0)}$, Connaissant $x^{(k)}$, $d^{(k-1)}$ et $g^{(k)}$, et si $g^{(k)} \neq 0$, on choisit la direction de descente

$$d^{(k)} = -g^{(k)} + \beta_{k-1} d^{(k-1)}, \tag{3.11}$$

et

$$x^{(k+1)} = x^{(k)} + \alpha_k d^{(k)}, \tag{3.12}$$

de manière à minimiser la fonction à deux variables

$$f : (\alpha, \beta) \mapsto J(x^{(k)} + \alpha(g^{(k)} + \beta d^{(k-1)}) = J(x^{(k)} + \alpha d^{(k)}).$$

Donc ici le pas α_k est optimal et on a la relation d'orthogonalité (voir lemme (3.2.1)) suivante

$$(d^{(k-1)}, g^{(k)}) = 0. \tag{3.13}$$

Si $g^{(k)} \neq 0$, d'après (3.11), on aura aussi $d^{(k)} \neq 0$ et par conséquent

$$\alpha_k = -\frac{(Ax^{(k)} - b, d^{(k)})}{(Ad^{(k)}, d^{(k)})} = -\frac{(g^{(k)}, -g^{(k)} + \beta_{k-1} d^{(k-1)})}{(Ad^{(k)}, d^{(k)})} = \frac{\|g^{(k)}\|_2^2}{(Ad^{(k)}, d^{(k)})}.$$

Il suffit donc de déterminer la direction de descente $d^{(k)}$. Ceci revient à chercher β_{k-1}. Sachant que cette dernière doit vérifier

$$\begin{aligned}\frac{\partial f}{\partial \beta}(\alpha_k, \beta_{k-1}) &= \alpha_k(\nabla J(x^{(k)} + \alpha_k d^{(k)}), d^{(k-1)}) \\ &= \alpha_k(\nabla J(x^{(k)} + \alpha_k(-g^{(k)} + \beta_{k-1} d^{(k-1)})), d^{(k-1)}) = 0.\end{aligned}$$

et que $\nabla J(x) = Ax - b$. on déduit d'abord

$$\beta_{k-1} = \frac{(Ad^{(k-1)}, g^{(k)})}{(Ad^{(k-1)}, d^{(k-1)})},$$

puis que les directions $d^{(k)}$ et $d^{(k-1)}$ sont $A-$conjugués,

$$(Ad^{(k)}, d^{(k-1)}) = 0.$$

En remplaçant $Ad^{(k-1)}$ par $\dfrac{g^{(k)}-g^{(k-1)}}{\alpha_{k-1}}$, on déduit que

$$(Ad^{(k-1)},g^{(k)})=\frac{\|g^{(k)}\|_2^2}{\alpha_{k-1}} \text{ et que } (Ad^{(k-1)},d^{(k-1)})=\frac{\|g^{(k-1)}\|_2^2}{\alpha_{k-1}}.$$

On tire alors que

$$\beta_{k-1}=\frac{\|g^{(k)}\|_2^2}{\|g^{(k-1)}\|_2^2}.$$

D'où l'algorithme du gradient conjugué :

1. On choisit $x^{(0)}\in\mathbb{R}^n$. On prend $g^{(0)}=Ax^{(0)}-b=-d^{(0)}$.
2. Pour $k\geq 0$, on calcule

$$\left|\begin{array}{ll} \alpha_k=\dfrac{\|g^{(k)}\|_2^2}{(Ad^{(k)},d^{(k)})}, & x^{(k+1)}=x^{(k)}+\alpha_k d^{(k)},\\ g^{(k+1)}=Ax^{(k+1)}-b & =g^{(k)}+\alpha_k Ad^{(k)},\\ \beta_k=\dfrac{\|g^{(k+1)}\|_2^2}{\|g^{(k)}\|_2^2}, & d^{(k+1)}=-g^{(k+1)}+\beta_k d^{(k)}. \end{array}\right.$$

On montre alors :

Proposition 3.2.4 *Si A est symétrique définie positive, alors l'algorithme du gradient conjugué converge en au plus n itérations vers le minimum de $J(x):=\frac{1}{2}(Ax,x)-(b,x)$ sur $\mathbb{R}^n$.*

Preuve. Pour $k=0$, si $g^{(0)}=0$, alors $Ax^{(0)}=b$ et $\bar{x}=x^{(0)}$. Sinon on a :

$$d^{(0)}=-g^{(0)}=-Ax^{(0)}+b=-\nabla J(x^{(0)}),$$

et

$$\alpha_0=-\frac{(g^{(0)},d^{(0)})}{(Ad^{(0)},d^{(0)})},\ x^{(1)}=x^{(0)}+\alpha_0 d^{(0)},\ g^{(1)}=Ax^{(1)}-b$$

vérifiant

$$(g^{(1)},g^{(0)})=(g^{(1)},d^{(0)})=(d^{(1)},Ad^{(0)})=0.$$

Supposons maintenant, pour $1\leq k<n$, l'hypothèse de récurrence

$$\left|\begin{array}{lll} (g^{(k)},g^{(j)})=0 & \text{pour} & 0\leq j<k,\\ (g^{(k)},d^{(j)})=0 & \text{pour} & 0\leq j<k,\\ (d^{(k)},Ad^{(j)})=0 & \text{pour} & 0\leq j<k, \end{array}\right. \tag{3.14}$$

Si $g^{(k)}=0$, alors $\bar{x}=x^{(k)}$ et l'algorithme converge à l'itération k. Si $g^{(k)}\neq 0$, on construit $x^{(k+1)}, g^{(k+1)}$ et $d^{(k+1)}$ par l'algorithme du gradient conjugué.

Vérifions l'hypothèse de récurrence à l'ordre $(k+1)$

$$\left|\begin{array}{ll} (g^{(k+1)}, g^{(j)}) = 0 & \text{pour} \quad 0 \leq j < k+1, \\ (g^{(k+1)}, d^{(j)}) = 0 & \text{pour} \quad 0 \leq j < k+1, \\ (d^{(k+1)}, Ad^{(j)}) = 0 & \text{pour} \quad 0 \leq j < k+1, \end{array}\right.$$

D'après (3.13), et par construction on a

$$(g^{(k+1)}, d^{(k)}) = 0,$$

et pour $j < k$, on a

$$(g^{(k+1)}, d^{(j)}) = (g^{(k+1)}, d^{(j)}) - \underbrace{(g^{(k)}, d^{(j)})}_{=0} = (g^{(k+1)} - g^{(k)}, d^{(j)}) = \alpha_k (Ad^{(k)}, d^{(j)}) = 0.$$

On a

$$g^{(k)} = -d^{(k)} + \beta_{k-1} d^{(k-1)}.$$

Donc, pour $j \leq k$,

$$(g^{(k+1)}, g^{(j)}) = (g^{(k+1)}, d^{(j)}) - \beta_j (g^{(k+1)}, d^{(j-1)}) = 0.$$

Or, pour $j \leq k$ on a $g^{(j+1)} = g^{(j)} + \alpha_j Ad^{(j)}$ et donc

$$\alpha_j (g^{(k+1)}, Ad^{(j)}) = (g^{(k+1)}, g^{(j+1)}) - (g^{(k+1)}, g^{(j)}) = 0.$$

Sachant que $g^{(j)} \neq 0$, donc $\alpha_j \neq 0$ et par conséquent $(g^{(k+1)}, Ad^{(j)}) = 0$.

Comme

$$d^{(k+1)} = -g^{(k+1)} + \beta_k d^{(k)},$$

alors, pour $j \leq k$, on a

$$(d^{(k+1)}, Ad^{(j)}) = (g^{(k+1)}, Ad^{(j)}) + \beta_k (d^{(k)}, Ad^{(j)}) = 0,$$

et la récurrence est vérifiée.

Si $g^{(k)} \neq 0$, pour $k = 0, \ldots, n-1$, alors $(d^{(0)}, d^{(1)}, \ldots, d^{(n-1)})$ est une famille $A-$conjuguée de n vecteurs, donc base de $\mathbb{R}^n$ et le vecteur $g^{(n)}$ sera orthogonal à $d^{(0)}, d^{(1)}, \ldots, d^{(n-1)}$. Nécessairement $g^{(n)} = 0$ et par suite $x^{(n)} = \bar{x}$. L'algorithme ainsi converge à l'itération exactement n.

Remarque 3.2.4 *Les directions de descente sont telles que la base $(d^{(0)}, d^{(1)}, \ldots, d^{(n-1)})$ est obtenue par le procédé d'orthogonalité de Gram-Shmidt, adapté au produit scalaire (Ax, y), à partir de la base $(-g^{(0)}, \ldots, -g^{(n-1)})$.*

□

3.2.6 Vitesse de convergence

En général une suite $(x^{(k)})$ qui converge vers $\bar{x}$, vérifiant $x^{(k)} \neq \bar{x}$ pour tout entier k, est dite d'ordre de convergence au moins $p \in \mathbb{R}_+^*$ si

$$\|x^{(k+1)} - \bar{x}\| = O(\|x^{(k)} - \bar{x}\|^p).$$

En particulier, s'il existe $\gamma \in]0,1[$ tel que

$$\lim_{k \to +\infty} \frac{\|x^{(k+1)} - \bar{x}\|}{\|x^{(k)} - \bar{x}\|^p} = \gamma,$$

alors la convergence est dite exactement d'ordre p. La convergence est dite linéaire si $p = 1$ et quadratique si $p = 2$.

Les méthodes de descente du gradient à pas fixe et à pas optimal sont à convergence au moins linéaire et on pourra montrer que

$$\|x^{(k+1)} - \bar{x}\| \leq \frac{\mathrm{Cond}_2(A) - 1}{\mathrm{Cond}_2(A) + 1} \|x^{(k)} - \bar{x}\|, \tag{3.15}$$

où $\|.\|$ désigne la norme euclidienne pour la méthode du gradient à pas fixe et la norme définie par $\|x\| = \|x\|_A = (Ax, x)^{\frac{1}{2}}$ pour la méthode du gradient à pas optimal.

Pour la méthode du gradient conjugué, et tant que $g^{(k)} = Ax^{(k)} - b \neq 0$, on peut montrer la majoration suivante :

$$\|x^{(k+1)} - \bar{x}\|_A \leq 2 \left(\frac{\sqrt{\mathrm{Cond}_2(A)} - 1}{\sqrt{\mathrm{Cond}_2(A)} + 1} \right)^k \|x^{(0)} - \bar{x}\|_A. \tag{3.16}$$

3.2.7 Méthodes du gradient et préconditionnement

Des majorations (3.15) et (3.16), on constate que l'erreur $x^{(k)} - \bar{x}$, obtenue au cours des itérations des méthodes du gradient, diminue d'autant plus rapide que $\mathrm{Cond}_2(A)$ est petit.

On rappelle aussi que les méthodes du gradient sont aussi des méthodes numériques pour résoudre un système linéaire $Ax = b$, pour $A \in M_n(\mathbb{R})$ symétrique définie positive. Ainsi, la matrice du système préconditionné doit aussi être symétrique définie positive. Le préconditionnement de ce système consiste alors à appliquer les méthodes du gradient pour le problème de minimisation

$$\min_{y \in \mathbb{R}^n} \tfrac{1}{2}(\tilde{A}y, y) - (\tilde{b}, y), \tag{P_c}$$

avec $\tilde{A} = PAP^T$ et $\tilde{b} = Pb$, où P est une matrice inversible bien choisie de sorte que $\tilde{A}$ soit bien conditionnée. La nouvelle matrice $\tilde{A}$ est aussi symétrique définie positive et la solution $\bar{x}$ du système $Ax = b$ est donnée par $\bar{x} = P^T \bar{y}$, pour $\bar{y}$ solution optimale de (P_c).

Chapitre 4

Optimisation avec contraintes linéaires

Lorsque un minimum $\bar{x}$ d'une fonction $J : \mathbb{R}^n \to \mathbb{R}$ doit satisfaire certaines conditions, on parle de problème de minimisation contraint ou avec contraintes. Dans ce cas on cherche à calculer le minimum d'une fonction J non pas sur $\mathbb{R}^n$, mais sur un ensemble fermé K de $\mathbb{R}^n$ et notre problème sera de type

$$\min_{x \in K} J(x) \qquad (P).$$

Ce chapitre comporte une étude théorique et numérique de problème (P). On commence par donner des résultats d'existence et d'unicité, ainsi que des conditions d'optimalités pour un ensemble K convexe puis pour un problème à contraintes linéaires. Ensuite, pour résoudre numériquement (P), on propose la méthode du gradient projeté pour une fonction objectif quadratique et pour K un ensemble convexe quelconque, puis la méthode d'Uzawa pour un problème quadratique avec contraintes d'inégalité.

4.1 Problèmes d'optimisations sous contraintes

Rappelons d'abord qu'un ensemble K est dit fermé si toute suite convergente d'éléments de K, sa limite est aussi dans K.

Un exemple simple de de problème d'optimisation avec (ou sous) contraintes est de déterminer dans le plan le point $M(x, y)$ de $\mathbb{R}^2$ le plus proche possible à un point donné (x_0, y_0) et qui appartient à une droite donnée d'équation $ax + by = c$, donc chercher le couple (x, y) solution de

$$\min_{ax+by=c} (x - x_0)^2 + (y - y_0)^2.$$

Ici l'ensemble K est la droite du plan d'équation $ax + by + c = 0$.

On appelle problème de minimisation sous contraintes d'égalités linéaires si

$$K := \{x \in \mathbb{R}^n \text{ tel que } Cx = d\},$$

où C est une matrice de $\mathcal{M}_{m,n}(\mathbb{R})$ et $d \in \mathbb{R}^m$.
Pour les problèmes avec contraintes d'inégalités linéaires l'ensemble des contraintes est

$$K := \{x \in \mathbb{R}^n \text{ tel que } Bx \leq c\},$$

où B est une matrice de $\mathcal{M}_{p,n}(\mathbb{R})$ et $c \in \mathbb{R}^p$.

On parle de problème de minimisation sous contraintes linéaires si les contraintes sont d'égalités et d'inégalités linéaires.

4.1.1 Existence et unicité de minimum

On rappelle que

Définition 4.1.1 *Une partie K de $\mathbb{R}^n$ est dite convexe si :*

$$tx + (1-t)y \in K \ \ \forall\, (x,y) \in K \times K, \ \forall \ \ t \in [0,1].$$

Autrement dit K est une partie convexe si elle contient tout segment d'extrémités deux quelconques de ses points.

Remarques 4.1.1

1. *La notion de convexité et de stricte convexité d'une fonction J sur une partie convexe K, ainsi que leurs caratérisation lorsque J est différentiable, sont identiques à celles données dans la définition* (3.1.1) *et dans les propositions* (3.1.3) *et* (3.1.4) *qui s'appliquent pour tout $x, y \in K$ au lieu de $\mathbb{R}^n$.*
2. *Lorsque K est convexe et la fonction à minimiser J est convexe, le problème de minimisation (P) est dit convexe.*

Comme pour le problème sans contraintes, on a

Proposition 4.1.1
Soient K un fermé non vide de $\mathbb{R}^n$ et $J : K \to \mathbb{R}$ continue. On suppose que J est infinie à l'infini **ou** *que K est borné, alors le problème de minimisation (P) admet au moins une solution. Si de plus K est convexe et J est strictement convexe sur K, ce minimum est unique.*

Preuve. La démonstration est similaire au cas sans contraintes, où on prend pour l'existence une suite minimisante $(x^{(k)})$ qui est aussi bornée. Comme K est fermé, donc $\bar{x}$ la limite d'une sous suite de $(x^{(k)})$ appartient aussi à K. Lorsque le problème est strictement convexe et il admet un minimum, et comme dans le cas non contraint, l'unicité est immédiate. □

Projection sur un convexe

Soit K une partie non vide de $\mathbb{R}^n$ et x un vecteur de $\mathbb{R}^n$ qui n'appartient pas à K. On cherche à définir la distance de ce vecteur x à K, c.à.d, le réel qui réalise la distance minimale de x à tous les points de K. Cette distance est elle finie ? et existe- il $\bar{x} \in K$ qui réalise cette distante ? La réponse à ces deux questions n'est pas évidente sans aucune hypothèse sur K. Cependant et lorsque K est convexe et fermé on a :

Proposition 4.1.2 *Soit K une partie convexe fermée non vide de $\mathbb{R}^n$ et x un point de $\mathbb{R}^n$. Alors il existe un unique point de K, noté $P_K(x)$ tel que :*

$$\begin{cases} P_K(x) \in K \\ \|x - P_K(x)\|_2 \leq \|y - x\|_2, \ \forall\, y \in K \end{cases} \tag{4.1}$$

$P_K(x)$ est appelé projection de point x sur le convexe fermé K.

Preuve. On pose, pour $y \in K$

$$J(y) = \frac{1}{2}\|y - x\|_2^2 = \frac{1}{2}(y - x, y - x) = \tfrac{1}{2}\|y\|_2^2 - (y, x) + \tfrac{1}{2}\|x\|_2^2.$$

La fonction J est strictement convexe et différentiable avec $\nabla J(y) = y - x$. D'après le théorème d'existence et d'unicité, le problème admet une unique solution. □

Exemple 4.1.1 *Si $K = [0, +\infty[$, alors $P_K(x) = \begin{cases} x & si\, x \geq 0 \\ 0 & si\, x < 0 \end{cases}$. On note aussi $P_K(x) = x_+$.*

4.1.2 Condition d'optimalité

Soit $\bar{x}$ une solution optimale de problème simple suivant :

$$\min_{x \in [-1,1]} ax,$$

pour $a \in \mathbb{R}$. Si $a \geq 0$, alors la fonction $J : x \mapsto ax$ atteint son minimum en $\bar{x} = -1$ vérifiant donc $J'(\bar{x}) \leq 0$. Si $a < 0$, alors $\bar{x} = 1$ et dans ce cas $J'(\bar{x}) \geq 0$. Dans les deux cas, $\bar{x}$ vérifie

$$J'(\bar{x})(x - \bar{x}) \geq 0, \ \forall\, x \in [-1, 1]. \tag{4.2}$$

Dans la proposition qui suit, on généralise la condition d'optimalité (4.2) pour un ensemble de contraintes convexe K quelconque.

Proposition 4.1.3 *Soit $J : \mathbb{R}^n \to \mathbb{R}$ une fonction différentiable en $\bar{x}$ et soit K un convexe fermé de $\mathbb{R}^n$.*

Si $\bar{x}$ réalise un minimum de J sur K, alors

$$(\nabla J(\bar{x}), x - \bar{x}) \geq 0 \ \forall\, x \in K. \tag{4.3}$$

Si de plus la fonction J est convexe sur K, alors $\bar{x}$ réalise un minimum de J sur K, si et seulement si

$$(\nabla J(\bar{x}), x - \bar{x}) \geq 0 \ \forall\, x \in K. \tag{4.4}$$

L'inéquation (4.3) *et connue sous le nom d'inéquation d'Euler.*

Preuve.

1. Soit $x \in K$, on a alors pour $0 < t < 1$, $\bar{x} + t(x - \bar{x}) = tx + (1-t)\bar{x} \in K$, puisque K est convexe. Comme $\bar{x}$ est un minimum de J sur K, alors

$$J(\bar{x} + t(x - \bar{x})) - J(\bar{x}) \geq 0.$$

Par conséquent

$$\lim_{t \to 0^+} \frac{J(\bar{x} + t(x - \bar{x}) - J(\bar{x})}{t} = (\nabla J(\bar{x}), x - \bar{x}) \geq 0.$$

2. On suppose que J est convexe sur K et soit $\bar{x} \in K$ vérifiant l'inéquation d'Euler. De la convexité de J on déduit que

$$\begin{aligned} J(x) &\geq J(\bar{x}) + (\nabla J(\bar{x}), x - \bar{x}) \\ &\geq J(\bar{x}),\ \forall\ x \in K \end{aligned}$$

Par suite $\bar{x}$ est une solution de (P).

□

Exercice 4.1.1 *Montrer que la fonction projection sur le convexe fermé K est caractérisée par*

$$(x - P_K(x), y - P_K(x)) \leq 0\ \forall\ y \in K. \tag{4.5}$$

et qu'elle vérifie

$$\|P_K(x) - P_K(y)\|_2 \leq \|x - y\|_2\ \forall\ x, y\ \in \mathbb{R}^n. \tag{4.6}$$

Cette application projection est donc continue.

Exercice 4.1.2 *Montrer que si*

$$K = [0, +\infty|^p = \{(x_i)_{1 \leq i \leq p}, \textit{tel que } x_i \geq 0 \quad i = 1, ..., p\},$$

alors $P_K(x) = (\max(0, x_i))_{1 \leq i \leq p}$.

Cas de contraintes d'égalité linéaires

Commençons par le cas d'un problème de minimisation avec contraintes d'égalités où on considère

$$\min_{Cx = d} J(x), \tag{4.7}$$

pour $C \in \mathcal{M}_{m,n}(\mathbb{R})$ et $d \in \mathbb{R}^m$.

Proposition 4.1.4 *On suppose que J est différentiable en $\bar{x}$. Si le vecteur $\bar{x}$ est solution de (4.7), il existe alors $\bar{\lambda} \in \mathbb{R}^m$ tel que*

$$\nabla J(\bar{x}) + C^T \bar{\lambda} = 0. \tag{4.8}$$

Le multiplicateur $\bar{\lambda}$ est unique si la matrice C est de rang m. ($rg\,(C) = m$).

Preuve.
Il est clair que l'ensemble des contraintes $K := \{x \in \mathbb{R}^n \text{ tel que } Cx = d\}$ est un convexe fermé de $\mathbb{R}^n$. Comme J est différentiable en $\bar{x}$ qui est un minimum de J sur K, d'après la condition d'optimalité (4.1.3), $\bar{x}$ vérfie

$$(\nabla J(\bar{x}), x - \bar{x}) \geq 0 \ \ \forall \ \ x \in K$$

Or si $x \in K$, alors $Cx = d = C\bar{x}$ et par conséquent $C(x - \bar{x}) = 0$. Ainsi $x - \bar{x} \in \ker C$ où $\ker C$ est le noyau de C.

Inversement, si $y \in \ker C$, alors $y = x - \bar{x}$ pour $x \in K$. On a alors

$$(\nabla J(\bar{x}), y) \geq 0 \ \ \forall \ \ y \in \ker C.$$

Comme $\ker C$ est un sous espace vectoriel, on en déduit que

$$(\nabla J(\bar{x}), y) = 0 \ \ \forall \ y \in \ker C.$$

Ceci est équivalent à $\nabla J(\bar{x}) \in (\ker C)^{\perp}$. Or, on sait que $(\ker C)^{\perp} = \mathrm{Im}(C^T)$, il existe donc $\bar{\lambda} \in \mathbb{R}^m$ tel que

$$\nabla J(\bar{x}) + C^T \bar{\lambda} = 0.$$

L'unicité de ce multiplicateur $\bar{\lambda}$ est évidente si C est surjective (de rang m) puisque cette dernière propriété est équivalente à $\ker C^T = 0$.

Remarque 4.1.1 *Si $C = (c_{ij})_{1 \leq i \leq m, 1 \leq j \leq n} \in \mathcal{M}_{m,n}(\mathbb{R})$ et $d = (d_1, ..., d_m)^T \in \mathbb{R}^m$, la contrainte $Cx = d$ implique que x est solution d'un système linéaire de m équations et à n inconnues, s'écrivant donc de la forme*

$$\left\{ \begin{array}{rcccccccc} g_1(x) := & c_{11}x_1 & + & ... & + & c_{1n}x_n - d_1 & = & 0 \\ \vdots & & & & \vdots & & \vdots & \\ g_m(x) := & c_{m1}x_1 & + & ... & + & c_{mn}x_n - d_m & = & 0 \end{array} \right.$$

où $C_i = (c_{i1}, ..., c_{in})^T, i = 1, ..., m$ représentent les lignes de la matrice C.

Si $\bar{\lambda} = \sum_{i=1}^{m} \lambda_i \tilde{e}_i$, avec $(\tilde{e}_i)_{1 \leq i \leq m}$ est la base canonique de $\mathbb{R}^m$, alors

$$C^T \bar{\lambda} = \sum_{i=1}^{m} \lambda_i C^T e_i = \sum_{i=1}^{m} \lambda_i C_i.$$

La condition d'optimalité s'écrit alors

$$\nabla J(\bar{x}) + \sum_{i=1}^{m} \lambda_i C_i = \nabla J(\bar{x}) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\bar{x}) = 0$$

Cas de contraintes d'inégalité linéaires

Soit le problème

$$\min_{Bx \leq c} J(x), \tag{4.9}$$

Pour $B \in \mathcal{M}_{np}(\mathbb{R})$ et $c \in \mathbb{R}^p$.

On adopte ici les notations suivantes :

$$\forall\, \mu = (\mu_1, ..., \mu_p) \in \mathbb{R}^p,\ \mu \geq 0 \text{ si et seulement si } \ \mu_j \geq 0,\ \ \forall j = 1, ..., p.$$

Proposition 4.1.5 *Si $\bar{x}$ réalise un minimum de J sur $K = \{x \in \mathbb{R}^n,$ tel que $Bx = c$ et J est différentiable en $\bar{x}$, il existe alors $\bar{\mu} \in \mathbb{R}^p$ tel que*

$$\bar{\mu} \geq 0,\ \nabla J(\bar{x}) + B^T\bar{\mu} = 0 \ \ et\ \ \bar{\mu}_j(B\bar{x} - c)_j = 0,\ \textit{pour tout}\ 1 \leq j \leq p. \tag{4.10}$$

Ce vecteur $\bar{\mu}$ est appelé aussi **multiplicateur de Lagrange** *et il est unique si la matrice B est de rang p ($rg(B) = p$).*

Preuve. On se limitera dans la démonstration pour une matrice B de rang p où pour tout vecteur $y \in \mathbb{R}^p$ il existe $x \in \mathbb{R}^n$ tel que $y = Bx$.

Soit $d = B\bar{x}$. Alors $d \leq c$ et tout vecteur x vérifiant $Bx = d$ est un vecteur de K. Clairement, $\bar{x}$ est aussi solution du problème à contraintes égalités linéaires

$$\min_{Bx=d} J(x). \tag{4.11}$$

Il existe donc $\bar{\mu} \in \mathbb{R}^p$ tel que

$$\nabla J(\bar{x}) + B^T\bar{\mu} = 0.$$

Montrons que $\bar{\mu}_j \geq 0$ et que $\bar{\mu}_j(B\bar{x} - d)_j = 0$, pour tout $1 \leq j \leq p$.
Pour $\varepsilon > 0$ on a $d_j - \varepsilon \leq d_j \leq c_j$ et pour $j \in \{1, ..., p\}$, si x est tel que $Bx = d - \varepsilon e_j$, pour e_j j-ième vecteur de la base canonique de $\mathbb{R}^p$, alors x est dans K.
D'après l'inéquation d'Euler, on a

$$(\nabla J(\bar{x}), x - \bar{x}) \leq 0.$$

En remplaçant $\nabla J(\bar{x})$ par $-B^T\bar{\mu}$, on obtient

$$-(\bar{\mu}, Bx - B\bar{x}) = \varepsilon\bar{\mu}_j \geq 0.$$

Par conséquent $\bar{\mu}_j \geq 0$.
Soit maintenant $j \in \{1, ..., p\}$ tel que $(B\bar{x} - c)_j < 0$. Alors si $\varepsilon > 0$ suffisamment petit, tel que $d_j + \varepsilon \leq c_j$ et si y solution de $By = d + \varepsilon e_j$, alors $y \in K$ et on a

$$(\bar{\mu}, By - B\bar{x}) = -\varepsilon\bar{\mu}_j \geq 0.$$

Le réel positif μ_j est aussi négatif, donc nul. □

4.1.3 Cas de contraintes d'égalités et d'inégalités linéaires

Soit le problème

$$\min_{x\in K} J(x)$$

où

$$K := \{x \in \mathbb{R}^n \text{ tel que } Cx = d \text{ et } Bx \leq c\},$$

pour

$$C \in \mathcal{M}_{m,n}(\mathbb{R}), d \in \mathbb{R}^m, B \in \mathcal{M}_{p,n}(\mathbb{R}), c \in \mathbb{R}^p.$$

Proposition 4.1.6 *Si $\bar{x}$ réalise un minimum de J sur K et J est différentiable en $\bar{x}$, il existe $(\bar{\lambda}, \bar{\mu})^T \in \mathbb{R}^m \times \mathbb{R}^p$,tel que*

$$\nabla J(\bar{x}) + C^T\bar{\lambda} + B^T\bar{\mu} = 0 \text{ et } \bar{\mu} \geq 0,\ \mu_j(B\bar{x} - c)_j = 0,\ j = 1, ..., n.$$

Preuve. Il suffit d'appliquer le cas avec contraintes d'inégalités linéaires en remarquant que la contrainte d'égalité $Cx = d$ est équivalente à $Cx \leq d$ et $-Cx \leq -d$. □

4.1.4 Problème quadratique avec contraintes linéaires

Problème quadratique avec contraintes d'égalité

Soit le problème

$$\min_{Cx=d} \frac{1}{2}(Ax, x) - (b, x), \tag{4.12}$$

pour $A \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique, $C \in \mathcal{M}_{m,n}(\mathbb{R})$, et $b \in \mathbb{R}^n$ et $d \in \mathbb{R}^m$.

Proposition 4.1.7 *Si A est symétrique définie positive, alors le problème* (4.12) *admet une solution unique $\bar{x}$ caractérisée par :*

$$\exists \bar{\lambda} \in \mathbb{R}^m \text{ tel que } \begin{pmatrix} A & C^T \\ C & 0 \end{pmatrix} \begin{pmatrix} \bar{x} \\ \bar{\lambda} \end{pmatrix} = \begin{pmatrix} b \\ d \end{pmatrix}. \tag{4.13}$$

Preuve.
Il s'agit d'un problème de minimisation dont la fonction objectif est infinie à l'infini, strictement convexe et différentiable, donc admettant un minimum unique $\bar{x}$. Comme il s'agit d'un problème de minimisation avec contraintes d'égalités linéaires et d'après (4.8), alors il existe un multiplicateur de Lagrange $\bar{\lambda} \in \mathbb{R}^m$ tel que

$$\nabla J(\bar{x}) + C^T\bar{\lambda} = A\bar{x} - b + C^T\bar{\lambda} = 0.$$

Le problème (4.12) est un problème convexe, donc $\bar{x}$ solution de (4.12) si et seulement si il existe $\bar{\lambda} \in \mathbb{R}^p$ tel que

$$\begin{cases} A\bar{x} + C^T\bar{\lambda} &= b \\ C\bar{x} &= d \end{cases}$$

vérifiant donc

$$\begin{pmatrix} A & C^T \\ C & 0 \end{pmatrix} \begin{pmatrix} \bar{x} \\ \bar{\lambda} \end{pmatrix} = \begin{pmatrix} b \\ d \end{pmatrix}.$$

□

Remarque 4.1.2 *Si A est symétrique semi définie positive, alors le problème* (4.12) *peut ne pas admettre de solution comme il peut avoir une ou plusieurs solutions. Comme il s'agit d'un problème convexe, alors $\bar{x}$ est une solution de* (4.12) *si et seulement si elle vérifie* (4.13)*.*

Exemple 4.1.2 *Soit le problème*

$$\min_{x \in K} J(x) \tag{4.14}$$

pour

$$J(x) = \tfrac{1}{2}x_1^2 + \tfrac{1}{2}x_2^2 + \tfrac{1}{2}x_3^2 - x_1x_2 - x_1 - 2x_2 - x_3$$

et

$$K = \{(x_1, x_2, x_3)^T \in \mathbb{R}^n \text{ tel que } x_1 + x_2 - x_3 = 1, 2x_3 + x_2 = 2\}.$$

Il s'agit d'un problème quadratique de type (4.12)*, pour*

$$A = \begin{pmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \text{ et } b = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}.$$

La matrice A ici est semi-définie positive, (ses valeurs propres sont 0, 1 et 2). Si on note $c_1 = \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}$ et $c_2 = \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix}$ les deux vecteurs représentant les deux lignes de la matrice C, alors

$$C^T\bar{\lambda} = \bar{\lambda}_1 c_1 + \bar{\lambda}_2 c_2.$$

un vecteur $\bar{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$ est solution de ce problème si et seulement si il existe $\bar{\lambda} = (\lambda_1, \lambda_2)^T \in \mathbb{R}^2$ tel que

$$\begin{cases} x_1 - x_2 + \lambda_1 = 1 \\ -x_1 + x_2 + \lambda_1 + \lambda_2 = 2 \\ x_3 - \lambda_1 + 2\lambda_2 = 1 \\ x_1 + x_2 - x_3 = 2 \\ x_2 + 2x_3 = 1 \end{cases}$$

Ce système admet comme solution $\bar{x} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$ et $\bar{\lambda} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ Donc $\bar{x} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$ est la solution cherchée.

Problème quadratique avec contraintes d'inégalité

On considère le problème

$$\min_{Bx \leq c} \frac{1}{2}(Ax, x) - (b, x), \tag{4.15}$$

pour $A \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique, $B \in \mathcal{M}_{p,n}(\mathbb{R})$, et $b \in \mathbb{R}^n$ et $c \in \mathbb{R}^p$.

Proposition 4.1.8 *Si A est symétrique définie positive, alors le problème* (4.15) *admet une solution unique $\bar{x}$ si et seulement si*

$$\exists\, \bar{\mu} \in \mathbb{R}^p,\ \bar{\mu} \geq 0 \text{ tel que } \begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} \bar{x} \\ \bar{\mu} \end{pmatrix} = \begin{pmatrix} b \\ c \end{pmatrix} \text{ et } \bar{\mu}_j (B\bar{x} - c)_j = 0, \text{ pour tout } 1 \leq j \leq p.$$

Preuve.
Similaire au cas avec contraintes d'égalités où on utilise ici (4.1.5). □

Remarque 4.1.3 *On peut vérifier facilement que ce multiplicateur $\bar{\lambda}$ est solution de*

$$\max_{\substack{\lambda \in \mathbb{R}^p \\ \lambda \geq 0}} \mathcal{L}(\bar{x}, \lambda)$$

où

$$\mathcal{L}(x, \lambda) = \tfrac{1}{2}(Ax, x) - (b, x) + (Bx - c, \lambda)$$

est appelé le Lagrangien du problème quadratique (4.15).

Exercice 4.1.3 *Soit*

$$K = \mathbb{R}^p_+ = \{x \in \mathbb{R}^p, x \geq 0\}.$$

Montrer que $P_K(x) = \max(0, x) = (\max(x_i, 0))_{1 \leq i \leq p}$

4.2 Quelques algorithmes

Dans cette section on donnera deux méthodes numériques pour calculer la solution d'un problème d'optimisation avec contraintes.

On se limitera dans ce chapitre à étudier la méthode du gradient projeté pour un problème quadratique pour K quelconque et la méthode d'Uzawa pour un problème quadratique avec contraintes d'inégalité.

4.2.1 Méthode du gradient projeté

On rappelle que dans le cas du problème sans contraintes, la méthode du gradient à pas fixe α s'écrit, pour $x^{(0)}$ donné,

$$x^{(k+1)} = x^{(k)} - \alpha \nabla J(x^{(k)}).$$

Lorsque qu'on minimise sur une partie K supposée fermée convexe, on n'est pas sur qu'à chaque étape k, l'itéré $x^{(k)}$ reste dans K même si on part d'un $x^{(0)}$ admissible. Pour cette raison, on choisit

$$x^{(k+1)} = P_K(x^{(k)} - \alpha \nabla J(x^{(k)})), \quad x^{(0)} \in K,$$

si P_k est calculable, où P_k désigne la projection sur le fermé convexe K. L'algorithme du gradient projeté s'écrit :

1. On choisit $x^{(0)} \in K$
2. Pour $k \geq 0$, on calcule $\left| \begin{array}{c} d^{(k)} = -\nabla J(x^{(k)}), \\ x^{(k+1)} = P_K(x^{(k)} + \alpha d^{(k)}) \end{array} \right.$

Proposition 4.2.1 *Soit A une matrice symétrique définie positive et K est un convexe fermé non vide de $\mathbb{R}^n$. Si $0 < \alpha < \dfrac{2}{\rho(A)}$, alors la suite générée par la méthode du gradient projeté converge vers $\bar{x}$ réalisant le minimum de $J(x) = \frac{1}{2}(Ax, x) - (b, x)$ sur K.*

Preuve.
Commençons par montrer que si $\bar{x}$ est la solution unique du problème (P), et si J est différentiable en $\bar{x}$, alors, pour tout $\alpha > 0$, on a

$$\bar{x} = P_K(\bar{x} - \alpha \nabla J(\bar{x})).$$

En effet, comme $\bar{x}$ est un minimum de J sur le convexe fermé K, alors d'après l'inéquation d'Euler, on a pour tout $x \in K$,

$$(\nabla J(\bar{x}), x - \bar{x}) \geq 0,$$

donc, si $\alpha > 0$, pour tout $x \in K$,

$$(\bar{x} - (\bar{x} - \alpha \nabla J(\bar{x}), x - \bar{x}) = \alpha(\nabla J(\bar{x}), x - \bar{x}) \geq 0.$$

De la propriété caractérisant la projection sur un convexe (voir (4.5)) vient

$$\bar{x} = P_K(\bar{x} - \alpha \nabla J(\bar{x})).$$

Montrons que $(x^{(k)})$ converge vers $\bar{x}$. On a pour $k \geq 0$,

$$\begin{array}{rcl} \|x^{(k+1)} - \bar{x}\|_2 & = & \|P_K(x^{(k)} - \alpha \nabla J(x^{(k)})) - P_K(\bar{x} - \alpha \nabla J(\bar{x}))\|_2 \\ & \leq & \|(x^{(k)} - \alpha \nabla J(x^{(k)})) - (\bar{x} - \alpha \nabla J(\bar{x})\|_2 \\ & = & \|(I - \alpha A)(x^{(k)} - \bar{x})\|_2 \\ & \leq & \|I - \alpha A\|_2 \|x^{(k)} - \bar{x}\|_2 \end{array}.$$

Par récurrence, on obtient alors

$$\|x^{(k)} - \bar{x}\|_2 \leq \|I - \alpha A\|_2^k \|x^{(0)} - \bar{x}\|_2 = \rho(I - \alpha A)^k \|x^{(0)} - \bar{x}\|_2,$$

car $I - \alpha A$ est symétrique.
Si $0 < \rho < \frac{2}{\rho(A)}$, alors $\rho(I - \alpha A) < 1$ et donc la suite $(x^{(k)})$ converge vers $\bar{x}$. □

Remarque 4.2.1 *La méthode du gradient projeté devient intéressante si la projection sur le convexe K est facile à calculer. C'est le cas par exemple d'une projection sur $\mathbb{R}^p_+$ ou sur un pavé de $\mathbb{R}^p$. En général, le calcul de la projection sur K est délicat. D'où la difficulté de la mise en oeuvre de cette méthode malgré son apparente simplicité.*

4.2.2 Méthode d'Uzawa

On rappelle que pour une matrice A symétrique définie positive, si $\bar{x}$ est le minimum de

$$\min_{Bx \leq c} \tfrac{1}{2}(Ax, x) - (b, x),$$

alors il existe $\bar{\lambda} \in \mathbb{R}^p_+$, tel que

$$A\bar{x} + B^T\bar{\lambda} = b \text{ et } \lambda_j(B\bar{x} - c)_j = 0,\ j = 1, ..., p$$

et que ce multiplicateur $\bar{\lambda}$ est solution de

$$\max_{\lambda \in \mathbb{R}^p_+} \mathcal{L}(\bar{x}, \lambda)$$

où

$$\mathcal{L}(x, \lambda) = \tfrac{1}{2}(Ax, x) - (b, x) + (Bx - c, \lambda).$$

Comme $x \mapsto \mathcal{L}(x, \bar{\lambda})$ est convexe et elle vérifie

$$\frac{\partial \mathcal{L}}{\partial x}(\bar{x}, \bar{\lambda}) = A\bar{x} - b + B^T\bar{\lambda} = 0,$$

donc $\bar{x}$ est aussi solution de

$$\min_{x \in \mathbb{R}^n} \mathcal{L}(x, \bar{\lambda}).$$

Ainsi, $\forall x \in \mathbb{R}^n, \forall \lambda \in \mathbb{R}^p_+$, on a

$$\mathcal{L}(\bar{x}, \lambda) \leq \mathcal{L}(\bar{x}, \bar{\lambda}) \leq \mathcal{L}(x, \bar{\lambda}).$$

On dit que $(\bar{x}, \bar{\lambda})$ est un **point selle** de $\mathcal{L}$.

Le principe de la méthode d'Uzawa consiste à calculer numériquement ce point selle. On se donne $(x^{(0)}, \lambda^{(0)})$, et si à l'itération k on connaît $(x^{(k)}, \lambda^{(k)})$, on commence par calculer $x^{(k+1)}$ solution de $\min_{x \in \mathbb{R}^n} \mathcal{L}(x, \lambda^{(k)})$ sur $\mathbb{R}^n$ solution donc de système $Ax = b - B^T\lambda^{(k)}$, puis calculer

$$\lambda^{(k+1)} = P_{\mathbb{R}^p_+}(\lambda^{(k)} + \alpha \frac{\partial \mathcal{L}}{\partial \lambda}(x^{(k+1)}, \lambda^{(k)})) = \max(0, \lambda^{(k)} + \alpha(Bx^{(k)} - c)),$$

itéré obtenu par la méthode du gradient projeté, de pas α, appliquée au problème dual $\max_{\lambda \in \Lambda} \mathcal{L}(x^{(k)}, \lambda)$.

Algorithme d'Uzawa pour un problème quadratique avec contraintes inégalités

Si (P) est le problème

$$\min_{Bx\leq c} \tfrac{1}{2}(Ax,x)-(b,x),$$

alors l'algorithme de la méthode d'Uzawa de pas α s'écrit :

1. On choisit $(x^{(0)},\lambda^{(0)})\in\mathbb{R}^n\times\mathbb{R}^p_+$ et $\alpha>0$.
2. pour $k\geq 0$, on calcule
   $x^{(k+1)}$ solution de $Ax=b-B^T\lambda^{(k)}$
   $\lambda^{(k+1)}=\max(0,\lambda^{(k)}+\alpha(Bx^{(k+1)}-c))$

On peut utiliser une des méthodes numériques directes ou itératives pour déterminer $x^{(k+1)}$

Proposition 4.2.2 *Soit A une matrice symétrique définie positive et soit λ_1 sa plus petite valeur propre. Alors si $0<\alpha<\dfrac{2\lambda_1}{\|B\|_2^2}$, la suite $(x^{(k)})$ générée par l'algorithme d'Uzawa converge vers $\bar{x}$ solution du problème $\min_{Bx\leq c}\frac{1}{2}(Ax,x)-(b,x)$.*

Preuve. On a

$$A\bar{x}-b+B^T\bar{\lambda}=0$$

et

$$Ax^{(k+1)}-b+B^T\lambda^{(k)}=0.$$

Donc

$$A(x^{(k+1)}-\bar{x})=-B^T(\lambda^{(k)}-\bar{\lambda}).$$

Si on applique le produit scalaire avec $x^{(k+1)}-\bar{x}$, on obtient

$$(A(x^{(k+1)}-\bar{x}),x^{(k+1)}-\bar{x})=-(B^T(\lambda^{(k)}-\bar{\lambda}),x^{(k+1)}-\bar{x})=-(\lambda^{(k)}-\bar{\lambda},B(x^{(k+1)}-\bar{x})).$$

Ainsi

$$\lambda_1\|x^{(k+1)}-\bar{x}\|_2^2+(\lambda^{(k)}-\bar{\lambda},B(x^{(k+1)}-\bar{x}))\leq 0. \tag{4.16}$$

Or

$$\|\lambda^{(k+1)}-\bar{\lambda}\|_2=\|P_{\mathbb{R}^p_+}(\lambda^{(k)}+\alpha(Bx^{(k+1)}-c)-P_{\mathbb{R}^p_+}(\bar{\lambda}+\alpha(B\bar{x}-c)\|_2\leq\|\lambda^{(k)}-\bar{\lambda}+\alpha B(x^{(k+1)}-\bar{x})\|_2.$$

$$\|\lambda^{(k+1)}-\bar{\lambda}\|_2^2\leq\|\lambda^{(k)}-\bar{\lambda}\|_2^2+2\alpha(\lambda^{(k)}-\bar{\lambda},B(x^{(k+1)}-\bar{x}))+\alpha^2\|B(x^{(k+1)}-\bar{x})\|_2^2.$$

D'après (4.16), on déduit que

$$\|\lambda^{(k+1)}-\bar{\lambda}\|_2^2\leq\|\lambda^{(k)}-\bar{\lambda}\|_2^2-2\alpha\lambda_1\|x^{(k+1)}-\bar{x}\|_2^2+\alpha^2\|B\|_2^2\|x^{(k+1)}-\bar{x}\|_2^2 \tag{4.17}$$

Si on choisit α tel que $0 < \alpha < \dfrac{2\lambda_1}{\|B\|_2^2}$, alors $\beta = \alpha(2\lambda_1 - \alpha\|B\|_2^2) > 0$ et

$$\|\lambda^{(k+1)} - \bar{\lambda}\|_2^2 \leq \|\lambda^{(k)} - \bar{\lambda}\|_2^2 - \beta\|x^{(k+1)} - \bar{x}\|_2^2. \tag{4.18}$$

Ainsi la suite $(\|\lambda^{(k)} - \bar{\lambda}\|_2^2)$ est décroissante, donc convergente et par conséquent $(\|x^{(k+1)} - \bar{x}\|_2^2)$ converge vers 0 puisque

$$\beta\|x^{(k+1)} - \bar{x}\|_2^2 \leq \|\lambda^{(k)} - \bar{\lambda}\|_2^2 - \|\lambda^{(k+1)} - \bar{\lambda}\|_2^2.$$

□

Exercice 4.2.1 *Ecrire la méthode d'Uzawa pour un problème quadratique avec contraintes égalités.*

Références

[**1**] R. Bessi, Optimisation quadratique. ENIT, 2010.

[**2**] P.G. Ciarlet, Introduction à l'analyse numérique matricielle et à l'optimisation. Edition Masson, 1982.

[**3**] P.G. Ciarlet, B. Miara, J.M. Thomas, Exercices d'analyse numérique matricielle et d'optimisation avec solutions. Edition Masson, 1991.

[**4**] P. Ciarlet, H. Zidani, Optimisation quadratique. Cours AO101, ENSTA, 2011.

[**5**] H. El Fekih, K. Griba, T. Hadhri, Cours d'analyse numérique matricielle. ENIT, 1996.