Laurence Lamoulie
Chrysostome Baskiotis

ANALYSE NUMÉRIQUE

Fascicule 2

Algèbre linéaire

Année 2009 – 2010

INTRODUCTION

Dans la deuxième partie de ce cours nous présentons des méthodes de résolution des systèmes linéaires ainsi que des méthodes d'inversion matricielle.

1

MÉTHODES DE RÉSOLUTION DE SYSTÈMES LINÉAIRES

1.1 Introduction

L'étude des méthodes de résolution des systèmes linéaires est une étape obligatoire dans un cours d'analyse numérique. En effet, presque tous les calculs passent par la résolution d'un système. On en rencontre dans la discrétisation de problèmes aux limites, dans les problèmes d'approximation par exemple.

La résolution de systèmes n'est plus une opération destinée à être menée à la main. Du fait de l'apparition des ordinateurs, et du développement de méthodes qui leur sont adaptées, on peut envisager de résoudre de grands systèmes. Si la matrice est pleine, c'est à dire contient peu de zéros, on pourra "seulement" résoudre des systèmes à quelques centaines de milliers d'inconnues. Si la matrice est creuse, c'est à dire contient beaucoup de zéros, la résolution de systèmes à plusieurs millions d'inconnues est possible.

La résolution d'un système ne doit pas être comprise comme l'obtention de la solution exacte en un nombre fini d'étapes. Cette vision ne concerne que les méthodes dites "directes". On est souvent amené à accepter un compromis : obtenir une solution approchée en un nombre fini

d'étapes, sachant que la solution exacte serait obtenue en un nombre infini d'étapes, chose impossible à mettre en oeuvre. Ce compromis est le principe même des méthodes itératives. C'est devenu la solution la plus répandue pour résoudre les grands systèmes.

En plus de la taille de la matrice, ses propriétés sont aussi un élément déterminant : certaines assurent la convergence a priori de méthodes itératives. Dans le cas où elles ne sont pas vérifiées par une matrice, ou bien quand on ne pas peut démontrer qu'elles le sont, la méthode itérative est choisie aux risques et périls de l'utilisateur. C'est un risque à éviter, surtout dans des applications industrielles, parfois sensibles. Des navettes spatiales ont explosé pour moins que ça...

La résolution des systèmes linéaires est donc à la fois une question de théorie et une question de pratique : il faut choisir le bon algorithme, s'assurer que les conditions sont réunies pour l'utiliser, qu'il sera numériquement stable et assez rapide pour les besoins de la cause.

Il faut repenser la résolution de systèmes dans le cadre d'une mise en oeuvre informatique : nous allons voir sur deux exemples simples que les méthodes utilisables à la main ne sont pas, en général, adaptées à la résolution de systèmes en machine, notamment pour des questions de temps de calcul. Bien sur, de nombreux logiciels proposent des bibliothèques de résolution de systèmes linéaires, qui s'appuient sur des bibliothèques publiques d'algèbre linéaire de base (BLAS). Ces dernières ont été précieuses dans les progrès du calcul scientifique matriciel ; elles permettent d'effectuer de façon optimale les opérations algébriques de base. Les principaux logiciels sont soit du domaine public (LAPACK= Linear Algebra PACKage par exemple), soit des logiciels commerciaux (NAG = Numerical Algorithms Group, IMSL = International Mathematical and Statistical Library, MATLAB,...). Mais en tout état de cause, la possession d'une voiture de course ne permet pas d'avancer quand on ne sait pas la conduire...

1.1.1 Exemple 1 : Économie : analyse d'entrées-sorties

On veut déterminer l'équilibre entre la demande et l'offre de certains biens. Dans le modèle de production considéré, $m \geq n$ usines produisent n produits différents. Elles doivent faire face à une demande interne (l'entrée) nécessaire au fonctionnement propre des usines, ainsi qu'à une demande externe (la sortie) provenant des consommateurs.

La principale hypothèse du modèle de Leontieff (1930) [1] est que le modèle de production est linéaire, c'est à dire que la sortie est proportionnelle à l'entrée utilisée. Sous cette hypothèse, l'activité des usines est entièrement décrite par deux matrices : la matrice d'entrée $\mathbf{C} = (c_{ij}) \in \mathbb{R}^{n\times m}$ et la matrice de sortie $\mathbf{P} = (p_{ij}) \in \mathbb{R}^{n\times m}$. Le coefficient c_{ij} (resp. p_{ij}) représente la quantité du i^{ème} bien absorbé (resp. produit) pour la j^{ème} usine sur une période fixée. La matrice $\mathbf{A} = \mathbf{P} - \mathbf{C}$ est appelée matrice d'entrée-sortie : un a_{ij} positif (resp. négatif) désigne la quantité du i^{ème} bien produit (resp. absorbé) par la j^{ème} usine. Enfin, on peut raisonnablement supposer que le système de production satisfait à la demande du marché, qu'on peut représenter par un vecteur $b = (b_i) \in \mathbb{R}^n$ (vecteur de la demande finale). La composante b_i représente la quantité du i^{ème} bien absorbé sur le marché. L'équilibre est atteint lorsque le vecteur $x = (x_i) \in \mathbb{R}^m$ représentant

1. Wassily Leontieff a reçu en 1973 le prix Nobel d'économie pour ses travaux

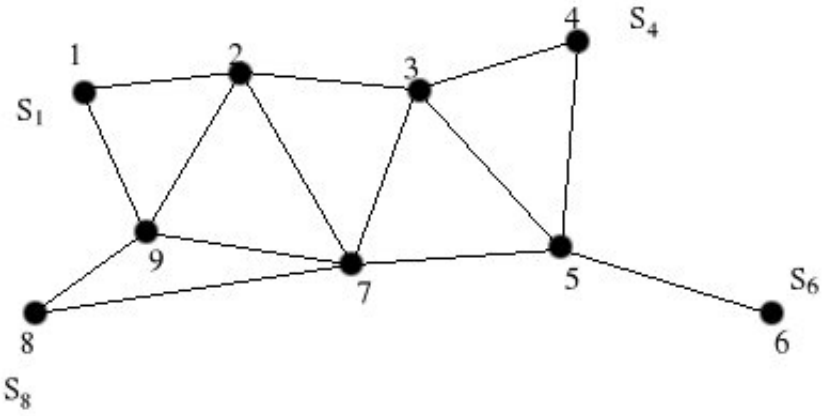
la production totale est égal à la demande totale, c'est à dire

$$\mathbf{Ax} = \mathbf{b}, \text{ où } \mathbf{A} = \mathbf{P} - \mathbf{C}$$

1.1.2 Exemple 2 : Résolution d'un problème de réseaux

Un réseau est un ensemble de nœuds P_i et d'arêtes $E_{i,j}$ reliant certains de ces nœuds :

- lignes électriques,
- canalisations d'eaux, égouts,...



Dans chaque arête circule un fluide ; à chaque nœud est associé un potentiel. L'intensité (ou le débit) du fluide est proportionnelle à la différence de potentiel entre les deux extrémités de l'arête où il circule ; c'est la loi d'Ohm pour les circuits électriques :

$$q_{i,j} = k_{i,j}(u_i - u_j)$$

Une loi physique de conservation (de Kirchoff dans le cas électrique) impose un équilibre : la somme algébrique des intensités en chaque nœud est égale à la valeur de la source (ou du puits) qu'il figure.

Au nœud P_i, on a dans le cas du circuit électrique :

$$S_i = \sum_j q_{i,j} = \sum_j k_{i,j}(u_i - u_j)$$

Cette somme peut être étendue aux nœuds adjacents de P_i, les équations d'équilibre s'écrivent :

$$\begin{cases} S_1 = k_{1,2}(u_1 - u_2) + k_{1,9}(u_1 - u_9) \\ 0 = k_{2,1}(u_2 - u_1) + k_{2,9}(u_2 - u_9) + k_{2,7}(u_2 - u_7) + k_{2,3}(u_2 - u_3) \\ \cdots = \cdots \\ 0 = k_{9,1}(u_9 - u_1) + k_{9,2}(u_9 - u_2) + k_{9,7}(u_9 - u_7) + k_{9,8}(u_9 - u_8) \end{cases}$$

de sorte que l'équilibre du système est connu en résolvant le système linéaire

$$\mathbf{Au} = \mathbf{S}$$

avec une matrice $\mathbf{A}$ dont les coefficients non nuls sont représentés ci-dessous par une étoile :

$$\mathbf{A} = \begin{bmatrix} * & * & & & & & & * \\ * & * & * & & & * & & * \\ & * & * & * & & * & & \\ & & * & * & * & & & \\ & & & * & * & * & * & \\ & & & & * & * & & \\ & * & * & & * & & * & * \\ & & & & & & * & * \\ * & * & & & & & * & * \end{bmatrix}$$

Le second membre est défini par $S^T = (S_1, 0, 0, S_4, 0, S_6, 0, S_8, 0)$.

1.1.3 Comment résoudre ces systèmes

La première idée qui vient est de résoudre ce système en utilisant les formules de Cramer, dites aussi "méthode des déterminants". Ces formules fournissent une solution exacte du système linéaire $\mathbf{Ax} = \mathbf{b}$ dans $\mathbb{R}^n$ avec $\mathbf{x} = (x_1, x_2, ...x_n)^T$ donnée par :

$$x_i = \frac{\det \mathbf{A}^{(i)}}{\det \mathbf{A}}$$

où $A^{(i)}$ désigne la matrice obtenue en substituant dans A la colonne i par le second membre du système.

Il y a donc $N+1$ déterminants à calculer, donnés dans le cas général pour une matrice $\mathbf{A}$ quelconque par :

$$\det(A) = \sum_{\sigma \in P_n} \epsilon(\sigma) a_{1,\sigma(1)} \times a_{2,\sigma(2)} \cdots \times a_{n,\sigma(n)}$$

où P_n désigne l'ensemble des permutations de $\{1, ...n\}$ qui compte $n!$ éléments et $\epsilon(\sigma)$ est $+1$ ou -1, la signature de la permutation. Le coût du calcul d'un tel déterminant est donc tel que :

$$c(n) \sim n \times n! \text{ opérations élémentaires}$$

et le coût global de la méthode est donc

$$C(n) \sim n \times (n-1)! \text{ opérations}$$

Le coût de la résolution d'un système 100×100 peut alors être évalué par la formule de Stirling ($n! \sim n^{n+1/2} e^{-n} \sqrt{2\pi}$). Cela conduit à l'évaluation :

$$C(n) \sim 9.4 \times 10^{161}$$

opérations élémentaires. Sur un ordinateur qui réalise 10^9 opérations flottantes par seconde (1 gigaflop), on devra attendre la solution pendant environ 3×10^{145} années !

1.2 Les systèmes faciles à résoudre

Afin de construire des méthodes qui simplifient la résolution d'un système sans passer par des calculs inappropriés pour une machine, voyons deux cas de systèmes dont la résolution est simple à programmer.

1.2.1 Systèmes diagonaux

Si **A** est une matrice diagonale, c'est à dire si

$$\mathbf{A} = (a_{i,j})_{1 \leq i,j \leq n} \text{ avec } a_{i,j} = 0 \text{ si } i \neq j$$

le système $Ax = b$ est immédiatement résolu du fait que

$$x_i = \frac{1}{a_{ii}} b_i$$

L'algorithme correspondant est donné par :

Algorithme : `Matrice diagonale`

{On suppose que $A_{kk} \neq 0$ }
Pour i ← 1 **à** n **faire**
$x_i \leftarrow b_i / a_{ii}$
FinPour

Le coût est $c(n) = n$ opérations élémentaires

1.2.2 Systèmes triangulaires

La matrice **A** d'un système triangulaire supérieur est telle que

$$\mathbf{A} = (a_{i,j})_{1 \leq i,j \leq n} \text{ avec } a_{i,j} = 0 \text{ si } i > j$$

Comme A est inversible,

$$a_{i,i} \neq 0 \text{ si } 1 \leq i \leq n$$

Le système s'écrit

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1 \\ a_{22}x_2 + \ldots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{nn}x_n &= b_n \end{aligned}$$

On le résout en remontant :

$$\begin{aligned} x_n &= \frac{b_n}{a_{n,n}} \\ x_{n-1} &= (b_{n-1} - a_{n-1,n}x_n)/a_{n-1,n-1} \\ &\vdots \\ x_1 &= (b_1 - a_{12}x_2 - \ldots - a_{1,n}x_n)/a_{1,1} \end{aligned}$$

On obtient alors la solution par un algorithme de *substitution rétrograde*, dit aussi simplement *algorithme de remontée* :

Algorithme : Algorithme de remontée

$x_n \leftarrow b_n/a_{n,n}$
Pour i ← n-1 **à** 1 **par pas de** -1 **faire**
$x_i \leftarrow \left(b_i - \sum_{j=i+1}^{n} a_{i,j}x_j\right)/a_{i,i}$
FinPour

Le coût du calcul d'un $x_k = \left(b_k - a_{k,k+1}x_{k+1} - \ldots - a_{k,n}x_n\right)/a_{k,k}$ se décompose en :

- $(n-k)$ additions,
- $(n-k)$ multiplications,
- 1 division

Le coût total de remontée est donc de

$$\sum_{k=1}^{n}(n-k) = n^2 - \frac{n(n-1)}{2} \sim \frac{n^2}{2} \text{ additions + multiplications}$$

soit n^2 opérations élémentaires.

EXERCICE 1.1 *Établir l'algorithme de résolution d'un système triangulaire inférieur*

1.3 Méthodes directes

L'idée des méthodes directes est de remplacer la résolution d'un système du type $\mathbf{Ax} = \mathbf{b}$ dans lequel la matrice $\mathbf{A}$ est pleine, par un système ou plusieurs systèmes plus facile(s) à résoudre car creux (i.e. de matrice triangulaire ou diagonale). Le système diagonal serait idéal puisque c'est à la fois le plus rapide et le moins coûteux à résoudre. Mais diagonaliser $\mathbf{A}$, c'est rechercher ses éléments propres et déterminer $\mathbf{P}$ et $\mathbf{D}$ vérifiant $\mathbf{A} = \mathbf{PDP}^{-1}$. L'idée est séduisante mais malheureusement inapplicable car la recherche d'éléments propres est beaucoup plus difficile numériquement que la résolution d'un système. L'alternative est donc de remplacer $\mathbf{A}$ par le produit de deux matrices triangulaires, notées en général $\mathbf{L}$ et $\mathbf{U}$, respectivement triangulaire inférieure et triangulaire supérieure. En effet on résout alors successivement deux systèmes

triangulaires :

$$\mathbf{Ly} = \mathbf{b} \text{ puis } \mathbf{Ux} = \mathbf{y}$$

dont la solution $\mathbf{x}$ vérifie évidemment $\mathbf{Ax} = \mathbf{b}$. On va voir dans ce qui suit comment on s'y prend numériquement.

1.3.1 Elimination de Gauss sans recherche de pivot

1.3.1.1 Factorisation LU

Soit à résoudre le système : Trouver $\mathbf{x} \in \mathbb{R}^n$ tel que

$$\mathbf{Ax} = \mathbf{b}$$

pour $\mathbf{A} \in \mathbb{R}^{n\times n}$, et $\mathbf{b} \in \mathbb{R}^n$. On suppose que $\mathbf{A}$ est inversible.

Le but est de se ramener à un système triangulaire. On procède par étapes :

Étape 1 : Elimination de l'inconnue x_1 des lignes 2 à n

Sous l'hypothèse que $a_{11} \neq 0$ on peut l'utiliser pour éliminer l'inconnue x_1 des lignes 2 à n. Le terme a_{11} est appelé **pivot** et on note pour la suite :

$$\pi_1 = a_{11}$$

La ligne i devient alors :

$$0x_1 + (a_{i2} - \frac{a_{i1}}{\pi_1}a_{12})x_2 + \ldots + (a_{in} - \frac{a_{i1}}{\pi_1}a_{1n})x_n = b_i - \frac{a_{i1}}{\pi_1}b_1$$

ce que l'on écrit encore

$$a_{i2}^{(2)}x_2 + \ldots + a_{in}^{(2)}x_n = b_i^{(2)}, \forall i > 1$$

en posant

$$\begin{aligned} a_{i2}^{(2)} &= a_{i2} - \frac{a_{i1}}{\pi_1}a_{12} \\ &\vdots \\ a_{in}^{(2)} &= a_{in} - \frac{a_{i1}}{\pi_1}a_{1n} \\ b_i^{(2)} &= b_i - \frac{a_{i1}}{\pi_1}b_1 \end{aligned}$$

Si on pose aussi, pour $1 \leq j \leq n$

$$a_{1j}^{(2)} = a_{1j}, \forall 1 \leq j \leq n \text{ et } b_1^{(2)} = b_1$$

on a obtenu le système équivalent :

$$\mathbf{A}^{(2)}\mathbf{x} = \mathbf{b}^{(2)}$$

avec

$$\mathbf{A}^{(2)} = \begin{bmatrix} a_{11}^{(2)} & a_{12}^{(2)} & \dots & a_{1n}^{(2)} \\ 0 & a_{22}^{(2)} & \dots & a_{2n}^{(2)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(2)} & \dots & a_{nn}^{(2)} \end{bmatrix}$$

On voit que

$$\mathbf{A}^{(2)} = \mathbf{M}^{(1)}\mathbf{A} \text{ où } \mathbf{M}^{(1)} = \begin{bmatrix} 1 & 0 & \dots & \dots & 0 \\ -\frac{a_{21}}{a_{11}} & 1 & \ddots & & \vdots \\ \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ -\frac{a_{n1}}{a_{11}} & 0 & \dots & 0 & 1 \end{bmatrix}$$

Le système s'écrit alors

$$\mathbf{A}^{(2)} = \mathbf{M}^{(1)}\mathbf{A}, \text{ et } \mathbf{b}^{(2)} = \mathbf{M}^{(1)}\mathbf{b}$$

Tout se passe comme si on avait prémultiplié à gauche le système initial par $\mathbf{M}^{(1)}$.

Étape k : Elimination de l'inconnue x_k des lignes $k+1$ à n

On suppose que l'on a pu itérer le procédé ci-dessus $k-1$ fois, c'est que l'on n'a jamais rencontré de pivot nul :

$$\pi_i = a_{ii}^{(i)} \neq 0 \text{ pour } 1 \leq i \leq k-1$$

On obtient alors le système équivalent :

$$\mathbf{A}^{(k)}\mathbf{x} = \mathbf{b}^{(k)}$$

avec $\mathbf{A}^{(k)}$ de la forme

$$\mathbf{A}^{(k)} = \begin{bmatrix} a_{11}^{(k)} & a_{12}^{(k)} & \dots & \dots & \dots & \dots & \dots & a_{1n}^{(k)} \\ 0 & a_{22}^{(k)} & \ddots & & & & & a_{2n}^{(k)} \\ \vdots & \ddots & a_{33}^{(k)} & \ddots & & & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & & & \vdots \\ 0 & \dots & \dots & 0 & a_{k-1,k-1}^{(k)} & \dots & \dots & a_{k-1,n}^{(k)} \\ \vdots & & & \vdots & 0 & & & \\ \vdots & & & \vdots & \vdots & \widetilde{\mathbf{A}^{(k)}} & & \\ 0 & \dots & \dots & 0 & 0 & & & \end{bmatrix} \tag{1.3.1}$$

où $\widetilde{\mathbf{A}^{(k)}}$ est une matrice carrée d'ordre $n-k+1$:

$$\widetilde{\mathbf{A}^{(k)}} = \begin{bmatrix} a_{kk}^{(k)} & \dots & a_{kn}^{(k)} \\ \vdots & & \vdots \\ a_{nk}^{(k)} & \dots & a_{nn}^{(k)} \end{bmatrix} \tag{1.3.2}$$

Comme précédemment, on doit effectuer une hypothèse sur la valeur du pivot :

$$\pi_k = a_{kk}^{(k)} \neq 0$$

On peut alors introduire la matrice

$$\mathbf{M}^{(k)} = \begin{bmatrix} 1 & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & 1 & \ddots & & & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & & & \vdots \\ \vdots & \vdots & 0 & 1 & \ddots & & & \vdots \\ \vdots & & & -\frac{a_{k+1,k}^{(k)}}{\pi_k} & \ddots & \ddots & & \vdots \\ \vdots & & & \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & & & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \dots & -\frac{a_{n,k}^{(k)}}{\pi_k} & 0 & \dots & 0 & 1 \end{bmatrix}$$

(1.3.3)

Remarquons que l'inverse de $\mathbf{M}^{(k)}$ est $\mathbf{L}^{(k)}$:

$$\mathbf{L}^{(k)} = \begin{bmatrix} 1 & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & 1 & \ddots & & & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & & & \vdots \\ \vdots & \vdots & 0 & 1 & \ddots & & & \vdots \\ \vdots & & & \frac{a_{k+1,k}^{(k)}}{\pi_k} & \ddots & \ddots & & \vdots \\ \vdots & & & \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & & & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \dots & \frac{a_{n,k}^{(k)}}{\pi_k} & 0 & \dots & 0 & 1 \end{bmatrix} \tag{1.3.4}$$

Soit

$$\mathbf{A}^{(k+1)} = \mathbf{M}^{(k)}\mathbf{A}^{(k)} \text{ et } \mathbf{b}^{(k+1)} = \mathbf{M}^{(k)}\mathbf{b}^{(k)}$$

alors

$$\mathbf{A}^{(k+1)}\mathbf{x} = \mathbf{b}^{(k+1)}$$

et

$$\mathbf{A}^{(k+1)} = \begin{bmatrix} a_{11}^{(k+1)} & a_{12}^{(k+1)} & \dots & \dots & \dots & \dots & \dots & a_{1n}^{(k+1)} \\ 0 & a_{22}^{(k+1)} & \ddots & & & & & a_{2n}^{(k+1)} \\ \vdots & \ddots & a_{33}^{(k+1)} & \ddots & & & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & & & \vdots \\ 0 & \dots & \dots & 0 & a_{k,k}^{(k+1)} & \dots & \dots & a_{k,n}^{(k+1)} \\ \vdots & & & \vdots & 0 & & & \\ \vdots & & & \vdots & \vdots & \widetilde{A^{(k+1)}} & & \\ 0 & \dots & \dots & 0 & 0 & & & \end{bmatrix}$$

où $\widetilde{\mathbf{A}^{(k+1)}}$ est une matrice carrée d'ordre $n-k$:

$$\widetilde{\mathbf{A}^{(k+1)}} = \begin{bmatrix} a_{k+1,k+1}^{(k+1)} & \dots & a_{k+1n}^{(k+1)} \\ \vdots & & \vdots \\ a_{nk+1}^{(k+1)} & \dots & a_{nn}^{(k+1)} \end{bmatrix}$$

Après n-1 étapes : Si on itère $n-1$ fois, et si les pivots apparus sont tous non nuls, soit :

$$\pi_i = a_{ii}^{(i)} \neq 0 \text{ pour } 1 \leq i \leq n-1$$

on arrive au système triangulaire équivalent :

$$\mathbf{A}^{(n)}\mathbf{x} = \mathbf{b}^{(n)}$$

avec

$$\mathbf{A}^{(n)} = \begin{bmatrix} \pi_1 & * & * & \cdots & * \\ 0 & \pi_2 & * & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & * \\ 0 & \cdots & \cdots & 0 & \pi_n \end{bmatrix}$$

qui est inversible, si de plus

$$\pi_n \neq 0$$

Bilan

Si on appelle $\mathbf{L}$ la matrice triangulaire inférieure avec des 1 sur la diagonale, et $\mathbf{L}^{(k)}$ la matrice définie en (1.3.4)

$$\mathbf{L} = \mathbf{L}^{(1)} \cdots \mathbf{L}^{(n-1)}$$

et U la matrice triangulaire supérieure

$$\mathbf{U} = \mathbf{A}^{(n)}$$

on a

$$\mathbf{A} = \mathbf{LU}$$

On dit qu'on a effectué une **factorisation de Gauss** ou **factorisation LU** de $\mathbf{A}$.

REMARQUE 1.3.1 *Il est aussi possible avec le même algorithme d'obtenir la factorisation* $\mathbf{LU}$ *d'une matrice de* $\mathcal{M}_m(\mathbb{C})$*, avec* $m \geq n$*, si les pivots qui apparaissent sont non nuls.*

1.3.1.2 Coût de la méthode d'élimination de Gauss

On estime le coût de l'élimination de x_k. Rappelons qu'à l'étape $k-1$, on obtient la matrice $\mathbf{A}_k$ donnée en (1.3.1) comportant le bloc $\widetilde{\mathbf{A}^{(k)}}$ donné par (1.3.2) :

$$\widetilde{\mathbf{A}^{(k)}} = \begin{bmatrix} a_{kk}^{(k)} & \dots & a_{kn}^{(k)} \\ \vdots & & \vdots \\ a_{nk}^{(k)} & \dots & a_{nn}^{(k)} \end{bmatrix}$$

On construit tout d'abord $\mathbf{M}^{(k)}$ donné par (1.3.3) donnée par

$$\mathbf{M}^{(k)} = \begin{bmatrix} 1 & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & 1 & \ddots & & & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & & & \vdots \\ \vdots & \vdots & 0 & 1 & \ddots & & & \vdots \\ \vdots & & & -\frac{a_{k+1,k}^{(k)}}{\pi_k} & \ddots & \ddots & & \vdots \\ \vdots & & & \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & & & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \dots & -\frac{a_{n,k}^{(k)}}{\pi_k} & 0 & \dots & 0 & 1 \end{bmatrix}$$

il faut pour cela diviser par le pivot sur $n-k$ lignes. On effectue donc $n-k$ **divisions**.

On construit ensuite le produit $\mathbf{A}^{(k+1)} = \mathbf{M}^{(k)}\mathbf{A}^{(k)}$ (puisque notre objectif est de déterminer à la fin $\mathbf{A}^{(n)}$). Pour cela il faut effectuer $(n-k)^2$ **additions et multiplications.**

Au total :

$$\sum_{k=1}^{n}(n-k)^2 = \frac{1}{3}n(n-1)(n-\frac{1}{2}) \sim \frac{n^3}{3} \textbf{additions et multiplications.}$$

$$\sum_{k=1}^{n}(n-k) = \frac{1}{2}n(n-1) \textbf{ divisions}$$

1.3.1.3 Utilisation de la factorisation LU

Calcul de déterminant

Une utilisation de la factorisation $\mathbf{LU}$ est le calcul de déterminant de $\mathbf{A}$: en effet, si $\mathbf{A}$ admet une factorisation $\mathbf{LU}$, on a

$$\det \mathbf{A} = \det(\mathbf{LU}) = \det \mathbf{L}. \det \mathbf{U}$$

Or $\mathbf{L}$ est triangulaire à diagonale unité donc de déterminant 1, ce qui donne finalement

$$\det \mathbf{A} = \det \mathbf{U} = \prod_{i=1}^{n} \pi_i$$

La factorisation $\mathbf{LU}$ permet donc de calculer le déterminant de $\mathbf{A}$ avec une complexité de l'ordre de $\frac{n^3}{3}$ additions et multiplications, au lieu de $n!$ avec la formule du déterminant !

Résolution de systèmes linéaires

N'oublions pas notre objectif initial : résoudre un système linéaire. Il est clair que la factorisation $\mathbf{LU}$ permet de résoudre successivement deux systèmes triangulaires :

$$\mathbf{Ly} = \mathbf{b} \text{ puis } \mathbf{Ux} = \mathbf{y}$$

dont la solution $\mathbf{x}$ vérifie $\mathbf{Ax} = \mathbf{b}$.

Au delà de ce point, si on dispose de plusieurs systèmes linéaires de seconds membres différents mais de même matrice $\mathbf{A}$:

$$\mathbf{Ax} = \mathbf{b}_i \text{ pour } i = 1, ...I$$

il suffit de calculer une seule fois les matrices $\mathbf{L}$ et $\mathbf{U}$ et de les stocker. On peut alors effectuer autant de descentes et de remontées qu'on a de systèmes pour les résoudre tous, sans pour autant refaire la factorisation LU.

Pour I systèmes à résoudre, la complexité est de l'ordre de $\frac{n^3}{3} + I.n^2$additions et multiplications plutôt que $I\frac{n^3}{3}$.

1.3.1.4 Et le stockage...

Pour une matrice de taille n, on remarque que le nombre de termes à connaître pour disposer des matrices $\mathbf{L}$ et $\mathbf{U}$ n'est que de n^2, puisque la diagonale unité de L n'est pas à stocker. De ce fait on peut utiliser la place mémoire disponible dans $\mathbf{A}$ pour y enregistrer les termes de $\mathbf{L}$ et $\mathbf{U}$.

Cette remarque explique que l'algorithme qui suit "écrase" la matrice $\mathbf{A}$ par sa décomposition $\mathbf{LU}$.

1.3.1.5 Algorithme

La factorisation $\mathbf{LU}$ présentée au paragraphe (1.3.1.1) n'est pas implémentée selon la méthode décrite : il existe une algorithme permettant de calculer directement les termes des matrices $\mathbf{L}$ et $\mathbf{U}$. On notera dans l'algorithme ci-dessous que l'ordre de caclul des coefficients n'est pas indifférent.

Voici l'algorithme réalisant la factorisation $\mathbf{LU}$ d'une matrice $\mathbf{A}$ et stockant cette factorisation dans la place mémoire occupée par $\mathbf{A}$: (la matrice $\mathbf{A}$ est perdue, on dit qu'on écrase $\mathbf{A}$).

Algorithme : `Factorisation` **LU**

```
Pour j ← 1 à n-1
  Pour i ← j+1 à n
     A(i,j) = A(i,j)/A(j,j) //construction de la j-ième colonne de L
     Pour k ← j+1 à n
       A(i,k) = A(i,k) - A(i,j) × A(j,k) //actualisation des n-j-1 dernières lignes de A
     FinPour
  FinPour
FinPour
```

1.3.1.6 Exercice

EXERCICE 1.2 *On considère la matrice de Vandermonde*

$$\mathbf{A} = (a_{ij}) \text{ avec } a_{ij} = x_i^{j-1}, i, j = 1, ...n$$

où les x_i *sont* n *abscisses distinctes.*

(1) Programmer la factorisation **LU** *de la matrice* $\mathbf{A}$ *pour des valeurs de* n *comprises entre* 10 *et* 60 *par pas de* 10. *On pourra utiliser des* x_i *définis par* $x_i = i$, *pour* $i = 1, ...n$.

(2) Calculer et représenter graphiquement le nombre d'opérations effectuées.

(3) Retrouver l'ordre de complexité de la méthode. On pourra utiliser l'option de la commande `plot2d logflag` *qui permet d'afficher des échelles logarithmiques, ainsi que la commande* `reglin` *fournissant les coefficients* a *et* b *de l'équation de la droite de régression linéaire appliquée à un nuage de points.*

1.3.1.7 CNS d'existence d'un factorisation LU

THÉORÈME 1.3.1 *Soit* $\mathbf{A}$ *une matrice de* $\mathcal{M}_m(\mathbb{C})$. *Pour* $1 \leq p \leq n$, *on note* $\mathbf{A}_p$ *le bloc*

$$\mathbf{A}_p = \begin{bmatrix} a_{11} & \cdots & \cdots & a_{1p} \\ a_{21} & & & a_{2p} \\ \vdots & & & \vdots \\ a_{p1} & \cdots & \cdots & a_{pp} \end{bmatrix}$$

La matrice $\mathbf{A}$ *admet une factorisation*

$$\mathbf{A} = \mathbf{LU}$$

où $\mathbf{L}$ *est triangulaire inférieure avec des* 1 *sur la diagonale, et* $\mathbf{U}$ *est triangulaire supérieure et inversible si et seulement si tous les blocs* $\mathbf{A}_p, 1 \leq p \leq n$ *sont inversibles. De plus, cette factorisation est unique. De, plus, si* $\mathbf{A}$ *est réelle,* $\mathbf{L}$ *et* $\mathbf{U}$ *le sont aussi.*

On pourra trouver la démonstration détaillée de ce résultat dans [YA].

1.3.1.8 Exercices

EXERCICE 1.3 *Supposons qu'on résolve* $\mathbf{Ax} = \mathbf{b}$ *avec*

$$\mathbf{A} = \begin{bmatrix} 1 & 1-\varepsilon & 3 \\ 2 & 2 & 2 \\ 3 & 6 & 4 \end{bmatrix} \text{ et } \mathbf{b} = \begin{bmatrix} 5-\varepsilon \\ 6 \\ 13 \end{bmatrix}$$

(1) Déterminer pour quelles valeurs de ε la matrice $\mathbf{A}$ ne satisfait pas les hypothèse du théorème ci-dessus.

(2) Pour quelles valeurs de ε cette matrice est-elle singulière ?

(3) Est-il possible de calculer la factorisation dans ce cas ?

EXERCICE 1.4 *Montrer que la factorisation* $\mathbf{LU}$ *d'une matrice* $\mathbf{A}$ *peut être utilisée pour calculer la matrice inverse* $\mathbf{A}^{-1}$. *(On remarquera que la j-ième colonne de* $\mathbf{A}^{-1}$ *vérifie le système linéaire* $\mathbf{Ay}_j = \mathbf{e}_j$, $\mathbf{e}_j$ *étant le j-ième vecteur de la base canonique.)*

EXERCICE 1.5 *Considérons la matrice inversible*

$$\mathbf{A} = \begin{bmatrix} 1 & 1+0.5.10^{-15} & 3 \\ 2 & 2 & 20 \\ 3 & 6 & 4 \end{bmatrix}$$

(1) Effectuer la factorisation $\mathbf{LU}$ *de* $\mathbf{A}$ *à l'aide du programme de l'exercice .*

(2) Calculer le résidu $\mathbf{A} - \mathbf{LU}$, *et montrer qu'il vérifie :*

$$\mathbf{A} - \mathbf{LU} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 4 \end{bmatrix}$$

(3) Que peut-on en conclure ?

(4) Comparez le résultat que vous obtenez avec celui fourni par scilab avec la routine `lu`. *Analysez la différence.*

EXERCICE 1.6 *Soit la matrice* $\mathbf{A} \in \mathbb{R}^{n \times n}$. *On définit la matrice* $\mathbf{B} = [\mathbf{A}, \mathbf{I}_n] \in \mathbb{R}^{n \times 2n}$, *où* $\mathbf{I}_n$ *la matrice identité.*

Pour calculer l'inverse de $\mathbf{A}$ *on établit l'algorithme suivant*

```
Pour j ← 1 à n
  Pour i ← 1 à n
    Si  (i = j)
      Pour k ← 1 à n
        B'(j,k) = B(j,k)/B(j,j)
      FinPour
    Sinon
      Pour k ← 1 à n
        B'(i,k) = B(i,k) - B(i,j) * B'(j,k)
      FinPour
    FinSi
    B ← B'
  FinPour
FinPour
```

(1) *Montrer que le résultat est équivalent à multiplier à gauche* $\mathbf{B}$ *par une suite de matrices* $\mathbf{C}^{(i)}$ *que vous calculerez.*

(2) *Montrer qu'à la fin de l'algorithme les* n *premières colonnes de la matrice* $\mathbf{B}$ *forment la matrice identité* $\mathbf{I}_n$.

(3) *En déduire que les* n *dernières colonnes de* $\mathbf{B}$ *forment* $\mathbf{A}^{-1}$.

(4) *Application :* $\mathbf{A} = \begin{bmatrix} 2 & 4 & 2 \\ 1 & 0 & 3 \\ 3 & 1 & 2 \end{bmatrix}$ *en utilisant les programme* `inverMat`

1.4 Méthode de Cholesky

Dans le cas où la matrice $\mathbf{A}$ est hermitienne et définie positive, on peut toujours effectuer la factorisation décrite ci-dessus. de plus, on peut trouver une factorisation du type $\mathbf{A} = \mathbf{L}\mathbf{L}^*$ moins gourmande en place mémoire.

1.4.1 Existence de la factorisation

THÉORÈME 1.4.1 *Si* $\mathbf{A}$ *est hermitienne et définie positive, alors* $\mathbf{A}$ *admet une unique factorisation* $\mathbf{LU}$ *où* $\mathbf{L}$ *est triangulaire inférieure avec des* 1 *sur la diagonale et* $\mathbf{U}$ *est triangulaire supérieure et inversible.*

DÉMONSTRATION. Si $\mathbf{A}$ est hermitienne et définie positive, ses blocs $\mathbf{A}_p, 1 \leq p \leq n$ (voir théorème ci-dessus) le sont aussi, et on peut donc appliquer le théorème. ∎

THÉORÈME 1.4.2 *Si* $\mathbf{A}$ *est hermitienne et définie positive, alors il existe une unique matrice* $\mathbf{L}$ *triangulaire inférieure et inversible, avec des coefficients positifs sur la diagonale telle que*

$$\mathbf{A} = \mathbf{L}\mathbf{L}^*$$

Cette factorisation porte le nom de Cholesky (colonel de l'armée de Napoléon). De plus si $\mathbf{A}$ *est réelle, symétrique et définie positive,* $\mathbf{L}$ *est réelle.*

On pourra trouver la démonstration dans [YA].

REMARQUE 1.4.1 *Il est important de noter que les matrices* $\mathbf{L}$ *dans les factorisation* $\mathbf{LU}$ *et de Cholesky sont différentes.*

1.4.2 Algorithme

Algorithme : Algorithme de Cholesky

$L(1,1) \leftarrow \sqrt{A(1,1)}$ // Construction de l_{11}
Pour i ← 2 **à** n
  Pour j ← 1 **à** i-1
    $L(i,j) \leftarrow \frac{1}{L(j,j)} \left(A(i,j) - \sum_{k=1}^{j-1} L(i,k) \times L(j,k) \right)$
  FinPour
  $L(i,i) \leftarrow \left(a(i,i) - \sum_{k=1} i-1 L^2(i,k) \right)^{1/2}$
FinPour

1.4.3 Complexité

On montre facilement que l'on effectue :

- $\frac{n^3}{6}$ additions + multiplications
- $\frac{1}{2}n(n-1)$ divisions
- n évaluations de racines carrées.

REMARQUE 1.4.2 *L'intérêt de la méthode de Cholesky réside dans le fait qu'elle demande une place mémoire deux fois inférieure à celle de la factorisation LU, puisqu'on ne stocke que* $\mathbf{L}$ *et que sa complexité est aussi deux fois moindre.*

1.5 Elimination de Gauss avec recherche de pivot partiel

La méthode de Gauss peut-être bloquée si on tombe sur un pivot nul. C'est pour cela qu'a été mise au point la méthode dite "de Gauss avec pivotage partiel". Elle consiste à échanger les lignes du système lorsqu'on tombe sur un pivot nul, ce qui permet de continuer sans problème.

Elle repose sur l'utilisation de matrices de permutation, dont on rappelle ci-dessous la définition :

DÉFINITION 1.5.1 *On appelle permutation de {1,...,n} une bijection de {1,...,n} sur {1,...,n}.*

DÉFINITION 1.5.2 *Soit* σ *une permutation de {1,...,n}, on associe à* σ *une matrice* $\mathbf{P}$ *dite de permutation d'ordre* n *par*

$$P_{ij} = \delta_{\sigma(i)j}$$

c'est à dire

$$P_{ij} = 1 \text{ si } j = \sigma(i)$$
$$P_{ij} = 0 \text{ si } j \neq \sigma(i)$$

Les matrices de permutation ont des propriétés intéressantes, qui font que leur utilisation pour effectuer les pivotages conduit au résultat suivant :

THÉORÈME 1.5.1 *Soit* $\mathbf{A} \in \mathcal{M}_m(\mathbb{R})$, *inversible. Alors il existe :*
- *une matrice de permutation* $\mathbf{P}$,
- *une matrice triangulaire inférieure* $\mathbf{L}$, *avec des* 1 *sur la diagonale,*
- *une matrice triangulaire supérieure* $\mathbf{U}$ *inversible,*

telles que

$$\mathbf{PA} = \mathbf{LU}$$

REMARQUE 1.5.1 *Le pivotage partiel ne sert pas seulement à garantir l'obtention d'une factorisation. Il garantit aussi une meilleure stabilité que celle obtenue par la méthode* $\mathbf{LU}$ *dans le cas de matrices mal conditionnées. On pourra s'en convaincre en examinant le cas du système* $\mathbf{Ax} = \mathbf{b}$, *constitué de la matrice* $\mathbf{A}$ *:*

$$\mathbf{A} = \begin{bmatrix} 10^{-9} & 1 \\ 1 & 1 \end{bmatrix}$$

et du second membre $\mathbf{b}$ *:*

$$\mathbf{b} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

Sa solution est

$$\mathbf{x} = \left[\frac{1}{1-10^{-9}}, \frac{1-2.10^{-9}}{1-10^{-9}} \right]^\top \simeq [1,1]^\top$$

Si on le résout sur une machine à 8 chiffres significatifs, la factorisation calculée est

$$\mathbf{U} = \begin{bmatrix} 10^{-9} & 1 \\ 0 & -10^{-9} \end{bmatrix} \text{ et } \mathbf{L} = \begin{bmatrix} 1 & 0 \\ 10^{-9} & 1 \end{bmatrix}$$

ce qui donne la solution

$$\mathbf{x} = [0,1]^\top$$

Si on utilise le pivotage partiel, on obtient :

$$\mathbf{U} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \text{ et } \mathbf{L} = \begin{bmatrix} 1 & 0 \\ 10^{-9} & 1 \end{bmatrix}$$

ce qui donne la solution

$$\mathbf{x} = [1,1]^\top$$

En exploitant cette remarque, on aboutit à la méthode de Gauss dite de *pivot maximal*. Elle consiste à prendre, à chaque étape, comme pivot l'élément diagonal qui est le plus grand en

valeur absolue afin de limiter les erreurs d'arrondi inévitables dues à un résultat de division trop faible (pertes de chiffres significatifs).

EN SUBSTANCE

- ***La factorisation LU d'une matrice A consiste à calculer une matrice triangulaire inférieure L et une matrice triangulaire supérieure U telles que $\mathrm{A} = \mathrm{LU}$.***
- ***La factorisation LU, quand elle existe, n'est pas unique. Cependant, on peut la rendre unique en se donnant des conditions supplémentaires, par exemple en fixant les valeurs des éléments diagonaux de L à 1. Ceci s'appelle factorisation de Gauss.***
- ***La factorisation de Gauss existe et est unique si et seulement si les mineurs principaux de A d'ordre 1 à $n-1$ sont non nuls (autrement, au moins un pivot est nul).***
- ***Quand on trouve un pivot nul, un nouveau pivot peut être obtenu en échangeant des lignes (ou colonnes) convenablement choisies. C'est la stratégie du pivot.***
- ***Le calcul de la factorisation de Gauss nécessite de l'ordre de $2n^3/3$ opérations en général, et seulement de l'ordre de n opérations dans le cas d'un système tridiagonal.***
- ***Pour les amtrices symétriques définies positives, on peut utiliser la factorisation de Cholesky $\mathrm{A} = \mathrm{HH}^T$, où H est une matrice triangulaire inférieure. Le coût de calcul est de l'ordre de $n^3/3$ opérations.***
  La sensibilité du résultat aux perturbations des données dépend du conditionnement de la matrice du système : la solution calculée peut être imprécise quand la matrice est mal conditionnée.

1.6 Bibliographie

Les ouvrages ci-dessous sont disponibles sous forme de fichier téléchargeable sur le site du cours ou sur Arel

[CB] **Algèbre matricielle numérique**, *Claude Brezinski*

[YA] **Algèbre linéaire et analyse numérique matricielle**, *Yves Achdou*, téléchargeable à l'adresse http ://www.ann.jussieu.fr/~achdou/files/teaching/linalg/book.pdf

[AH1]**Analyse numérique matricielle, Cours de 3ème année**, *Alain Huard*, téléchargeable à partir du site de l'INSA Toulouse www-gmm.insa-toulouse.fr

[AH2]**Analyse numérique des grands problèmes linéaires, Cours de 4ème année**, *Alain Huard*, téléchargeable à partir du site de l'INSA Toulouse www-gmm.insa-toulouse.fr

Les ouvrages ci-dessous sont disponibles en librairie

[QS] **Calcul scientifique, Cours, exercices corrigés et illustrations en Matlab et Octave**, Alfio Quarteroni, Fausto Saleri, Springer, 2206.

[AF] **Analyse numérique pour Ingénieurs**, André Fortin, Presses Internationales Polytechnique, 2001.

2

ALGÈBRE LINÉAIRE ET PERTURBATIONS

Nous présentons, au chapitre suivant, des méthodes itératives de résolution d'un système d'équations linéaires et d'inversion d'une matrice. Avant ces méthodes, nous essaierons d'évaluer l'influence sur la solution d'un système linéaire d'une perturbation du second ou du premier membre. Pour établir cette évaluation il est utile de pouvoir évaluer de combien est différente une matrice d'une autre. Par exemple savoir la différence entre $\mathbf{A} \cdot \mathbf{A}^{-1}$, où $\mathbf{A}^{-1}$ est issue d'un calcul numérique, et la matrice identité $\mathbf{I}$. Nous commendonc par la notion de la norme qui permet de répondre à ce dernier problème.

2.1 Normes vectorielles et matricielles

Si nous voulons estimer la proximité entre deux vecteurs ou deux matrices nous devons utiliser un concept analogue à celui de la différence entre deux scalaires. Ce concept est la *norme*

qui est une fonction avec valeurs dans $\mathbb{R}_+$. Dans cette section nous faisons un rappel rapide des normes vectorielles et nous introduisons les normes matricielles.

2.1.1 Normes vectorielles

Si nous voulons qu'à partir du concept de la norme, nous pouvons dériver celui de la métrique, c'est-à-dire de la distance euclidienne entre deux points de l'espace, on doit imposer à la norme d'un vecteur les propriétés suivantes :

- $\|\mathbf{x}\| \geq 0$ et $\|\mathbf{x}\| = 0$ si et seulement si $\mathbf{x} = \mathbf{0}$
- $\|\alpha\mathbf{x}\| = \mid \alpha \mid \|\mathbf{x}\|$
- $\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$
- $\|\mathbf{x}-\mathbf{y}\| \geq \mid \|\mathbf{x}\| - \|\mathbf{y}\| \mid$

La définition la plus générale d'une telle norme est donnée par

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} \mid x_i \mid^p \right)^{1/p} \; ; \; p \geq 1 \tag{2.1.1}$$

Les normes les plus utilisées sont issues de la formule précédente pour $p = 1, 2$ et ∞ :

- *Norme 1* : $\|\mathbf{x}\|_1 = \sum_{i=1}^{n} \mid x_i \mid$.
- *Norme 2 ou euclidienne* : $\|\mathbf{x}\|_2 = \|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^\top \mathbf{x}}$
- *Norme ∞ ou max* : $\|\mathbf{x}\|_\infty = \max_{i=1,\ldots n} \mid x_i \mid$

De plus, nous avons deux inégalités très utiles pour les calculs

- Inégalité de Hölder

$$\sum_{i=1}^{n} \mid x_i y_i \mid \leq \|\mathbf{x}\|_{\ell_p} \|\mathbf{y}\|_{\ell_q} = \left(\sum_{i=1}^{n} \mid x_j \mid^p \right)^{1/p} \left(\sum_{i=1}^{n} \mid y_j \mid^q \right)^{1/q}$$

- Inégalité de Cauchy (qui est l'inégalité de Hölder pour $p = q = 2$)

$$\sum_{i=1}^{n} \mid x_i y_i \mid \leq \sqrt{\sum_{i=1}^{n} \mid x_j \mid} \cdot \sqrt{\sum_{i=1}^{n} \mid y_j \mid}$$

et les deux théorèmes suivants :

THÉORÈME 2.1.1 *Chaque norme $\|\cdot\|$ de $\mathbb{R}^n$ est uniformément continue par rapport à la métrique $\max_i \mid x_i - y_i \mid$ sur $\mathbb{R}^n$.*

DÉMONSTRATION. Du fait que $\mid \|\mathbf{x}+\eta\| - \|\eta\| \mid \leq \|\mathbf{x}\|$ et si on note $\eta = \sum_{i=1}^{n} \eta_i \mathbf{e}_i$, où $\eta = [\eta_1, \ldots, \eta_n]^\top$ et $[\mathbf{e}_1, \ldots, \mathbf{e}_n]$ la base canonique de $\mathbb{R}^n$, alors nous avons

$$\|\eta\| \leq \sum_{i=1}^{n} \mid \eta_i \mid \|\mathbf{e}_i\| \leq \max_i \mid \eta_i \mid \sum_{j=1}^{n} \|e_j\| = M \max_i \mid \eta_i \mid \text{, avec } M = \sum_{j=1}^{n} \|\mathbf{e}_j\|$$

Donc pour chaque $\varepsilon > 0$ nous pouvons choisir η tel que $\max_i \mid \eta_i \mid \leq \frac{\varepsilon}{M}$, i.e. indépendamment de $\mathbf{x}$ et avoir l'inégalité

$$\mid \|\mathbf{x}+\eta\| - \|\eta\| \mid \leq \varepsilon$$

Par conséquent $\|\bullet\|$ est uniformement continue. ∎

THÉORÈME 2.1.2 *Toutes les normes $\|\cdot\|$ de $\mathbb{R}^n$ sont équivalentes en ce sens que pour chaque couple de normes n_1 et n_2 il existe des constantes positives c et C telles que*

$$c \cdot n_2(\mathbf{x}) \leq n_1(\mathbf{x}) \leq C \cdot n_2(\mathbf{x}) \quad \forall \mathbf{x} \in \mathbb{R}^n$$

DÉMONSTRATION. Considérons l'ensemble $S = \{\mathbf{x} \in \mathbb{R}^n \ / \ n_2(\mathbf{x}) \leq 1\}$. Il s'agit d'un ensemble compact et du fait que $n_2(\mathbf{x})$ est continue d'après le théorème précédent, nous avons que les quantités $c = \min_{\mathbf{x} \in S} n_1(\mathbf{x}) > 0$ et $C = \max_{\mathbf{x} \in S} n_2(\mathbf{x}) > 0$ existent. Donc pour tout $\mathbf{x} \neq 0$ on a $\frac{\mathbf{x}}{\|\mathbf{x}\|} \in S$ et par conséquent

$$c \leq n_1\left(\frac{\mathbf{x}}{\|\mathbf{x}\|}\right) = \frac{\mathbf{1}}{\|\mathbf{x}\|} n_1(\mathbf{x}) \leq C$$

d'où on obtient le résultat cherché. ∎

Le choix de la norme est un élément très important pour un problème d'optimisation de la forme

$$\min_{\mathbf{x} \in \mathbb{R}^n} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_{\ell_p}$$

Nous pouvons montrer que des valeurs de p qui sont proches de 1 donnent des algorithmes plus stables que des valeurs de p proches de 2.

2.1.2 Norme matricielle

Une *norme matricielle* est une application $\|\cdot\| : R^{m \times n} \to R_+$ qui a les propriétés suivantes :

(1) $\|\mathbf{A}\| \geq 0$ *et* $\|\mathbf{A}\| = 0$ si et seulement si $\mathbf{A} = \mathbf{0}$.

(2) $\|\alpha\mathbf{A}\| = \mid \alpha \mid \|\mathbf{A}\| \ \forall \alpha \in \mathbb{R}$.

(3) $\|\mathbf{A}+\mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$

Si, de plus, elle vérifie la relation

(4) $\|\mathbf{A}\mathbf{B}\| \leq \|\mathbf{A}\| \, \|\mathbf{B}\|$

on dit qu'elle est une *norme sous-multiplicative*.

Nous pouvons remarquer que toute norme vectorielle peut induire une norme matricielle à l'aide de la relation

$$\|\mathbf{A}\| = \sup_{\substack{\mathbf{x} \in \mathbb{R}^n \\ \|\mathbf{x}\|=1}} \|\mathbf{A}\mathbf{x}\| = \sup_{\substack{\mathbf{x} \in \mathbb{R}^n \\ \mathbf{x} \neq \mathbf{0}}} \frac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{x}\|}$$

Cette norme est appelée *norme subordonnée* (à la norme vectorielle correspondante) et elle sera notée par

$$\mathrm{lub}\,(\mathbf{A}) = \sup_{\substack{\mathbf{x}\in\mathbb{R}^n \\ \mathbf{x}\neq\mathbf{0}}} \frac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{x}\|}$$

Pour prouver que lub est une norme, considérons une valeur p telle que $1 \leq p \leq +\infty$. À chaque vecteur $\mathbf{x} \in \mathbb{R}^n$ on fait correspondre un réel, qui est la norme vectorielle $\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |\, x_i \,|^p\right)^{1/p}$.

Donc $\|\mathbf{A}\|_p = \sup_{\substack{\mathbf{x}\in\mathbb{R}^n \\ \mathbf{x}\neq\mathbf{0}}} \frac{\|\mathbf{A}\mathbf{x}\|_p}{\|\mathbf{x}\|_p}$ a les propriétés d'une norme matricielle, à savoir

- $\|\mathbf{A}\|_p \geq 0$ et, par convention, $\|\mathbf{A}\|_p = 0$ si et seulement si $\mathbf{x} = \mathbf{0}$.
- $\|\alpha\mathbf{A}\|_p = \sup_{\substack{\mathbf{x}\in\mathbb{R}^n \\ \mathbf{x}\neq\mathbf{0}}} \frac{\|\alpha\mathbf{A}\mathbf{x}\|_p}{\|\mathbf{x}\|_p} = |\,\alpha\,| \sup_{\substack{\mathbf{x}\in\mathbb{R}^n \\ \mathbf{x}\neq\mathbf{0}}} \frac{\|\mathbf{A}\mathbf{x}\|_p}{\|\mathbf{x}\|_p} = |\,\alpha\,|\,\|\mathbf{A}\|_p$
- $\|\mathbf{A}+\mathbf{B}\|_p = \sup_{\substack{\mathbf{x}\in\mathbb{R}^n \\ \mathbf{x}\neq\mathbf{0}}} \frac{\|(\mathbf{A}+\mathbf{B})\,\mathbf{x}\|_p}{\|\mathbf{x}\|_p} \leq \sup_{\substack{\mathbf{x}\in\mathbb{R}^n \\ \mathbf{x}\neq\mathbf{0}}} \left(\frac{\|\mathbf{A}\mathbf{x}\|_p}{\|\mathbf{x}\|_p} + \frac{\|_p\mathbf{B}\mathbf{x}\|_p}{\|\mathbf{x}\|_p}\right) = \|\mathbf{A}\|_p + \|\mathbf{B}\|_p$

De plus, pour les normes subordonnées, nous avons que $\mathrm{lub}\,(\mathbf{I}) = 1$, où $\mathbf{I}$ la matrice identité.

Notons que géométriquement $\mathrm{lub}\,(\mathbf{A})$ exprime la quantité maximale que la norme du point image $\mathbf{A}\mathbf{x}$ peut excéder la norme du point de départ $\mathbf{x}$.

De ce qui précède nous pourrions conclure – abusivement – que pour toute valeur de p entre 1 et $+\infty$ nous pouvons déterminer une norme. En réalité il n'en est rien et nous pouvons déterminer des normes subordonnées pour les valeurs $p = 1, p = 2$ et $p = \infty$ seulement. On a ainsi les trois normes subordonnées :

- *Norme 1 – somme des colonnes* : $\|\mathbf{A}\|_1 = \sup_{\substack{\mathbf{x}\in\mathbb{R}^n \\ \|\mathbf{x}\|_1=1}} \|\mathbf{A}\mathbf{x}\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^{n} |\, a_{ij} \,|$
- *Norme 2 ou de Schur* : $\|\mathbf{A}\|_2 = \sup_{\substack{\mathbf{x}\in\mathbb{R}^n \\ \|\mathbf{x}\|_2=1}} \|\mathbf{A}\mathbf{x}\|_2 = \left(\sum_{i,j=1}^{n} |\, a_{ij} \,|^2\right)^{1/2}$
- *Norme infinie – somme des lignes* : $\|\mathbf{A}\|_\infty = \sup_{\substack{\mathbf{x}\in\mathbb{R}^n \\ \|\mathbf{x}\|_\infty=1}} \|\mathbf{A}\mathbf{x}\|_\infty = \max_{i=1,\ldots,n} \sum_{j=1}^{n} |\, a_{ij} \,|$

Une norme matricielle $\|\mathbf{A}\|$ est *consistante* avec une norme vectorielle $\|\mathbf{x}\|$ si $\|\mathbf{A}\mathbf{x}\| \leq \|\mathbf{A}\| \cdot \|\mathbf{x}\|$. Notons que la norme de Schur est consistante avec la norme euclidienne.

Les normes consistantes et subordonnées sont sous-multiplicatives, i.e.

$$\|\mathbf{A}\mathbf{B}\| \leq \|\mathbf{A}\|\,\|\mathbf{B}\| \tag{2.1.2}$$

Une norme qui n'est pas induite par une norme vectorielle est la norme de Frobenius donnée

par

$$\|\mathbf{A}\|_F = \left(\sum_{i=1}^{m} \sum_{,j=1}^{n} | a_{ij} |^2 \right)^{1/2} = \sqrt{\mathrm{tr}\,(\mathbf{A}^\top \mathbf{A})};\ \mathbf{A} \in \mathbb{R}^{m \times n}$$

où par tr on a noté la trace de la matrice.

Pour la norme de Frobenius, nous avons $\|\mathbf{A}\|_F = \sqrt{n}$

On termine cette section par un résultat important :

PROPRIÉTÉ 2.1.1 *Soit $\|\bullet\|$ une norme quelconque. Alors le rayon spectrale d'une matrice carrée $\mathbf{A}$ est au plus égal à la norme de cette matrice*

$$\varrho(\mathbf{A}) \leq \|\mathbf{A}\| \tag{2.1.3}$$

2.1.3 Exercices

EXERCICE 2.1 *Calculer les normes 1, 2 et ∞ du vecteur $\mathbf{x} = [1, 0, 1, -4]^\top$.*

EXERCICE 2.2 *Calculer les normes 1, 2 et ∞ de la matrice*

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & 1 \\ 2 & 3 & 2 \\ 1 & 2 & 2 \end{bmatrix}$$

EXERCICE 2.3 *(1)* $\forall x \in R^n : \|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_p \leq n^{\frac{1}{p}} \|\mathbf{x}\|_\infty$

(2) En déduire que $\lim_{p \to +\infty} \|\mathbf{x}\|_p = \|\mathbf{x}\|_\infty$

EXERCICE 2.4 *Considérons une norme vectorielle $\|\cdot\|$ et la norme subordonée* $\mathrm{lub}\,(\mathbf{A})$ *que l'on notera $\|\mathbf{A}\|$. Montrer que si $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$, alors*

(1) $\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$

(2) $\|\mathbf{A} \cdot \mathbf{B}\| \leq \|\mathbf{A}\| \, \|\mathbf{B}\|$

(3) $\|\mathbf{A}^n\| \leq \|\mathbf{A}\|^n$

EXERCICE 2.5 *Soit une matrice carrée $\mathbf{A} \in \mathbb{R}^{n \times n}$. Montrer que*

(1) $\mathbf{A}$ est orthogonale, c'est-à-dire $\mathbf{A}^\top = \mathbf{A}^{-1}$, si et seulement si $\|\mathbf{A}\mathbf{x}\|_2 = \|\mathbf{x}\|_2$, $\mathbf{x} \in \mathbb{R}^n$.

(2) $\|\mathbf{A}\|_2 = 1$ pour toute matrice orthogonale.

EXERCICE 2.6 *Montrer que*

(1) $\mathrm{lub}(\mathbf{A}) \leq \|\mathbf{A}\|$

(2) $\mathrm{lub}(\mathbf{AB}) \leq \mathrm{lub}(\mathbf{A}) \cdot \mathrm{lub}(\mathbf{B})$

(3) Calculer $\mathrm{lub}_\infty(\mathbf{A})$

(4) Calculer $\mathrm{lub}(\mathbf{A})$ *pour la norme euclidienne.*

EXERCICE 2.7 *Montrer que* $\mathrm{cond}(\mathbf{A}) \geq 1$.

2.2 Conditionnement d'une matrice

On sait qu'un problème est bien conditionné si une petite variation des entrées (données) n'entraîne pas une grande variation des sorties (résultats). Afin de caractériser la nature du conditionnement d'un problème, on calcule une valeur caractéristique de ce conditionnement qui s'appelle *nombre-condition* ou *conditionnement* du problème (cf. §1.7). Nous allons établir le conditionnement pour les matrices.

On sait que l'erreur relative de précision (cf **??**) est donnée par

$$\eta(x) = \frac{\Delta x}{x} = \frac{m(x) - x}{x} = \frac{m(x)}{x} - 1$$

Pour un scalaire $x \in \mathbb{R}$, la quantité $-\log_{10}\eta(x)$ correspond approximativement au nombre de digits significatifs de $m(x)$. Si on passe aux vecteurs, la quantité

$$\eta(\mathbf{x}) = \frac{m(\|\mathbf{x}\|_\infty) - \|\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty}$$

fournit une bonne approximation du nombre digits significatifs pour les composants de $\mathbf{x}$ qui ont des grandes valeurs absolues. Mais cette approximation se détériore pour les petites valeurs.

On définit le conditionnement d'une matrice $\mathbf{A}$ par la quantité

$$\mathrm{cond}(\mathbf{A}) = \|\mathbf{A}\| \cdot \left\|\mathbf{A}^{-1}\right\|$$

Le conditionnement peut aussi être noté par $\kappa(A)$.

La section suivante fournit une justification du choix que nous venons d'effectuer pour le conditionnement d'une matrice.

2.2.1 Exercice

EXERCICE 2.8 *Monter que* $\mathrm{cond}(\mathbf{A}^2) \leq (\mathrm{cond}(\mathbf{A}))^2$ *pour toute matrice carrée* $\mathbf{A}$.

2.3 Suite de matrices

La convergence d'une suite de matrices de format $(n \times m)$ est équivalente à la convergence de $n \cdot m$ suites scalaires formées par les termes de ces matrices. Nous avons le théorème suivant~ :

THÉORÈME 2.3.1 *Soit* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *une matrice carrée. Alors les propositions suivantes sont équivalentes :*

(1) $\lim_{k\to+\infty} \mathbf{A}^k = 0$

(2) $\lim_{k\to+\infty} \mathbf{A}^k\mathbf{x} = 0; \forall \mathbf{x} \in \mathbb{R}^n$

(3) $\rho(\mathbf{A}) < 1$

(4) $\|\mathbf{A}\| < 1$ *pour au moins une norme subordonnée.*

2.3.1 Exercice

EXERCICE 2.9 (THÉORÈME DE VON NEUMANN) .- *Soit* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *matrice carrée avec* $\rho(\mathbf{A}) < 1$. *alors*

(1) La matrice $\mathbf{I} - \mathbf{A}$ *est régulière.*

(2) $(\mathbf{I} - \mathbf{A})^{-1} = \lim_{n\to+\infty} \sum_{k=0}^{n} \mathbf{A}^k = \sum_{k=0}^{\infty} \mathbf{A}^k$.

Indication.- Si λ_i *est la valeur propre de* $\mathbf{A}$, *alors* $1 - \lambda_i$ *est la valeur propre correspondante de la matrice* $\mathbf{I} - \mathbf{A}$.

2.4 Bornes de l'erreur de la solution d'un système linéaire

Considérons le système d'équations linéaires

$$\mathbf{A}\mathbf{x} = \mathbf{b}; \quad \mathbf{x} \in \mathbb{R}^n, \mathbf{b} \in \mathbb{R}^m \tag{2.4.1}$$

La solution obtenue par des méthodes numériques est entachée d'erreurs d'arrondi. De cette façon, au lieu d'avoir la solution exacte, nous avons une solution approchée $\mathbf{x} + \Delta\mathbf{x}$, qui peut être vu comme une solution perturbée de la solution exacte. L'objectif ici est d'évaluer la borne supérieure de la perturbation $\Delta\mathbf{x}$. À cette analyse de l'erreur directe nous pouvons adjoindre une analyse de l'erreur inverse. On s'intéresse ainsi à la perturbation $\Delta\mathbf{x}$ de $\mathbf{A}$ et/ou à la perturbation $\Delta\mathbf{b}$ de $\mathbf{b}$ qui conduisent à la solution $\mathbf{x} + \Delta\mathbf{x}$ et on exprime $\Delta\mathbf{x}$ en fonction de $\Delta\mathbf{A}$ et $\Delta\mathbf{b}$.

On commence par un résultat qui montre l'importance de $\frac{1}{\text{cond}(\mathbf{A})}$.

THÉORÈME 2.4.1 *Si* $\mathbf{A}$ *est régulière et si*

$$\frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|} < \frac{1}{\text{cond}(\mathbf{A})} \tag{2.4.2}$$

alors la matrice $\mathbf{A}+\Delta\mathbf{A}$ *est aussi régulière.*

De ce théorème on en déduit que l'inverse du nombre-condition de la matrice « mesure » la distance qui sépare $\mathbf{A}$ d'une matrice singulière. Donc si le nombre-condition d'une matrice $\mathbf{A}$ est grand, alors cette matrice est proche de la singularité.

2.4.1 Perturbations de $\mathbf{b}$

Soit le système $\mathbf{Ax} = \mathbf{b}$. Si la place du vecteur $\mathbf{b}$ nous avons le vecteur perturbé $\mathbf{b} + \mathbf{\Delta b}$, la solution $\mathbf{x}$ est aussi perturbée et elle devient $\mathbf{x} + \mathbf{\Delta x}$. Nous avons donc le système linéaire :

$$\mathbf{A}(\mathbf{x} + \mathbf{\Delta x}) = \mathbf{b} + \mathbf{\Delta b},\ \mathbf{A} \in \mathbb{R}^{n \times n}$$

On a $\mathbf{A\Delta x} = \mathbf{\Delta b}$ et donc

$$\mathbf{\Delta x} = \mathbf{A}^{-1}\mathbf{\Delta b}$$

d'où

$$\|\mathbf{\Delta x}\| \leq \left\|\mathbf{A}^{-1}\right\| \|\mathbf{\Delta b}\| \tag{2.4.3}$$

Puisque $\mathbf{Ax} = \mathbf{b}$, on a $\|\mathbf{b}\| \leq \|\mathbf{A}\| \|\mathbf{x}\|$ et par conséquent

$$\|\mathbf{\Delta x}\| \|\mathbf{b}\| \leq \|\mathbf{A}\| \left\|\mathbf{A}^{-1}\right\| \|\mathbf{x}\| \|\mathbf{\Delta b}\|$$

et si $\|\mathbf{b}\| \neq 0$, on obtient

$$\frac{\|\mathbf{\Delta x}\|}{\|\mathbf{x}\|} \leq \kappa(\mathbf{A}) \frac{\|\mathbf{\Delta b}\|}{\|\mathbf{b}\|} \tag{2.4.4}$$

Nous constatons ainsi que si $\kappa(\mathbf{A}) \gg 1$, une petite perturbation du vecteur $\mathbf{b}$ peut provoquer une grande perturbation de la solution.

2.4.2 Perturbations de $\mathbf{A}$

Si la matrice $\mathbf{A}$ est perturbée et on a à sa place $\mathbf{A} + \mathbf{\Delta A}$, alors le système linéaire devient :

$$(\mathbf{A} + \mathbf{\Delta A})(\mathbf{x} + \mathbf{\Delta x}) = \mathbf{b} \tag{2.4.5}$$

d'où on a

$$\mathbf{A\Delta x} = -\mathbf{\Delta A}(\mathbf{x} + \mathbf{\Delta x}) \tag{2.4.6}$$

et finalement

$$\mathbf{\Delta x} = -\mathbf{A}^{-1}\mathbf{\Delta A}(\mathbf{x} + \mathbf{\Delta x}) \tag{2.4.7}$$

En utilisant les normes, on a

$$\|\mathbf{\Delta x}\| \leq \left\|\mathbf{A}^{-1}\right\| \|\mathbf{\Delta A}\| (\|\mathbf{x}\| + \|\mathbf{\Delta x}\|) \quad \Rightarrow \quad \|\mathbf{\Delta x}\| \left(1 - \left\|\mathbf{A}^{-1}\right\| \|\mathbf{\Delta A}\|\right) \leq \left\|\mathbf{A}^{-1}\right\| \|\mathbf{\Delta A}\| \|\mathbf{x}\|$$

ce qui donne pour l'erreur sur $\mathbf{x}$:

$$\frac{\|\mathbf{\Delta x}\|}{\|\mathbf{x}\|} \leq \frac{\left\|\mathbf{A}^{-1}\right\| \|\mathbf{\Delta A}\|}{1 - \|\mathbf{A}^{-1}\| \|\mathbf{\Delta A}\|} = \frac{\|\mathbf{A}\| \left\|\mathbf{A}^{-1}\right\|}{1 - \|\mathbf{A}^{-1}\| \|\mathbf{\Delta A}\|} \cdot \frac{\|\mathbf{\Delta A}\|}{\|\mathbf{A}\|} \tag{2.4.8}$$

c'est-à-dire

$$\frac{\|\mathbf{\Delta x}\|}{\|\mathbf{x}\|} \leq \kappa(\mathbf{A}) \cdot \frac{\|\mathbf{\Delta A}\|}{\|\mathbf{A}\|} \tag{2.4.9}$$

si $\left\|\mathbf{A}^{-1}\right\| \|\mathbf{\Delta A}\| \ll 1$.

Nous pouvons donc répéter la remarque de la fin du paragraphe précédent.

2.4.3 Perturbations de $\mathbf{A}$ et de $\mathbf{b}$

Si $\mathbf{A}$ et $\mathbf{b}$ subissent des perturbations, le système linéaire devient :

$$(\mathbf{A}+\mathbf{\Delta A})(\mathbf{x}+\mathbf{\Delta x})=\mathbf{b}+\mathbf{\Delta b} \tag{2.4.10}$$

On a

$$\mathbf{\Delta A}\cdot\mathbf{x}+\mathbf{A}\cdot\mathbf{\Delta x}+\mathbf{\Delta A}\cdot\mathbf{\Delta x}=\mathbf{\Delta b}$$

En négligeant les termes du 2e ordre, on obtient

$$\mathbf{\Delta x}=-\mathbf{A}^{-1}\cdot\mathbf{\Delta A}\cdot\mathbf{x}+\mathbf{A}^{-1}\cdot\mathbf{\Delta b}$$

En utilisant la norme on a :

$$\|\mathbf{\Delta x}\|\leq\left\|\mathbf{A}^{-1}\right\|\cdot\|\mathbf{\Delta A}\|\cdot\|\mathbf{x}\|+\left\|\mathbf{A}^{-1}\right\|\cdot\|\mathbf{\Delta b}\|$$

d'où

$$\frac{\|\mathbf{\Delta x}\|}{\|\mathbf{x}\|}\leq\left\|\mathbf{A}^{-1}\right\|\cdot\|\mathbf{\Delta A}\|+\left\|\mathbf{A}^{-1}\right\|\cdot\frac{\|\mathbf{\Delta b}\|}{\|\mathbf{x}\|}\leq\frac{\left\|\mathbf{A}^{-1}\right\|\cdot\|\mathbf{A}\|\cdot\|\mathbf{\Delta A}\|}{\|\mathbf{A}\|}+\frac{\left\|\mathbf{A}^{-1}\right\|\cdot\|\mathbf{A}\|\cdot\|\mathbf{\Delta b}\|}{\|\mathbf{A}\|\cdot\|\mathbf{x}\|}$$

et finalement

$$\frac{\|\mathbf{\Delta x}\|}{\|\mathbf{x}\|}\leq\kappa(\mathbf{A})\cdot\left(\frac{\|\mathbf{\Delta A}\|}{\|\mathbf{A}\|}+\frac{\|\mathbf{\Delta b}\|}{\|\mathbf{b}\|}\right) \tag{2.4.11}$$

De trois formules (2.4.4), (2.4.9) et (2.4.11) on conclut que pour avoir une petite modification $\mathbf{\Delta x}$ du vecteur $\mathbf{x}$ il n'est pas suffisant que les erreurs relatives $\frac{\|\mathbf{\Delta A}\|}{\|\mathbf{A}\|}$ et $\frac{\|\mathbf{\Delta b}\|}{\|\mathbf{b}\|}$ soient petites. Il faut, de plus, que la matrice $\mathbf{A}$ soit bien conditionnée.

2.4.4 Exercices

EXERCICE 2.10 *Soit le système* $\mathbf{Ax}=\mathbf{b}$ *et supposons que nous avons une solution approchée* $\widetilde{\mathbf{x}}=\mathbf{x}+\mathbf{\Delta x}$. *Dans ce cas on* $\mathbf{A}\widetilde{\mathbf{x}}=\mathbf{b}+\mathbf{\Delta b}$.

(1) Montrer que $\|\mathbf{\Delta x}\|\leq\left\|\mathbf{A}^{-1}\right\|\cdot\|\mathbf{\Delta b}\|$

(2) Calculer une majoration de la norme de l'erreur relative $\frac{\|\mathbf{\Delta x}\|}{\|\mathbf{x}\|}$.

EXERCICE 2.11 *Si* $\mathbf{A}$ *matrice carrée avec* $\|\mathbf{A}\|<1$ *, alors*

(1) $(\mathbf{I}+\mathbf{A})^{-1}$ *existe*

(2) $\left\|(\mathbf{I}+\mathbf{A})^{-1}\right\|\leq\frac{1}{1-\|\mathbf{A}\|}$

EXERCICE 2.12 *Soit* $\mathbf{A}$ *matrice carrée, régulière et* $\mathbf{B}=\mathbf{A}(\mathbf{I}+\mathbf{C})$ *avec* $\|\mathbf{C}\|<1$ *. Soient aussi* $\mathbf{x}$ *et* $\mathbf{\Delta x}$ *définis par* $\mathbf{Ax}=\mathbf{b}$ *et* $\mathbf{B}(\mathbf{x}+\mathbf{\Delta x})=\mathbf{b}$ *. Supposons que* $\operatorname{cond}(\mathbf{A})\frac{\|\mathbf{B}-\mathbf{A}\|}{\|\mathbf{A}\|}<1$. *Montrer que*

(1) $\dfrac{\|\mathbf{\Delta x}\|}{\|\mathbf{x}\|} \leq \dfrac{\|\mathbf{C}\|}{1-\|\mathbf{C}\|}$

(2) $\dfrac{\|\mathbf{\Delta x}\|}{\|\mathbf{x}\|} \leq \dfrac{\text{cond}\,(\mathbf{A})}{1-\text{cond}\,(\mathbf{A})\cdot\frac{\|\mathbf{B}-\mathbf{A}\|}{\|\mathbf{A}\|}} \cdot \dfrac{\|\mathbf{B}-\mathbf{A}\|}{\|\mathbf{A}\|}$

EXERCICE 2.13 *Soit* $\mathbf{A}$ *matrice carrée regulière et soit* $\mathbf{B}_0$ *son inverse calculé. Supposons que* $\|\mathbf{I}-\mathbf{AB}_0\| = \rho < 1$.

$\Delta\mathbf{A}_0 = \mathbf{I}-\mathbf{AB}_0$
$\mathbf{B}_1 = \mathbf{B}_0\,(\mathbf{I}+\Delta\mathbf{A}_0)$ *et* $\Delta\mathbf{A}_1 = \mathbf{I}-\mathbf{AB}_1$
$\mathbf{B}_2 = \mathbf{B}_1\,(\mathbf{I}+\Delta\mathbf{A}_1)$ *et* $\Delta\mathbf{A}_2 = \mathbf{I}-\mathbf{AB}_2$
.....

(1) *Montrer que* $\mathbf{B}_m = \mathbf{A}^{-1}\left(\mathbf{I}-\Delta\mathbf{A}_0^{2^m}\right)$

(2) *Montrer que* $\left\|\mathbf{B}_m-\mathbf{A}^{-1}\right\| \leq \|\mathbf{B}_0\|\,\dfrac{\rho^{2^m}}{1-\rho}$

(3) *En utilisant Scilab calculer l'inverse de la matrice*

$$\mathbf{A} = \begin{bmatrix} 1 & 0.42 & 0.54 & 0.66 \\ 0.42 & 1 & 0.32 & 0.44 \\ 0.54 & 0.32 & 1 & 0.22 \\ 0.66 & 0.44 & 0.22 & 1 \end{bmatrix}$$

et améliorer ce calcul en utilisant la procédure décrite ci-dessus.

Peut-on quantifier l'amélioration ainsi obtenue ?

2.5 Analyse active de l'erreur

Nous allons présenter, à l'aide d'un exemple, une méthode de calcul d'erreur d'un algorithme, qui se fait en même temps que le déroulement de l'algorithme.

Rappelons d'abord, la relation (1.4.4) d'un nombre-machine avec le nombre réel qu'il représente :

$$m\,(x) = x\,(1+\eta)\,, \quad |\,\eta\,| \leq \text{eps} \tag{2.5.1}$$

Cette relation peut aussi s'écrire

$$m\,(x) = \frac{x}{(1+\eta)}, \quad |\,\eta\,| \leq \text{eps} \tag{2.5.2}$$

Supposons que nous voulons faire la somme des nombres x_k; $k = 1,\ldots,n$. Notons la i-ième somme partielle $s_k = x_i + \cdots + x_k$. Le calcul de cette somme par l'ordinateur se fait selon l'itération

$$s_k = s_{k-1} + x_k \tag{2.5.3}$$

et donne le résultat

$$m(s_k) = s_k + e_k, \text{ avec } \mid e_k \mid \leq \text{eps} \tag{2.5.4}$$

Nous avons aussi d'après (5.1)

$$m(x_k) = x_k + \eta_k x_k, \text{ avec } \mid \eta_k \mid \leq \text{eps} \tag{2.5.5}$$

et d'après (5.2)

$$m(s_k) = \frac{s_k}{1+\varepsilon_k}, \text{ avec } \mid \varepsilon_k \mid \leq \text{eps}$$

ce qui compte tenu de (5.3.3) s'écrit

$$(1+\varepsilon_k)\, m(s_k) = m(s_{k-1}) + m(x_k) \tag{2.5.6}$$

En développant et en utilisant de nouveau (5.3.3) et (2.5.5), on obtient

$$\begin{aligned} s_k + e_k + \varepsilon_k m(s_k) &= m(s_{k-1}) + m(x_k) \\ &= m(s_{k-1}) + x_k + \eta_k x_k \\ &= s_{k-1} + e_{k-1} + x_k + \eta_k x_k \end{aligned}$$

En tenant compte de (2.5.3) on a que

$$e_k = e_{k-1} + \eta_k x_k - \varepsilon_k m(s_k) \tag{2.5.7}$$

ce qui donne

$$\mid e_k \mid \leq \mid e_{k-1} \mid + \text{eps} \cdot \mid x_k \mid + \text{eps} \cdot \mid m(s_k) \mid \tag{2.5.8}$$

On pose

$$\delta_k = \delta_{k-1} + \mid x_k \mid + \mid m(s_k) \mid, \quad \delta_0 = 0 \text{ (car } e_0 = 0) \tag{2.5.9}$$

et on a finalement :

$$\mid e_k \mid \leq \text{eps} \cdot \delta_k \tag{2.5.10}$$

Cette méthode a été proposée par Wilkinson qui l'a nommée *running error analysis* et que nous traduisons par *analyse active de l'erreur*. L'idée fondamentale est que, quand on exécute une opération arithmétique (addition, soustraction, multiplication, division et extraction de la racine caréé) notée $\oplus$, l'erreur peut s'écrire

$$\mid x \oplus y - m(x \oplus y) \mid \leq \text{eps} \cdot \mid m(x \oplus y) \mid \tag{2.5.11}$$

et dans la mesure où on connaît $m(x \oplus y)$, on peut calculer la borne supérieure de cette erreur.

2.6 Produits vectoriels

Soient deux vecteurs $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ et formons leur produit intérieur $s_n = \mathbf{x}^\top \mathbf{y}$. Notons par $s_k = x_1 y_1 + \cdots + x_k y_k$ la k-ième somme partielle. En utilisant le modèle (5.1), on obtient

$$\begin{aligned} m(s_1) &= m(x_1 y_1) = x_1 y_1 (1+\eta_1) \\ m(s_2) &= m(m(s_1) + x_2 y_2) = (m(s_1) + x_2 y_2 (1+\eta_2))(1+\eta_3) \\ &= (x_1 y_1 (1+\eta_1) + x_2 y_2 (1+\eta_2))(1+\eta_3) \\ &= x_1 y_1 (1+\eta_1)(1+\eta_3) + x_2 y_2 (1+\eta_2)(1+\eta_3) \end{aligned}$$

avec $|\,\eta_k\,| \leq \text{eps}$. Si on suppose que approximativement $1+\eta_k \simeq 1 \pm \eta, \quad \forall k$, alors on a :

$$m(s_3) = x_1 y_1 (1 \pm \eta)^3 + x_2 y_2 (1 \pm \eta)^3 + x_3 y_3 (1 \pm \eta)^2$$

d'où, en généralisant :

$$m(s_n) = x_1 y_1 (1 \pm \eta)^n + x_2 y_2 (1 \pm \eta)^n + x_3 y_3 (1 \pm \eta)^{n-1} + \cdots + x_n y_n (1 \pm \eta)^2 \tag{2.6.1}$$

Pour simplifier cette expression on utilise le lemme suivant :

LEMME 2.6.1 *Si* $|\,\eta_k\,| \leq \text{eps}$ *et* $\rho_k = \pm 1$ *pour* $k = 1, \ldots, n$ *et si* $n \cdot \text{eps} < 1$, *alors*

$$\prod_{k=1}^{n} (1+\eta_k)^{\rho_k} = 1 + \theta_k$$

où

$$|\,\theta_k\,| \leq \gamma_k, \;\; \textit{avec} \;\; \gamma_k = \frac{n \cdot \text{eps}}{1 - n \cdot \text{eps}}$$

En appliquant ce lemme, la relation (2.6.1) s'écrit :

$$m(s_n) = x_1 y_1 (1+\theta_n) + x_2 y_2 \left(1+\theta'_n\right) + x_3 y_3 (1+\theta_{n-1}) + \cdots + x_n y_n (1+\theta_2) \tag{2.6.2}$$

Cette relation fournit l'erreur en retard du produit intérieur de deux vecteurs et elle peut être interprétée comme étant la somme exacte des données $x_1, \ldots, x_n$ et de données perturbées $y_1(1+\theta_n), \ldots, y_n(1+\theta_2)$ (ou, alternativement, on peut perturber x_i). Remarquons que chaque perturbation est bornée par γ_k, donc elle est négligable.

Ainsi nous avons pour le produit intérieur la relation

$$m\left(\mathbf{x}^\top \mathbf{y}\right) = (\mathbf{x} + \mathbf{\Delta} x)^\top \mathbf{y} = \mathbf{x}^\top (\mathbf{y} + \mathbf{\Delta} y), \quad \text{avec } \|\mathbf{\Delta x}\| \leq \gamma_n \|\mathbf{x}\|, \; \|\mathbf{\Delta y}\| \leq \gamma_n \|\mathbf{y}\| \tag{2.6.3}$$

qui montre que cette opération est stable si $n \cdot \text{eps} < 1$.

Nous pouvons aussi calculer une borne pour l'erreur en avance, en utilisant la dernière relation :

$$\left\| \mathbf{x}^\top \mathbf{y} - \mathbf{m}\left(\mathbf{x}^\top \mathbf{y}\right) \right\| \leq \gamma_n \sum_{k=1}^{n} |\,x_k y_k\,|$$

Pour le produit extérieur $\mathbf{A} = \mathbf{x}\mathbf{y}^\top$, avec $\mathbf{A} = (a_{ij})$, on a $\mathbf{m}(a_{ij}) = x_i y_j (1+\eta_{ij})$, où $|\,\eta_{ij}\,| \leq \text{eps}$. Donc

$$m(\mathbf{A}) = \mathbf{x}\mathbf{y}^\top + \Delta, \quad \text{avec } \|\Delta\| \leq \text{eps} \cdot \left\| \mathbf{x}\mathbf{y}^\top \right\| \tag{2.6.4}$$

2.6.1 Exercice

EXERCICE 2.14 *Soient* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ *avec* $n = 2m$, *nombre pair. Montrer que l'algorithme suivant pour le calcul de* $\mathbf{x}^\top \mathbf{y}$

- $s_1 \leftarrow \mathbf{x}(1:m)^\top \mathbf{y}(1:m)$
- $s_2 \leftarrow \mathbf{x}(m+1:n)^\top \mathbf{y}(m+1:n)$

– $s \leftarrow s_1 + s_2$

réduit la borne supérieure de l'erreur en avance.

Généraliser ce résultat en décomposant en k sous-vecteurs les vecteurs $\mathbf{x}$ et $\mathbf{y}$.

Pour quelle valeur de n cette décomposition est réalisable ?

2.7 Multiplication matricielle

Soient deux matrices $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{B} \in \mathbb{R}^{n\times p}$ et soit $\mathbf{C} = \mathbf{A}\cdot\mathbf{B}$ leur produit. Exprimons $\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_m \end{bmatrix}$ où $\mathbf{a}_i$ est un vecteur-ligne. De même on a $\mathbf{B} = \begin{bmatrix} \mathbf{b}_1 & \cdots & \mathbf{b}_p \end{bmatrix}$, où $\mathbf{b}_1$ est un vecteur colonne. Nous avons $c_{ij} = \mathbf{a}_i\mathbf{b}_j$. Donc $m(c_{ij}) = (\mathbf{a}_i + \mathbf{\Delta a}_i)\mathbf{b}_j$ avec $\|\mathbf{\Delta a}_i\| \leq \gamma_n \|\mathbf{a}_i\|$. Si on note $\mathbf{C} = \begin{bmatrix} \mathbf{c}_1 & \cdots & \mathbf{c}_p \end{bmatrix}$ les colonnes de la matrice $\mathbf{C}$, nous avons

$$m(\mathbf{c}_j) = (\mathbf{A} + \mathbf{\Delta A})\mathbf{b}_j avec \|\mathbf{\Delta A}\| \leq \gamma_n \|\mathbf{A}\|$$

qui donne l'erreur en retard pour une colonne de la matrice $\mathbf{C}$.

Pour l'erreur en avance, nous avons la borne supérieure suivante :

$$\|\mathbf{C} - \mathbf{m}(\mathbf{C})\| \leq \gamma_n \|\mathbf{A}\| \|\mathbf{B}\| \tag{2.7.1}$$

2.7.1 Exercice

EXERCICE 2.15 *Soient $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$ deux matrices régulières. Montrer que*

$$m(\mathbf{AB}) = (\mathbf{A} + \mathbf{\Delta} A)\mathbf{B}$$

et évaluer la borne supérieure de $\|\mathbf{\Delta A}\|$.

Même chose si on suppose que la matrice $\mathbf{B}$ est perturbée d'une matrice $\mathbf{\Delta B}$.

2.8 Complexité

En dehors de la précision des calculs d'un algorithme, un autre paramètre important pour le choix d'un algorithme est sa *complexité*, c'est-à-dire le nombre d'opérations arithmétiques que son exécution nécessite. Cette mesure doit être indépendante de l'ordinateur sur lequel s'exécutera l'algorithme. On introduit donc, une opération arithmétique abstraite appelée *opération flottante standarisée* et notée flop.

On convient que les opérations d'addition, soustraction et multiplication repésentent 1 flop. La division et l'extraction de la racine carrée entre 10 et 30 flop$s$. Le calcul des fonctions exponentielles et trigonométriques 50 flop$s$.

Dans plusieurs cas d'évaluation de la complexité d'un algorithme, on cherche à avoir une idée qualitative du nombre d'opérations et non pas le nombre exact de ces opérations. On utilise dans ce cas la notion de l'*ordre de la complexité*, dont la définition est la suivante :

DÉFINITION 2.8.1 *Soit un algorithme dont le nombre d'opérations dépend d'un paramètre n. On dit que la complexité $C(n)$ de l'algorithme est de l'ordre $f(n)$ s'il existe deux constantes a et b telles que*

$$C(n) \leq bf(n), \quad \forall n \geq a \tag{2.8.1}$$

Cette complexité est exprimée en utilisant la O-notation de Landau [1]

$$C(n) = O(f(n)) \tag{2.8.2}$$

Les algorithmes, en fonction de leur complexité, sont partagées en différentes classes, dont nous donnons ci-après les plus importantes pour l'analyse numérique.

Ordre	**Classe**	**Exemple de O (f (n))**
$O(1)$	constante	$a \in \mathbb{R}_+$
$O(\log n)$	logarithmic	$a \log n$
$O(n)$	linéaire	$a_1 n + a_0$
$O(n^2)$	quadratique	$a_2 n^2 + a_1 n + a_0$
$O(n^3)$	cubique	$a_3 n^3 + a_2 n^2 + a_1 n + a_0$
$O(n^m)$	polynomiale	$a_m n^m + \cdots + a_0$
$O(c^n)$	exponentielle	$c^{bn} + \ldots$
$O(n!)$	factorielle	$qn!$

2.8.1 Exercices

EXERCICE 2.16 *Supposons que nous avons un algorithme pour multiplier deux matrices $n \times n$ en $O(n^\omega)$ opérations avec $2 < \omega < 3$. Montrer que si $\mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{B} \in \mathbb{R}^{n \times p}$, alors $\mathbf{AB}$ nécessite $O\left(n_1^{\omega-2} n_2 n_3\right)$ opérations, où $n_1 = \min\{m, n, p\}$ et n_2, n_3 sont les deux autres dimensions.*

EXERCICE 2.17 *Soit la matrice triangulaire supérieure par blocs :*

$$\mathbf{C} = \begin{bmatrix} \mathbf{I} & \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{B} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{bmatrix}$$

Calculer son inverse.

En déduire que nous pouvons calculer la multiplication des matrices $\mathbf{A}, \mathbf{B}$ en utilisant l'inverse de la matrice $\mathbf{C}$.

2.9 Multiplication rapide des matrices

Soient deux matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$ et soit $\mathbf{C} = \mathbf{A} \cdot \mathbf{B}$. L'algorithme de multiplication

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$$

nécessite n^3 multiplications et $n^2(n-1)$ additions. Donc la complexité de la multiplication matricielle des matrices carrées est de l'ordre de $O\left(n^3\right)$.

1. La définition de $O-$notation de Landau est la suivante : Soit $f : \mathbb{R} \to \mathbb{R}$. S'il existe une constante C et $\varepsilon > 0$ tels que $| f(x) | \leq C | x |^p ; \forall | x | < \varepsilon$, alors on dit que f est de classe $O(x^p)$ qu'on devait écrire $f(x) \in O(x^p)$ et que, par abus de notation, on écrit $f(x) = O(x^p)$.

La question posée concerne la possibilité d'avoir une complexité $O(n^\omega)$ avec $\omega < 3$, c'est-à-dire la possibilité d'avoir des algorithmes rapides pour la multiplication. Il y a plusieurs réponses à cette question. Nous présentons les deux premières historiquement.

Winograd en 1967 a utilisé, pour une dimension n paire, la formule

$$\mathbf{x}^\top \mathbf{y} = \sum_{k=1}^{n/2} (x_{2k-1} + y_{2k})(x_{2k} + y_{2k-1}) - \sum_{k=1}^{n/2} x_{2k-1}x_{2k} - \sum_{k=1}^{n/2} y_{2k-1}y_{2k} \tag{2.9.1}$$

Si cette relation est appliquée au produit matriciel, alors les deuxième et troisième termes sont calculés une seule fois pour chaque ligne et chaque colonne. La complexité reste la même mais au lieu de faire n^3 multiplications, on fera moitié moins.

En 1969 Strassen a présenté une nouvelle technique de calcul du produit matriciel de deux matrices fondée sur la stratégie de diviser et régner. Cet algorithme se decompose en deux phases :

- Phase 1 : (diviser) Le problème est decomposé en deux ou plus sous-problèmes de dimension identique et qui peuvent être résolus indépendamment les uns des autres.
- Phase 2 : (régner) Les solutions des sous problèmes sont arrangées pour former une solution pour le problème initial.

Nous allons présenter cet algorithme dans le cas simple de deux matrices $\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$ et $\mathbf{B} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}$. Si $\mathbf{A} \cdot \mathbf{B} = \mathbf{C} = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}$, on a

$$\begin{aligned} c_{11} &= p_1 + p_4 - p_5 + p_7 \\ c_{12} &= p_3 + p_5 \\ c_{21} &= p_2 + p_4 \\ c_{22} &= p_1 + p_3 - p_2 + p_6 \end{aligned} \tag{2.9.2}$$

où on a posé

$$\begin{aligned} p_1 &= (a_{11} + a_2)(b_{11} + b_{22}) \\ p_2 &= (a_{21} + a_{22})\, b_{11} \\ p_3 &= a_{11}(b_{12} - b_{22}) \\ p_4 &= a_{22}(b_{21} - b_{11}) \\ p_5 &= (a_{11} + a_{12})\, b_{22} \\ p_6 &= (a_{21} - a_{11})(b_{11} + b_{12}) \\ p_7 &= (a_{12} - a_{22})(b_{21} + b_{22}) \end{aligned} \tag{2.9.3}$$

Cet algorithme requiert 25 $\mathrm{flop}s$ (7 multiplications et 18 additions/soutrctions) tandis l'algorithme classique nécessite 12 $\mathrm{flop}s$ (8 multiplications et 4 additions). Donc l'algorithme de Strassen devient intéressant quand la format de la matrice $n \times n$ est grand, car les formules (2.9.2) et (2.9.3) restent valables si à la place des scalaires on a des matrices. Dans ce cas on partage les matrices initiales en quatre sous-matrices de dimension identique. Chaque sous-matrice ainsi obtenue peut, à son tour, être décomposée en quatre sous-matrices et ainsi de suite, de façon recursive. Si on considère que $n = 2^p$ et que $n_0 \times n_0$ est le format des matrices à l'arrêt de la recursivité, avec $n_0 = 2^r$, on a que l'algorithme de Strassen effectue $M(k,r) = 7^{k-r}8^r$ multiplications et $A(k,r) = 4^r(2^r+5)7^{k-r} - 6 \cdot 4^k$ additions. La somme $M(k,r) + A(k,r)$ est minimale pour $r = 3$. Dans ce cas on $M(k,r) + A(k,r) = 512 \cdot 7^{k-2} - 6 \cdot 4^k < 4 \cdot 7^k = 2 \cdot 2^{\log_2 7^k} = 4\left(2^k\right)^{\log_2 7} =$

$4n^{\log_2 7} \simeq 4n^{2.807}$.

2.9.1 Exercice

EXERCICE 2.18 *On note par $S_n(n_0)$ par le nombre d'opérations de la méthode de Strassen appliquée à une matrice $n \times n$ quand la procédure recursive s'arrête aux matrices $n_0 \times n_0$.*

Supposons que n et n_0 sont des puissances de 2.

Pour n grand estimer $S_n(8)/S_n(n)$ et $S_n(1)/S_n(8)$ et expliquer la signification de ces quantités.

2.10 Préconditionnement d'une matrice

Considérons un système linéaire $\mathbf{Ax} = \mathbf{b}$ Si la matrice $\mathbf{A}$ est mal conditionnée, il est conseillé, avant de procéder à la résolution du système, d'effectuer un *préconditionnement* de la matrice $\mathbf{A}$ afin que la matrice résultante après cette opération soit mieux conditionnée. Nous présentons dans ce paragraphe une telle méthode de préconditionnement.

Considérons un système linéaire

$$\mathbf{A}_0\mathbf{x} = \mathbf{b},\ \mathbf{A} \in \mathbb{R}^{n\times n} \tag{2.10.1}$$

avec $\mathbf{A}_0$ matrice mal conditionnée. Supposons que nous avons calculer l'inverse $\mathbf{B}_0$ de cette matrice et nous avons que la quantité $\|\mathbf{A}_0 \cdot \mathbf{B}_0 - \mathbf{I}\|$ est supérieure à un seuil donné, c'est-à-dire que l'inversion s'est mal deroulée. On construit le système préconditionné

$$\mathbf{B}_0\mathbf{A}_0\mathbf{x} = \mathbf{B}_0\mathbf{b} \tag{2.10.2}$$

qui est équivalent au système initial (2.10.1). On pose $\mathbf{A}_1 = \mathbf{B}_0\mathbf{A}_0$ et (2.10.2) devient

$$\mathbf{A}_1\mathbf{x} = \mathbf{B}_0\mathbf{b} \tag{2.10.3}$$

Si on suppose que $\mathbf{B}_0$ est régulière, on peut calculer l'inverse de $\mathbf{A}_1$ et soit $\mathbf{B}_1$ cette matrice inverse. Ainsi le produit $\mathbf{B}_1\mathbf{B}_0$ est une nouvelle approximation de l'inverse de $\mathbf{A}_0$. Si cette approximation n'est pas bonne, nous pouvons continuer en calculant une nouvelle matrice $\mathbf{B}_2$ qui est l'inverse de la matrice $\mathbf{A}_2 = \mathbf{B}_1\mathbf{B}_0\mathbf{A}_0$ et ainsi de suite. Nous obtenons ainsi la suite des matrices régulières $\mathbf{B}_i$ telles que :

$$\mathbf{B}_k\mathbf{B}_{k-1}\cdots\mathbf{B}_1\mathbf{B}_0 \simeq \mathbf{A}_0^{-1} \tag{2.10.4}$$

Si la matrice inverse obtenue est « assez proche » de $\mathbf{A}^{-1}$, on arrête les itérations.

Nous pouvons aussi utiliser l'expansion polynômiale de von Neumann pour l'inverse d'une matrice. Supposons que la matrice $\mathbf{A}$ peut se décomposer en deux matrices selon la formule :

$$\mathbf{A}_0 = \mathbf{P} - \mathbf{Q} \tag{2.10.5}$$

avec $\mathbf{P}$ matrice régulière. Posons

$$\mathbf{G} = \mathbf{P}^{-1}\mathbf{Q} \tag{2.10.6}$$

et faisons l'hypothèse que le rayon spectral de $\mathbf{G}$ est, en module, inférieur à 1 : $\mid \rho(\mathbf{G}) \mid < 1$. Alors l'expansion polynômiale de von Neumann est donnée par la relation :

$$\mathbf{A}_0^{-1} = \left(\mathbf{I} + \mathbf{G} + \mathbf{G}^2 + \cdots + \mathbf{G}^{m-1}\right)\mathbf{P}^{-1} \tag{2.10.7}$$

Pour appliquer cette expansion dans notre cas, on constate qu'à chaque étape du calcul nous avons

$$\mathbf{A}_i = \mathbf{B}_i^{-1} - \mathbf{E}_i;\ i = 0, 1, \ldots, k \tag{2.10.8}$$

où $\mathbf{B}_i$ est l'inverse approchée de $\mathbf{A}_i$ et $\mathbf{E}_i$ est la matrice d'erreur de l'approximation. Afin d'exploiter une expansion polynômiale de von Neumann, on doit décompose $\mathbf{A}_0$ en deux matrices. Si on compare avec (2.10.8) on en conclut que pour la première étape on doit avoir

$$\mathbf{A}_0 = \mathbf{B}_0^{-1} - \mathbf{E}_0 \tag{2.10.9}$$

Ainsi la formule (2.10.7) s'applique pour $\mathbf{G} = \mathbf{G_0} = \mathbf{B_0}\mathbf{E_0}$, $\mathbf{P^{-1}} = \mathbf{B_0}$ et à condition que $\mid \rho(\mathbf{B_0}\mathbf{E_0}) \mid < 1$.

L'expansion de Neumman (2.10.7) peut aussi se mettre sous la forme

$$\mathbf{A}_0^{-1} = \left(\mathbf{I} + \mathbf{G}^{2^s}\right)\left(\mathbf{I} + \mathbf{G}^{2^{s-1}}\right)\cdots\left(\mathbf{I} + \mathbf{G}^2\right)(\mathbf{I} + \mathbf{G})\mathbf{P}^{-1} \tag{2.10.10}$$

qui, dans notre cas, donne

$$\mathbf{A}_0^{-1} = \left(\mathbf{I} + \mathbf{G}_0^{2^s}\right)\left(\mathbf{I} + \mathbf{G}_0^{2^{s-1}}\right)\cdots\left(\mathbf{I} + \mathbf{G}_0^2\right)(\mathbf{I} + \mathbf{G}_0)\mathbf{B}_0 \tag{2.10.11}$$

à condition que $\mid \rho(\mathbf{G}_0) \mid < 1$.

Comme

$$\mathbf{B}_1 = (\mathbf{B}_0\mathbf{A}_0)^{-1} = \mathbf{A}_0^{-1}\mathbf{B}_0^{-1} \tag{2.10.12}$$

on obtient, à cause de la relation (2.10.11)

$$\mathbf{B}_1 = \left(\mathbf{I} + \mathbf{G}_0^{2^s}\right)\left(\mathbf{I} + \mathbf{G}_0^{2^{s-1}}\right)\cdots\left(\mathbf{I} + \mathbf{G}_0^2\right)(\mathbf{I} + \mathbf{G}_0) \simeq \mathbf{I} + \mathbf{G}_0 \tag{2.10.13}$$

ce qui permet d'amorcer la prochaine étape du calcul de l'expansion.

2.10.1 Exercices

EXERCICE 2.19 *Établir un algorithme pour la méthode de préconditionnement exposée ci-dessus.*

EXERCICE 2.20 *Soit le système linéaire*

$$\begin{bmatrix} 5 - \frac{1}{68} & 7 & 6 & 5 \\ 7 & 10 & 8 & 7 \\ 6 & 8 & 10 & 9 \\ 5 & 7 & 9 & 10 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 23 \\ 32 \\ 33 \\ 31 \end{bmatrix}$$

avec réponse exacte $\mathbf{x} = \begin{bmatrix} 1.00063979 & 1 & 1 & 1 \end{bmatrix}$.

Vérifier, en utilisant Silab, la qualité de la résolution directe du système et améliorer cette qualité en utilisant l'algorithme de l'exercice précédent et des fonctions de Scilab.

2.11 Inversion par perturbation des matrices singulières

Considérons un système linéaire

$$\mathbf{A}_0\mathbf{x} = \mathbf{b},\ \mathbf{A}_0 \in \mathbb{R}^{n\times n}$$

et supposons que la matrice $\mathbf{A}_0$ est mal conditionnée, c'est-à-dire le conditionnement $\kappa(\mathbf{A}_0)$ a une très grande valeur, voire elle est singulière, c'est-à-dire $\kappa(\mathbf{A}_0)$ est infini. Nous envisageons de faire subir à la matrice $\mathbf{A}_0$ des petites perturbations afin de la rendre bien conditionnée. Il va de soi que ces perturbations ne doivent pas modifier de façon importante la solution du système. Nous obtenons ainsi la matrice perturbée

$$\mathbf{A}(\varepsilon) = \mathbf{A}_0 + \varepsilon\mathbf{A}_1 + \varepsilon^2\mathbf{A}_2 + \cdots, \quad \varepsilon \in \mathbb{R},\ |\,\varepsilon\,| < \varepsilon_{\max} \tag{2.11.1}$$

On suppose que $\mathbf{A}(\varepsilon)$ est inversible dans le disque $0 < |\,\varepsilon\,| < \varepsilon_{\max}$. Ainsi la matrice inverse peut être calculée en utilisant un développement en séries de Laurent :

$$\mathbf{A}^{-1}(\varepsilon) = \frac{1}{\varepsilon^s}\left(\mathbf{X}_0 + \varepsilon\mathbf{X}_1 + \varepsilon^2\mathbf{X}_2 + \cdots\right) \tag{2.11.2}$$

En utilisant le fait que $\mathbf{A}(\varepsilon)\,\mathbf{A}^{-1}(\varepsilon) = \mathbf{I}$, et en groupant les coefficients de même puissance pour ε, on obtient

$$\begin{aligned}
\mathbf{A}_0\mathbf{X}_0 &= \mathbf{0}\\
\mathbf{A}_0\mathbf{X}_1 + \mathbf{A}_1\mathbf{X}_0 &= \mathbf{0}\\
&\vdots\ \ \vdots\ \ \vdots\\
\mathbf{A}_0\mathbf{X}_s + \cdots + \mathbf{A}_s\mathbf{X}_0 &= \mathbf{I}\\
\mathbf{A}_0\mathbf{X}_{s+1} + \cdots + \mathbf{A}_{s+1}\mathbf{X}_0 &= \mathbf{0}\\
&\vdots\ \ \vdots\ \ \vdots
\end{aligned} \tag{2.11.3}$$

Ce système infini d'équations linéaires détermine de façon unique les coefficients de la série de Laurent. Pour la démonstration de ce théorème voir K. E. Avrachenkov, pp. 17-18.

Pour calculer la solution, il faut d'abord déterminer la valeur de s. Pour ce faire on construit de façon itérative les matrices

$$\mathbf{B}^{(k)} = \begin{bmatrix}
\mathbf{A}_0 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0}\\
\mathbf{A}_1 & \mathbf{A}_0 & \mathbf{0} & \cdots & \mathbf{0}\\
\mathbf{A}_2 & \mathbf{A}_1 & \mathbf{A}_0 & \cdots & \mathbf{0}\\
\vdots & \vdots & \vdots & \ddots & \vdots\\
\mathbf{A}_k & \mathbf{A}_{k-1} & \mathbf{A}_{k-2} & \cdots & \mathbf{A}_0
\end{bmatrix} \tag{2.11.4}$$

Selon le test de Sain et Massey la valeur de s est la valeur minimale de k pour laquelle on a

$$\operatorname{rang}\mathbf{B}^{(k)} = \operatorname{rang}\mathbf{B}^{(k-1)} + n \tag{2.11.5}$$

Pour le calcul des matrices $\mathbf{X}_i$ on procède par s étapes de réduction. À l'étape ℓ on forme le

système d'équations

$$\begin{array}{rcl} \mathbf{A}_0^{(\ell)}\mathbf{X}_0^{(\ell)} & = & \mathbf{R}_0^{(\ell)} \\ \mathbf{A}_0^{(\ell)}\mathbf{X}_1^{(\ell)} + \mathbf{A}_1^{(\ell)}\mathbf{X}_0^{(\ell)} & = & \mathbf{R}_1^{(\ell)} \\ \vdots \quad \vdots & & \vdots \\ \mathbf{A}_0^{(\ell)}\mathbf{X}_{s-\ell}^{(\ell)} + \cdots + \mathbf{A}_{s-\ell}^{(\ell)}\mathbf{X}_0^{(\ell)} & = & \mathbf{R}_{s-\ell}^{(\ell)} \end{array} \tag{2.11.6}$$

On remarque que si $\ell = 0$, nous avons le système initial (2.11.3) et par conséquent on a

$$\begin{array}{l} \mathbf{R}_i^{(0)} = \mathbf{0}, \text{ pour } i = 0, \ldots, s-1 \\ \mathbf{R}_s^{(0)} = \mathbf{I} \\ \mathbf{A}_i^{(0)} = \mathbf{A}_i, \text{ pour } i = 0, \ldots, s \end{array} \tag{2.11.7}$$

Pour les matrices $\mathbf{A}_i^{(\ell)}$ et $\mathbf{R}_i^{(\ell)}$, et $\ell = 1, \ldots, s$ on a les formules suivantes :

$$\begin{array}{l} \mathbf{A}_i^{(\ell)} = \mathbf{M}^{(\ell)}\mathbf{U}_i^{(\ell)}\mathbf{Q}^{(\ell)}, \; i = 0, \ldots, s-1 \\ \mathbf{R}_i^{(\ell)} = \mathbf{M}^{(\ell)} \left(-\sum_{j=0}^{i} \mathbf{U}_{i-j}^{(\ell)} \mathbf{A}_0^{(\ell-1)+} \mathbf{R}_j^{(\ell-1)} + \mathbf{R}_{i+1}^{(\ell-1)} \right), \; i = 0, \ldots, s-1 \end{array} \quad , \; \ell = 1, \ldots, s \tag{2.11.8}$$

avec

$$\begin{array}{ll} \mathbf{Q}^{(\ell)} \in \mathbb{R}^{n \times m}, & \text{matrice dont les colonnes forment une base} \\ & \text{pour le noyau droit de } \mathbf{A}_0^{(\ell-1)} \\ & \text{et où } m = n - \operatorname{rang}\left(\mathbf{A}_0^{(\ell-1)}\right) \\ \mathbf{M}^{(\ell)} \in \mathbb{R}^{m \times n}, & \text{matrice dont les lignes forment une base} \\ & \text{pour le noyau gauche de } \mathbf{A}_0^{(\ell-1)} \\ \mathbf{U}_0^{(\ell)} & = \mathbf{A}_1^{(\ell-1)}, \quad \mathbf{U}_i^{(\ell)} = \mathbf{A}_{i+1}^{(\ell-1)} - \sum_{j=1}^{i} \mathbf{A}_j^{(\ell-1)} \mathbf{A}_0^{(\ell-1)} \mathbf{U}_{i-j}^{(\ell)}, \\ & i = 0, \ldots, s-1 \\ \mathbf{A}^{+} & \text{la pseudoinverse de la matrice} \mathbf{A} (cf. infra) \end{array} \tag{2.11.9}$$

Après s étapes on obtient le système final d'équations réduites :

$$\mathbf{A}_0^{(s)}\mathbf{X}_0^{(s)} = \mathbf{R}_0^{(s)} \tag{2.11.10}$$

On peut démontrer (voir K. E. Avrachenkov, pp. 21-23) que ce système et le système initial donné par (2.11.3) sont équivalents, donc la matrice $\mathbf{A}_0^{(s)}$ est régulière et, par conséquent

$$\mathbf{X}_0^{(s)} = \left[\mathbf{A}_0^{(s)}\right]^{-1} \mathbf{R}_0^{(s)} \tag{2.11.11}$$

Pour obtenir donc la solution $\mathbf{X}_0 = \mathbf{X}_0^{(0)}$, il faut procéder à une recursion en arrière selon le schéma :

$$\mathbf{X}_0^{(\ell-1)} = \mathbf{A}_0^{(\ell-1)+}\mathbf{R}_0^{(\ell-1)} + \mathbf{Q}^{(\ell)}\mathbf{X}_0^{(\ell)}, \; \ell = s, \ldots, 1 \tag{2.11.12}$$

En posant

$$\mathbf{R}_i^{(0)} = \begin{cases} -\sum_{j=1}^{k} \mathbf{A}_{i+j}\mathbf{X}_{k-j}, & \text{si } i = 0, \ldots, s-1 \\ & ; \; k = 1, 2, \ldots \\ \mathbf{I} - \sum_{j=1}^{k} \mathbf{A}_{i+j}\mathbf{X}_{k-j}, & \text{si } i = s \end{cases} \tag{2.11.13}$$

nous pouvons calculer les coefficients de la série de Laurent :

$$\mathbf{X}_k = \sum_{i=0}^{s} \mathbf{G}_{0i}^{(s)} \cdot \mathbf{R}_i^{(0)};\ k = 1, 2, \ldots \tag{2.11.14}$$

avec $\mathbf{G}_{0i}^{(s)} \in \mathbb{R}^{n \times n}$ le bloc $(0, j)$ de la matrice $\mathbf{B}^{(k)+}$.

Remarquons que la formule (2.11.14) est une généralisation de l'expansion de von Neumann. En effet dans cette dernière expansion on suppose que la matrice $\mathbf{A}_0$ est régulière. Donc $s = 0$ et $\mathbf{G}^{(0)} = \mathbf{A}_0^{-1}$ et la relation (2.11.14) donne :

$$\mathbf{X}_0 = -\mathbf{A}_0^{-1}, \quad \mathbf{X}_k = -\mathbf{A}_0^{-1} \sum_{j=1}^{k} \mathbf{A}_j \mathbf{X}_{k-j};\ k = 1, 2, \ldots \tag{2.11.15}$$

qui est l'expansion de von Neumann.

2.11.1 Exercices

EXERCICE 2.21 *Établir un algorithme qui permet d'inverser une matrice singulière, en lui faisant subir des petites perturbations.*

EXERCICE 2.22 *Utiliser l'algorithme de l'exercice précédent et des fonctions de Scilab pour calculer l'inverse de la matrice*

$$\mathbf{A}_0 = \begin{bmatrix} 1 & 2 & 1 \\ -1 & 1 & 0 \\ 0 & 3 & 1 \end{bmatrix}$$

avec matrice de perturbation

$$\mathbf{A}_1 = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ -1 & 0 & 1 \end{bmatrix}$$

2.12 Références

Les livres et articles qui sont pris en compte pour la redaction de ces notes sur l'Algèbre Linéaire et les Perturbations sont les suivants :

J. H. WILKINSON : *Rounding errors in algebraic processes*, Dover, 1994
J. H. WILKINSON : *The algebraic eigenvalue problem*, Clarendon Pr., 1965
F. R. GANTMACHER : *Théorie des matrice, tome 1, Théorie générale*, Dunod, 1966
V. N. FADDEEVA : *Computational methods of linear algebra*, Dover, 1959
N. J. HIGHAM : *Accurancy and stability of numerical algorithms*, Siam, 1996
J. STOER, R. BULIRSCH : *Introduction to numerical analysis*, Second ed., Springer-Verlag, 1992
K. E. AVRACHENKOV : Analytic perturbation theory and its applications, Un. South Australia, 1999

2.A APPENDICE.- BREF RAPPEL DE L'ALGÈBRE LINÉAIRE

Considérons un quelconque problème de l'algèbre linéaire dont la solution est un vecteur $\mathbf{x} = [x_1 \cdots x_n]^\top \in U \subset \mathbb{R}^n$. Le calcul de cette solution peut se faire en analyse numérique par un algorithme itératif ϕ selon le schéma suivant :

$$\begin{array}{l} \mathbf{x}(0) \text{ valeur de départ choisie au hasard.} \\ \mathbf{x}(k+1) = \phi(\mathbf{x}(k)) \ ; \ k = 0, 1, \ldots \end{array} \tag{2.A.1}$$

L'algorithme ϕ est donc une fonction de U dans lui-même et k est le numéro d'itération.

L'application du schéma $(2.A.1)$ permet d'obtenir une approximation de la vraie solution. Nous dirons que l'algorithme itératif *converge* vers la solution si $\|\mathbf{x}(k+1) - \mathbf{x}(k)\| < \textit{seuil}$, où $\|\bullet\|$ norme établie selon la métrique usuelle.

Le schéma $(2.A.1)$ peut être réalisé selon deux méthodes :

– *parallèle* : Toutes les coordonnées du vecteur $\mathbf{x}(k+1)$ sont calculées simultanément :

$$\forall i = 1, \ldots, n \ : \ x_i(k+1) = \phi_i(x_1(k), \ldots x_n(k))$$

– *séquentielle* : Pour le calcul d'une coordonnée du vecteur $\mathbf{x}(k+1)$ on tient compte des valeurs des coordonnées déjà calculées :

$$\forall i = 1, \ldots, n \ : \ x_i(k+1) = \phi_i(x_1(k+1), \ldots, x_{i-1}(k+1), x_i(k), \ldots, x_n(k))$$

Notons que l'ordre des indices pour le calcul peut être quelconque et même aléatoire.

En algèbre linéaire il y a aussi des algorithme à résolution directe. On peut considérer que dans ce cas il s'agit des algorithmes séquentiels et dont le nombre d'itérations est égal à l'ordre du problème.

Essayons maintenant de formaliser la notion de la convergence d'un algorithme. À l'algorithme ϕ on peut associer l'ensemble de ses points fixes $\phi^\infty(U)$ défini par

$$\phi^\infty(U) = \{\mathbf{u} \in U \quad | \quad \phi(\mathbf{u}) = \mathbf{u}\}$$

Un point fixe $\mathbf{u}^\circ$ est dit *attractif* s'il admet un voisinage $V(\mathbf{u}^\circ) \subset U$ tel que

$$\forall \mathbf{u} \in V(\mathbf{u}^\circ), \quad \lim_{k \to \infty} \phi^k(\mathbf{u}) = \mathbf{u}^\circ.$$

De la même fa, on dira qu'un sous-ensemble U° est attractif si

$$\forall \mathbf{u}^\circ \in U^\circ \quad \exists V(\mathbf{u}^\circ) \quad : \forall \mathbf{u} \in V(\mathbf{u}^\circ), \quad \lim_{k \to \infty} \phi^k(\mathbf{u}) \in U^\circ.$$

Tout l'art de l'analyse numérique pour la résolution d'un problème dans un espace U et dont l'ensemble des solutions forme un ensemble qui sera noté U^*, est de trouver un algorithme ϕ tel

que $\phi^{\infty}(U)$ soit « proche » et attractif([2]). On définit aussi le *bassin d'attraction* associé à un point fixe $\mathbf{u}^{\circ}$. C'est l'ensemble de toutes les solutions initiales $\mathbf{u}(0)$ qui conduisent à une suite $\phi^{k}(\mathbf{u}(0))$ convergente vers $\mathbf{u}^{\circ}$. Un tel sous-ensemble est généralement très difficile à décrire, sauf lorsque que tout U est bassin d'attraction d'un unique point fixe.

On dira qu'un algorithme *converge* si la suite de valeurs qu'il engendre tend vers une limite dans l'espace considéré. Bien sûr pour que cette convergence soit « intéressante » la limite doit être la solution du problème. En d'autres termes, l'algorithme converge si ses points fixes attractifs sont candidats pour être solution du problème posé et leurs bassins d'attraction recouvrent la totalité de la région de U où nous avons défini notre problème([3]).

La *vitesse de convergence* d'un algorithme mesure le taux de la décroissance vers zéro de la distance entre les valeurs engendrées et leur limite. Par exemple soit, dans $\mathbb{R}^n$, la limite $\mathbf{u}^*$ vers laquelle converge la suite $(\mathbf{u}(k))_{k\in\mathbb{N}}$ (engendrée par un algorithme ϕ donné) et $\|\bullet\|$ la norme euclidienne :

DÉFINITION 2.A.1 (Vitesse de Convergence).
- *Si* $\lim\limits_{k\to\infty}\sup\dfrac{\|\mathbf{u}(k+1)-\mathbf{u}^*\|}{\|\mathbf{u}(k)-\mathbf{u}^*\|}\leq\alpha<1$, *alors on dit que la convergence est* linéaire *et* α *est le* taux de convergence associé. *(On remarque qu'il s'agit ici de la convergence des suites géométriques.)*
- *Si* $\lim\limits_{k\to\infty}\sup\dfrac{\|\mathbf{u}(k+1)-\mathbf{u}^*\|}{\|\mathbf{u}(k)-\mathbf{u}^*\|}=0$, *alors on dit que la convergence est* super-linéaire.
- *Si* $\exists\gamma>1$ *tel que* $\lim\limits_{k\to\infty}\sup\dfrac{\|\mathbf{u}(k+1)-\mathbf{u}^*\|}{\|\mathbf{u}(k)-\mathbf{u}^*\|^{\gamma}}\leq M<+\infty$, *alors la convergence est* super-linéaire d'ordre γ *et, en particulier, si* $\gamma=2$, *on parle de vitesse de* convergence quadratique.

Ces définitions ont intrinsèquement un caractère asymptotique et on ne peut, en général, démontrer que la convergence est super-linéaire ou quadratique, que dans un voisinage de la limite recherchée $\mathbf{u}^*$ (loin de $\mathbf{u}^*$, il se peut très bien que l'algorithme concerné converge lentement – ou pas du tout –, bien que la vitesse de convergence asymptotique soit en théorie quadratique). D'autre part, deux algorithmes ayant la même vitesse de convergence asymptotique peuvent très bien nécessiter des temps de calculs très différents si le nombre d'opérations exécutées à chaque itération est très différent.

Cette définition est néanmoins très utile pour évaluer la qualité de la limite obtenue par un algorithme. Par exemple, supposons, pour fixer les idées, que la vitesse de convergence asymptotique d'un algorithme soit quadratique dans un voisinage de $\mathbf{u}^*$, avec $M=100$. Il faut donc avoir pour une itération k l'inégalité

$$\frac{\|\mathbf{u}(k+1+i)-\mathbf{u}^*\|}{\|\mathbf{u}(k+i)-\mathbf{u}^*\|}\leq 100,\qquad\forall i=0,1,...$$

2. La notion de proximité peut être comprise comme une inclusion dans l'un sens ou dans l'autre - ce qui n'est bien sûr pas la même chose sur le plan du résultat. On pourrait même demander à avoir l'égalité entre $\phi^{\infty}(U)$ et U^*, mais ce serait très exigeant et peu efficace.

3. Les points fixes non attractifs ne présentent aucun intérêt car leur rencontre, au cours d'itérations, ne peut être que fortuite. Par conséquent nous ne pouvons pas fonder sur ces points des shémas itératifs qui convergent.

Si, pour l'itération k, on a $\|\mathbf{u}(k)-\mathbf{u}^*\| \leq 10^{-3}$ c'est-à-dire on est dans un voisinage de la solution, alors on vérifie aisément que $\|\mathbf{u}(k+1)-\mathbf{u}^*\| \leq 10^{-4}$, $\|\mathbf{u}(k+2)-\mathbf{u}^*\| \leq 10^{-6}$, $\|\mathbf{u}(k+3)-\mathbf{u}^*\| \leq 10^{-10}$, etc. L'erreur d'approximation passe donc, en trois itérations, de 10^{-3} à 10^{-10}.

Nous présentons dans les chapitres suivants différentes méthodes de résolution des problèmes d'algèbre linéaire numérique. Lors de la présentation de ces méthodes la terminologie et les résultats fondamentaux de l'algèbre linéaire seront considérés comme connus. Nous nous bornerons donc de lister les principaux résultats qui seront utilisés par la suite.

Définitions et fondements théoriques des applications linéaires

Dans tout le cours de l'analyse numérique, les espaces vectoriels seront considérés sur le corps des réels $\mathbb{R}$.

(1) On appelle application linéaire (ou opérateur linéaire) de l'espace vectoriel U dans l'espace vectoriel V une application f de U dans V vérifiant la propriété suivante :

$$\forall \mathbf{u}, \mathbf{u}' \in E, \forall \alpha, \beta \in \mathbb{R} \ : \ f(\alpha \mathbf{u} + \beta \mathbf{u}') = \alpha f(\mathbf{u}) + \beta f(\mathbf{u}')$$

(2) Une application linéaire de U dans V est appelée *forme linéaire*.

(3) Une application linéaire de U dans lui-même est appelée *endomorphisme*.

(4) Une application linéaire bijective de U dans V est appelée *isomorphisme*.

(5) Une application linéaire bijective de E dans lui-même est appelée *automorphisme*.

(6) Soit f une application linéaire bijective (ou isomorphisme) de U sur V. Alors f est inversible et l'application inverse f^{-1} est un isomorphisme de V sur U.

(7) Soit f une application linéaire de U dans V.
- Si f est injective, l'image d'une famille libre de U est une famille libre de V.
- Si f est surjective, l'image d'une famille génératrice de E est une famille génératrice de V.
- Si f est bijective, l'image d'une base de U est une base de F.

(8) Soit f un opérateur linéaire de U dans V. On définit :
- *L'image* de f : $R(f) = \{y \in V / \exists x \in U \text{ avec } y = f(x)\}$, c'est-à-dire $R(f)$ est composé de tous les vecteurs de V qui peuvent se représenter sous la forme $f(\mathbf{x})$ avec $\mathbf{x} \in U$. On a $R(f) = f(E)$ et $R(f)$ est un sous-espace vectoriel de V.
- *Le noyau* de f : $N(f) = \{x \in U / f(x) = 0\}$, c'est-à-dire $N(f)$ est composé de tous les vecteurs $\mathbf{x} \in U$ qui s'annulent par f. $N(f)$ est un sous-espace vectoriel de U.

(9) Pour l'application linéaire f de U dans V nous avons $\dim(N(f)) + \dim(R(f)) = \dim(U)$.

(10) Si f est injective, alors $N(f) = \{0\}$.

Définitions et fondements théoriques des matrices

Considérons deux espaces vectoriels $U \subseteq \mathbb{R}^m$ avec base $(\mathbf{u}_i)_{i=1,\ldots,m}$ et $V \subseteq \mathbb{R}^n$ avec base $(\mathbf{v}_j)_{j=1,\ldots,n}$. Soit $f : U \to V$ une application linéaire. L'image de f est engendrée par les images des vecteurs de la base de U, $f(\mathbf{u}_i); i = 1, \ldots, m$. En effet, soit $\mathbf{x} \in U$. Alors l'image de $\mathbf{x}$ par f

s'écrit $f(\mathbf{x}) = \sum_{i=1}^{m} f(\mathbf{u}_i)\mathbf{x}$. Avec les composants des vecteurs $f(\mathbf{u}_i)$ on peut former le tableau

$$\mathbf{F} = \begin{bmatrix} f_{11} & \cdots & f_{1,m} \\ \vdots & \ddots & \vdots \\ f_{n1} & \cdots & f_{nm} \end{bmatrix}, \text{ où } f_{ji} \text{ est la j-ième composante du vecteur } f(\mathbf{u}_i)$$

Le tableau $\mathbf{F}$ s'appelle *la matrice de l'application linéaire* f et elle décrit complètement f, à savoir si $\mathbf{x} \in U$, alors $f(\mathbf{x}) = \mathbf{F}\mathbf{x}$.

(1) Une matrice est dite *régulière* si son inverse existe. Sinon elle est *singulière*.

(2) Le *rang* d'un matrice est la dimension de la plus grande sous–matrice carrée qui est régulière. Si $f : U \to V$ application linéaire, alors $\text{rang}(\mathbf{F}) = \dim(V)$, où $\mathbf{F}$ est la matrice de f.

(3) Une *matrice de permutation* $\mathbf{P}$ est une matrice dont chaque ligne et chaque colonne a un élément égal à 1 et tous les autres sont nuls. Elle s'appelle matrice de permutation car si on la multiplie avec une autre matrice provoque la permutation de deux lignes ou de deux colonnes de cette dernière. De plus nous avons $\det \mathbf{P} = (-1)^{k_\mathbf{P}}$, où $k_\mathbf{P}$ est le nombre de lignes de $\mathbf{P}$ qui sont différentes des lignes correspondantes de la matrice identité $\mathbf{I}$.

(4) Une matrice carrée $\mathbf{A}$ est *orthogonale* si $\mathbf{A}^\top \mathbf{A} = \mathbf{I}$, d'où $\mathbf{A}^{-1} = \mathbf{A}^\top$.

(5) Une matrice carrée $\mathbf{A}$ est *diagonalement dominante* si $\mid a_{ii} \mid \geq \sum_{\substack{k=1 \\ k \neq i}}^{n} \mid a_{ik} \mid ; i = 1, \ldots, n$ et il existe au moins un indice i_0 tel que $\mid a_{i_0 i_0} \mid > \sum_{\substack{k=1 \\ k \neq i_0}}^{n} \mid a_{i_0 k} \mid$.

La définition est valable si à la place des lignes on utilise les colonnes à la formule précédente.
La matrice est *strictement diagonalement dominante* si dans la formule précédente nous avons des inégalités strictes. Une telle matrice est régulière.

(6) Une matrice $\mathbf{A}$ est *symétrique* si $\mathbf{A} = \mathbf{A}^\top$.
Une matrice symétrique $\mathbf{A}$ est *définie positive* si

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} > 0 ; \forall \mathbf{x} \in \mathbb{R}^n$$

Elle est *semi-définie positive* si

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} \geq 0 ; \forall \mathbf{x} \in \mathbb{R}^n$$

(7) **Factorisation LU**.- Toute matrice régulière $\mathbf{A}$ peut se factoriser de manière unique selon le produit

$$\mathbf{A} = \mathbf{L}\mathbf{U}$$

avec

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \ell_{21} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \ell_{n1} & \ell_{n2} & \cdots & 1 \end{bmatrix} ; \mathbf{R} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ 0 & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & r_{nn} \end{bmatrix}$$

(8) Pour toute matrice $\mathbf{A}$ régulière, il existe une matrice de permutation $\mathbf{P}$, telle que

$$\mathbf{PA} = \mathbf{LR}$$

Dans ce cas nous avons aussi

$$\det \mathbf{A} = (-1)^{k_{\mathbf{P}}} \det \mathbf{R} = (-1)^{k_{\mathbf{P}}} r_{11} r_{22} \cdots r_{nn}$$

(9) Une matrice strictement diagonalement dominante $\mathbf{A}$ peut se factoriser de manière unique selon le produit

$$\mathbf{A} = \mathbf{LDL}^{\top}$$

avec $\mathbf{L}$ matrice triangulaire inférieure et $\mathbf{D}$ une matrice diagonale.

(10) Une matrice définie positive $\mathbf{A}$ peut se factoriser de manière unique selon le produit

$$\mathbf{A} = \mathbf{LL}^{\top}$$

avec $\mathbf{L}$ matrice triangulaire inférieure avec $\ell_{ii} > 0$ pour tout i.

(11) Soit une matrice $\mathbf{A} \in \mathbb{R}^{n \times n}$ regulière. Un réel $\lambda \in \mathbb{R}$ est une *valeur propre* de $\mathbf{A}$ s'il existe un vecteur $\mathbf{x} \in \mathbb{R}^n$, avec $\mathbf{x} \neq \mathbf{0}$, tel que $\mathbf{Ax} = \lambda \mathbf{x}$. Le vecteur $\mathbf{x}$ est appelé *vecteur propre* de $\mathbf{A}$ associé à la valeur propre λ.

(12) *Rayon spectral* d'une matrice $\mathbf{A}$: $\rho(\mathbf{A}) = \max\{ \mid \lambda \mid / \lambda \in \mathbb{R}, \lambda \text{ valeur propre de } \mathbf{A} \}$.

(13) **Théorème de Gershgorin**.- Toutes les valeurs propres de la matrice $\mathbf{A}$ se trouvent dans l'union des cercles de Gershgorin

$$C_i = \left\{ z \in \mathbb{R} \;\middle/\; \mid z - a_{ii} \mid \leq \sum_{\substack{k=1 \\ k \neq i}}^{n} \mid a_{ik} \mid \right\} \; ; \; i = 1, \ldots, n$$

(14) **Quotient de Rayleigh**.- Si $\mathbf{A}$ est une matrice carrée, le quotient de Rayleigh est

$$R[\mathbf{x}] = \frac{\mathbf{x}^{\top} \mathbf{A} \mathbf{x}}{\mathbf{x}^{\top} \mathbf{x}} \; ; \; \mathbf{x} \in \mathbb{R}^n \; , \; \mathbf{x} \neq \mathbf{0}$$

Si $\mathbf{A}$ est une matrice carrée symétrique, alors

$$\mid R[\mathbf{x}] \mid \leq \mid \lambda_{\max} \mid \quad \forall \mathbf{x} \in \mathbb{R}^n$$

où $\lambda_{\max}$ est la plus grande, en module, valeur propre de $\mathbf{A}$.

(15) À tout vecteur $\mathbf{x} = [x_1, x_2, \cdots, x_n]^{\top} \in \mathbb{R}^n$ on associe la matrice–colonne $[\mathbf{x}]$ de format $(n, 1)$ et dont la i-ième ligne est la i-ième coordonnée du vecteur $\mathbf{x}$ dans la base canonique de $\mathbb{R}^n$. Si on note par $\mathcal{M}(n, 1)$ l'ensemble de matrices-colonne de format $(n \times 1)$, nous pouvons avoir une application linéaire entre $\mathbb{R}^n$ et $\mathcal{M}(n, 1)$ qui est, en plus, bijective. $\mathbb{R}^n$ et $\mathcal{M}(n, 1)$ sont donc isomorphes.

(16) Soit $\mathbf{A}$ une matrice de format (m, n), le *noyau* de $\mathbf{A}$ est l'ensemble $N(\mathbf{A}) = \{\mathbf{x} \in \mathbb{R}^n / \mathbf{Ax} = \mathbf{0}\}$. $N(\mathbf{A})$ est un sous–espace vectoriel de $\mathbb{R}^n$.

(17) Soit $\mathbf{A}$ une matrice de format (m,n), l'*image* de $\mathbf{A}$ est l'ensemble $R(\mathbf{A}) = \{\mathbf{y} \in \mathbb{R}^m / \mathbf{x} \in \mathbb{R}^n$ et $\mathbf{A}\mathbf{x} = \mathbf{y}\}$. $R(\mathbf{A})$ est un sous–espace vectoriel de $\mathbb{R}^m$ engendrée par les colonnes de $\mathbf{A}$.

(18) On appelle *rang d'une matrice* $\mathbf{A}$, la dimension de l'image de $\mathbf{A}$: $\text{rang}(\mathbf{A}) = \dim R(\mathbf{A})$.

(19) Soit $\mathbf{A}$ une matrice carree d'ordre n, alors $\text{rang}(\mathbf{A}) = n \quad \Leftrightarrow \quad \mathbf{A}$ est inversible.

(20) **Théorème noyau–image ou théorème du rang.-** Soit $\mathbf{A}$ une matrice de format (m,n). Alors : $n = \dim R(\mathbf{A}) + \dim N(\mathbf{A}) = \text{rang}(\mathbf{A}) + \dim N(\mathbf{A})$.

(21) Une matrice de rang–colonne plein est une matrice $\mathbf{A}$ de format (m,n) dont les colonnes forment une famille libre de $\mathbb{R}^m$. Donc comme $n \leq m$ et que les colonnes de $\mathbf{A}$ constituent une base de $R(\mathbf{A})$, la dimension de $R(\mathbf{A})$ est égale au nombre de colonnes de $\mathbf{A}$. Par conséquent $\dim N(\mathbf{A}) = 0$, c'est-à-dire que $N(\mathbf{A}) = \{\mathbf{0}\}$.

(22) Soit $\mathbf{A}$ une matrice de format (m,n). L'application $f_{\mathbf{A}}$ de $\mathbb{R}^n$ dans $\mathbb{R}^m$ définie par :

$$f_{\mathbf{A}} : \mathbf{x} \in \mathbb{R}^n \to \mathbf{y} \in \mathbb{R}^m \text{ telle que } [\mathbf{y}] = \mathbf{A}\,[\mathbf{x}]$$

est une application linéaire de $\mathbb{R}^n$ dans $\mathbb{R}^m$ appelée *application linéaire canonique* associée à la matrice $\mathbf{A}$.

(23) Soient $\mathbf{A}$ une matrice de format m,n et $f_{\mathbf{A}}$ l'application linéaire associée à la matrice $\mathbf{A}$, alors
- $N(f_{\mathbf{A}}) = N(\mathbf{A})$;
- $R(f_{\mathbf{A}}) = R(\mathbf{A})$;
- le rang de l'application linéaire $f_{\mathbf{A}}$ est égal au rang de la matrice $\mathbf{A}$.

3

MÉTHODES ITÉRATIVES

Pour la redaction de ce chapitre, nous nous sommes fortement inspirés du livre de A. Quarteroni et al. [QSS] cité en bibliographie.

3.1 Introduction

Une méthode itérative consiste à construire une suite de vecteurs $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, ..., \mathbf{x}^{(k)}, ...$ qui convergera vers la solution du système linéaire $\mathbf{Ax} = \mathbf{b}$ à résoudre. La notion de convergence de cette suite est traitée grâce aux normes vectorielles qui ont été abordées dans le chapitre précédent. Les conditions de convergence de ces méthodes vont requérir l'emploi de majorations faisant appel aux normes matricielles, également abordées dans le chapitre précédent.

Les méthodes itératives donnent en théorie, la solution $\mathbf{x}$ d'un système linéaire après un nombre fini d'itérations. A chaque pas, elles nécessitent le calcul du résidu du système, qui permet de décider quand on estime avoir atteint la solution espérée. Dans le cas d'une matrice pleine, leur coût est donc de l'ordre de n^2 opérations à chaque itération, alors que le coût des méthodes directes est, en tout et pour tout, de l'ordre de $2/3n^3$. Les méthodes itératives peuvent devenir compétitives si elles convergent en un nombre d'itérations indépendant de n, ou croissant sous-linéairement avec n. Elles sont utilisées soit pour la résolution de systèmes linéaires de grande taille, soit lorsqu'on dispose d'une solution approchée du système que l'on désire améliorer.

Il est à noter que les méthodes itératives sont sensibles au conditionnement de la matrice du système, et que l'on peut donc être amené à utiliser des techniques de préconditionnement, telles que vues au chapitre précédent.

3.2 Convergence des méthodes itératives

L'idée de base des méthodes itératives est de construire une suite convergente de vecteurs $\mathbf{x}^{(k)}$ telle que

$$\mathbf{x} = \lim_{k \to +\infty} \mathbf{x}^{(k)} \tag{3.2.1}$$

où $\mathbf{x}$ est la solution de

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{3.2.2}$$

En pratique, le calcul devrait être interrompu à la première itération n pour laquelle $\left\|\mathbf{x}^{(n)} - \mathbf{x}\right\| < \varepsilon$, où ε est une tolérance fixée et $\|.\|$ une norme vectorielle donnée. Mais comme la solution exacte n'est évidemment pas connue, il faudra trouver un critère d'arrêt plus commode (voir section 3.4).

Considérons pour commencer les méthodes itératives de la forme :

$$\begin{cases} \mathbf{x}^{(0)} \text{ donné} \\ \mathbf{x}^{(k+1)} = \mathbf{B}\mathbf{x}^{(k)} + \mathbf{f}; \quad k \geq 0 \end{cases} \tag{3.2.3}$$

où $\mathbf{B}$ désigne une matrice carrée $n \times n$ appelée *matrice d'itération* et où $\mathbf{f}$ est un vecteur dépendant de $\mathbf{b}$ (second membre du système à résoudre).

DÉFINITION 3.2.1 *Une méthode itérative de la forme (3.2.3) est dite consistante avec (3.2.2) si* $\mathbf{f}$ *et* $\mathbf{B}$ *sont tels que* $\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{f}$, $\mathbf{x}$ *étant la solution de (3.2.2), ou de faéquivalente, si* f *et* $\mathbf{B}$ *satisfont :*

$$\mathbf{f} = (\mathbf{I} - \mathbf{B})\mathbf{A}^{-1}\mathbf{b}$$

Si on note

$$\mathbf{e}^{(k)} = \mathbf{x}^{(k)} - \mathbf{x} \tag{3.2.4}$$

l'erreur à l'itération k, la condition (3.2.1) revient à $\lim_{k \to +\infty} \mathbf{e}^{(k)} = \mathbf{0}$ pour toute valeur initiale $\mathbf{x}^{(0)}$.

REMARQUE 3.2.1 *La seule propriété de consistance ne suffit pas à assurer la convergence d'une méthode itérative, comme le montre l'exemple suivant :*

EXEMPLE 3.2.1 *On veut résoudre le système linéaire* $2\mathbf{I}\mathbf{x} = \mathbf{b}$ *avec la méthode itérative*

$$\mathbf{x}^{(k+1)} = -\mathbf{x}^{(k)} + \mathbf{b}$$

qui est clairement consistante. Cette suite n'est pas convergente pour une donnée initiale arbitraire. Si par exemple $x^{(0)} = 0$, *la méthode donne* $\mathbf{x}^{(2k)} = \mathbf{0}$, $\mathbf{x}^{(2k+1)} = \mathbf{b}$; $k = 0, 1, ...$
En revanche si $\mathbf{x}^{(0)} = \frac{1}{2}\mathbf{b}$ *la méthode est convergente.*

THÉORÈME 3.2.1 *Si la méthode (3.2.3) est consistante, la suite de vecteurs* $\mathbf{x}^{(k)}$ *de (3.2.3) converge vers la solution de (3.2.2) pour toute donnée initiale* $\mathbf{x}^{(0)}$ *si et seulement si* $\rho(\mathbf{B}) < 1$.

D'après (3.2.4), et grâce à l'hypothèse de consistance, on a $\mathbf{e}^{(k+1)} = \mathbf{B}\mathbf{e}^{(k)}$, d'où

$$\mathbf{e}^{(k+1)} = \mathbf{B}^k \mathbf{e}^{(0)}; \ \forall k = 0, 1, \dots \tag{3.2.5}$$

Il en résulte donc que $\lim_{k \to +\infty} \mathbf{B}^k \mathbf{e}^{(0)} = \mathbf{0}$ pour tout $\mathbf{e}^{(0)}$ si et seulement si $\rho(\mathbf{B}) < 1$.

Réciproquement, supposons que $\rho(\mathbf{B}) > 1$, alors il existe au moins une valeur propre $\lambda(\mathbf{B})$ de module plus grand que 1. Soit $\mathbf{e}^{(0)}$ un vecteur propre associé à λ, alors $\mathbf{B}\mathbf{e}^{(0)} = \lambda \mathbf{e}^{(0)}$. Comme $|\lambda| > 1$, $\mathbf{e}^{(k)}$ ne peut pas tendre vers 0 quand $k \to +\infty$

REMARQUE 3.2.2 *On a déjà vu que pour une tolérance fixée aussi petite que l'on veut, il existe toujours une norme matricielle telle que la norme d'une matrice* $\mathbf{A}$ *soit arbitrairement proche du rayon spectral de* $\mathbf{A}$. *De plus on a toujours* $\rho(\mathbf{A}) \leq \|\mathbf{A}\|$. *La condition de convergence* $\rho(\mathbf{B}) < 1$ *entraîne donc immédiatement que la condition* $\|\mathbf{B}\| < 1$, *pour une norme matricielle consistante arbitraire, est suffisante pour que la méthode converge.*

Il est raisonnable de penser que la convergence est d'autant plus rapide que $\rho(\mathbf{B})$ est petit. Une estimation de $\rho(\mathbf{B})$ peut donc fournir une bonne indication sur la convergence de l'algorithme. En effet le théorème suivant (cf. [SB]) établit une relation entre l'erreur de la solution obtenue après k itération et le rayon spectral :

THÉORÈME 3.2.2 *Pour la méthode (3.2.3) les erreurs* $\mathbf{e}^{(k)} = \mathbf{x}^{(k)} - \mathbf{x}$ *satisfont à*

$$\sup_{\mathbf{e}^{(0)} \neq \mathbf{0}} \lim_{k \to \infty} \sup \sqrt{\frac{\left\|\mathbf{e}^{(k)}\right\|}{\left\|\mathbf{e}^{(0)}\right\|}} = \rho(\mathbf{B})$$

Les itérations définies en (3.2.3) sont un cas particulier des méthodes itératives de la forme

$$\begin{aligned} \mathbf{x}^{(0)} &= \mathbf{f}_0(\mathbf{A}, \mathbf{b}) \\ \mathbf{x}^{(n+1)} &= \mathbf{f}_{n+1}(\mathbf{x}^{(n)}, \mathbf{x}^{(n-1)}, \mathbf{x}^{(n-m)}, \mathbf{A}, \mathbf{b}) \ \text{ pour } n \geq m \end{aligned}$$

où les $\mathbf{f}_i$ sont des fonctions et les $\mathbf{x}^{(m)}, \dots, \mathbf{x}^{(1)}$ des vecteurs donnés. Le nombre de pas dont dépend l'itération courante (ici $m+1$) s'appelle *ordre de la méthode*. Si les fonctions $\mathbf{f}_i$ sont indépendantes de i, la méthode est dite *stationnaire*. Elle est *instationnaire* dans le cas contraire. Enfin, si $\mathbf{f}_i$ dépend linéairement de $\mathbf{x}^{(0)}, ...\mathbf{x}^{(m)}$, la méthode est dite *linéaire*, autrement elle est *non linéaire*.

Au regard de ces définitions, les algorithmes considérés jusqu'à présent sont donc des méthodes **itératives linéaires, stationnaires du premier ordre**.

3.3 Méthodes itératives linéaires

Une technique générale pour définir une méthode itérative linéaire consistante est basée sur la décomposition de $\mathbf{A}$, aussi appelée *splitting*, sous la forme $\mathbf{A} = \mathbf{P} - \mathbf{N}$, où $\mathbf{P}$ est une matrice inversible.

On se donne $\mathbf{x}^{(0)}$, et on calcule $\mathbf{x}^{(k)}$ pour $k \geq 1$, en résolvant le système

$$\mathbf{P}\mathbf{x}^{(k+1)} = \mathbf{N}\mathbf{x}^{(k)} + \mathbf{b};\ k \geq 0 \tag{3.3.1}$$

Selon le formalisme introduit en (3.2.3), on peut écrire la matrice d'itération $\mathbf{B} = \mathbf{P}^{-1}\mathbf{N}$, et $\mathbf{f} = \mathbf{P}^{-1}\mathbf{b}$ puisqu'on peut écrire de faéquivalente (3.3.1) comme

$$\mathbf{x}^{(k+1)} = \mathbf{P}^{-1}\mathbf{N}\mathbf{x}^{(k)} + \mathbf{P}^{-1}\mathbf{b};\ k \geq 0$$

On peut aussi écrire

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \mathbf{P}^{-1}\mathbf{r}^{(k)};\ k \geq 0 \tag{3.3.2}$$

où

$$\begin{aligned}
\mathbf{P}^{-1}\mathbf{r}^{(k)} &= \mathbf{P}^{-1}\mathbf{b} + \mathbf{P}^{-1}\mathbf{N}\mathbf{x}^{(k)} - \mathbf{x}^{(k)} \\
&= \mathbf{P}^{-1}\mathbf{b} + (\mathbf{P}^{-1}\mathbf{N} - \mathbf{I})\mathbf{x}^{(k)} \\
&= \mathbf{P}^{-1}\left[\mathbf{b} + (\mathbf{N} - \mathbf{P})\mathbf{x}^{(k)}\right] \\
&= \mathbf{P}^{-1}\left[\mathbf{b} - \mathbf{A}\mathbf{x}^{(k)}\right]
\end{aligned}$$

donc

$$\mathbf{r}^{(k)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(k)} \tag{3.3.3}$$

$\mathbf{r}^{(k)}$ désigne le résidu à l'itération k. La relation (3.3.2) montre qu'on doit résoudre un système linéaire de matrice $\mathbf{P}$ à chaque itération. En plus d'être inversible, $\mathbf{P}$ doit donc être facile à inverser afin de minimiser le coût de calcul. On peut d'ailleurs remarquer que si $\mathbf{P} = \mathbf{A}$ et $\mathbf{N} = \mathbf{0}$, la méthode (3.3.2) converge en une itération et est équivalente à une méthode directe.

Deux résultats garantissent la convergence de (3.3.2) sous des hypothèses convenables concernant le splitting de $\mathbf{A}$.

THÉORÈME 3.3.1 *Soit $\mathbf{A} = \mathbf{P} - \mathbf{N}$, avec $\mathbf{A}$ et $\mathbf{P}$ symétriques définies positives. Si la matrice $2\mathbf{P} - \mathbf{A}$ est définie positive, alors la méthode (3.3.2) est convergente pour toute donnée initiale $\mathbf{x}^{(0)}$ et*

$$\rho(\mathbf{B}) = \|\mathbf{B}\|_A = \|\mathbf{B}\|_P < 1$$

De plus, la convergence est monotone pour les normes $\|.\|_P$ et $\|.\|_A$, i.e.

$$\left\|\mathbf{e}^{(k+1)}\right\|_P \leq \left\|\mathbf{e}^{(k)}\right\|_P \text{ et } \left\|\mathbf{e}^{(k+1)}\right\|_A \leq \left\|\mathbf{e}^{(k)}\right\|_A;\ k = 0, 1, \ldots$$

THÉORÈME 3.3.2 *Soit $\mathbf{A} = \mathbf{P} - \mathbf{N}$ avec $\mathbf{A}$ symétrique définie positive. Si la matrice $\mathbf{P} + \mathbf{P}^\top - \mathbf{A}$ est définie positive, alors $\mathbf{P}$ est inversible, la méthode itérative (3.3.2) converge de manière monotone pour la norme $\|.\|_A$,et $\rho(\mathbf{B}) \leq \|\mathbf{B}\|_A < 1$.*

3.3.1 Méthodes de Jacobi, Gauss-Seidel et relaxation

On rappelle que l'on veut résoudre le système linéaire $\mathbf{Ax} = \mathbf{b}$, dans lequel $\mathbf{A}$ est une matrice carrée. Pour cela on décompose $\mathbf{A}$ sous la forme $\mathbf{A} = \mathbf{M} - \mathbf{N}$. Alors $\mathbf{Ax} = \mathbf{b} \Leftrightarrow \mathbf{Mx} = \mathbf{Nx} + \mathbf{b}$. Si $\mathbf{M}$ est une matrice régulière (i.e. inversible) on définit la méthode itérative par :

$$\mathbf{Mx}^{(k+1)} = \mathbf{Nx}^{(k)} + \mathbf{b} \Leftrightarrow \mathbf{x}^{(k+1)} = \mathbf{M}^{-1}\mathbf{Nx}^{(k)} + \mathbf{M}^{-1}\mathbf{b} \tag{3.3.4}$$

REMARQUE 3.3.1 *Cette formulation est bien consistante puisque sa solution $\mathbf{x}^*$ vérifie (si elle existe) $\mathbf{x}^* = \mathbf{M}^{-1}\mathbf{Nx}^* + \mathbf{M}^{-1}\mathbf{b}$. Or ceci équivaut à $\mathbf{Mx}^* = \mathbf{Nx}^* + \mathbf{b} \Leftrightarrow (\mathbf{M} - \mathbf{N})\mathbf{x}^* = \mathbf{b} \Leftrightarrow \mathbf{Ax}^* = \mathbf{b}$ donc $\mathbf{x}^*$ est le vecteur solution du système initial. On dit aussi qu'il s'agit d'un problème de point fixe.*

3.3.1.1 Méthodes non relaxées

La matrice $\mathbf{A}$ peut s'écrire aussi $\mathbf{A} = \mathbf{D} - \mathbf{E} - \mathbf{F} = \mathbf{D} + \mathbf{L} + \mathbf{U}$ (choix dépendant des auteurs), où $\mathbf{D}$ est la diagonale de $\mathbf{A}$, $-\mathbf{E} = \mathbf{L}$ sa partie triangulaire inférieure et $-\mathbf{F} = \mathbf{U}$ sa partie triangulaire supérieure, i.e. : $\mathbf{A} = \begin{bmatrix} \ddots & & -\mathbf{F} \\ & \mathbf{D} & \\ -\mathbf{E} & & \ddots \end{bmatrix} = \begin{bmatrix} \ddots & & \mathbf{U} \\ & \mathbf{D} & \\ \mathbf{L} & & \ddots \end{bmatrix}$.

La décomposition sous la forme $\mathbf{A} = \mathbf{M} - \mathbf{N}$ doit être établie à partir de celle sous la forme $\mathbf{A} = \mathbf{D} - \mathbf{E} - \mathbf{F} = \mathbf{D} + \mathbf{L} + \mathbf{U}$: suivant les matrices $\mathbf{D}$, $\mathbf{E}$ et $\mathbf{F}$ (ou $\mathbf{D}$, $\mathbf{L}$ et $\mathbf{U}$) que l'on associe dans $\mathbf{M}$ et $\mathbf{N}$ on obtient les méthodes de Jacobi, Gauss-Seidel et Richardson :

Méthode	Décomposition $\mathbf{A} = \mathbf{M} - \mathbf{N}$	Matrice $\mathbf{M}^{-1}\mathbf{N}$ de (3.3.4)	Description d'une itération
Jacobi	$\mathbf{A} = \underbrace{\mathbf{D}}_{\mathbf{M}} - \underbrace{(\mathbf{E}+\mathbf{F})}_{\mathbf{N}}$	$\mathbf{J} = \mathbf{M}^{-1}\mathbf{N} = \mathbf{D}^{-1}(\mathbf{E}+\mathbf{F}) = \mathbf{I} - \mathbf{D}^{-1}\mathbf{A}$	$\mathbf{Dx}^{(k+1)} = (\mathbf{E}+\mathbf{F})\mathbf{x}^{(k)} + \mathbf{b}$
Gauss-Seidel	$\mathbf{A} = \underbrace{(\mathbf{D}-\mathbf{E})}_{\mathbf{M}} - \underbrace{\mathbf{F}}_{\mathbf{N}}$	$\mathbf{G} = \mathbf{M}^{-1}\mathbf{N} = (\mathbf{D}-\mathbf{E})^{-1}\mathbf{F}$	$(\mathbf{D}-\mathbf{E})\mathbf{x}^{(k+1)} = \mathbf{Fx}^{(k)} + \mathbf{b}$
Richardson	$\mathbf{A} = \underbrace{\mathbf{I}}_{\mathbf{M}} - \underbrace{(\mathbf{I}-\mathbf{A})}_{\mathbf{N}}$	$\mathbf{R} = \mathbf{M}^{-1}\mathbf{N} = \mathbf{I} - \mathbf{A}$	$\mathbf{x}^{(k+1)} = (\mathbf{I}-\mathbf{A})\mathbf{x}^{(k)} + \mathbf{b}$

Si on utilise la décomposition $\mathbf{A} = \mathbf{D} + \mathbf{L} + \mathbf{U}$, on pour les méthodes de Jacobi et de Gauss-Seidel le tableau suivant :

Méthode	Décomposition $\mathbf{A} = \mathbf{M} - \mathbf{N}$	Matrice $\mathbf{M}^{-1}\mathbf{N}$ de (3.3.4)	Description d'une itération
Jacobi	$\mathbf{A} = \underbrace{\mathbf{D}}_{\mathbf{M}} - \underbrace{(\mathbf{L}+\mathbf{U})}_{\mathbf{N}}$	$\mathbf{J} = \mathbf{M}^{-1}\mathbf{N} = -\mathbf{D}^{-1}(\mathbf{L}+\mathbf{U})$	$\mathbf{Dx}^{(k+1)} = -(\mathbf{L}+\mathbf{U})\mathbf{x}^{(k)} + \mathbf{b}$
Gauss-Seidel	$\mathbf{A} = \underbrace{(\mathbf{D}+\mathbf{L})}_{\mathbf{M}} - \underbrace{\mathbf{U}}_{\mathbf{N}}$	$\mathbf{G} = \mathbf{M}^{-1}\mathbf{N} = (\mathbf{D}+\mathbf{L})^{-1}\mathbf{U}$	$(\mathbf{D}+\mathbf{L})\mathbf{x}^{(k+1)} = \mathbf{Ux}^{(k)} + \mathbf{b}$

La description des itérations est directement liée à la formulation (3.3.4) et peut être retrouvée facilement.

REMARQUE 3.3.2 *La méthode de Richardson est ici donnée pour mémoire car elle n'a que peu d'intérêt numérique. Cela est notamment dû au fait que le rayon spectral de la matrice d'itération, $\mathbf{R}$, n'est pas très bon en général.*

3.3.1.2 Méthodes relaxées

Des performances des méthodes non relaxées, parfois peu satisfaisantes, a rapidement découlé l'émergence de méthodes plus efficaces, dites *relaxées*. L'idée de base est assez simple : Ces méthodes consistent à reprendre les différents choix du paragraphe précédent en pondérant par un facteur ω (ou μ suivant les auteurs) les facteurs de (3.3.4). On remplace l'itéré calculé selon l'une des méthodes précédentes ($\widehat{\mathbf{x}}^{(k+1)}$ dans (3.3.5)) par $\mathbf{x}^{(k+1)}$ qui fait intervenir $\mathbf{x}^{(k)}$ via la formule :

$$\mathbf{x}^{(k+1)} = \omega\widehat{\mathbf{x}}^{(k+1)} + (1-\omega)\mathbf{x}^{(k)} \tag{3.3.5}$$

On utilise en fait une combinaison convexe de l'itéré au rang k et de l'itéré au rang $k+1$ pour minimiser les risques de divergence ou d'oscillation.
Chaque méthode vue au paragraphe précédent peut être relaxée, on obtient alors :

Méthode relaxée	Description d'une itération
Jacobi relaxé	$\mathbf{x}^{(k+1)} = \left[(1-\omega)\mathbf{I} + \omega\mathbf{D}^{-1}(\mathbf{E}+\mathbf{F})\right]\mathbf{x}^{(k)} + \omega\mathbf{D}^{-1}\mathbf{b}$
Gauss-Seidel relaxé	$\mathbf{x}^{(k+1)} = (\frac{\mathbf{D}}{\omega} - \mathbf{E})^{-1}\left[(\frac{1}{\omega}-1)\mathbf{D} + \mathbf{F}\right]\mathbf{x}^{(k)} + (\mathbf{D}-\omega\mathbf{E})^{-1}\mathbf{b}$
Richardson relaxé	$\mathbf{x}^{(k+1)} = (1-\omega)\mathbf{x}^{(k)} + \omega(\mathbf{I}-\mathbf{A})\mathbf{x}^{(k)} + \omega\mathbf{b} = (\mathbf{I}-\omega\mathbf{A})\mathbf{x}^{(k)} + \omega\mathbf{b}$

De même pour la décomposition $\mathbf{A} = \mathbf{D} + \mathbf{L} + \mathbf{U}$, on a le tableau suivant :

Méthode relaxée	Description d'une itération
Jacobi relaxé	$\mathbf{x}^{(k+1)} = (1-\omega)\mathbf{x}^{(k)} - \omega\mathbf{D}^{-1}(\mathbf{L}+\mathbf{U})\mathbf{x}^{(k)} + \omega\mathbf{D}^{-1}\mathbf{b}$
Gauss-Seidel relaxé	$\mathbf{x}^{(k+1)} = (1-\omega)(\mathbf{I}+\omega\mathbf{D}^{-1}\mathbf{L})^{-1}\mathbf{x}^{(k)} - \omega(\mathbf{I}+\omega\mathbf{D}^{-1}\mathbf{L})^{-1}\mathbf{D}^{-1}\mathbf{U}\mathbf{x}^{(k)} + \omega(\mathbf{D}-\omega\mathbf{E})^{-1}\mathbf{D}^{-1}\mathbf{b}$

Si on considère l'exemple de la méthode de Jacobi, pour lequel la $i-$ème composante de l'itéré $\mathbf{x}^{(k+1)}$ est obtenue par :

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\left[b_i - \sum_{j=1, j\neq i}^{n} a_{ij}x_j^{(k)}\right];\ i = 1, \ldots, n \tag{3.3.6}$$

on obtient ainsi facilement la $i-$ème composante de l'itéré $\mathbf{x}^{(k+1)}$ de la méthode de Jacobi relaxée, en appliquant la technique de pondération expliquée ci-dessus :

$$x_i^{(k+1)} = \frac{\omega}{a_{ii}}\left[b_i - \sum_{j=1, j\neq i}^{n} a_{ij}x_j^{(k)}\right] + (1-\omega)x_i^{(k)};\ i = 1, \ldots, n \tag{3.3.7}$$

La littérature abonde en dénominations variées pour les méthodes de relaxation, signifiant parfois toutes la même chose. Ainsi la méthode de Jacobi relaxée est parfois dite *méthode de sur-relaxation*, ou *méthode JOR*, pour *Jacobi over relaxation.*

REMARQUE 3.3.3 *Il faut noter que la méthode de Jacobi relaxée est consistante pour $\omega \neq 0$ et que pour $\omega = 1$, elle coïncide avec la méthode de Jacobi.*

Si on considère maintenant le cas de la méthode de Gauss-Seidel, pour lequel la $i$ème composante de l'itéré $x^{(k+1)}$ est obtenue par :

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\left[b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\right], i = 1, \ldots n \tag{3.3.8}$$

on obtient ainsi facilement la $i$ème composante de l'itéré $x^{(k+1)}$ de la méthode de Gauss-Seidel relaxée, en appliquant la technique de pondération expliquée ci-dessus :

$$x_i^{(k+1)} = \frac{\omega}{a_{ii}} \left[b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right] + (1-\omega) x_i^{(k)};\ i = 1, \ldots, n \tag{3.3.9}$$

Cette méthode est aussi qualifiée de *méthode de sur-relaxation successive* ou méthode *SOR* (pour *successive over relaxation*)

REMARQUE 3.3.4 *Il faut noter que la méthode de Gauss-Seidel relaxée est consistante pour $\omega \neq 0$ et que pour $\omega = 1$, elle coïncide avec la méthode de Gauss-Seidel. Si $\omega \in]0,1[$, la méthode est dite de sous-relaxation, par contre si $\omega > 1$, elle est qualifiée de sur-relaxation.*

3.3.2 Résultats de convergence pour les méthodes de Jacobi et Gauss-Seidel

Il existe des cas où l'on peut établir des résultats de convergence a priori pour les méthodes examinées à la section précédente. Voici deux résultats dans ce sens :

THÉORÈME 3.3.3 *Si $\mathbf{A}$ est une matrice à diagonale strictement dominante, les méthodes de Jacobi et de Gauss-Seidel sont convergentes.*

THÉORÈME 3.3.4 *Si $\mathbf{A}$ et $2\mathbf{D} - \mathbf{A}$ sont symétriques définies positives, alors la méthode de Jacobi est convergente et $\rho(\mathbf{J}) = \|\mathbf{J}\|_A = \|\mathbf{J}\|_D$.*

Dans le cas de la méthode de Jacobi relaxée, on peut se passer de la condition sur $2\mathbf{D} - \mathbf{A}$:

THÉORÈME 3.3.5 *Quand A est symétrique définie positive, la méthode de Jacobi relaxée est convergente si $0 < \omega < 2/\rho(\mathbf{D}^{-1}\mathbf{A})$*

En ce qui concerne la méthode de Gauss-Seidel, on a le résultat suivant :

THÉORÈME 3.3.6 *Quand $\mathbf{A}$ est symétrique définie positive, la méthode de Gauss-Seidel converge de manière monotone pour la norme $\|.\|_A$.*

Enfin, si la matrice $\mathbf{A}$ est tridiagonale symétrique définie positive, on peut montrer que la méthode de Jacobi est convergente et que

$$\rho(\mathbf{G}) = \rho^2(\mathbf{J})$$

où $\mathbf{G}$ et $\mathbf{J}$ représentent respectivement les matrices d'itérations de la méthode de Gauss-Seidel et de celle de Jacobi.

Dans ce cas, la méthode de Gauss-Seidel converge plus rapidement que celle de Jacobi.

La littérature regorge de résultats de convergence établis sur des familles de matrices correspondant souvent à la résolution d'équations aux dérivées partielles provenant de la physique. Il serait trop long de les citer ou même de les répertorier toutes. Il faut retenir que dans le cas général de matrices quelconques, il n'existe pas de résultat assurant de la convergence des méthodes précitées. Démontrer la convergence pour une matrice ne vérifiant pas des propriétés classiques peut donc devenir très technique, mais appliquer la méthode itérative sans s'être assuré de la convergence peut aussi être assez risqué...

L'exemple ci-dessous montre que l'on ne peut tirer aucune conclusion a priori sur la convergence des méthodes de Jacobi et de Gauss-Seidel.

EXEMPLE 3.3.1 *Considérons les sytèmes 3×3 de la forme $\mathbf{A}_i\mathbf{x} = \mathbf{b}_i$. On choisit $\mathbf{b}_i$ de manière à ce que la solution du système soit le vecteur unité, et les matrices $\mathbf{A}_i$ sont données par :*

$$\mathbf{A}_1 = \begin{bmatrix} 3 & 0 & 4 \\ 7 & 4 & 2 \\ -1 & 1 & 2 \end{bmatrix} \quad \mathbf{A}_2 = \begin{bmatrix} -3 & 3 & -6 \\ -4 & 7 & -8 \\ 5 & 7 & -9 \end{bmatrix}$$

$$\mathbf{A}_3 = \begin{bmatrix} 4 & 1 & 1 \\ 2 & -9 & 0 \\ 0 & -8 & -6 \end{bmatrix} \quad \mathbf{A}_4 = \begin{bmatrix} 7 & 6 & 9 \\ 4 & 5 & -4 \\ -7 & -3 & 8 \end{bmatrix}$$

On peut vérifier que la méthode de Jacobi ne converge pas pour $\mathbf{A}_1$ ($\rho(\mathbf{J}) = 1.33$), contrairement à celle de Gauss-Seidel. C'est exactement le contraire qui se produit pour $\mathbf{A}_2$($\rho(\mathbf{G}) = 1.1$). La méthode de Jacobi converge plus lentement que celle de Gauss-Seidel pour la matrice $\mathbf{A}_3$ ($\rho(\mathbf{J}) = 0.44$ et $\rho(\mathbf{G}) = 0.018$), alors que la la méthode de Jacobi est plus rapide pour $\mathbf{A}_4$ ($\rho(\mathbf{J}) = 0.64$ et $\rho(\mathbf{G}) = 0.77$).

Concluons par un dernier résultat :

THÉORÈME 3.3.7 *Si la méthode de Jacobi converge alors la méthode JOR converge pour $0 < \omega \leq 1$.*

3.3.3 Résultats de convergence pour la méthode de relaxation

Lorsque rien n'est précisé par ailleurs, la méthode diet de relaxation est celle de Guass-Seidel relaxée.

Sans hypothèse particulière sur $\mathbf{A}$, on peut déterminer les valeurs de ω pour lesquelles la méthode SOR ne peut pas converger :

THÉORÈME 3.3.8 *On a $\rho(\mathbf{G}_\omega) \geq |\omega - 1|, \forall \omega \in \mathbb{R}$. La méthode SOR diverge donc pour tout $\omega \leq 0$ ou $\omega \geq 2$.*

Notons que du tableau donné en section (3.3.1.2) on déduit que la matrice d'itération de la méthode relaxée est donnée par :

$$\mathbf{G}_\omega = (\frac{\mathbf{D}}{\omega} - \mathbf{E})^{-1} \left[(\frac{1}{\omega} - 1)\mathbf{D} + \mathbf{F} \right] \tag{3.3.10}$$

que l'on peut aussi exprimer sous la forme équivalente :

$$\mathbf{G}_\omega = \omega\mathbf{D}^{-1}(\mathbf{I}-\omega\mathbf{D}^{-1}\mathbf{E})^{-1}\frac{\mathbf{D}}{\omega}\left[(1-\omega)\mathbf{I}+\omega\mathbf{D}^{-1}\mathbf{F}\right] \tag{3.3.11}$$

soit

$$\mathbf{G}_\omega = (\mathbf{I}-\omega\mathbf{D}^{-1}\mathbf{E})^{-1}\left[(1-\omega)\mathbf{I}+\omega\mathbf{D}^{-1}\mathbf{F}\right] \tag{3.3.12}$$

Si $\{\lambda_i\}$ désigne l'ensemble des valeurs propres de la matrice d'itération de SOR, alors

$$\left|\prod_{i=1}^{n}\lambda_i\right| = \det\left[(1-\omega)\mathbf{I}+\omega\mathbf{D}^{-1}\mathbf{F}\right] = |1-\omega|^n$$

Par conséquent, au moins une valeur propre λ_i est telle que $|\lambda_i| > |1-\omega|$. Pour avoir convergence, il est donc nécessaire que $|1-\omega| < 1$, c'est à dire que $0 < \omega < 2$.

Si on suppose $\mathbf{A}$ symétrique définie positive, la condition nécessaire $0 < \omega < 2$ devient suffisante pour avoir convergence. On a en effet le résultat suivant :

THÉORÈME 3.3.9 *(Propriété d'Ostrowski)*
Si $\mathbf{A}$ *est symétrique définie positive, alors la méthode SOR converge si et seulement si* $0 < \omega < 2$. *De plus sa convergence est monotone pour la norme* $\|.\|_A$.

Enfin :

THÉORÈME 3.3.10 *Si* $\mathbf{A}$ *est à diagonale dominante stricte, alors la méthode SOR converge si* $0 < \omega \leq 1$.

Les résultats ci-dessus montrent que SOR converge plus ou moins vite selon le choix du paramètre de relaxation ω. On ne peut donner de réponse satisfaisante à la question du choix du paramètre optimal ω pour lequel le taux de convergence est le plus grand, que dans le cas de matrices particulières. On pourra consulter les ouvrages cités en référence pour plus de détails.

3.4 Test d'arrêt

Dans cette section, nous abordons le problème de l'estimation de l'erreur induite par une méthode itérative (au sens défini dans (3.2.4)). En particulier, on cherche à évaluer le nombre d'itérations $k_{\min}$ nécessaire pour que la norme de l'erreur divisée par celle de l'erreur initiale soir inférieur à un ε fixé.

En pratique, une estimation a priori de $k_{\min}$ peut être obtenue à partir de (3.2.3) qui donne la vitesse à laquelle $\left\|\mathbf{e}^{(k)}\right\| \to 0$ quand k tend vers l'infini. D'après (3.2.5) on obtient :

$$\frac{\left\|\mathbf{e}^{(k)}\right\|}{\left\|\mathbf{e}^{(0)}\right\|} \leq \left\|\mathbf{B}^k\right\|$$

Ainsi $\|\mathbf{B}^k\|$ donne une estimation du facteur de réduction de la norme de l'erreur après k itérations. Typiquement, on poursuit les itérations jusqu'à ce que

$$\left\|\mathbf{e}^{(k)}\right\| \leq \varepsilon \left\|\mathbf{e}^{(0)}\right\| \text{ avec } \varepsilon < 1 \tag{3.4.1}$$

Des considérations sur les propriétés des matrices d'itérations et les normes, conduisent à établir que

$$k_{\min} \simeq -\frac{\log(\varepsilon)}{R(\mathbf{B})} \tag{3.4.2}$$

où $R(\mathbf{B})$ est le taux de convergence asymptotique défini par

$$R(\mathbf{B}) = -\log \rho(\mathbf{B})$$

Cette dernière estimation est plutôt optimiste, mais elle présente en plus le désavantage de nécessiter le calcul de $\rho(\mathbf{B})$, qui peut constituer un problème en lui-même.

De ce fait, plutôt que des estimations a priori du nombre d'itérations nécessaires, on préfère en général utiliser un indicateur facilement évaluable au cours des itérations. On donne ci-après deux exemples.

3.4.1 Un test d'arrêt basé sur l'incrément

D'après la relation de récurrence sur l'erreur $\mathbf{e}^{(k+1)} = \mathbf{B}\mathbf{e}^{(k)}$, on a

$$\left\|e^{(k+1)}\right\| \leq \|B\| \left\|e^{(k)}\right\| \tag{3.4.3}$$

En utilisant la relation de récurrence sur l'erreur, on a :

$$\begin{aligned}
\mathbf{e}^{(k+1)} &= \mathbf{B}\mathbf{e}^{(k)} \\
&= \mathbf{B}\left(\mathbf{e}^{(k)} + \mathbf{e}^{(k+1)} - \mathbf{e}^{(k+1)}\right) \\
&= \mathbf{B}\left(\mathbf{e}^{(k+1)} + \mathbf{x}^{(k)} - \mathbf{x} - \mathbf{x}^{(k+1)} + \mathbf{x}\right) \\
&\mathbf{B}\mathbf{e}^{(k+1)} - \mathbf{B}\left(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right)
\end{aligned}$$

d'où

$$\left\|\mathbf{e}^{(k+1)}\right\| \leq \|\mathbf{B}\| \left(\left\|\mathbf{e}^{(k+1)}\right\| + \left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\|\right)$$

et donc

$$\left\|\mathbf{x} - \mathbf{x}^{(k+1)}\right\| \leq \frac{\|\mathbf{B}\|}{1 - \|\mathbf{B}\|} \left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\| \tag{3.4.4}$$

En particulier, en prenant $k = 0$ dans (3.4.4) et en appliquant la formule de récurrence (3.4.3) on obtient aussi l'inégalité :

$$\left\|\mathbf{x} - \mathbf{x}^{(k+1)}\right\| \leq \frac{\|\mathbf{B}\|^{k+1}}{1 - \|\mathbf{B}\|} \left\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right\|$$

qu'on peut utiliser pour estimer le nombre d'itérations nécessaires à satisfaire la condition $\left\|\mathbf{e}^{(k+1)}\right\| \leq \varepsilon$, pour une tolérance ε donnée.

En pratique, on peut estimer $\|\mathbf{B}\|$ comme suit : puisque

$$\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} = -\left(\mathbf{x} - \mathbf{x}^{(k+1)}\right) + \left(\mathbf{x} - \mathbf{x}^{(k)}\right) = \mathbf{B}\left(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right)$$

la quantité $\|\mathbf{B}\|$ est minorée par

$$c = \frac{\delta_{k+1}}{\delta_k}$$

où

$$\delta_{k+1} = \left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\|$$

En rempla$\|\mathbf{B}\|$ par c, le membre de droite de (3.4.4) suggère d'utiliser l'indicateur suivant pour $\left\|\mathbf{e}^{(k+1)}\right\|$

$$\epsilon^{(k+1)} = \frac{\delta_{k+1}^2}{\delta_k - \delta_{k+1}} \tag{3.4.5}$$

Il faut prendre garde au fait qu'avec l'approximation utilisée pour $\|\mathbf{B}\|$, on ne peut pas voir $\epsilon^{(k+1)}$ comme un majorant de $\left\|\mathbf{e}^{(k+1)}\right\|$. Néanmoins $\epsilon^{(k+1)}$ fournit souvent une indication raisonnable du comportement de l'erreur.

3.4.2 Tests d'arrêt fondés sur le résidu

Un autre critère d'arrêt consiste à tester si $\left\|\mathbf{r}^{(k)}\right\| \leq \varepsilon$, pour une tolérance ε fixée. Comme

$$\left\|\mathbf{x} - \mathbf{x}^{(k)}\right\| = \left\|\mathbf{A}^{-1}\mathbf{b} - \mathbf{x}^{(k)}\right\| = \left\|\mathbf{A}^{-1}\mathbf{r}^{(k)}\right\| \leq \|\mathbf{A}^{-1}\| \left\|\mathbf{r}^{(k)}\right\| \leq \|\mathbf{A}^{-1}\| \varepsilon$$

on doit prendre $\varepsilon \leq \frac{\delta}{\|\mathbf{A}^{-1}\|}$ pour que l'erreur soit inférieure à δ.

Il est en général plus judicieux de considérer un résidu normalisé : on interrompt les itérations quand $\left\|\mathbf{r}^{(k)}\right\| / \left\|\mathbf{r}^{(0)}\right\| \leq \varepsilon$ ou bien quand $\left\|\mathbf{r}^{(k)}\right\| / \|\mathbf{b}\| \leq \varepsilon$ (ce qui correspond au choix $\mathbf{x}^{(0)} = \mathbf{0}$). Dans ce dernier cas, le test d'arrêt fournit le contrôle suivant de l'erreur relative

$$\frac{\left\|\mathbf{x} - \mathbf{x}^{(k)}\right\|}{\|\mathbf{x}\|} \leq \frac{\|\mathbf{A}^{-1}\| \left\|\mathbf{r}^{(k)}\right\|}{\|\mathbf{x}\|} \leq \kappa(\mathbf{A}) \frac{\left\|\mathbf{r}^{(k)}\right\|}{\|\mathbf{b}\|} \leq \kappa(\mathbf{A}).\varepsilon$$

On retrouve l'influence du conditionnement de la matrice du système, qui même si la tolérance est faible, pour entraîner une erreur importante sur la solution. Avec la technique de préconditionnement par une matrice $\mathbf{P}$, le critère précédent devient

$$\frac{\left\|\mathbf{P}^{-1}\mathbf{r}^{(k)}\right\|}{\left\|\mathbf{P}^{-1}\mathbf{r}^{(0)}\right\|} \leq .\varepsilon$$

ce qui permet de s'affranchir de l'influence du conditionnement.

En plus, dans [CB] est donné la majoration suivante pour le résidu :

$$\left\|\mathbf{x}-\mathbf{x}^{(k)}\right\| \leq \frac{K}{1-K}\left\|\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}\right\|$$

avec $K = \|\mathbf{I}-\mathbf{CA}\|$ où $\mathbf{C} = \begin{cases} \mathbf{D}^{-1}, & \text{si méthode de Jacobi} \\ (\mathbf{D}+\mathbf{L})^{-1}, & \text{si méthode de Gauss-Seidel} \end{cases}$

En conclusion et au delà des tests d'arrêt, il y a deux possibilités pour arrêter un algorithme itératif de résolution d'un système d'équations :

(1) Dépassement du'un nombre d'itérations défini d'avance, et

(2) le résidu devient inférieure à un seuil défini d'avance.

3.5 Exercices

EXERCICE 3.1 *Soient $\mathbf{A} \in \mathbb{R}^{2\times 2}$ et $\mathbf{b} \in \mathbb{R}^2$. La solution du système $\mathbf{Ax} = \mathbf{b}$ s'interprète gométriquement comme le point d'intersection de deux droites*

$$(D_1) \ : \ a_{11}x_1 + a_{12}x_2 = b_1$$
$$(D_2) \ : \ a_{21}x_1 + a_{22}x_2 = b_2$$

On suppose que a_{11} et a_{22} sont non nuls.

(1) Calculer les matrices d'itération des méthodes de Jacobi et de Gauss-Seidel associées à ce système.

(2) Calculer les rayons spectraux de ces matrices. Que remarque-t-on ?

(3) Calculer les rayons spectraux des matrices d'itération des méthodes de Jacobi et de Gauss-Seidel associées au système obtenu en permutant les équations ci-dessus.

EXERCICE 3.2 *(1) Interpréter géométriquement les méthodes de Jacobi et de Gauss-Seidel appliquées à $\mathbf{Ax} = \mathbf{b}$. Pour cette dernière, on notera que le vecteur $x_{k+1} \in \mathbb{R}^2$ est solution du système triangulaire*

$$\begin{cases} a_{11}x_{k+1,1} + a_{12}x_{k,2} = b_1 \\ a_{21}x_{k+1,1} + a_{22}x_{k+1,2} = b_2 \end{cases}$$

EXERCICE 3.3 *Soient*

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & -1/4 & -1/4 \\ 0 & 1 & -1/4 & -1/4 \\ -1/4 & -1/4 & 1 & 0 \\ -1/4 & -1/4 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{I}_2 & \mathbf{K} \\ \mathbf{K} & \mathbf{I}_2 \end{bmatrix}$$

et

$$\mathbf{b} = \frac{1}{2}\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

(1) *Calculer les matrices d'itération des méthodes de Jacobi, Gauss-Seidel et de relaxation associées à* $\mathbf{A}$. *On les notera* $\mathbf{J}$, $\mathbf{G}$ *et* $\mathbf{G}_\omega$.

(2) *Donner l'expression en fonction de* $k, \mathbf{b}$ *et* $\mathbf{K}$ *des itérés de la résolution du système par la méthode de Jacobi et de Gauss-Seidel quand le point initial est l'origine. Il est recommandé de tirer parti de la structure par blocs de la matrice.*

(3) *Calculer les rayons spectraux de* $\mathbf{J}$ *et* $\mathbf{G}$.

(a) *Montrer que si* λ *est valeur propre de* $\mathbf{G}_\omega$ *alors* $\lambda = 1-\omega$ *ou bien* λ *est racine de l'équation*

$$\lambda^2 - \left(2(1-\omega) + \frac{\omega^2}{4}\right)\lambda + (1-\omega)^2 = 0$$

(b) *Calculer* $\rho(\mathbf{G}_\omega)$ *en distinguant les cas où les racines de l'équation précédente sont réelles ou non.*

(c) *Trouver la valeur de* ω *qui rend* $\rho(\mathbf{G}_\omega)$ *minimum.*

EXERCICE 3.4 *Soit* $a \in \mathbb{R}$ *et*

$$\mathbf{A} = \begin{bmatrix} 1 & a & a \\ a & 1 & a \\ a & a & 1 \end{bmatrix}$$

Montrer que $\mathbf{A}$ *est symétrique définie positive si et seulement si* $-1/2 < a < 1$ *et que la méthode de Jacobi converge si et seulement si* $-1/2 < a < 1/2$.

EXERCICE 3.5 *Considérons le système d'équations linéaires*

$$\mathbf{A}\mathbf{x} = \mathbf{b}$$

Pour la résolution on applique une méthode itérative

$$\begin{aligned} \mathbf{x}(0) &\in \mathbb{R}^n \\ \mathbf{P}\mathbf{x}(k+1) &= \mathbf{N}\mathbf{x}(k) + \mathbf{b} \end{aligned}$$

avec $\mathbf{A} = \mathbf{P} - \mathbf{N}$ *et* $\rho\left(\mathbf{P}^{-1}\mathbf{N}\right) < 1$.

À cause des erreurs de calcul on a pour la solution itérative

$$(\mathbf{P} + \Delta\mathbf{P}_{k+1}) \cdot \widehat{\mathbf{x}}(k+1) = \mathbf{N}\widehat{\mathbf{x}}(k) + \mathbf{b} + \Delta\mathbf{b}_k$$

où on a noté par $\widehat{\mathbf{x}}$ *la représentation machine de* $\mathbf{x}$, *c'est-à-dire le nombre-machine* $m(\mathbf{x})$.

(1) *Montrer que*

$$\mathbf{P} \cdot \widehat{\mathbf{x}}(k+1) = \mathbf{N}\widehat{\mathbf{x}}(k) + \mathbf{b} - \mathbf{d}_k$$

et évaluer $\mathbf{d}_k$.

(2) *Supposons que* $\widehat{\mathbf{x}}(0) = \mathbf{x}(0)$. *Montrer que*

$$\widehat{\mathbf{x}}(k+1) = \left(\mathbf{P}^{-1}\mathbf{N}\right)^{k+1}\widehat{\mathbf{x}}(0) + \sum_{l=0}^{k}\left(\mathbf{P}^{-1}\mathbf{N}\right)^{l}\mathbf{P}^{-1}(\mathbf{b} - \mathbf{d}_{k-l})$$

(3) *Posons* $\|\Delta\mathbf{P}_{k+1}\| \leq c\|\mathbf{P}\|$, $\delta_k = c(\|\mathbf{P}\|\,\|\widehat{\mathbf{x}}(k+1)\| + \|\mathbf{N}\|\,\|\widehat{\mathbf{x}}(k)\| + \|\mathbf{b}\|)$, *d'où* $\|\mathbf{d}_k\| \leq \delta_k$. *Considérons* $\mathbf{x}^*$ *le point fixe et soit* $\mathbf{e}(k+1) = \mathbf{x}^* - \widehat{\mathbf{x}}(k)$ *l'erreur de la solution calculée à chaque itération par rapport à la vraie solution. Montrer que*

$$\|\mathbf{e}(k+1)\| \leq \left\|\left(\mathbf{P}^{-1}\mathbf{N}\right)^{k+1}\mathbf{e}(0)\right\| + \sum_{l=0}^{k}\left\|\left(\mathbf{P}^{-1}\mathbf{N}\right)^{l}\mathbf{P}^{-1}\right\|\delta_{k-l}$$

(4) *Posons* $\theta = \sup_k \frac{\|\widehat{\mathbf{x}}(k)\|}{\|\mathbf{x}^*\|}$ *et* $\varepsilon(\mathbf{A}) = \arg\min\left\{\varepsilon \cdot \sum_{l=0}^{k}\left\|\left(\mathbf{P}^{-1}\mathbf{N}\right)^{l}\mathbf{P}^{-1}\right\| - \sum_{l=0}^{k}\left(\mathbf{P}^{-1}\mathbf{N}\right)^{l}\mathbf{P}^{-1} \geq 0\right\}$. *Remarquons que nous avons* $\sum_{l=0}^{k}\left(\mathbf{P}^{-1}\mathbf{N}\right)^{l}\mathbf{P}^{-1} \backsimeq \mathbf{A}^{-1}$, *car* $\mathbf{A} = \mathbf{P} - \mathbf{N}$ $=\mathbf{A} = \mathbf{P}\left(\mathbf{I} - \mathbf{P}^{-1}\mathbf{N}\right)$ *d'où* $\mathbf{A}^{-1} = \left(\mathbf{I} - \mathbf{P}^{-1}\mathbf{N}\right)^{-1}\mathbf{P}^{-1} = \sum_{l=0}^{\infty}\left(\mathbf{P}^{-1}\mathbf{N}\right)^{l}\mathbf{P}^{-1}$.

Calculer $\|\mathbf{e}(k+1)\|$.

(5) *Calculer* $\|\mathbf{e}(k+1)\|$ *pour l'itération de Jacobi.*

3.6 Bibliographie

Les ouvrages ci-dessous sont disponibles sous forme de fichier téléchargeable sur le site du cours ou sur Arel

[CB] **Algèbre matricielle numérique**, *Claude Brezinski*

[YA] **Algèbre linéaire et analyse numérique matricielle**, *Yves Achdou*, téléchargeable à l'adresse http ://www.ann.jussieu.fr/~achdou/files/teaching/linalg/book.pdf

[AH1]**Analyse numérique matricielle, Cours de 3ème année**, *Alain Huard*, téléchargeable à partir du site de l'INSA Toulouse www-gmm.insa-toulouse.fr

[AH2]**Analyse numérique des grands problèmes linéaires, Cours de 4ème année**, *Alain Huard*, téléchargeable à partir du site de l'INSA Toulouse www-gmm.insa-toulouse.fr

Les ouvrages ci-dessous sont disponibles en librairie

[QS] **Calcul scientifique, Cours, exercices corrigés et illustrations en Matlab et Octave**, Alfio Quarteroni, Fausto Saleri, Springer, 2206.

[QSS] **Numerical Mathematics**, Alfio Quarteroni, Riccardo Sacco, Fausto Saleri, Springer, 2000.

[AF] **Analyse numérique pour Ingénieurs**, André Fortin, Presses Internationales Polytechnique, 2001.

[AD] **Analyse numérique matricielle, Cours exercices et corrigés**, Luca Amodei, Jean-Pierre Dedieu, Dunod, 2008.

[CD] **Numerical Analysis**, S. D. Conte, Carl de Boor, McGraw-Hill, 1980

[SB] **Introduction to numerical ananlysis**, Second Edition, J. Stoer, B. Bulirsch, Springer-Verlag, 1993

Pour des logiciels relatifs aux méthodes itératives, on peut se référer à

http ://www.netlib.org/ est un dépôt des programmes d'analyse numérique. Il contient aussi TOMS (Transactions on Mathematical Software).

4

MÉTHODES DE DESCENTE

Dans l'esprit des méthodes itératives construites au chapitre précédent, consistant comme on l'a dit en une suite de vecteurs $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, ..., \mathbf{x}^{(k)}, ...$ qui convergera vers la solution du système linéaire $\mathbf{Ax} = \mathbf{b}$ à résoudre, on a cherché à construire des méthodes plus efficaces. En effet, au siècle dernier sont apparus des moyens de calcul permettant d'envisager la résolution de problèmes liés à la physique, notamment par discrétisation d'équations aux dérivées partielles, dans des domaines aussi variés que l'aéronautique, la météorologie ou la médecine. On en verra un exemple dans la première partie de ce chapitre, qui justifie notamment des inconvénients des méthodes vues au chapitre consacré aux méthodes directes, et ayant nécessité des progrès en la matière.

Pour autant, les progrès effectués dans le domaine ne doivent pas faire oublier au lecteur que les méthodes directes restent dans certains cas intéressantes et efficaces.

4.1 Un exemple d'application

Dans l'intention d'expliquer le contexte d'utilisation des méthodes d'optimisation, nous donnons ci-après un exemple tiré des problématiques d'écoulement (thermique, aéronautique ou mécanique des fluides en général).

On se place dans un espace bidimensionnel homéomorphe à $\mathbb{R}^2$ et on considère le problème modèle dit *de Dirichlet* : Trouver la fonction $u : \mathbb{R}^2 \to \mathbb{R}$ satisfaisant l'équation et les conditions aux limites suivantes :

$$\begin{aligned} -\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} &= f(x,y), \qquad 0 < x,y < 1 \\ u(x,y) &= 0 \qquad \text{pour } (x,y) \in \partial\Omega \end{aligned} \tag{4.1.1}$$

où $\partial\Omega$ désigne la frontière du carré unité $\Omega = \{(x,y)/\ 0 < x,y < 1\} \subset \mathbb{R}^2$. On suppose que la fonction f est continue sur $\Omega \cup \partial\Omega$. Pour résoudre ce problème, on peut utiliser la *méthode des différences finies*, qui consiste à remplacer les opérateurs de différenciation par des quotients de type taux d'accroissement pris en des points situés sur une grille recouvrant le domaine $\Omega \cup \partial\Omega$. On introduit des domaines discrétisés Ω_h et $\partial\Omega_h$ définis par :

$$\begin{aligned} \Omega_h &= \{(x_i,y_i)/\ i,j = 1,2,...N\} \\ \partial\Omega_h &= \{(x_i,0),(x_i,1),(0,y_i),(1,y_i)/\ i,j = 0,1,2,...N+1\} \end{aligned}$$

sur lesquels on définit des points

$$\begin{aligned} x_i &= ih \qquad y_i = jh \qquad i,j = 0,1,2,...N+1 \\ \text{pour } h &= \frac{1}{N+1},\ N \geq 1 \text{ et } N \in \mathbb{N} \end{aligned}$$

En notant $\widetilde{u}$ la solution approchée évaluée aux points de la discrétisation, et en utilisant la notation abrégée

$$\widetilde{u}(x_i,y_i) = \widetilde{u}_{ij} \qquad i,j = 0,1,2,...N+1$$

on montre que la solution approchée vérifie

$$\begin{aligned} 4\widetilde{u}_{ij} - \widetilde{u}_{i-1,j} - \widetilde{u}_{i+1,j} - \widetilde{u}_{i,j-1} - \widetilde{u}_{i,j+1} &= h^2 f_{ij} + h^2 \tau_{ij}, \qquad i,j = 1,2,...N \\ \widetilde{u}_{0j} &= \widetilde{u}_{N+1,j} = \widetilde{u}_{i0} = \widetilde{u}_{i,N+1} = 0, \qquad i = 0,1,...,N+1 \end{aligned}$$

où f_{ij} désigne $f(x_i,y_i)$ et τ_{ij} désigne l'erreur commise par approximation du premier et du second membre de l'équation (4.1.1).

Sous l'hypothèse que le paramètre h soit suffisamment petit, on peut montrer que la solution approchée $\widetilde{u}$ converge vers la solution exacte u et qu'elle est solution du système d'inconnue z suivant :

$$\begin{aligned} 4z_{ij} - z_{i-1,j} - z_{i+1,j} - z_{i,j-1} - z_{i,j+1} &= h^2 f_{ij}, \qquad i,j = 1,2,...N \\ z_{0j} &= z_{N+1,j} = z_{i0} = z_{i,N+1} = 0, \qquad i = 0,1,...,N+1 \end{aligned} \tag{4.1.2}$$

On a donc à résoudre un système à N^2 inconnues (égales aux valeurs de la fonction z aux points intérieurs de la grille), dont le second membre comporte les valeurs $f(x_i,y_i)$. Si on note les inconnues dans un vecteur (c'est à dire que l'on met bout à bout les inconnues correspondant aux points en colonnes dans la grille) et le second membre par :

$$\begin{aligned} \mathbf{z} &= ([z_{11},z_{21},...,z_{N1},z_{12},z_{22},...,z_{N2},...z_{NN}]^\top \\ \mathbf{b} &= h^2([f_{11},f_{21},...,f_{N1},f_{12},f_{22},...,f_{N2},...f_{NN}]^\top \end{aligned}$$

alors le système (4.1.2) est équivalent au système matriciel $\mathbf{Az} = \mathbf{b}$ dont la matrice $\mathbf{A}$ est de dimension $N^2 \times N^2$ et définie par :

$$\mathbf{A} = \left[\begin{array}{cccc}
\begin{matrix} 4 & -1 & & \\ -1 & 4 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{matrix} & \begin{matrix} -1 & & & \\ & -1 & & \\ & & \ddots & \\ & & & -1 \end{matrix} & & \\
\begin{matrix} -1 & & & \\ & -1 & & \\ & & \ddots & \\ & & & -1 \end{matrix} & \begin{matrix} 4 & -1 & & \\ -1 & 4 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{matrix} & \begin{matrix} -1 & & & \\ & -1 & & \\ & & \ddots & \\ & & & -1 \end{matrix} & \\
 & \begin{matrix} -1 & & & \\ & -1 & & \\ & & \ddots & \\ & & & -1 \end{matrix} & \begin{matrix} 4 & -1 & & \\ -1 & 4 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{matrix} & \begin{matrix} -1 & & & \\ & -1 & & \\ & & \ddots & \\ & & & -1 \end{matrix} \\
 & & \begin{matrix} -1 & & & \\ & -1 & & \\ & & \ddots & \\ & & & -1 \end{matrix} & \begin{matrix} 4 & -1 & & \\ -1 & 4 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{matrix}
\end{array}\right]$$

et peut donc aussi s'écrire en tirant parti de sa structure par blocs sous la forme :

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & & 0 \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \ddots & \\ & \ddots & \ddots & \mathbf{A}_{N-1,N} \\ 0 & & \mathbf{A}_{N,N-1} & \mathbf{A}_{NN} \end{bmatrix}$$

On observe que la matrice $\mathbf{A}$ est symétrique et très creuse puisque ses termes non nuls sont concentrés sur 5 vecteurs, ou autrement dit que chaque ligne comportant 5 valeurs non nulles sur N^2 valeurs, et en fait trois valeurs différentes seulement du fait de la symétrie. Sa largeur de bande dépend de h, le paramètre de discrétisation des points sur la grille, ce qui implique que plus on met de points sur la grille, plus la bande est large.

Lors de résolutions numériques, on cherche évidemment à tirer parti de la structure creuse de la matrice. On voit tout de suite que la méthode de Gauss n'est pas très adaptée du point de vue du stockage, car en introduisant des zéros sous la diagonale, on garde la structure de bande mais on accroît le nombre de termes non nuls en remplissant l'espace entre les lignes de valeurs non nulles. Par contre le nombre d'opérations pour effectuer une décomposition de Choleski ($\mathbf{A} = \mathbf{LL}^\top$ puisque $\mathbf{A}$ est symétrique), c'est à dire essentiellement le calcul de $\mathbf{L}$, est d'environ $\frac{N^2}{4}$.

Les méthodes de Jacobi et celle de Gauss-Seidel, quant à elles, requièrent à chaque itération $5N^2$ opérations (une opération est une multiplication ou division plus une addition) au lieu de N^4 pour une matrice pleine, à condition de coder une fonction de multiplication matrice-vecteur optimisée tenant compte de la structure bande.

Les méthodes itératives sont donc plus intéressantes du point de vue du stockage mais

moins du point de vue du temps de calcul. De plus on peut montrer que le conditionnement de la matrice $\mathbf{A}$ pour la norme 2 vaut $cond_2(\mathbf{A}) = \frac{4}{\pi^2 h^2}$ (cf [SB] pour le détail des calculs) et est donc inversement proportionnel à h, ce qui a pour effet de ralentir la vitesse de convergence des méthodes itératives au fur et à mesure que h augmente.

Ces considérations ont donc occasionné l'apparition de différentes méthodes dans l'objectif d'échapper aux inconvénients ci-dessus :

- les méthodes itératives par blocs, tenant compte de la structure de la matrice, et dont nous ne parlerons pas ici,
- des méthodes itératives moins sensibles au conditionnement de la matrice du système, mais conservant les avantages de stockage, auxquelles nous allons consacrer la suite de ce chapitre.

4.2 Résolution d'un système linéaire : un problème d'optimisation

Remarquons que la résolution d'un système d'équations linéaires

$$\mathbf{A}\mathbf{x} = \mathbf{b}\ ;\ \mathbf{A} \in \mathbb{R}^{n\times n}, \mathbf{x} \in \mathbb{R}^n \tag{4.2.1}$$

peut aussi être vu comme un problème de minimisation d'une fonctionnelle dans le cas où la matrice $\mathbf{A}$ est symétrique, définie positive.

En effet, considérons la fonctionnelle

$$J(\mathbf{x}) = \frac{1}{2}\left(\mathbf{x}^\top \mathbf{A}\mathbf{x}\right) - \mathbf{x}^\top \mathbf{b}$$

Pour résoudre le système $(4.2.1)$ on cherche à calculer un vecteur $\mathbf{x} \in \mathbb{R}^n$ tel que

$$\mathbf{A}\mathbf{x} - \mathbf{b} = \mathbf{0}$$

Si on prend le gradient de la fonctionnelle on a

$$\nabla J(\mathbf{x}) = \frac{1}{2}\left(\mathbf{A}^\top + \mathbf{A}\right)\mathbf{x} - \mathbf{x}^\top \mathbf{A} = \mathbf{A}\mathbf{x} - \mathbf{b}$$

Par conséquant la solution $\mathbf{x}$ du système est aussi solution de $\nabla J(\mathbf{x}) = 0$, c'est-à-dire que la solution $\boldsymbol{x}$ minimise la valeur de la fonctionnelle $J(\mathbf{x})$. Il s'agit donc d'un problème d'optimisation sans contraintes.

Dans ce chapitre on commence par présenter les algorithmes qui permettent de trouver une solution au problème de l'optimisation et on termine en appliquant ces algorithmes dans le cas de la résolution d'un système d'équations linéaires.

Formellement pour un problème d'optimisation on considère une fonctionnelle $J : U \subseteq \mathbb{R}^n \to \mathbb{R}$ [1] On suppose que J est partout définie sur U.

Le problème de la minimisation de la fonctionnelle J consiste donc au calcul d'un élément $\mathbf{x}^* \in \mathbb{R}$ qui minimise J, c'est-à-dire tel que

$$J(\mathbf{x}^*) \leq J(\mathbf{x})\ \forall \mathbf{x} \in \mathbb{R}$$

1. Une fonctionnelle est une fonction au sens classique du terme à ceci près que son argument peut être une autre fonction, c'est-à-dire une fonctionnelle peut être une fonction de fonction.

On peut aussi avoir un problème de maximisation et, en règle générale, on parle des problèmes d'optimisation sans contraintes.

Lors de la résolution d'un problème d'optimisation se posent essentiellement deux questions :

(1) Quelle méthode doit-on utiliser pour calculer $\mathbf{x}^*$?

(2) Comment s'assurer que la valeur calculée est un minimum global et non pas local ?

4.3 Outils mathématiques

Dans la suite du chapitre on utilisera les trois notions suivantes relatives à la fonctionnelle J :

Gradient de J Le *gradient* de J est

$$\nabla J = \left[\frac{\partial J}{\partial x_1}, \frac{\partial J}{\partial x_2}, \cdots, \frac{\partial J}{\partial x_n}\right]$$

Hessienne de J La *matrice hessienne* de J est donnée par $\mathbf{H}(J) = \nabla^\top J \cdot \nabla J$ à savoir

$$\mathbf{H}(J) = \begin{bmatrix} \frac{\partial^2 J}{\partial^2 x_1} & \frac{\partial^2 J}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 J}{\partial x_1 \partial x_n} \\ \frac{\partial^2 J}{\partial x_2 \partial x_1} & \frac{\partial^2 J}{\partial x_2^2} & \cdots & \frac{\partial^2 J}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 J}{\partial x_{n1} \partial x_1} & \frac{\partial^2 J}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 J}{\partial^2 x_n} \end{bmatrix}$$

Notons que le déterminant de la hessienne s'appelle le *hessien*.

Norme de l'énergie C'est le scalaire $\|\mathbf{x}\|_{\mathbf{A}}^2 = \mathbf{x}^\top \mathbf{A}\mathbf{x}$ où l'indice $\mathbf{A}$ se refère à la matrice $\mathbf{A}$.

Les deux dernières notions permettent d'introduire un ordre non complet, au sens des formes quadratiques, dans l'ensemble de matrices symétriques de dimension $(n \times n)$. En effet on pose $\mathbf{A} \leq \mathbf{B}$ si et seulement si $\|\mathbf{x}\|_{\mathbf{A}}^2 \leq \|\mathbf{x}\|_{\mathbf{B}}^2$.

Nous avons aussi besoin de la notion de la convexité :

- Un ensemble E est *convexe* si pour tout $\mathbf{x}, \mathbf{y} \in E$ on a que le segment fermé $[\mathbf{x}, \mathbf{y}]$ est dans U.
- Une fonctionnelle J définie sur un ensemble U convexe, est *convexe* si

$$J(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda J(\mathbf{x}) + (1-\lambda) J(\mathbf{y}) \ ; \ 0 < \lambda < 1, \forall \mathbf{x}, \mathbf{y} \in \mathbf{U}$$

  Notons que la fonctionnelle est concave si la relation de l'inégalité est dans l'autre sens.
- Si J est *fortement convexe* de rapport a, alors

$$\forall \mathbf{u}, \mathbf{v} \in U : \quad J(\mathbf{v}) \geq J(\mathbf{u}) + \nabla J(\mathbf{u})^\top (\mathbf{v} - \mathbf{u}) + \frac{a}{2} \|\mathbf{u} - \mathbf{v}\|^2$$

Nous avons deux conditions pour l'optimalité d'un élément de U :

- Une condition nécsessaire : Si $\mathbf{x}^* \in U$ est un extremum de J et J est dérivable en $\mathbf{x}^*$, alors

$$\nabla J(\mathbf{x}^*) = 0$$

  Le contraire n'est pas obligatoirement vrai.

– Une condition suffisante : Soit U convexe et deux fois dérivable sur U. Soit $\mathbf{x}^* \in U$ un élément tel que $\nabla J(\mathbf{x}^*) = 0$. Si $\nabla^2 J(\mathbf{x})$ est symétrique, definie positive pour tout $\mathbf{x} \in U$, alors $\mathbf{x}^*$ minimise J sur U.

Notons que si $\nabla^2 J(\mathbf{x})$ est symétrique, définie négative, alors $\mathbf{x}^*$ maximise J sur U.

Notons aussi que la convexité forte implique immédiatement l'existence et l'unicité d'une solution optimale.

Nous avons la typologie suivante des points de U selon l'optimisation de J :

– $\mathbf{x}^* \in U$ est un point critique de J si elle est dérivable sur ce point et $\nabla J(\mathbf{x}) = 0$.

– un point critique $\mathbf{x}^* \in U$ de J est non dégénéré si $J \in \mathcal{C}^2$ sur une boule ouverte de centre $\mathbf{x}^*$ dans U et la hessienne $\mathbf{H}(\mathbf{x}^*)$ est regulière.

Nous terminons cette section par la définition du gradient lipschitzien

– Si J est à gradient lipschitzien de constante L, on a

$$\forall \mathbf{u}, \mathbf{v} \in U, \quad J(\mathbf{v}) - J(\mathbf{u}) \leq \nabla J(\mathbf{u})^T (\mathbf{v} - \mathbf{u}) + \frac{N}{2} \|\mathbf{v} - \mathbf{u}\|^2$$

4.4 Méthode de descente : formulation générale

Nous commençons par la définition du minimum local.

DÉFINITION 4.4.1 *Un minimum local de J sur U est un vecteur $\mathbf{x}^*$ tel qu'il existe une boule $B(\mathbf{x}^*, r)$ de centre $\mathbf{x}^*$ et rayon $r > 0$, avec*

$$\forall \mathbf{x} \in B(\mathbf{x}^*, r) \cap U : \ J(\mathbf{x}) \geq J(\mathbf{u}^*)$$

DÉFINITION 4.4.2 *Un minimum global de J sur U est un vecteur $\mathbf{x}^*$ tel que*

$$\forall \mathbf{x} \in U : \qquad J(\mathbf{x}) \geq J(\mathbf{x}^*)$$

Cette distinction étant faite, notons que sauf situation très favorable (il n'existe qu'un seul minimum ou bien on est dans le cadre de l'optimisation linéaire, quadratique à matrice positive ou optimisation convexe), nous nous contenterons d'un minimum local de J.

Pour trouver un minimum, les méthodes présentées dans la suite partent du point de vue intuitif suivant : on connaît un point $\mathbf{x}$ ainsi que la valeur de $J(\mathbf{x})$. En se déplaçant localement à partir de $\mathbf{x}$, on peut déterminer si J augmente ou diminue. Le déplacement le plus simple étant la ligne droite (dans un espace euclidien), on effectuera donc des petites excursions à partir du $\mathbf{x}$ vers différentes directions $\mathbf{d}$, en essayant d'améliorer la valeur du critère J.

DÉFINITION 4.4.3 *On appelle direction admissible en $\mathbf{x}$ un vecteur $\mathbf{d}$ (direction) le long duquel on pourra se déplacer en partant de $\mathbf{x}$ tout en restant dans U, c'est-à-dire tel qu'il existe un $h > 0$ de sorte que nous ayons $[\mathbf{x}, \mathbf{x} + h\mathbf{d}] \subset U$. On notera $D(\mathbf{x})$ l'ensemble des directions admissibles en $\mathbf{x}$.*

DÉFINITION 4.4.4 *Un vecteur $\mathbf{d} \in D(\mathbf{x})$ est une direction de descente pour une fonctionnelle J au point $\mathbf{x} \in U$ si pour tout réel $\gamma > 0$ il existe un $h \in]0, \gamma[$ tel que $J(\mathbf{x} + h\mathbf{d}) \leq J(\mathbf{x})$.*

Notons que si $\mathbf{x}$ est un minimum local de J, il n'existe aucune direction de descente pour J au point $\mathbf{x}$.

L'idée de la méthode de descente est de partir d'un point $\mathbf{x}^{(0)} \in U$ tel que $\nabla J\left(\mathbf{x}^{(0)}\right) \neq 0$. Pour calculer le prochain point $\mathbf{x}^{(1)} \in U$ on utilise une direction de descente $\mathbf{d}$ au point $\mathbf{x}^{(0)}$ et on a $\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + h\mathbf{d}$ avec $J\left(\mathbf{x}^{(1)}\right) \leq J\left(\mathbf{x}^{(0)}\right)$. De cette façon on réduit la valeur de la fonctionnelle J.

La méthode de descente se caractérise par deux choix :

- Le choix de la direction de la descente.
  - *Méthode de Cauchy* : $\mathbf{d} = -\nabla J(\mathbf{x})$. Elle donne naissance aux algorithmes de gradient qui minimisent $\nabla^{\top} J(\mathbf{x})\,\mathbf{d} = -\|\nabla J(\mathbf{x})\|^2$ qui est la dérivée de la fonctionnelle $J(\mathbf{x} + h\mathbf{d})$. Cette méthode est fondée sur le fait qu'en utilisant la formule de Taylor au premier ordre, on peut approcher J au voisinage de $\mathbf{x}^{(0)}$ par la fonction

$$J\left(\mathbf{x}^{(1)}\right) = J\left(\mathbf{x}^{(0)}\right) + \nabla J\left(\mathbf{x}^{(0)}\right) \cdot \left(\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right) + \mathcal{O}\left(\left\|\nabla J\left(\mathbf{x}^{(0)}\right)\right\|\right)$$

    Considérons la droite passant par $\mathbf{x}^{(0)}$ le long du gradient de J en $\mathbf{x}^{(0)}$:

$$\mathbf{x}(h) = \mathbf{x}^{(0)} - h \times \nabla J\left(\mathbf{x}^{(0)}\right)$$

    Si h est choisi positif, alors, en posant $\mathbf{x}^{(1)} = \mathbf{x}(h)$

$$J\left(\mathbf{x}^{(1)}\right) = J\left(\mathbf{x}^{(0)}\right) - h\left\|\nabla J\left(\mathbf{x}^{(0)}\right)\right\|^2 + \mathcal{O}\left(h\left\|\nabla J\left(\mathbf{x}^{(0)}\right)\right\|\right)$$

    et si h est suffisamment petit, on aura

$$J\left(\mathbf{x}^{(1)}\right) < J\left(\mathbf{x}^{(0)}\right)$$

    En d'autres termes, *la direction opposée à celle du gradient est une direction de descente*. C'est la meilleure direction *localement* (c'est-à-dire pour h petit). De ce fait, la méthode de Cauchy fournit à chaque itération une direction de descente. Par contre sa convergence est lente.
  - *Méthode de Newton* : $d = -\mathbf{H}^{-1}(J(\mathbf{x}))\,\nabla J(\mathbf{x})$, qui est une direction de descente si $\mathbf{H}(J(\mathbf{x}))$ est définie positive. De ce fait la méthode de Newton ne fournit pas obligatoirement une direction de descente. Par contre, si elle converge, sa convergence est plus rapide que celle de la méthode de Cauchy.
- Choix du pas h qui doit être défini de sorte que le nombre d'itérations soit minimal. Il y a deux techniques :
  - *Pas fixe*.- On choisit un pas fixe pour l'ensemble des itérations.
  - *Pas optimal*.- Le pas est choisi à chaque itération de sorte que la fonction $J\left(\mathbf{x}^{(0)} + \gamma\mathbf{d}\right)$ soit minimale.

4.5 Algorithme du gradient à pas fixe

Cet algorithme utilise la méthode de Cauchy avec un pas h constant et fixé d'avance.

ALGORITHME DU GRADIENT À PAS FIXE (DE PLUS GRANDE PENTE)

– En partant de $\mathbf{x}^{(0)}$, on calcule la direction de descente

$$\mathbf{d}_0 = -\nabla J\left(\mathbf{x}^{(0)}\right)^\top$$

et le nouveau point $\mathbf{x}^{(1)}$

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + h \times \mathbf{d}_0$$

– À l'étape k, connaissant $\mathbf{x}^{(k)}$, on calcule la direction de la descente

$$\mathbf{d}_k = -\nabla J\left(\mathbf{x}^{(k)}\right)^\top$$

et le nouveau point $\mathbf{x}^{(k+1)}$ par

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + h \times \mathbf{d}_k$$

– On décide d'arrêter l'algorithme lorsqu'un test de convergence

$$\left\|\nabla J\left(\mathbf{x}^{(k)}\right)\right\| < \theta = \text{seuil}$$

est vérifié.

Pour la convergence, nous avons le théorème suivant :

THÉORÈME 4.5.1 *Supposons que*

$\mathbf{H}_1$ *la fonctionnelle J est de classe $\mathcal{C}$ dans U ;*

$\mathbf{H}_2$ *l'ensemble* $U_0 = \left\{\mathbf{x} \in U \;\middle|\; J(\mathbf{x}) \leq J\left(\mathbf{x}^{(0)}\right)\right\} \subset U$ *est fermé ;*

$\mathbf{H}_3$ $\forall \mathbf{x} \in U_0$ *on a* $c \cdot I \leq \nabla^2 J(\mathbf{x}) \leq C \cdot I$, *où I la matrice identité, et*

$\mathbf{H}_4$ *Le pas h est choisi tel que* $h < \frac{2}{C}$.

Alors l'algorithme du gradient à pas fixe converge vers un minimum local $\mathbf{x}^* \in U$ *non dégénéré.*
De plus, nous avons

$$\left\|\mathbf{x}^* - \mathbf{x}^{(0)}\right\| \leq \frac{\left\|\nabla J\left(\mathbf{x}^{(0)}\right)\right\|}{c}$$

4.6 Algorithme du gradient à pas variable

Cet algorithme effectue à chaque itération, étant donnée la direction de descente $\mathbf{d} = \nabla J(\mathbf{x})^\top$, le calcul d'un pas h qui minimise la fonctionnelle $J(\mathbf{x} + h\mathbf{d})$.

ALGORITHME DU GRADIENT À PAS OPTIMAL

– En partant de $\mathbf{x}^{(0)}$, on calcule la direction de descente

$$\mathbf{d_0} = -\nabla J\left(\mathbf{x}^{(0)}\right)^\top$$

le pas optimal

$$h_0 = \arg\min_{h>0} J\left(\mathbf{x}^{(0)} - h \times \nabla J\left(\mathbf{x}^{(0)}\right)^\top\right)$$

et le nouveau point $\mathbf{x}^{(1)}$

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + h \times \mathbf{d_0}$$

– À l'étape k, connaissant $\mathbf{x}^{(k)}$, on calcule la direction de la descente

$$\mathbf{d_k} = -\nabla J\left(\mathbf{x}^{(k)}\right)^\top$$

le pas optimal

$$h_k = \arg\min_{h>0} J\left(\mathbf{x}^{(k)} - h \times \nabla J\left(\mathbf{x}^{(k)}\right)^\top\right)$$

et le nouveau point $\mathbf{x}^{(k+1)}$ par

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + h \times \mathbf{d_k}$$

– On décide d'arrêter l'algorithme lorsqu' un test de convergence

$$\left\|\nabla J\left(\mathbf{x}^{(k+1)}\right)\right\| < \theta = \text{seuil}$$

est vérifié.

4.6.1 Convergence de l'algorithme

Habituellement on prend $\theta = 10^{-6} \times \left\|\nabla J\left(\mathbf{x}^{(0)}\right)\right\|$. Ce test évite d'accumuler des itérations qui n'apportent plus rien à la qualité de la solution trouvée. Il ne donne en revanche aucune garantie pour l'optimalité de la solution trouvée. En particulier il faut vérifier que le point $\mathbf{x}^{(k+1)}$ correspondant est bien un minimum, car il se peut qu'il soit un point selle.

FAIT 4.1 *L'algorithme du gradient à pas optimal a les propriétés suivantes :*
- *Deux directions de recherche successives sont orthogonales. En effet, dans le calcul de h_k, si la minimisation est exacte (c'est à dire si l'on trouve exactement le meilleur h), la condition nécessaire d'optimalité de la recherche linéaire s'écrit*

$$\frac{d}{dh}J\left(\mathbf{x}^{(k)}-h\times\nabla J\left(\mathbf{x}^{(k)}\right)^{\top}\right)|_{h=h_k}=-\nabla J\left(\mathbf{x}^{(k+1)}\right)\cdot\nabla J\left(\mathbf{x}^{(k)}\right)^{\top}=0.$$

- *Dans le cas où J est quadratique, de la forme $J(\mathbf{u})=\frac{1}{2}\mathbf{u}^{\top}\mathbf{A}\mathbf{u}-\mathbf{b}^{\top}\mathbf{u}$ avec $\mathbf{A}$ symétrique définie positive, on peut calculer facilement le paramètre h_k. Il est donné par l'equation $\nabla J\left(\mathbf{x}^{(k+1)}\right)\cdot\nabla J\left(\mathbf{x}^{(k)}\right)^{\top}=0$ ou encore*

$$\left(\mathbf{A}\mathbf{x}^{(k+1)}-\mathbf{b}\right)^{\top}\left(\mathbf{A}\mathbf{x}^{(k)}-\mathbf{b}\right)=\left(\mathbf{A}\left[\mathbf{x}^{(k)}-h_k\left(\mathbf{A}\mathbf{x}^{(k)}-\mathbf{b}\right)\right]-\mathbf{b}\right)^{\top}\left(\mathbf{A}\mathbf{x}^{(k)}-\mathbf{b}\right)=0.$$

Nous considérons maintenant la convergence de la méthode du gradient à pas optimal dans le cas quadratique. Nous avons le théorème suivant :

THÉORÈME 4.6.1 *Si $J(\mathbf{x})=\frac{1}{2}\mathbf{x}^{\top}\mathbf{A}\mathbf{x}-\mathbf{b}^{\top}\mathbf{x}$ avec $\mathbf{A}$ symétrique définie positive, la méthode du gradient à pas optimal converge vers l'optimum unique $\mathbf{x}^*=\mathbf{A}^{-1}\mathbf{b}$. De plus*

$$\left\|\mathbf{x}^{(k+1)}-\mathbf{x}^*\right\|<\left\|\mathbf{x}^{(k+1)}-\mathbf{x}^*\right\|_{\mathbf{A}}^2\left(\frac{\kappa(\mathbf{A})-1}{\kappa(\mathbf{A})+1}\right)^2$$

où $\kappa(\mathbf{A})$ est le conditionnement de $\mathbf{A}$ relativement à la norme 2.

Nous avons aussi le lemme et le théorème suivants :

LEMME 4.6.1 (INÉGALITÉ DE KANTOROVICH) *Si $\mathbf{A}$ est une matrice symétrique définie positive d'ordre n, avec $\kappa(\mathbf{A})=\frac{\lambda_+}{\lambda}_-$, où$\lambda_+$ et λ_- la plus grande et la plus petite respectivement valeurs propres, alors :*

$$\frac{\|\mathbf{x}\|^2}{\|\mathbf{x}\|_{\mathbf{A}}^2\times\|\mathbf{x}\|_{\mathbf{A}^{-1}}^2}\geq\frac{4\kappa(\mathbf{A})}{(1+\kappa(\mathbf{A}))^2}$$

THÉORÈME 4.6.2 *Si J est continûment différentiable et fortement convexe, alors l'algorithme de gradient à pas optimal converge vers l'unique optimum.*

4.7 Méthode de Newton

Le principe de la méthode de Newton pour l'optimisation est de minimiser successivement les approximations au second ordre de la fonctionnelle J :

ALGORITHME DE NEWTON

– Soit $\mathbf{x}^{(0)} \in U$. Par un développement de Taylor au second ordre au voisinage de $\mathbf{x}^{(0)}$, on obtient :

$$J^0(\mathbf{x}) = J\left(\mathbf{x}^{(0)}\right) + \nabla J\left(\mathbf{x}^{(0)}\right)\left(\mathbf{x}-\mathbf{x}^{(0)}\right) + \frac{1}{2}\left(\mathbf{x}-\mathbf{x}^{(0)}\right)^{\top} J''\left(\mathbf{x}^{(0)}\right)\left(\mathbf{x}-\mathbf{x}^{(0)}\right)$$

On minimise la fonctionnelle quadratique $J^0(\mathbf{x})$, ce qui fournit un vecteur $\mathbf{x}^{(1)}$ qui est solution du système linéaire

$$\mathbf{H}\left(J\left(\mathbf{x}^{(0)}\right)\right)\mathbf{x}^{(1)} = \mathbf{H}\left(J\left(\mathbf{x}^{(0)}\right)\right)\mathbf{x}^{(0)} - \nabla J\left(\mathbf{x}^{(0)}\right)$$

et qui s'écrit comme suit : $\mathbf{x}^{(1)} = \mathbf{x}^{(0)} - \left[\mathbf{H}\left(J\left(\mathbf{x}^{(0)}\right)\right)\right]^{-1} \nabla J\left(\mathbf{x}^{(0)}\right)^{\top}$

– À l'itération k, on construit J^k, approximation quadratique de J au voisinage de $\mathbf{x}^{(k)}$, que l'on minimise pour obtenir $\mathbf{x}^{(k+1)}$, défini par

$$\mathbf{H}\left(J\left(\mathbf{x}^{(k)}\right)\right)\delta_k = -\nabla J\left(\mathbf{x}^{(k)}\right), \quad \mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \delta_k. \tag{4.7.1}$$

On notera le parti pris qui consiste à ne pas écrire $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \left[\mathbf{H}\left(J\left(\mathbf{x}^{(k)}\right)\right)\right]^{-1} \nabla J\left(\mathbf{x}^{(k)}\right)^{\top}$ dans l'algorithme. En effet, cette syntaxe sous-entendrait que l'on procède à l'inversion $\mathbf{H}\left(J\left(\mathbf{x}^{(k)}\right)\right)$ ce qui n'est absolument pas nécessaire.

4.7.1 Propriétés de l'algorithme de Newton.

L'algorithme de Newton est la généralisation multi-dimensionnelle de la méthode Newton-Raphson, appliquée à la recherche des racines de $G(\mathbf{x}) = \nabla J(\mathbf{x})$. On pourrait démontrer la convergence dans tout voisinage d'un minimum local, ainsi qu'une vitesse de convergence quadratique. Cette méthode fonctionne très bien pour des problèmes de petites dimensions (quelques dizaines de variables), lorsque le calcul de la hessienne $\mathbf{H}(J)$ est facile. Dans les autres cas, on préferera souvent une méthode de gradient conjugué (cf. infra) ou une méthode de quasi-Newton[2].

Dans le cas quadratique, elle fournit évidemment la solution du problème en une itération : $\mathbf{x}^{(1)} = \mathbf{A}^{-1}\mathbf{b}$. Mais ceci est bien sûr une illusion puisqu'il reste à résoudre le système linéaire $\mathbf{A}\mathbf{x} = \mathbf{b}$, c'est à dire $\mathbf{A}\mathbf{x}^{(1)} = \mathbf{b}$ qui constitue le plus gros du travail dans ce cas. Dans le cas général, pour résoudre le système $\mathbf{H}\left(J\left(\mathbf{x}^{(k)}\right)\right)\delta_k = -\nabla J\left(\mathbf{x}^{(k)}\right)$, lorsque $\mathbf{H}\left(J\left(\mathbf{x}^{(k)}\right)\right)$ est définie positive, la factorisation $\mathbf{LU}$ est la mieux adaptée.

La méthode de Newton peut aussi bien fournir des maxima locaux ou des points-selle, puisqu'elle cherche seulement à satisfaire la condition nécessaire d'optimalité $\nabla J(\mathbf{x}) = 0$. En d'autres termes, la direction de Newton, δ_k n'est pas forcément une direction de descente ! Si $\mathbf{H}\left(J\left(\mathbf{x}^{(k)}\right)\right)$ est définie positive, alors $\nabla J\left(\mathbf{x}^{(k)}\right)\delta_k = -\left\|\nabla J\left(\mathbf{x}^{(k)}\right)\right\|^2_{\mathbf{H}\left(J\left(\mathbf{x}^{(k)}\right)\right)} < 0$ pour $\nabla J\left(\mathbf{x}^{(k)}\right) \neq 0$. Si $\mathbf{H}\left(J\left(\mathbf{x}^{(k)}\right)\right)$ n'est pas définie positive, il existe des modifications qui permettent d'avoir une

2. Cette méthode dépasse le cadre de ce cours

convergence globale alliant les avantages de la vitesse de convergence quadratique près de la solution et ceux d'une méthode de descente.

4.8 Méthode de gradient conjugué

Soit J la fonctionnelle quadratique $J(\mathbf{u}) = \frac{1}{2}\mathbf{u}^\top \mathbf{A}\mathbf{u} - \mathbf{b}^\top \mathbf{u}$. Nous aurons besoin de la définition suivante :

DÉFINITION 4.8.1 *$\mathbf{A}$ étant une matrice symétrique donnée, deux directions $\mathbf{d}_0$ et $\mathbf{d}_1$ sont dites conjuguées par rapport à $\mathbf{A}$ si $(\mathbf{d}_0)^\top \mathbf{A}\mathbf{d}_1 = 0$.*

Le lemme suivant est aussi utile.

LEMME 4.8.1 *Si $\mathbf{A}$ est symétrique définie positive, et si les vecteurs non nuls $(\mathbf{d}_1, \cdots, \mathbf{d}_k)$ sont conjugués deux à deux par rapport à $\mathbf{A}$, alors ils forment une famille libre.*

Ce résultat s'applique d'une manière très astucieuse : si l'on note $\mathbf{x}^*$ la solution (unique) de $\mathbf{A}\mathbf{x} = \mathbf{b}$, et si l'on dispose d'une famille de n vecteurs conjugués par rapport à $\mathbf{A}$, alors, la famille étant une base de $\mathbb{R}^n$, on peut exprimer $\mathbf{x}^*$ dans cette base :

$$\mathbf{x}^* = \sum_{j=0}^{n-1} h_j \mathbf{d}_j$$

On calcule aisément les coefficients h_i en multipliant cette équation à gauche par $\mathbf{d}_i^\top \mathbf{A}$, d'où

$$\mathbf{d}_i^\top \mathbf{A}\mathbf{x}^* = h_i \times \left(\mathbf{d}_i^\top \mathbf{A}\mathbf{d}_i\right) = h_i \times \|\mathbf{d}_i\|_{\mathbf{A}}^2$$

Comme $\mathbf{A}\mathbf{x}^* = \mathbf{b}$, on peut donc calculer les h_i uniquement à partir des données du problème, $\mathbf{A}$ et $\mathbf{b}$:

$$h_i = \frac{\mathbf{d}_i^\top \mathbf{b}}{\|\mathbf{d}_i\|_{\mathbf{A}}^2}.$$

Nous avons ainsi mis en evidence une méthode directe de calcul de $\mathbf{x}^*$. De plus, cette méthode nécessite au plus n étapes (trouver les directions conjugués $\mathbf{d}_i$) pour la fonctionnelle quadratique J. La seule difficulté est donc de construire successivement des directions conjuguées. Comme on l'a vu plus haut, les vecteurs propres de $\mathbf{A}$ constituent une solution, mais c'est une solution onéreuse. Essayons de mettre en œuvre une méthode plus simple, qui optimise J en même temps qu'elle construit de nouvelles directions conjuguées :

- Fixons un $\mathbf{x}^{(0)}$, et choisissons $\mathbf{d}_0 = \nabla J(\mathbf{x}^{(0)})^\top$.
- Notons $\mathbf{x}^{(1)}$ le vecteur obtenu par application de la méthode du gradient à pas optimal, le long de $\mathbf{x}^{(0)}$:

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} - h_0 \mathbf{d}_0$$

Remarquons que h_0 a été calculé plus haut. Il est donné par,

$$h_0 = \frac{\left\|\nabla J\left(\mathbf{x}^{(0)}\right)\right\|^2}{\left\|\nabla J\left(\mathbf{x}^{(0)}\right)\right\|_{\mathbf{A}}^2}$$

ce que l'on écrira

$$h_0 = \frac{\mathbf{d}_0^\top \nabla J\left(\mathbf{x}^{(0)}\right)}{\|\mathbf{d}_0\|_{\mathbf{A}}^2}$$

On dispose maintenant de $\nabla J(\mathbf{x}^{(1)})$ (orthogonal à $\nabla J(\mathbf{x}^{(0)})$).

– Recherchons une nouvelle direction $\mathbf{d}_1$ comme combinaison, linéaire de $\mathbf{d}_0$ et $\nabla J\left(\mathbf{x}^{(1)}\right)$, qui soit conjuguée de $\mathbf{d}_0$ par rapport à $\mathbf{A}$ [3].
Comme on désire que $\mathbf{d}_1$ et $\mathbf{d}_0$ soient conjugués, nous avons :

$$0 = \mathbf{d}_0^\top \mathbf{A}\mathbf{d}_1 = \mathbf{d}_0^\top \mathbf{A}\nabla J\left(\mathbf{x}^{(1)}\right)^\top - \beta_0 \mathbf{d}_0^\top \mathbf{A}\mathbf{d}_0 \Longrightarrow \beta_0 = \frac{\mathbf{d}_0^\top \mathbf{A}\nabla J\left(\mathbf{x}^{(1)}\right)^\top}{\mathbf{d}_0^\top \mathbf{A}\mathbf{d}_0}.$$

ce qui permet d'évaluer une nouvelle quantité β.

– Nous pouvons maintenant optimiser J le long de la direction $\mathbf{d}_1$, ce qui fournit

$$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} - h_1 \mathbf{d}_1, \qquad \text{avec } h_1 = \frac{\mathbf{d}_1^\top \cdot \nabla J\left(\mathbf{x}^{(1)}\right)}{\|\mathbf{d}_1\|_{\mathbf{A}}^2}$$

et on recherche une nouvelle direction d_2, de forme

$$\mathbf{d}_2 = \nabla J\left(\mathbf{x}^{(2)}\right) - \beta_1 \mathbf{d}_1$$

qui soit conjuguée par rapport à $\mathbf{d}_0$ et $\mathbf{d}_1$, etc.

Le fait que $\mathbf{d}_2$ soit conjuguée avec $\mathbf{d}_1$ s'impose naturellement, comme précédemment, en choisissant β_1 ad hoc :

$$\beta_1 = \frac{\mathbf{d}_1^\top \mathbf{A}\nabla J\left(\mathbf{x}^{(2)}\right)}{\mathbf{d}_1^\top \mathbf{A}\mathbf{d}_1}$$

Il est plus étonnant que $\mathbf{d}_2$ ainsi définie soit conjuguée avec $\mathbf{d}_0$. Cette propriété provient du fait que J est quadratique.

3. ces deux vecteurs sont les seuls dont on dispose, pour l'instant, et qui fournissent de l'information sur la fonctionnelle J

ALGORITHME DE GRADIENT CONJUGUÉ

– On choisit $\mathbf{x}^{(0)}$ et on calcule la direction de descente

$$\mathbf{d}_0 = \nabla J\left(\mathbf{x}^{(0)}\right)^\top$$

– À l'itération $k \geq 1$, on calcule la valeur du pas

$$h_k = \frac{\mathbf{d}_{k-1}^\top \cdot \nabla^\top J\left(\mathbf{x}^{(k-1)}\right)}{\mathbf{d}_{k-1}^\top \mathbf{H}\left(J\left(\mathbf{x}^{(k-1)}\right)\right)\mathbf{d}_{k-1}}$$

le nouveau point

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} - h_k \mathbf{d}_{k-1}$$

et la nouvelle direction, conjuguée aux précédentes

$$\mathbf{d}_k = \nabla J\left(\mathbf{x}^{(k)}\right) - \frac{\mathbf{H}\left(J\left(\mathbf{x}^{(k)}\right)\right)}{\mathbf{H}\left(J\left(\mathbf{x}^{(k-1)}\right)\right)}\mathbf{d}_k$$

– On décide d'arrêter l'algorithme lorsqu' un test de convergence

$$\left\|\nabla J\left(\mathbf{x}^{(k)}\right)\right\| < \theta = \text{seuil}$$

est vérifié.

Cette méthode est donc à peine plus compliquée à mettre en œuvre que la méthode du gradient à pas optimal, mais elle converge en n itérations lorsque J est quadratique. On peut aussi considérer la méthode du gradient conjugué comme une méthode de relaxation appliquée à J dans le système de coordonnées induit par les vecteurs propres de $\mathbf{A}$. La recherche de directions conjuguées est alors simplement une façon de décomposer J par rapport à ces nouvelles coordonnnées. L'algorithme de gradient conjugué admet plusieurs formulations différentes qui sont équivalentes pour les fonctions quadratiques mais peuvent avoir des caractéristiques différentes si J est plus générale.

4.9 Application à la résolution d'un système linéaire

Nous avons vu au début de ce chapitre que la solution d'un système linéaire $\mathbf{Ax} = \mathbf{b}$ peut être vue comme la minimisation d'une fonctionnelle J associée à $\mathbf{A}$ et $\mathbf{b}$. La résolution du système devient alors un problème d'optimisation qu'on peut traiter itérativement en utilisant les algorithmes de gradient. Nous donnons ci-après l'algorithme du gradient conjugué dans le cas d'un système linéaire.

Soit une matrice carrée $\mathbf{A}$ de dimension n, symétrique définie positive et $\mathbf{b}$ un vecteur fixé. On pose

$$J(\mathbf{x}) = \frac{1}{2}\left(\mathbf{x}^\top \mathbf{A}\mathbf{x}\right) - \left(\mathbf{b}^\top \mathbf{x}\right)$$

ALGORITHME DE DE GRADIENT CONJUGUÉ APPLIQUÉ À LA RÉSOLUTION D'UN SYSTÈME LINÉAIRE

(1) On choisit un vecteur initial $\mathbf{x}^{(0)} \in \mathbb{R}^n$. On calcule le résidu

$$\mathbf{r}_0 = \mathbf{A}\mathbf{x}^{(0)} - \mathbf{b}$$

et la direction

$$\mathbf{d}_0 = \mathbf{r}_0$$

– À l'itération $k \geq 1$, on calcule la valeur du pas

$$h_k = \frac{\mathbf{d}_{k-1}^\top \cdot \mathbf{r}_{k-1}}{\mathbf{d}_{k-1}^\top \mathbf{A}\mathbf{d}_{k-1}}$$

le nouveau point

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + h_k \mathbf{d}_{k-1}$$

le résidu

$$\mathbf{r}_k = \mathbf{r}_{k-1} - h_k \mathbf{A}\mathbf{d}_{k-1}$$

et la nouvelle direction, conjuguée aux précédentes

$$\mathbf{d}_k = \mathbf{r}_k - \frac{(\mathbf{A}\mathbf{d}_{k-1})^\top \cdot \mathbf{r}_k}{(\mathbf{A}\mathbf{d}_{k-1})^\top \cdot \mathbf{d}_{k-1}}$$

– On décide d'arrêter l'algorithme lorsqu' un test de convergence

$$\|\mathbf{r}_k\| < \theta = \text{seuil}$$

est vérifié.

On peut démontrer que l'algorithme de gradient conjugé converge en n itérations si $\mathbf{A}$ est symétrique, définie positive. Néanmoins il faut faire attention aux erreurs dues à la précision de la machine. Plus la valeur de n est grande, plus les erreurs de calcul s'accumulent avec comme conséquence que les directions successives ne soient pas conjuguées. Dans ce cas il est possible que la méthode ne converge pas au bout de n itérations.

Notons, pour finir, que la contrainte d'avoir une matrice $\mathbf{A}$ symétrique, définie positive, afin que les méthodes de gradient puissent s'appliquer, n'est pas limitative. En effet, considérons le système (4.2.1)

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{4.9.1}$$

avec $\mathbf{A}$ matrice carrée, régulière, quelconque. On peut appliquer les méthodes de gradient au système

$$\mathbf{A}^\top \mathbf{A}\mathbf{y} = \mathbf{A}^\top \mathbf{b} \tag{4.9.2}$$

parce que la matrice $\mathbf{A}^\top \mathbf{A}$ est symétrique, définie positive. On obtient donc la solution

$$\mathbf{y} = \left(\mathbf{A}^\top \mathbf{A}\right)^{-1} \mathbf{A}^\top \mathbf{b} = \mathbf{A}^{-1}\mathbf{b} = \mathbf{x}$$

c'est-à-dire la solution du système (4.9.2) est la même que celle du système (4.2.1).

4.10 Exercice

EXERCICE 4.1 *Soit la fonctionnelle*

$$J(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} + \mathbf{x}^\top \mathbf{b}; \mathbf{A} \in \mathbb{R}^n$$

avec $\mathbf{A}$ *matrice symétrique définie positive avec valeurs propres* $\lambda_1 \geq \ldots \geq \lambda_n > 0$.

(1) Calculer $\mathbf{x}^*$ *le point qui minimise* $J(\mathbf{x})$*, ainsi que la valeur de* $J(\mathbf{x}^*)$.

(2) Calculer pour l'algorithme à pas fixe, la direction $\mathbf{d}_k$ *à la* k*-ième itération.*

(3) Calculer $\mathbf{x}^{(k+1)}$ *et* $J\left(\mathbf{x}^{(k+1)}\right)$ *pour l'algorithme à pas fixe.*

(4) Il s'agit maintenant de trouver un pas h *pour cette itération qui optimise le critère, c'est-à-dire qui minimise la valeur de* $J\left(\mathbf{x}^{(k+1)}\right)$.

Calculer la valeur de ce pas h.

(5) En utilisant la nouvelle valeur calculée de h*, évaluer le prochain point* $\mathbf{x}^{(k+1)}$ *et la valeur de* $J\left(\mathbf{x}^{(k+1)}\right)$ *en fonction de* $\mathbf{x}^{(k)}$ *et* $\mathbf{d}_k$.

(6) Montrer qu'on a pour le taux de convergence

$$\frac{J\left(\mathbf{x}^{(k+1)}\right) - J(\mathbf{x}^*)}{J\left(\mathbf{x}^{(k)}\right) - J(\mathbf{x}^*)} = 1 - \frac{1}{\beta}$$

où

$$\beta = \frac{\left(\mathbf{d}_k^\top \mathbf{A}\mathbf{d}_k\right)\left(\mathbf{d}_k^\top \mathbf{A}^{-1}\mathbf{d}_k\right)}{\left(\mathbf{d}_k^\top \mathbf{d}_k\right)^2}$$

(7) Sachant que

$$\beta \leq \frac{(\lambda_1 + \lambda_n)}{4\lambda_1\lambda_n}$$

montrer pour le taux de convergence l'inégalité suivante :

$$\frac{J\left(\mathbf{x}^{(k+1)}\right) - J(\mathbf{x}^*)}{J\left(\mathbf{x}^{(k)}\right) - J(\mathbf{x}^*)} \leq \left(\frac{\kappa(\mathbf{A}) - 1}{\kappa(\mathbf{A}) + 1}\right)^2$$

où $\kappa(\mathbf{A})$ *est le conditionnement de la matrice selon la norme 2.*

(8) Application.- Calculer une borne supérieure du taux de convergence pour le système

$$\begin{bmatrix} 20 & 5 \\ 20 & 2 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 14 \\ 6 \end{bmatrix}$$

Remarques.

4.11 Références

Les livres qui sont pris en compte pour la redaction de ce chapitre sont les suivants :

E. ANGELINI : Polycopié sur l'optimisation, ENST,
http ://perso.telecom-paristech.fr/ angelini/master_spsiv/optimization/
D. P. BERTSEKAS, J. N. TSITSIKLIS : *Parallel and distributed computation*, Prentice-Hall, 1989
C. BREZINSKI : *Projection methods fot systems of equations*, Elsevier, 1997
P. G. CIARLET : *Introduction à l'analyse numérique matricielle et à l'optimisation*, Masson, 1988
P. G. CIARLET, B. MIARA, J.M. THOMAS : *Exercices d'analyse numérique matricielle et d'optimisation*, 2e édition, Masson, 1987
F. R. GANTMACHER : *Théorie des matrice, tome 1, Théorie générale*, Dunod, 1966
C. D. MEYER : *Matrixanalysis and applied linear algebra*, SIAM, 2001
R. D. MILNE : *Applied functional analysis*, Pittman, 1980
A. QUARTERONI, R. SACCO, F. SALERI : *Numerical mathematics*, Springer, 2000

5

CHOIX DES MÉTHODES DE RÉSOLUTION DE SYSTÈMES LINÉAIRES ET PRÉCONDITIONNEMENT

5.1 Introduction

Le fait de disposer de différentes méthodes de résolution d'un système linéaire est évidemment un avantage, mais cela ne résout pas la question qui se pose au néophyte : quelle méthode choisir dans un cas précis ? Au risque de ne pas rassurer le lecteur, on peut affirmer qu'il n'existe pas de méthode meilleure que les autres dans l'absolu, il n'y a que des cas particuliers dans lesquels le choix devra s'orienter vers telle méthode plutôt que telle autre. L'ingénieur aura à charge de déterminer les caractéristiques particulières de la matrice ou de la solution cherchée, afin de choisir une méthode adéquate. En cela il pourra s'aider des points de repères fournis dans le premier paragraphe de ce chapitre. Dans les cas difficiles, ou simplement lorsqu'une amélioration de la convergence est nécessaire, il pourra avoir recours aux techniques de préconditionnement abordées dans le second paragraphe.

Notons que la caractérisation des propriétés d'une matrice fait appel à des notions d'algèbre linéaire, et que la maîtrise de cette matière sera donc particulièrement utile (encore une fois).

5.2 Ce qui se voit à l'œil nu

Avant d'aller plus loin, il semble bon de rappeler que la condition *sine qua non* dont on doit s'assurer avant tout, est que la matrice soit carrée et **régulière.** Dans tous les cas contraires (non carrée et/ou non régulière) on se reportera à la décomposition en valeurs singulières et aux méthodes de moindres carrés vues dans la suite de ce cours. On considérera donc ici dorénavant que la matrice du système est carrée et régulière.

Dans le choix de la méthode de résolution d'un système linéaire le premier paramètre à prendre en compte est la taille du système, n. Dans le cas des systèmes de petite taille, disons $n < 10$ pour fixer les idées, toutes les méthodes feront l'affaire en général, sauf cas particuliers rares. On peut indifféremment ou presque, choisir une méthode directe ou itérative, sachant que dans la mesure où on peut accéder à une solution "exacte" en un nombre fini d'opérations il ne faut pas s'en priver : on privilégiera donc les méthodes directes en général.

Dans le cas des systèmes de grande taille, il en va tout autrement. Le choix de la méthode doit tenir compte de différents facteurs :

- les propriétés de la matrice : symétrie, définie positivité, structure creuse ou pleine, conditionnement,
- les besoins de l'utilisateur : précision, rapidité, parallélisation,
- les moyens de calcul disponibles : accès mémoire, processeurs rapides,
- le temps de développement : de nombreuses librairies existent et il est souvent plus intéressant de bien choisir une méthode déjà programmée et testée que de développer son propre code.

Nous donnons ci-après quelques éléments de choix relatifs au premier point, qui relève du périmètre de l'analyse numérique. Le second et le troisième seront relatifs au contexte technique, le dernier au contexte du projet et à son budget.

5.2.1 Systèmes linéaires creux

De nombreux problèmes de résolution d'équations aux dérivées partielles conduisent à des matrices creuses. On en a donné un exemple dans le chapitre précédent. Comme on l'a déjà dit, lors de résolutions numériques, on cherche évidemment à tirer parti de la structure creuse de la matrice. On voit tout de suite que la méthode de Gauss n'est pas très adaptée du point de vue du stockage, car en introduisant des zéros sous la diagonale, on garde la structure de bande mais on accroît le nombre de termes non nuls en remplissant l'espace entre les lignes de valeurs non nulles. Par contre le nombre d'opérations pour effectuer une décomposition de Cholesky ($\mathbf{A} = \mathbf{L}\mathbf{L}^\top$ lorsque $\mathbf{A}$ est symétrique), c'est à dire essentiellement le calcul de $\mathbf{L}$, est d'environ $\frac{n^2}{4}$, où n l'ordre de la matrice.

Les méthodes de Jacobi et celle de Gauss-Seidel, quant à elles, requièrent à chaque itération $5n^2$ opérations (une opération est une multiplication ou division plus une addition) au lieu de

n^4 pour une matrice pleine, à condition de coder une fonction de multiplication matrice-vecteur optimisée tenant compte de la structure bande.

Les méthodes itératives sont donc plus intéressantes du point de vue du stockage mais moins du point de vue du temps de calcul.

5.2.1.1 Cas des largeurs de bande constantes

Matrices symétriques

On se limite dans un premier temps aux matrices symétriques car les problèmes qui mènent à la résolution de matrices bandes, sont souvent également générateurs de matrices symétriques. Si la largeur de bande est constante en fonction de la taille, et faible vis à vis de cette dernière, il est intéressant de rester sur une méthode de Cholesky ($\mathbf{A}$ est décomposée sous la forme $\mathbf{A} = \mathbf{L}\mathbf{D}\mathbf{L}^\top$), car on peut montrer facilement (cf exercice ci-dessous) que le profil de $\mathbf{A}$ est conservé (i.e. la largeur de bande de la partie inférieure de $\mathbf{A}$ est la même que celle de la partie inférieure de $\mathbf{L}$). Par contre il est à noter que dans le cas de matrices non symétriques, la largeur de bande de $\mathbf{L}$ ne se transmet à $\mathbf{U}$.

Du fait de la propriété de conservation de la largeur de bande, on gagnera sur le stockage et sur le nombre d'opérations. On aura intérêt dans ce cas, à utiliser un stockage particulier, dit "stockage profil". Il fait appel à la notion de profil, définie ci-après :

Definition Le profil d'une matrice symétrique $\mathbf{A}$ est $\{(i,j), 1 \leq i \leq n, j_i \leq j \leq i\}$ où j_i est l'indice de la colonne du premier élément non nul de la ligne i.

L'intérêt de ce rangement réside dans le fait que le profil de $\mathbf{L}$ est inclus dans celui de $\mathbf{A}$. Ainsi, si l'on réserve une quantité de place mémoire suffisante pour stocker le profil de $\mathbf{A}$, on pourra ranger au fur et à mesure les coefficients de la matrice $\mathbf{L}$ dans cette place. Les éléments diagonaux de $\mathbf{L}$ valant 1 n'ayant pas besoin d'être stockés, les éléments de $\mathbf{D}$ seront rangés à la place des coefficients diagonaux de $\mathbf{A}$. On constate alors le double avantage de ce rangement : d'une part on fait une grande économie de place mémoire (à condition que le profil de $\mathbf{A}$ soit assez petit), d'autre part la gestion des données reste assez simple : il suffit de stocker le profil de $\mathbf{A}$ sous forme d'un tableau, la taille et la forme de ce tableau n'évoluant pas au cours de la factorisation.

EXERCICE 5.1 *En exprimant le terme général L_{ij} de la matrice $\mathbf{L}$ de la décomposition de Cholesky en fonction de celui de $\mathbf{A}$, de termes de $\mathbf{D}$ et des termes L_{ik} et L_{jk} pour k inférieur à j, montrer par récurrence que si $\mathbf{A}$ est tridiagonale alors la matrice L l'est aussi.*

Matrices non symétriques

Definition Soit $\mathbf{A}$ une matrice de taille n. On appelle largeur de bande inférieure (resp. supérieure) de la matrice $\mathbf{A}$, l'entier q (resp. p) tel que

$$\forall i = 1,...n, \forall j = 1,..,i-q, \qquad A_{ij} = 0$$
$$\forall j = 1,..,n, \forall i = 1,...j-p, \qquad A_{ij} = 0$$

REMARK 1 *Si la matrice* $\mathbf{A}$ *est symétrique alors la largeur de bande inférieure est égale à la largeur de bande supérieure*

On peut montrer aussi dans le cas de matrices non symétriques, que le profil se conserve :

THÉORÈME 5.2.1 *On suppose que* $\mathbf{A}$ *est une matrice de taille* n*, possédant une factorisation* $\mathbf{LU}$*. Si* $\mathbf{A}$ *est de largeur de bande supérieure* q*, et de largeur de bande inférieure* p*, alors* $\mathbf{U}$ *est de largeur de bande* q *et* $\mathbf{L}$ *de largeur de bande* p*.*

On pourra trouver la démonstration de ce théorème dans [GV].

Il est important de noter que dans le cas où $n \gg p$ et $n \gg q$ alors on peut facilement établir un algorithme de décomposition $\mathbf{LU}$ revenant à $2npq$ opérations.

EXERCICE 5.2 *Modifier l'algorithme de la factorisation* $\mathbf{LU}$ *pour tenir compte du cas d'une matrice* $\mathbf{A}$ *de largeur de bande supérieure* q*, et de largeur de bande inférieure* p*.*
De même, modifier les algorithmes de descente et de remontée permettant la résolution du système $\mathbf{Ax} = \mathbf{b}$*.*

5.2.1.2 Cas des largeurs de bande non constantes

Nous nous limitons au cas de matrices symétriques, possédant de nombreuses valeurs nulles mais ne présentant pas de structure bande. Dans ce cas, il serait évidemment maladroit de ne pas tenir compte des coefficients nuls, mais la notion de bande ne le permet pas. On utilise alors par exemple une représentation des valeurs non nulles sous forme de graphe, puis une renumérotation des sommets du graphe permet de transformer la matrice en une matrice de profil minimal.

À la matrice $\mathbf{A}$ de taille n, associons un graphe $\mathbf{G} = g(\mathbf{A})$ défini par :
- Les sommets de $\mathbf{G}$ sont numérotés de 1 à n.
- Les sommets i et j de $\mathbf{G}$ sont reliés si et seulement si A_{ij} est non nul.

Inversement, nous pouvons associer à tout graphe $\mathbf{G}$ une matrice A telle que $\mathbf{G} = g(\mathbf{A})$.

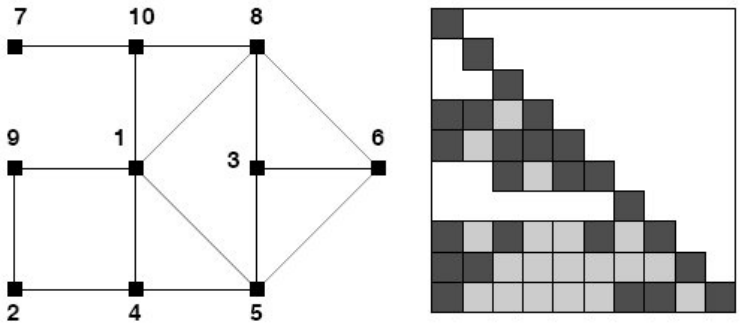


FIGURE 5.1 – Numérotation initiale

Le graphe $g(\mathbf{A})$ traduit en fait les relations entre les inconnues du système : les inconnues x_i et x_j sont en effet reliés si $A_{ij} = 0$, c'est-à-dire si les sommets i et j sont reliés. Si nous renumérotons les sommets sans changer la structure du graphe, nous ne changeons pas le système,

mais seulement l'ordre dans lequel apparaissent les inconnues. Il faut donc trouver une numérotation optimale du graphe, c'est-à-dire une numérotation dont la matrice associée a un profil de dimension minimale.

L'algorithme suivant est dû à Cuthill et Mc Kee, il permet de trouver de manière heuristique une bonne numérotation du graphe.

Algorithme (Cuthill-Mc Kee direct)

(1) On choisit un premier sommet.

(2) On numérote ses voisins.

(3) On numérote les voisins non encore numérotés du numéro 2, puis du numéro 3, . . .

À chaque étape, lorsqu'il y a plusieurs sommets à numéroter, on numérote en premier les sommets qui ont le moins de voisins non encore numérotés.

Cet algorithme consiste en fait en un parcours en largeur du graphe.

Chaque sommet du graphe est visité une et une seule fois. (Voir Fig. 2)

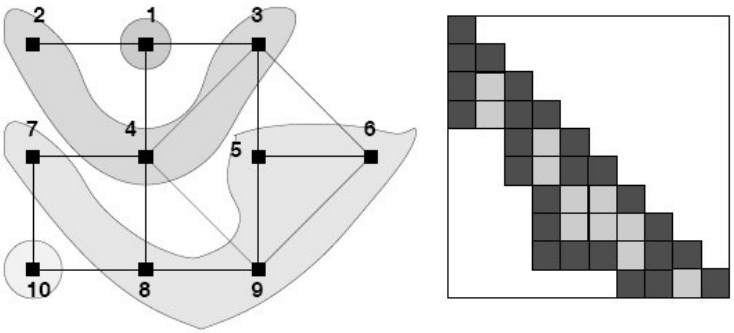


FIGURE 5.2 – 11 zéros dans le profil

5.2.2 Systèmes avec matrices pleines

Dans le cas de matrices pleines, de grande taille puisqu'on a déjà traité le cas des matrices de taille réduite, le nombre d'opérations à effectuer rend l'utilisation des méthodes directes très délicate :

- D'une part, le nombre d'opérations va occasionner des erreurs de calculs qui pénalisent le résultat
- D'autre part, il est nécessaire de faire tous les calculs jusqu'au dernier pour pouvoir obtenir la solution : il est impossible d'obtenir une solution approchée, mais si cette dernière suffirait aux besoins de l'utilisateur.

De ce fait, les méthodes itératives sont en général privilégiées dans ce cas. Ce n'est pas pour autant si simple. De nombreux problèmes peuvent se poser :

(1) La matrice peut ne pas être symétrique, ni définie positive, on verra au paragraphe (5.2.2.1) une piste de solution dans ce cas

(2) La matrice peut avoir un mauvais conditionnement, on verra dans le paragraphe (5.3) quelques techniques destinées à palier ce problème

Pour ces différentes raisons, une importante littérature est développée sur ces sujets. Face au problème de la résolution d'un système linéaire, et confronté à des exigences de performances ou de qualité de résultat, l'ingénieur devra envisager plusieurs solutions tirées de l'état de l'art, et faire preuve d'une bonne dose d'esprit critique.

5.2.2.1 Méthode de l'équation normale

Dans l'objectif de résoudre le système linéaire $\mathbf{Ax} = \mathbf{b}$ quand la matrice $\mathbf{A}$ n'est pas symétrique, on peut résoudre le système équivalent :

$$\mathbf{A}^\top \mathbf{Ax} = \mathbf{A}^\top \mathbf{b} \tag{5.2.1}$$

qui présente une matrice à la fois symétrique et définie positive.Le système est connu sous le nom de *systèmes d'équations normales* associé au système $\mathbf{Ax} = \mathbf{b}$. Il est associé au problème de minimisation au sens des moindres carrés :

$$\text{Trouver le minimum de } \|\mathbf{b} - \mathbf{Ax}\|_2$$

On peut remarquer que l'équation $(5.2.1)$ est utilisée pour résoudre les problème aux moindres carrés pour des systèmes surdéterminés, c'est à dire pour des matrices rectangulaires de taille nxm, $m < n$.

Une alternative à cette méthode est de poser $\mathbf{x} = \mathbf{A}^\top \mathbf{u}$ et de résoudre le problème en u :

$$\mathbf{AA}^\top \mathbf{u} = \mathbf{b} \tag{5.2.2}$$

Une fois que la solution u est obtenue, il suffit de la prémultiplier par A^T pour obtenir x.

On remarque qu'à partir du moment où la matrice du système est symétrique définie positive, on peut utiliser par exemple un algorithme de gradient conjugué pour résoudre le système. Néanmoins, il serait trompeur de considérer qu'on a trouvé dans cette astuce une excellente idée. En effet, dans le cas où la matrice A est dotée d'un mauvais conditionnement, celui de $A^T A$ est encore pire. Remarquons que le conditionnement en norme 2 de la matrice $A^T A$ est donné par

$$\kappa_2(\mathbf{A}^\top \mathbf{A}) = \left\|\mathbf{A}^\top \mathbf{A}\right\|_2 \left\|\left(\mathbf{A}^\top \mathbf{A}\right)^{-1}\right\|_2$$

Or $\left\|\mathbf{A}^\top \mathbf{A}\right\|_2 = \rho(\mathbf{A}^\top \mathbf{A})$ car $\mathbf{A}^\top \mathbf{A}$ est symétrique. Mais par ailleurs $\|\mathbf{A}\|_2 = \sqrt{\rho(\mathbf{A}^\top \mathbf{A})}$ dans le cas général, donc en utilisant le même raisonnement pour le second membre :

$$\kappa_2(\mathbf{A}^\top \mathbf{A}) = \|\mathbf{A}\|_2^2 \left\|\mathbf{A}^{-1}\right\|_2^2 = \kappa_2^2(\mathbf{A})$$

On a donc montré que le conditionnement de $\mathbf{A}^\top \mathbf{A}$ est le carré de celui de $\mathbf{A}$.

Dans le cas où la matrice $\mathbf{A}$ est bien conditionnée, la méthode de l'équation normale peut donc s'avérer une bonne solution. Par contre dans le cas de conditionnements moyens ou mauvais, les calculs des itérés peuvent conduire à la divergence ou à une propagation des erreurs telle qu'on n'obtient aucune solution satisfaisante. On est donc encore conduit à chercher à améliorer le conditionnement de la matrice d'un système.

5.2.2.2 Alternative à la méthode de l'équation normale

Il existe différentes alternatives à la méthode ci-dessus. Citons par exemple, le système équivalent à $\mathbf{Ax} = \mathbf{b}$:

$$\begin{bmatrix} \mathbf{I} & \mathbf{A} \\ \mathbf{A}^\top & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{r} \\ \mathbf{x} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix}$$

qui, avec $\mathbf{r} = \mathbf{b} - \mathbf{Ax}$ et $A^T r = 0$ conduit bien sûr à résoudre le système initial. La méthode est équivalente à la résolution d'un problème de minimisation sous contrainte :

$$\text{Trouver le minimum de } \|\mathbf{r} - \mathbf{b}\|_2^2 \text{ sous la contrainte } \mathbf{A}^\top \mathbf{r} = \mathbf{0}$$

Un autre système symétrique est obtenu par :

$$\begin{bmatrix} \mathbf{0} & \mathbf{A} \\ \mathbf{A}^\top & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{Ax} \\ \mathbf{x} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ \mathbf{A}^\top \mathbf{b} \end{bmatrix}$$

On remarque néanmoins que le système n'est en général pas plus facile à résoudre que le système initial, et que des méthodes comme le gradient conjugué présentent les mêmes inconvénients vis à vis du conditionnement.

5.3 Préconditionnement

L'idée du préconditionnement est simple : il s'agit de trouver une matrice $\mathbf{C}$ telle que $\mathbf{CA}$ soit mieux conditionnée que $\mathbf{A}$, et telle que le produit $\mathbf{CA}$ soit peu coûteux à calculer. L'idéal serait de choisir $\mathbf{C} = \mathbf{A}^{-1}$ mais c'est évidemment trop coûteux, on choisit donc d'approcher l'inverse de $\mathbf{A}$. Dans le cas où on veut utiliser le gradient conjugué ou ses dérivés, il est essentiel de conserver les propriétés de symétrie et de définie positivité de $\mathbf{A}$ pour le produit $\mathbf{CA}$, afin de conserver la convergence.

5.3.1 Décomposition de $\mathbf{A}$

Les préconditionnements les plus simples sont basés sur les méthodes linéaires, donc sur une décomposition de $\mathbf{A}$.

Le préconditionnement diagonal, dit aussi de Jacobi, consiste à choisir $\mathbf{C} = \mathbf{D}^{-1}$ où $\mathbf{D}$ est la matrice dont la diagonale est égale à celle de $\mathbf{A}$. Malgré sa simplicité, c'est un préconditionneur efficace, et très peu coûteux. On ne peut donc pas s'en passer dans tous les cas où il peut être d'une quelconque utilité.

Le précontionnement SSOR consiste à choisir $\mathbf{C} = (\mathbf{D} + \omega\mathbf{L})\mathbf{D}^{-1}(\mathbf{D} + \omega\mathbf{U})$ avec $\mathbf{A} = \mathbf{D} + \mathbf{L} + \mathbf{U}$, où $\mathbf{D}$ est diagonale, $\mathbf{L}$ triangulaire inférieure, $\mathbf{U}$ triangulaire supérieure. En général on choisit $\omega = 1$.

Ces deux préconditionneurs sont symétriques définis positifs, dès que $\mathbf{A}$ l'est.

5.3.2 Préconditionneur polynômial

Nous allons reprendre la méthode du préconditionnement présenté à la section 2.10 du chapitre 2. On suppose que la résolution de $\mathbf{Ax} = \mathbf{b}$ n'est possible que de façon approchée. En conséquence on recherche une solution $\widetilde{\mathbf{x}}$ qui diffère de $\mathbf{x}$ d'une erreur $\mathbf{e}$:

$$\mathbf{x} = \widetilde{\mathbf{x}} + \mathbf{e}$$

De $\mathbf{Ax} = \mathbf{b}$ on tire

$$-\mathbf{Ae} = \mathbf{b} - \mathbf{A}\widetilde{x} = \mathbf{r}$$

Si on recherche $\mathbf{z}$ solution de

$$\mathbf{Az} = \mathbf{r}$$

on dispose d'une connaissance de l'erreur $\mathbf{e}$, mais ceci revient à résoudre le système initial. Pour contourner cette difficulté, on approche $\mathbf{A}$ par $\mathbf{M}$, et on résout à la place

$$\mathbf{Mz} = \mathbf{r} \tag{5.3.1}$$

L'une des solutions est d'appliquer p étapes d'une méthode stationnaire sous la forme

$$\begin{aligned} \mathbf{M}_1\mathbf{z}^{(k+1)} &= \mathbf{N}_1\mathbf{z}^{(k)} + \mathbf{r}, \\ z^{(0)} &= 0 \end{aligned}$$

Si on pose $\mathbf{G} = \mathbf{M}_1^{-1}\mathbf{N}_1$, alors :

$$\mathbf{z} = \mathbf{z}^{(p)} = (\mathbf{I} + \mathbf{G} + \mathbf{G}^2 + ... + \mathbf{G}^{p-1})\mathbf{M}_1^{-1}\mathbf{r}$$

On peut alors poser, par analogie avec l'équation (5.3.1) :

$$\mathbf{M}^{-1} = (\mathbf{I} + \mathbf{G} + \mathbf{G}^2 + ... + \mathbf{G}^{p-1})\mathbf{M}_1^{-1}$$

Bien entendu, la fait que $\mathbf{M}$ soit symétrique, définie positive contraint le choix de $\mathbf{M}_1, \mathbf{N}_1$ et p. Le fait que le préconditionneur $\mathbf{M}$ soit un polynôme de la matrice $\mathbf{G}$, donne son nom à la méthode.

Nous pouvons rapprocher cette méthode avec les méthodes itératives de résolution de systèmes linéaires. En effet nous avons considéré, au chapitre 3, la matrice $\mathbf{A}$ comme étant la somme de trois matrices $\mathbf{M} = \mathbf{L} + \mathbf{D} + \mathbf{U}$, où $\mathbf{L}$ matrice triangulaire inférieure, $\mathbf{D}$ matrice diagonale et $\mathbf{U}$ matrice triangulaire supérieure. Dans ce cas, nous pouvons obtenir la décomposition

$$\mathbf{A} = \mathbf{M} - \mathbf{N}$$

et, selon la méthode utilisée, nous avons

- soit $\mathbf{M} = \mathbf{D}$ et $\mathbf{N} = -(\mathbf{L} + \mathbf{U})$, pour la méthode Jacobi. Le preconditionneur $\mathbf{M}^{-1} = \mathbf{D}^{-1}$ il est facile à calculer, mais il n'est pas très puissant.
- soit $\mathbf{M} = (\mathbf{D} + \mathbf{L})$ et $\mathbf{N} = -\mathbf{U}$, pour la méthode de Gauss-Seidel.

En posant $\mathbf{G} = \mathbf{M}^{-1}\mathbf{N}$, l'expansion polynômiale donne

$$\mathbf{A}_0^{-1} = \left(\mathbf{I} + \mathbf{G} + \mathbf{G}^2 + \cdots\right)\mathbf{M}^{-1} \tag{5.3.2}$$

Pour pouvoir faire un calcul fini, il faut que $\mid \rho(\mathbf{G}) \mid < 1$. Dans ce cas, on a

$$\mathbf{A}_0^{-1} \approx \left(\mathbf{I} + \mathbf{G} + \mathbf{G}^2 + \cdots + \mathbf{G}^p\right) \mathbf{M}^{-1}$$

Ce type de préconditionneur est intéressant du point de vue de la vectorisation ou parallélisation du code. Il a donc rencontré un vif succès dès l'apparition des machines correspondantes.

5.3.3 Factorisation incomplète

Les méthodes itératives sont souvent appliquées à des matrices creuses, dont une grande partie des coefficients sont nuls. Comme on l'a vu quand la matrices n'ont pas une forme bande, l'inconvénient des méthodes directes (et donc des préconditionneurs basées dessus) est de remplir les places occupées par des zéros, et donc d'augmenter le coût de stockage induit par une factorisation. L'idée des factorisations incomplètes, est de limiter le remplissage, et de ne pas nécessiter trop d'opérations, donc de ne pas engendrer trop d'erreurs numériques. On obtient alors

$$\mathbf{A} = \mathbf{LU} + \mathbf{R}$$

et on choisit pour préconditionneur

$$\mathbf{C} = \mathbf{U}^{-1}\mathbf{L}^{-1}$$

On parle dans ce cas de méthode ILU(k) : ILU pour Incomplete LU factorization, et le paramètre k indique combien de zéros de la matrice initiale ont été remplacés par des coefficients non nuls dans le préconditionneur.

Dans le cas général, on utilise la notion de profil P tel que :

$$P \subset \{(i,j)/i \neq j, 1 \leq i,j \leq n\}$$

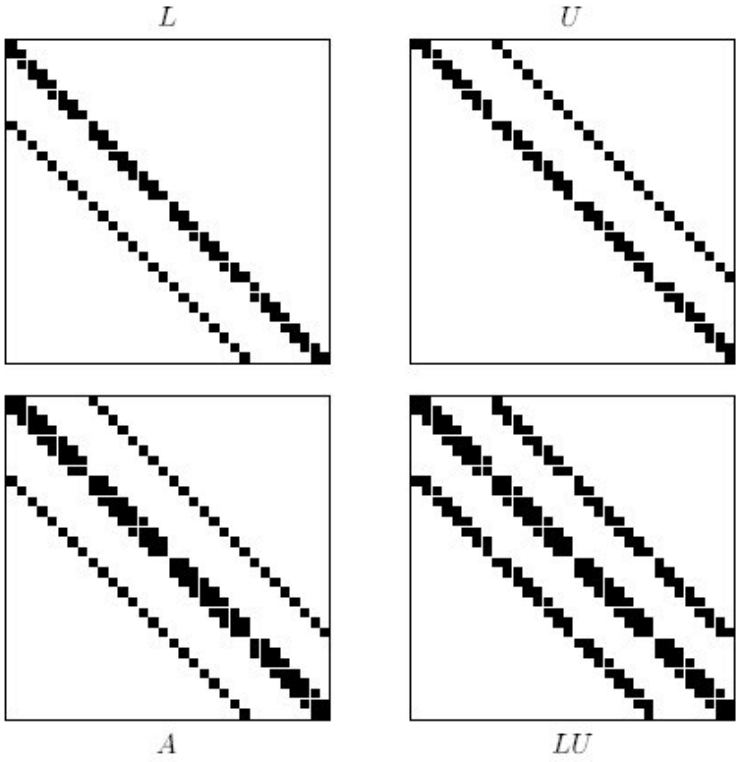
qui représente l'ensemble des couples (i,j) tels que pour une matrice quelconque $\mathbf{M}$: $M_{ij} = 0$. On adapte alors l'algorithme de la décomposition LU pour ne calculer que les éléments hors du profil :

[bl](0,)A▼ [l](-)B◀ [angleA=-90,angleB=180,linearc= 0.5, nodesep=-2pt, linecolor= blue]AB

```
Pour k = 1,...n-1 Faire
    Pour i = k+1,...n et si (i,k) ∉ P Faire
        a_ik = a_ik / a_kk
        Pour j = k+1,...n et si (i,j) ∉ P Faire
            a_ij = a_ij - a_ik * a_kj
        FinPour
    FinPour
FinPour
```

En pratique, il serait assez coûteux et maladroit de tester pour chaque couple (i,j) ou (i,k) s'il appartient au profil. On procède donc différemment en pratique.

L'une des solutions pour choisir le profil P de la factorisation incomplète, est de prendre précisément celui de la matrice A. Pour une matrice pentadiagonale, on aura donc la configuration suivante :



C'est dans des situations de ce type que la factorisation incomplète de Cholesky a été établie pour la première fois.

5.3.4 Inverse approché

On recherche ici une matrice C qui minimise $\|I - CA\|$ ou $\|I - AC\|$ pou rune norme à préciser. Ce préconditionnement peut s'avérer très efficace, mais couteux à calculer. En outre les conditions d'existence de C sont mal définie dans le cas où A n'est pas symétrique.

5.3.5 Multigrilles et multiniveaux

Les méthodes multigrilles et multiniveaux sont des méthodes itératives qui peuvent être utilisées en soi. Mais comme pou les méthodes linéaires de type Gauss-Seidel, il est possible de définir un préconditionnement à partir de ces méthodes. Leur intérêt est de réduire notablement le conditionnement de A lorsqu'elle provient de la résolution par discrétisation d'un problème d'équations aux dérivées partielles.

5.4 Gradient conjugué préconditionné

On a vu que dans de nombreux cas, la méthode itérative sera préférée à une méthode directe, et dans ce cas, c'est très souvent le Gradient Conjugué qui sera choisi. Dans les méthodes de préconditionnement vues au paragraphe (5.3) on ne présuppose pas quelle méthode est choisie, on

peut donc appliquer n'importe quelle méthode de préconditionnement associée à une méthode de résolution. Dans les faits, et toujours dans le but d'améliorer l'algorithme utilisé, on a intégré le préconditionnement à l'algorithme de gradient conjugué. Il en découle de nombreuses variantes, que nous ne détaillerons pas toutes ici.

Considérons un système linéaire $\mathbf{Ax} = \mathbf{b}$, de matrice symétrique définie positive. L'idée du gradient conjugué préconditionné est d'appliquer le gradient conjugué à un système transformé :

$$\widetilde{A}\widetilde{x}=\widetilde{b}$$

où $\mathbf{C}$ est une matrice symétrique définie positive telle que

$$\begin{aligned}\widetilde{A} &= \mathbf{C}^{-1}\mathbf{A}\mathbf{C}^{-1}\\ \widetilde{x} &= \mathbf{C}\mathbf{x}\\ \widetilde{b} &= \mathbf{C}^{-1}\mathbf{b}\end{aligned}$$

On doit bien sûr choisir $\mathbf{C}$ de telle façon que $\widetilde{\mathbf{A}}$ ait un meilleur conditionnement que $\mathbf{A}$. Pour des raisons qui s'éclairciront dans la suite, on doit également prévoir que $\mathbf{C}^2$ soit simple à évaluer.

On peut mettre l'algorithme du gradient conjugué sous la forme suivante :

[bl](0,)A▼ [l](-)B◀ [angleA=-90,angleB=180,linearc= 0.5, nodesep=-2pt, linecolor= blue]AB

Soit $x^{(0)}$ donné dans $\mathbb{R}^n$
$k = 0$
Calculer $\mathbf{r}^{(0)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(0)}$
Tant que $\mathbf{r}^{(k)} \neq \mathbf{0}$ Faire
  $k = k + 1$
  Si $k = 1$ alors
    $\mathbf{p}^{(1)} = \mathbf{r}^{(0)}$
  sinon
    $\beta_k = -\left(\mathbf{p}^{(k-1)}\right)^\top \mathbf{A}\mathbf{r}^{(k-1)} / \left(\mathbf{p}^{(k-1)}\right)^\top \mathbf{A}\left(\mathbf{p}^{(k-1)}\right)$
    $\mathbf{p}^{(k)} = \mathbf{r}^{(k-1)} + \beta_k \mathbf{p}^{(k-1)}$
  FinSi
  $\alpha_k = \left(\mathbf{p}^{(k)}\right)^\top \mathbf{r}^{(k-1)} / \left(\mathbf{p}^{(k)}\right)^\top \mathbf{A}\left(\mathbf{p}^{(k)}\right)$
  $\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + \alpha_k \mathbf{p}^{(k)}$
  $\mathbf{r}^{(k)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(k)}$
FinTantQue
$\mathbf{x} = \mathbf{x}^{(k)}$

En remarquant que l'on peut calculer les résidus récursivement, on peut remplacé $\mathbf{r}^{(k)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(k)}$ par $\mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - \alpha_k \mathbf{A}\mathbf{p}^{(k)}$. L'expression de β_k se transforme alors en $\beta_k = \left(\mathbf{r}^{(k)-1}\right)^\top \mathbf{r}^{(k-1)} \Big/ \left(\left(\mathbf{r}^{(k-2)}\right)^\top \left(\mathbf{r}^{(k-2)}\right)\right)$ et α_k peut être remplacé par $\vdash \alpha_k = \left(\mathbf{r}^{(k-1)}\right)^\top \mathbf{r}^{k-1} \Big/ \left(\left(\mathbf{p}^{(k)}\right)^\top \mathbf{A}\mathbf{p}^{(k)}\right)$. Si on applique l'algorithme ainsi transformé au système $\widetilde{A}\widetilde{x}=\widetilde{b}$ on obtient :

[bl](0,)A▼ [l](-)B◀ [angleA=-90,angleB=180,linearc= 0.5, nodesep=-2pt, linecolor= blue]AB

Soit $\widetilde{\mathbf{x}}^{(0)}$ donné dans $\mathbb{R}^n$
$k = 0$
Calculer $\widetilde{r}^{(0)} = \widetilde{b} - \widetilde{A}\widetilde{x}^{(0)}$
Tant que $\widetilde{r}^{(k)} \neq \mathbf{0}$ Faire
  $k = k + 1$
  Si $k = 1$ alors
    $\widetilde{p}^{(1)} = \widetilde{r}^{(0)}$
  sinon
    $\beta_k = -\left(\mathbf{r}^{(k-1)}\right)^{\top} \widetilde{\mathbf{r}}^{(k-1)} \Big/ \left(\left(\widetilde{\mathbf{r}}^{(k-2)}\right)^{\top} \widetilde{\mathbf{r}}^{(k-2)}\right)$
    $\widetilde{\mathbf{p}}^{(k)} = \widetilde{\mathbf{r}}^{(k-1)} + \beta_k^{(k-1)} \widetilde{\mathbf{p}}^{(k-1)}$
  FinSi
  $\alpha_k = \left(\widetilde{\mathbf{r}}^{(k-1)}\right)^{T} \widetilde{\mathbf{r}}^{(k-1)} \Big/ \left(\left(\widetilde{\mathbf{p}}^{(k)}\right)^{\top} \mathbf{C}^{-1}\mathbf{A}\mathbf{C}^{-1} \widetilde{\mathbf{p}}^{(k)}\right)$
  $\widetilde{\mathbf{x}}^{(k)} = \widetilde{\mathbf{x}}^{(k-1)} + \alpha_k \widetilde{\mathbf{p}}^{(k)}$
  $\widetilde{\mathbf{r}}^{(k)} = \mathbf{r}^{(k-1)} - \alpha_k \mathbf{C}^{-1}\mathbf{A}\mathbf{C}^{-1} \widetilde{\mathbf{p}}^{(k)}$
FinTantQue
$\widetilde{\mathbf{x}} = \widetilde{\mathbf{x}}^{(k)}$

On récupère en sortie un $\widetilde{\mathbf{x}}^{(k)}$ qui est une approximation de $\widetilde{\mathbf{x}}$, lequel permet de calculer $\mathbf{x}$ par $\mathbf{x} = \mathbf{C}^{-1}\widetilde{\mathbf{x}}$. Néanmoins pour éviter de calculer explicitement $\mathbf{C}^{-1}$, on peut définir des intermédiaires de calcul incluant $\mathbf{C}^{-1}$:

$$\begin{aligned} \widetilde{\mathbf{p}}^{(k)} &= \mathbf{C}\mathbf{p}^{(k)} \\ \widetilde{\boldsymbol{x}}^{(k)} &= \mathbf{C}\mathbf{x}^{(k)} \\ \widetilde{\mathbf{r}}^{(k)} &= \mathbf{C}\mathbf{r}^{(k)} \end{aligned}$$

En choisissant alors comme préconditionneur $\mathbf{M} = \mathbf{C}^2$ (qui reste définie positive) et en notant $\mathbf{z}^{(k)}$ la solution de $\mathbf{M}\mathbf{z}^{(k)} = \mathbf{r}^{(k)}$ alors on aboutit à l'**algorithme du gradient conjugué préconditionné** :

[bl](0,)A▼ [l](-)B◀ [angleA=-90,angleB=180,linearc= 0.5, nodesep=-2pt, linecolor= blue]AB

Soit $\mathbf{x}^{(0)}$ donné dans $\mathbb{R}^n$
$k = 0$
Calculer $\mathbf{r}^{(0)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(0)}$
Tant que $\mathbf{r}^{(k)} \neq \mathbf{0}$ Faire
  Résoudre $\mathbf{M}\mathbf{z}^{(k)} = \mathbf{r}^{(k)}$
  $k = k + 1$
  Si $k = 1$ alors
    $\mathbf{p}^{(1)} = \mathbf{z}^{(0)}$
  sinon
    $\beta_k = -\left(\mathbf{r}^{(k-1)}\right)^{\top} \mathbf{z}^{(k-1)} \Big/ \left(\left(\mathbf{r}^{(k-2)}\right)^{\top} \mathbf{z}^{(k-2)}\right)$
    $\mathbf{p}^{(k)} = \mathbf{z}^{(k-1)} + \beta_k \mathbf{p}^{(k-1)}$
  FinSi
  $\alpha_k = \left(\mathbf{r}^{(k-1)}\right)^{\top} \mathbf{z}^{(k-1)} \Big/ \left(\left(\mathbf{p}^{(k)}\right)^{T} \mathbf{A}\mathbf{p}^{(k)}\right)$
  $\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + \alpha_k \mathbf{p}^{(k)}$
  $\mathbf{r}^{(k)} = \mathbf{r}^{(k-1)} - \alpha_k \mathbf{A}\mathbf{p}^{(k)}$
FinTantQue
$\mathbf{x} = \mathbf{x}^{(k)}$

Le choix de la matrice de préconditionnement peut être fait selon les principes énoncés au paragraphe (5.3)

5.5 Raffinement itératif

Dans le cas où les solutions obtenues par résolution du système, préconditionné ou pas, sont entachées d'erreurs numériques, il est possible de recourir à un post-traitement du vecteur solution. On peut supposer, suivant le principe de l'analyse régressive, que la perturbation due aux erreurs d'arrondi, ne porte que sur le second membre on a par exemple réalisé en machine une factorisation :

$$\mathbf{L}_c\mathbf{U}_c = \mathbf{A}+\mathbf{E}$$

et on a résolu exactement le système

$$(\mathbf{A}+\mathbf{E})\mathbf{x}_c = \mathbf{b}$$

On vérifie généralement dans ce cas que le résidu calculé $\mathbf{r}_c = \mathbf{b}-\mathbf{A}\mathbf{x}_c$ n'est pas nul.

Posons $\mathbf{x}^{(0)} = \mathbf{x}_c$ et $\mathbf{r}^{(0)} = \mathbf{r}_c$. Ces valeurs initialisent un processus récursif. On calcul alors pour tout $k \geq 0$:

$$\begin{aligned}(\mathbf{A}+\mathbf{E})\,\mathbf{e}^{(k+1)} &= \mathbf{r}^{(k)}\\ \mathbf{x}^{(k+1)} &= \left(\mathbf{x}^{(k)}+\mathbf{e}^{(k+1)}\right)_c\\ \mathbf{r}^{(k+1)} &= \left(\mathbf{b}-\mathbf{A}\mathbf{x}^{(k+1)}\right)_c\end{aligned}$$

En supposant que l'erreur d'arrondi porte essentiellement sur la résolution des systèmes linéaires, on pose :

$$\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)} = (\mathbf{A}+\mathbf{E})^{-1}\,\mathbf{r}^{(k)}$$

$$\mathbf{e}^{(k+1)} = (\mathbf{A}+\mathbf{E})^{-1}\,\mathbf{A}\left(\mathbf{x}-\mathbf{x}^{(k)}\right)$$

d'où on tire

$$\mathbf{x}-\mathbf{x}^{(k+1)} = \left(\mathbf{I}-(\mathbf{A}+\mathbf{E})^{-1}\,\mathbf{A}\right)\left(\mathbf{x}-\mathbf{x}^{(k)}\right)$$

Posons $\mathbf{G} = \mathbf{I}-(\mathbf{A}+\mathbf{E})^{-1}\,\mathbf{A} = \mathbf{I}-\left(\mathbf{I}+\mathbf{A}^{-1}\mathbf{E}\right)^{-1}$, on a donc obtenu pour tout $k \geq 1$:

$$\mathbf{x}-\mathbf{x}^{(k+1)} = \mathbf{G}\left(\mathbf{x}-\mathbf{x}^{(k)}\right) = \mathbf{G}^{k+1}\left(\mathbf{x}-\mathbf{x}^{(0)}\right)$$

Supposons alors que pour une norme induite, on ait $\left\|\mathbf{A}^{-1}\mathbf{E}\right\| = \alpha < 1$ alors on aura

$$\mathbf{G} = \sum_{k=1}^{\infty}(-\mathbf{A}^{-1}\mathbf{E})^k$$

de sorte que $\|\mathbf{G}\| \leq \frac{\alpha}{1-\alpha}$. On obtient alors

$$\left\|\mathbf{x}-\mathbf{x}^{(k)}\right\| \leq \left(\frac{\alpha}{1-\alpha}\right)^{k} \left\|\mathbf{x}-\mathbf{x}^{(0)}\right\|$$

L'expérience prouve qu'en deux itérations ce processus peut être utile, mais pas plus.

EXERCICE 5.3 *Déterminant et conditionnement*

(1) Calculez en fonction de n le déterminant et le conditionnement de la matrice carrée $\mathbf{A}$ d'ordre n définie par

$$\mathbf{A} = \begin{bmatrix} 1 & & & \\ & 10 & & \\ & & \ddots & \\ & & & 10 \end{bmatrix}$$

(2) De même, calculez en fonction de n le déterminant et le conditionnement de la matrice carrée $\mathbf{B}$ d'ordre n définie par :

$$\begin{cases} B_{ii} = 1, 1 \leq i \leq n \\ B_{i,i+1} = 2, 1 \leq i \leq n-1 \\ B_{ij} = 0 \text{ sinon} \end{cases}$$

(3) Concluez quant au lien qui peut exister entre conditionnement et déterminant. Expliquez votre point de vue dans le cas général.

EXERCICE 5.4 *Notion de préconditionnement*
Soient

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 10^{-6} \end{bmatrix} \text{ et } \Delta\mathbf{A} = \begin{bmatrix} 10^{-8} & 0 \\ 0 & 10^{-14} \end{bmatrix}$$

(1) Calculez le conditionnement de $\mathbf{A}$ en norme 2.

(2) Que vaut

$$\frac{\|\Delta\mathbf{A}\|_2}{\|\mathbf{A}\|_2}$$

Déduisez-en une estimation de $\frac{\|\Delta\mathbf{x}\|_2}{\|\mathbf{x}\|_2}$, où $\mathbf{x}$ et $\mathbf{x}+\Delta\mathbf{x}$ sont solutions des systèmes $\mathbf{A}\mathbf{x} = \mathbf{b}$ et $(\mathbf{A}+\Delta\mathbf{A})(\mathbf{x}+\Delta\mathbf{x}) = \mathbf{b}$ avec $\mathbf{b}$ un vecteur quelconque.

(3) Que vaut le conditionnement de $2\mathbf{A}$, $-4\mathbf{A}$, puis de $\alpha\mathbf{A}$ avec α un réel quelconque ?

(4) Soit

$$\mathbf{D} = \begin{bmatrix} 1 & 0 \\ 0 & 10^{6} \end{bmatrix}$$

Exprimez $\mathbf{D}\cdot\mathbf{A}$ et $\mathbf{D}\cdot\Delta\mathbf{A}$.

(5) Que vaut le conditionnement de $\mathbf{D}\cdot\mathbf{A}$ *en norme 2 ? Que vaut*

$$\frac{\|\mathbf{D}\cdot\Delta\mathbf{A}\|_2}{\|\mathbf{D}\cdot\mathbf{A}\|_2}?$$

(6) Déduisez-en une nouvelle estimation de $\frac{\|\Delta\mathbf{x}\|_2}{\|\mathbf{x}\|_2}$

(7) Quelle conclusion en tirez-vous ?

5.6 Préconditionnement et erreur de calcul

Lors de l'application d'un algorithme itératif à la solution du système $\mathbf{Ax}=\mathbf{b}$, la seule information disponible à chaque itération est le résidu $\mathbf{r}^{(k)}=\mathbf{b}-\mathbf{Ax}^{(k)}$. Par contre l'erreur $\mathbf{e}^{(k)}=\mathbf{x}-\mathbf{x}^{(k)}$ est inconnue. Du fait que $\mathbf{r}^{(k)}=\mathbf{Ae}^{(k)}$, on a la relation

$$\frac{1}{\kappa(\mathbf{A})}\cdot\frac{\left\|\mathbf{r}^{(k)}\right\|}{\left\|\mathbf{r}^{(0)}\right\|}\leq\frac{\left\|\mathbf{e}^{(k)}\right\|}{\left\|\mathbf{e}^{(0)}\right\|}\leq\kappa(\mathbf{A})\frac{\left\|\mathbf{r}^{(k)}\right\|}{\left\|\mathbf{r}^{(0)}\right\|}\leq\kappa^2(\mathbf{A})\frac{\left\|\mathbf{e}^{(k)}\right\|}{\left\|\mathbf{e}^{(0)}\right\|} \tag{5.6.1}$$

De cette relation on en conclut que si $\kappa(\mathbf{A})$ est proche de 1, le quotient des normes des résidus est relié au quotient des normes des erreurs et donc si le résidu décroît, l'erreur aussi décroît. Par contre si le coonditionnement est grand, le comportement de deux quotients peut être différent, l'un augmentant et l'autre diminuant. Ces remarques montrent l'importance d'avoir un conditionnement proche de 1 et, par voie de conséquence, l'importance du préconditionnement.

Nous pouvons aussi utiliser les relations (5.6.1) pour élaborer un critère d'arrêt d'une méthode itérative. Soit $\mathbf{x}+\boldsymbol{\Delta}\mathbf{x}$ la solution calculée. Elle est solution du système perturbé $(\mathbf{A}+\boldsymbol{\Delta}\mathbf{A})(\mathbf{x}+\boldsymbol{\Delta}\mathbf{x})=\mathbf{b}$ (on suppose que le second membre est sans erreur). On calcule le résidu $\mathbf{r}=\mathbf{b}-\mathbf{A}(\mathbf{x}+\boldsymbol{\Delta}\mathbf{x})$. Nous avons la relation

$$\frac{\|\mathbf{r}\|}{\|\mathbf{A}\|\cdot\|\mathbf{x}+\boldsymbol{\Delta}\mathbf{x}\|}\leq\frac{\|\mathbf{A}+\boldsymbol{\Delta}\mathbf{A}\|}{\|\mathbf{A}\|} \tag{5.6.2}$$

Par conséquent si la norme du résidu relatif $\frac{\|\mathbf{r}\|}{\|\mathbf{A}\|\cdot\|\mathbf{x}+\boldsymbol{\Delta}\mathbf{x}\|}$ est petite, le résultat est acceptable. Sinon, la solution est relative au système initial qui a subi une grande perturbation. Remarquons que si $\frac{\|\mathbf{r}\|}{\|\mathbf{A}\|\cdot\|\mathbf{x}+\boldsymbol{\Delta}\mathbf{x}\|}$ est proche de la précision de l'ordinateur, alors la solution a une bonne précision.

5.6.1 Exercices

EXERCICE 5.5 *Démontrer la relation* (5.6.1).

EXERCICE 5.6 *Démontrer la relation* (5.6.2)

5.7 Bibliographie

Les ouvrages ci-dessous sont disponibles sous forme de fichier téléchargeable sur le site du cours ou sur Arel

[CB] CLAUDE BREZINSKI *:* **Algèbre matricielle numérique**,

[YA] YVES ACHDOU *:* **Algèbre linéaire et analyse numérique matricielle**,, téléchargeable à l'adresse http ://www.ann.jussieu.fr/~achdou/files/teaching/linalg/book.pdf

[AH1] ALAIN HUARD *:* **Analyse numérique matricielle, Cours de 3ème année**, téléchargeable à partir du site de l'INSA Toulouse www-gmm.insa-toulouse.fr

[AH2] ALAIN HUARD *:* **Analyse numérique des grands problèmes linéaires, Cours de 4ème année**, téléchargeable à partir du site de l'INSA Toulouse www-gmm.insa-toulouse.fr

[YS] YOUSEF SAAD : **Iterative Methods for Sparse Linear Systems**, téléchargeable à l'adresse www.stanford.edu/class/cme324/saad.pdf

Les ouvrages ci-dessous sont disponibles en librairie

[QS] ALFIO QUARTERONI, FAUSTO SALERI : *Calcul scientifique, Cours, exercices corrigés et illustrations en Matlab et Octave*, Springer, 2206.

[AF] ANDRÉ FORTIN : *Analyse numérique pour Ingénieurs*, Presses Internationales Polytechnique, 2001.

[AD] LUCA AMODEI, JEAN-PIERRE DEDIEU : *Analyse numérique matricielle, Cours exercices et corrigés*, Dunod, 2008.

[SB] J. STOER, R. BULIRSCH : *Introduction to numerical Analysis*, Third Edition, Springer, 2000.

[GV] G. GOLUB, C. VAN LOAN : *Matrix Computations*, Third Edition, The Johns Hopkins University Press, 1996.

[DW] DAVID S. WATKINS : *Fundamentals of Matrix computation*, Second Edition, Wiley, 2002

[CBr] CLAUDE BREZINSKI : *Projection methods for systems of equations*, Elsevier, 1997

Table des matières