Université Aix Marseille

Licence de mathématiques

Cours d'Analyse numérique

Raphaèle Herbin

8 octobre 2014

Table des matières

Introduction

L'objet de l'analyse numérique est de concevoir et d'étudier des méthodes de résolution de certains problèmes mathématiques, en général issus de la modélisation de problèmes "réels", et dont on cherche à calculer la solution à l'aide d'un ordinateur.
Le cours est structuré en quatre grands chapitres :

- Systèmes linéaires
- Systèmes non linéaires
- Optimisation
- Equations différentielles.

On pourra consulter les ouvrages suivants pour ces différentes parties (ceci est une liste non exhaustive !) :

- A. Quarteroni, R. Sacco et F. Saleri, Méthodes Numériques : Algorithmes, Analyse et Applications, Springer 2006.
- P.G. Ciarlet, Introduction à l'analyse numérique et à l'optimisation, Masson, 1982, (pour les chapitre 1 à 3 de ce polycopié).
- M. Crouzeix, A.L. Mignot, Analyse numérique des équations différentielles, Collection mathématiques appliquées pour la maitrise, Masson, (pour le chapitre 4 de ce polycopié).
- J.P. Demailly, Analyse numérique et équations différentielles Collection Grenoble sciences Presses Universitaires de Grenoble
- L. Dumas, Modélisation à l'oral de l'agrégation, calcul scientifique, Collection CAPES/Agrégation, Ellipses, 1999.
- E. Hairer, polycopié du cours "Analyse Numérique", http ://www.unige.ch/ hairer/polycop.html
- J. Hubbard, B. West, Equations différentielles et systèmes dynamiques, Cassini.
- J. Hubbard et F. Hubert, Calcul Scientifique, Vuibert.
- P. Lascaux et R. Théodor, Analyse numérique matricielle appliquée à l'art de l'ingénieur, tomes 1 et 2, Masson, 1987
- L. Sainsaulieu, Calcul scientifique cours et exercices corrigés pour le 2ème cycle et les éécoles d'ingénieurs, Enseignement des mathématiques, Masson, 1996.
- M. Schatzman, Analyse numérique, cours et exercices, (chapitres 1,2 et 4).
- D. Serre, Les matrices, Masson, (2000). (chapitres 1,2 et 4).
- P. Lascaux et R. Theodor, Analyse numérique sappliquée aux sciences de l'ingénieur, Paris, (1994)
- R. Temam, Analyse numérique, Collection SUP le mathématicien, Presses Universitaires de France, 1970.

Et pour les anglophiles...

- M. Braun, Differential Equations and their applications, Springer, New York, 1984 (chapitre 4).
- G. Dahlquist and A. Björck, Numerical Methods, Prentice Hall, Series in Automatic Computation, 1974, Englewood Cliffs, NJ.
- R. Fletcher, Practical methods of optimization, J. Wiley, New York, 1980 (chapitre 3).
- G. Golub and C. Van Loan, Matrix computations, The John Hopkins University Press, Baltimore (chapitre 1).
- R.S. Varga, Matrix iterative analysis, Prentice Hall, Englewood Cliffs, NJ 1962.

Pour des rappels d'algègre linéaire :

- Poly d'algèbre linéaire de première année, P. Bousquet, R. Herbin et F. Hubert, http ://www.cmi.univ-mrs.fr/ herbin/PUBLI/L1alg.pdf

— Introduction to linear algebra, Gilbert Strang, Wellesley Cambridge Press, 2008

Ce cours a été rédigé pour la licence de mathématiques à distance (téléenseignement) du CTES de l'université d'Aix-Marseille. Chaque chapitre est suivi d'un certain nombre d'exercices. On donne ensuite des suggestions pour effectuer les exercices, puis des corrigés détaillés. Il est fortement conseillé d'essayer de faire les exercices d'abord sans ces indications, et de ne regarder les corrigés détaillés qu'une fois l'exercice achevé (même si certaines questions n'ont pas pu être effectuées), ceci pour se préparer aux conditions d'examen.

Chapitre 1

Systèmes linéaires

1.1 Objectifs

On note $\mathcal{M}_n(\mathbb{R})$ l'ensemble des matrices carrées d'ordre n. Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible, et $b \in \mathbb{R}^n$, on a comme objectif de résoudre le système linéaire $A\boldsymbol{x} = b$, c'est–à– dire de trouver $\boldsymbol{x}$ solution de :

$$\left\{ \begin{array}{l} \boldsymbol{x} \in \mathbb{R}^n \\ A\boldsymbol{x} = b \end{array} \right. \tag{1.1}$$

Comme A est inversible, il existe un unique vecteur $\boldsymbol{x} \in \mathbb{R}^n$ solution de (1.1). Nous allons étudier dans les deux chapitres suivants des méthodes de calcul de ce vecteur $\boldsymbol{x}$: la première partie de ce chapitre sera consacrée aux méthodes “directes" et la deuxième aux méthodes “itératives". Nous aborderons ensuite en troisième partie les méthodes de résolution de problèmes aux valeurs propres.
Un des points essentiels dans l'efficacité des méthodes envisagées concerne la taille des systèmes à résoudre. La taille de la mémoire des ordinateurs a augmenté de façon drastique de 1980 à nos jours.
Le développement des méthodes de résolution de systèmes linéaires est liée à l'évolution des machines informatiques. C'est un domaine de recherche très actif que de concevoir des méthodes qui permettent de profiter au mieux de l'architecture des machines (méthodes de décomposition en sous domaines pour profiter des architectures parallèles, par exemple).
Dans la suite de ce chapitre, nous verrons deux types de méthodes pour résoudre les systèmes linéaires : les méthodes directes et les méthodes itératives. Pour faciliter la compréhension de leur étude, nous commençons par quelques rappels d'algèbre linéaire.

1.2 Pourquoi et comment ?

Nous donnons dans ce paragraphe un exemple de problème dont la résolution numérique recquiert la résolution d'un système linéaire, et qui nous permet d'introduire des matrices que nous allons beaucoup étudier par la suite. Nous commençons par donner ci-après après quelques rappels succincts d'algèbre linéaire, outil fondamental pour la résolution de ces systèmes linéaires.

1.2.1 Quelques rappels d'algèbre linéaire

Quelques notions de base

Ce paragraphe rappelle des notions fondamentales que vous devriez connaître à l'issue du cours d'algèbre linéaire de première année. On va commencer par revisiter le **produit matriciel**, dont la vision combinaison linéaire de lignes est fondamentale pour bien comprendre la forme matricielle de la procédure d'élimination de Gauss.

Soient A et B deux matrices carrées d'ordre n, et $M = AB$. Prenons comme exemple d'illustration

$$A = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}, B = \begin{bmatrix} -1 & 0 \\ 3 & 2 \end{bmatrix} \text{ et } M = \begin{bmatrix} 5 & 4 \\ 3 & 2 \end{bmatrix}$$

On note $a_{i,j}$ $b_{i,j}$ et $m_{i,j}$, $i, j = 1, \dots n$ les coefficients respectifs de A, B et M. Vous savez bien sûr que

$$m_{i,j} = \sum_{k=1}^{n} a_{i,k} b_{k,j}. \tag{1.2}$$

Si on écrit les matrices A et B sous forme de lignes (notées $\boldsymbol{\ell}_i$) et colonnes (notées $\boldsymbol{c}_j$) :

$$A = \begin{bmatrix} \boldsymbol{\ell}_1(A) \\ \dots \\ \boldsymbol{\ell}_n(A) \end{bmatrix} \text{ et } B = \begin{bmatrix} \boldsymbol{c}_1(B) & \dots & \boldsymbol{\ell}_n(B) \end{bmatrix}$$

Dans nos exemples, on a donc

$$\boldsymbol{\ell}_1(A) = \begin{bmatrix} 1 & 2 \end{bmatrix}, \ \boldsymbol{\ell}_2(A) = \begin{bmatrix} 0 & 1 \end{bmatrix}, \ \boldsymbol{c}_1(B) = \begin{bmatrix} -1 \\ 3 \end{bmatrix} \ \boldsymbol{c}_2(B) = \begin{bmatrix} 0 \\ 2 \end{bmatrix}.$$

L'expression (1.2) s'écrit encore

$$m_{i,j} = \boldsymbol{\ell}_i(A) \boldsymbol{c}_j(B),$$

qui est le produit d'une matrice $1 \times n$ par une matrice $n \times 1$, qu'on peut aussi écrire sous forme d'un produit scalaire :

$$m_{i,j} = (\boldsymbol{\ell}_i(A))^t \cdot \boldsymbol{c}_j(B)$$

où $(\boldsymbol{\ell}_i(A))^t$ désigne la matrice transposée, qui est donc maintenant une matrice $n \times 1$ qu'on peut identifier à un vecteur de $\mathbb{R}^n$. C'est la technique "habituelle" de calcul du produit de deux matrices. On a dans notre exemple :

$$\begin{aligned} m_{1,2} = \boldsymbol{\ell}_1(A)\, \boldsymbol{c}_2(B) &= \begin{bmatrix} 1 & 2 \end{bmatrix} \begin{bmatrix} 0 \\ 2 \end{bmatrix}. \\ &= (\boldsymbol{\ell}_i(A))^t \cdot \boldsymbol{c}_j(B) = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 2 \end{bmatrix} \\ &= 4. \end{aligned}$$

Mais de l'expression (1.2), on peut aussi avoir l'expression des lignes et des colonnes de $M = AB$ en fonction des lignes de B ou des colonnes de A :

$$\boldsymbol{\ell}_i(AB) = \sum_{k=1}^{n} a_{i,k} \boldsymbol{\ell}_k(B) \tag{1.3}$$

$$\boldsymbol{c}_j(AB) = \sum_{k=1}^{n} b_{k,j} \boldsymbol{c}_k(A) \tag{1.4}$$

Dans notre exemple, on a donc :

$$\boldsymbol{\ell}_1(AB) = \begin{bmatrix} -1 & 0 \end{bmatrix} + 2 \begin{bmatrix} 3 & 2 \end{bmatrix} = \begin{bmatrix} 5 & 4 \end{bmatrix}$$

ce qui montre que la ligne 1 de AB est combinaison linéaire des lignes de B. Le colonnes de AB, par contre, sont des combinaisons linéaires de colonnes de A. Par exemple :

$$\boldsymbol{c}_2(AB) = 0 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + 2 \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \end{bmatrix}$$

Il faut donc retenir que dans un produit matriciel AB,

les colonnes de AB sont des combinaisons linéaires des colonnes de A
les lignes de AB sont des combinaisons linéaires des lignes de B.

Cette remarque est très importante pour la représentation matricielle de l'élimination de Gauss : lorqu'on calcule des systèmes équivalents, on effectue des combinaisons linéaires de lignes, et donc on multiplie à gauche par une matrice d'élimination.

Le tableau ci-dessous est la traduction littérale de "Linear algebra in a nutshell", par Gilbert Strang [1] Pour une matrice carrée A, on donne les caractérisations du fait qu'elle est inversible ou non.

A inversible	A non inversible
Les vecteurs colonne sont indépendants	Les vecteurs colonne sont liés
Les vecteurs ligne sont indépendants	Les vecteurs ligne sont liés
Le déterminant est non nul	Le déterminant est nul
$A\boldsymbol{x} = 0$ a une unique solution $\boldsymbol{x} = \mathbf{0}$	$A\boldsymbol{x} = \mathbf{0}$ a une infinité de solutions.
Le noyau de A est réduit à $\{\mathbf{0}\}$	Le noyau de A contient au moins un vecteur non nul.
$A\boldsymbol{x} = \boldsymbol{b}$ a une solution unique $x = A^{-1}\boldsymbol{b}$	$A\boldsymbol{x} = \boldsymbol{b}$ a soit aucune solution, soit une infinité.
A a n (nonzero) pivots	A a $r < n$ pivots
A est de rang maximal : rg(A) = n.	rg(A) = $r < n$
La forme totatement échelonnée R de A est la matrice identité	R a au moins une ligne de zéros.
L'image de A est tout $\mathbb{R}^n$.	L'image de A est strictement incluse dans $\mathbb{R}^n$.
L'espace $L(A)$ engendré par les lignes de A est tout $\mathbb{R}^n$.	$L(A)$ est de dimension $r < n$
Toutes les valeurs propres de A sont non nulles	Zero est valeur propre de A.
A^tA is symétrique définie positive [2]	A^tA n'est que semi- définie .

TABLE 1.1: Extrait de "Linear algebra in a nutshell", G. Strang

On rappelle pour une bonne lecture de ce tableau les quelques définitions suivantes :

Définition 1.1 (Pivot)**.** *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice carrée d'ordre n. On appelle pivot de A le premier élément non nul de chaque ligne dans la forme échelonnée de A obtenue par élimination de Gauss. Si la matrice est inversible, elle a donc n pivots (non nuls).*

Définition 1.2 (Valeurs propres)**.** *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice carrée d'ordre n. On appelle valeur propre de A tout $\lambda \in \mathbb{C}$ tel qu'il existe $\boldsymbol{x} \in \mathbb{C}^n$, $\boldsymbol{x} \neq 0$ tel que $A\boldsymbol{x} = \lambda\boldsymbol{x}$. L'élément $\boldsymbol{x}$ est appelé vecteur propre de A associé à λ.*

Définition 1.3 (Déterminant)**.** *Il existe une unique application, notée* $\det$ *de $\mathcal{M}_n(\mathbb{R})$ dans $\mathbb{R}$ qui vérifie les propriétés suivantes*

(D1) Le déterminant de la matrice identité est égal à 1.

(D2) Si la matrice $\tilde{A}$ est obtenue à partir de A par échange de deux lignes, alors $\det\tilde{A} = -\det A$.

1. Voir la page web de Strang `www.mit.edu/~gs` pour une foule d'informations et de cours sur l'algèbre linéaire.

(D3) Le déterminant est une fonction linéaire de chacune des lignes de la matrice A.

(D3a) (multiplication par un scalaire) si $\tilde{A}$ est obtenue à partir de A en multipliant tous les coefficients d'une ligne par $\lambda \in \mathbb{R}$, alors $\det(\tilde{A}) = \lambda \det(A)$.

(D3b) (addition) si $A = \begin{bmatrix} \boldsymbol{\ell}_1(A) \\ \vdots \\ \boldsymbol{\ell}_k(A) \\ \vdots \\ \boldsymbol{\ell}_n(A) \end{bmatrix}$, $\tilde{A} = \begin{bmatrix} \boldsymbol{\ell}_1(A) \\ \vdots \\ \tilde{\boldsymbol{\ell}}_k(A) \\ \vdots \\ \boldsymbol{\ell}_n(A) \end{bmatrix}$ *et* $B = \begin{bmatrix} \boldsymbol{\ell}_1(A) \\ \vdots \\ \boldsymbol{\ell}_k(A) + \tilde{\boldsymbol{\ell}}_k(A) \\ \vdots \\ \boldsymbol{\ell}_n(A) \end{bmatrix}$, *alors*

$$\det(B) = \det(A) + \det(\tilde{A}).$$

On peut déduire de ces trois propriétés fondamentales un grand nombre de propriétés importantes, en particulier le fait que $\det(AB) = \det A \, \det B$ *et que le déterminant d'une matrice inversible est le produit des pivots : c'est de cette manière qu'on le calcule sur les ordinateurs. En particulier on n'utilise jamais la formule de Cramer, beaucoup trop coûteuse en termes de nombre d'opérations.*

On rappelle que si $A \in \mathcal{M}_n(\mathbb{R})$ une matrice carrée d'ordre n, les valeurs propres sont les racines du **polynôme caractéristique** P_A de degré n, qui s'écrit :

$$P_A(\lambda)) = \det(A - \lambda I).$$

Matrices diagonalisables

Un point important de l'algèbre linéaire, appelé "réduction des endomorphismes" dans les programmes français, consiste à se demander s'il existe une base de l'espace dans laquelle la matrice de l'application linéaire est diagonale ou tout au moins triangulaire (on dit aussi trigonale).

Définition 1.4 (Matrice diagonalisable dans $\mathbb{R}$)**.** *Soit A une matrice réelle carrée d'ordre n. On dit que A est diagonalisable dans $\mathbb{R}$ s'il existe une base $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$ de $\mathbb{R}^n$ et des réels $\lambda_1, \ldots, \lambda_n$ (pas forcément distincts) tels que $A\boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i$ pour $i = 1, \ldots, n$. Les réels $\lambda_1, \ldots, \lambda_n$ sont les valeurs propres de A, et les vecteurs $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ sont les vecteurs propres associés.*

Vous connaissez sûrement aussi la diagonalisation dans $\mathbb{C}$: une matrice réelle carrée d'ordre n admet toujours n valeurs propres dans $\mathbb{C}$, qui ne sont pas forcément identiques. Une matrice est diagonalisable dans $\mathbb{C}$ s'il existe une base $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$ de $\mathbb{C}^n$ et des nombres complexes $\lambda_1, \ldots, \lambda_n$ (pas forcément distincts) tels que $A\boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i$ pour $i = 1, \ldots, n$. Ceci est vérifié si la dimension de chaque sous espace propre $E_i = \mathrm{Ker}(A - \lambda \mathrm{Id})$ (appelée multiplicité géométrique) est égale a la multiplicité algébrique de λ_i, c.à.d. son ordre de multiplicité en tant que racine du polynôme caractéristique.

Par exemple la matrice $A = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$ n'est pas diagonalisable dans $\mathbb{C}$ (ni évidemment, dans $\mathbb{R}$). Le polynôme caractéristique de A est $P_A(\lambda) = \lambda^2$, l'unique valeur propre est donc 0, qui est de multiplicité algébrique 2, et de multiplicité géométrique 1, car le sous espace propre associé à la valeur propre nulle est $F = \{\boldsymbol{x} \in \mathbb{R}^2 \, ; A\boldsymbol{x} = 0\} = \{\boldsymbol{x} = (0, t), t \in \mathbb{R}\}$, qui est de dimension 1.

Ici et dans toute la suite, comme on résout des systèmes linéaires réels, on préfère travailler avec la diagonalisation dans $\mathbb{R}$; cependant il y a des cas où la diagonalisation dans $\mathbb{C}$ est utile et même nécessaire (étude de stabilité des

systèmes diférentiels, par exemple). Par souci de clarté, nous préciserons toujours si la diagonalisation considérée est dans $\mathbb{R}$ ou dans $\mathbb{C}$.

Lemme 1.5. *Soit A une matrice réelle carrée d'ordre n, diagonalisable dans $\mathbb{R}$. Alors*

$$A = P\mathrm{d}iag(\lambda_1, \ldots, \lambda_n)P^{-1}, \text{‘}$$

où P est la matrice dont les vecteurs colonnes sont égaux aux vecteurs propres $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ associées au valeurs propres $\lambda_1, \ldots, \lambda_n$

DÉMONSTRATION – Par définition d'un vecteur propre, on a $A\boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i$ pour $i = 1, \ldots N$, et donc, en notant P la matrice dont les colonnes sont les vecteurs propres $\boldsymbol{u}_i$,

$$\begin{bmatrix} A\boldsymbol{u}_1 & \ldots & A\boldsymbol{u}_n \end{bmatrix} = A \begin{bmatrix} \boldsymbol{u}_1 & \ldots & \boldsymbol{u}_n \end{bmatrix} = AP$$

et donc

$$AP = \begin{bmatrix} \lambda_1\boldsymbol{u}_1 & \ldots & \lambda_n\boldsymbol{u}_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{u}_1 & \ldots & \boldsymbol{u}_n \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & \lambda_n \end{bmatrix} = P\mathrm{diag}(\lambda_1, \ldots, \lambda_n).$$

Notons que dans ce calcul, on a fortement utilisé la multiplication des matrices par colonnes, c.à.d.

$$\boldsymbol{c}_i(AB) = \sum_{j=1,n} a_{i,j}\boldsymbol{c}_j(B).$$

Remarquons que P lest aussi la matrice définie (de manière unique) par $P\boldsymbol{e}_i = \boldsymbol{u}_i$, où $(e_i)_{i=1,\ldots,n}$ est la base canonique de $\mathbb{R}^n$, c'est-à-dire que $(\boldsymbol{e}_i)_j = \delta_{i,j}$. La matrice P est appelée matrice de passage de la base $(\boldsymbol{e}_i)_{i=1,\ldots,n}$ à la base $(\boldsymbol{u}_i)_{i=1,\ldots,n}$; (il est bien clair que la i-ème colonne de P est constituée des composantes de $\boldsymbol{u}_i$ dans la base canonique $(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n)$.
La matrice P est inversible car les vecteurs propres forment une base, et on peut donc aussi écrire :

$$P^{-1}AP = \mathrm{d}iag(\lambda_1, \ldots, \lambda_n) \text{ ou } A = P\mathrm{d}iag(\lambda_1, \ldots, \lambda_n)P^{-1}.$$

■

La diagonalisation des matrices réelles symétriques est un outil qu'on utilisera souvent dans la suite, en particulier dans les exercices. Il s'agit d'un résultat extrêmement important.

Lemme 1.6 (Une matrice symétrique est diagonalisable dans $\mathbb{R}$)**.** *Soit E un espace vectoriel sur $\mathbb{R}$ de dimension finie : $dimE = n$, $n \in \mathbb{N}^*$, muni d'un produit scalaire i.e. d'une application*

$$\begin{aligned} E \times E &\to \mathbb{R}, \\ (x, y) &\to (x \mid y)_E, \end{aligned}$$

qui vérifie :

$$\begin{aligned} &\forall x \in E, (x \mid x)_E \geq 0 \text{ et } (x \mid x)_E = 0 \Leftrightarrow x = 0, \\ &\forall (x, y) \in E^2, (x \mid y)_E = (y \mid x)_E, \\ &\forall y \in E, \text{ l'application de } E \text{ dans } \mathbb{R}, \text{ définie par } x \to (x \mid y)_E \text{ est linéaire.} \end{aligned}$$

Ce produit scalaire induit une norme sur E, $\|x\| = \sqrt{(x \mid x)_E}$.
Soit T une application linéaire de E dans E. On suppose que T est symétrique, c.à.d. que $(T(x) \mid y)_E = (x \mid T(y))_E, \forall (x, y) \in E^2$. Alors il existe une base orthonormée $(\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n)$ de E (c.à.d. telle que $(\boldsymbol{f}_i, \boldsymbol{f}_j)_E = \delta_{i,j}$) et $(\lambda_1 \ldots \lambda_n) \in \mathbb{R}^n$ tels que $T(f_i) = \lambda_i \boldsymbol{f}_i$ pour tout $i \in \{1 \ldots n\}$.

Conséquence immédiate : Dans le cas où $E = \mathbb{R}^n$, le produit scalaire canonique de $x = (x_1, \dots, x_n)^t$ et $y = (y_1, \dots, y_n)^t$ est défini par $(x \mid y)_E = x \cdot y = \sum_{i=1}^n x_i y_i$. Si $A \in \mathcal{M}_n(\mathbb{R})$ est une matrice symétrique, alors l'application T définie de E dans E par : $T(x) = Ax$ est linéaire, et : $(Tx|y) = Ax \cdot y = x \cdot A^t y = x \cdot Ay = (x \mid Ty)$. Donc T est linéaire symétrique. Par le lemme précédent, il existe $(\boldsymbol{f}_1, \dots, \boldsymbol{f}_n)$ et $(\lambda_1 \dots \lambda_n) \in \mathbb{R}$ tels que $T\boldsymbol{f}_i = A\boldsymbol{f}_i = \lambda_i \boldsymbol{f}_i \, \forall\, i \in \{1, \dots, n\}$ et $f_i \cdot \boldsymbol{f}_j = \delta_{i,j}, \forall\, (i,j) \in \{1, \dots, n\}^2$.

Interprétation algébrique : Il existe une matrice de passage P de $(e_1, \dots, e_n)$ base canonique dans $(\boldsymbol{f}_1, \dots, \boldsymbol{f}_n)$ dont la première colonne de P est constituée des coordonnées de f_i dans $(e_1 \dots e_n)$. On a : $Pe_i = f_i$. On a alors $P^{-1}APe_i = P^{-1}A\boldsymbol{f}_i = P^{-1}(\lambda_i f_i) = \lambda_i e_i = diag(\lambda_1, \dots, \lambda_n)e_i$, où $diag(\lambda_1, \dots, \lambda_n)$ désigne la matrice diagonale de coefficients diagonaux $\lambda_1, \dots, \lambda_n$. On a donc :

$$P^{-1}AP = \begin{bmatrix} \lambda_i & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix} = D.$$

De plus P est orthogonale, *i.e.* $P^{-1} = P^t$. En effet,

$$P^tPe_i \cdot e_j = Pe_i \cdot Pe_j = (f_i|f_j) = \delta_{i,j} \quad \forall i, j \in \{1 \dots n\},$$

et donc $(P^tPe_i - e_i) \cdot e_j = 0 \quad \forall j \in \{1 \dots n\} \quad \forall i \in \{1, \dots n\}$. On en déduit $P^tPe_i = e_i$ pour tout $i = 1, \dots n$, *i.e.* $P^tP = PP^t = Id$.

DÉMONSTRATION *du lemme 1.6* Cette démonstration se fait par récurrence sur la dimension de E.

1ère étape.

On suppose $dimE = 1$. Soit $e \in E$, $e \neq 0$, alors $E = \mathbb{R}e = f_1$ avec $\boldsymbol{f}_1 = \dfrac{e}{\|e\|}$. Soit $T : E \to E$ linéaire symétrique, on a : $T\boldsymbol{f}_1 \in \mathbb{R}\boldsymbol{f}_1$ donc il existe $\lambda_1 \in \mathbb{R}$ tel que $T\boldsymbol{f}_1 = \lambda_1 \boldsymbol{f}_1$.

2ème étape.

On suppose le lemme vrai si $dimE < n$. On montre alors le lemme si $dimE = n$. Soit E un espace vectoriel normé sur $\mathbb{R}$ tel que $dimE = n$ et $T : E \to E$ linéaire symétrique. Soit φ l'application définie par :

$$\begin{aligned} \varphi : \quad & E \to \mathbb{R} \\ & x \to (Tx|x). \end{aligned}$$

L'application φ est continue sur la sphère unité $S_1 = \{x \in E| \ \|x\| = 1\}$ qui est compacte car $dimE < +\infty$; il existe donc $e \in S_1$ tel que $\varphi(x) \leq \varphi(e) = (Te \mid e) = \lambda$ pour tout $x \in E$. Soit $y \in E \setminus \{0\}$, et soit $t \in]0, \frac{1}{\|y\|}[$ alors $e + ty \neq 0$. On en déduit que :

$$\frac{e+ty}{\|e+ty\|} \in S_1 \text{ et donc } \varphi(e) = \lambda \geq \left(T\frac{(e+ty)}{\|e+ty\|} | \frac{e+ty}{\|e+ty\|} \right)_E$$

donc $\lambda(e + ty \mid e + ty)_E \geq (T(e + ty) \mid e + ty)$. En développant on obtient :

$$\lambda[2t(e \mid y) + t^2(y \mid y)_E] \geq 2t(T(e) \mid y) + t^2(T(y) \mid y)_E.$$

Comme $t > 0$, ceci donne :

$$\lambda[2(e \mid y) + t(y \mid y)_E] \geq 2(T(e) \mid y) + t(T(y) \mid y)_E.$$

En faisant tendre t vers 0^+, on obtient $2\lambda(e \mid y)_E \geq 2(T(e) \mid y)$, Soit $0 \geq (T(e) - \lambda e \mid y)$ pour tout $y \in E \setminus \{0\}$. De même pour $z = -y$ on a $0 \geq (T(e) - \lambda e|z)$ donc $(T(e) - \lambda e \mid y) \geq 0$. D'où $(T(e) - \lambda e \mid y) = 0$ pour tout $y \in E$. On en déduit que $T(e) = \lambda e$. On pose $\boldsymbol{f}_n = e$ et $\lambda_n = \lambda$.

Soit $F = \{x \in E; (x \mid e) = 0\}$, on a donc $F \neq E$, et $E = F \bigoplus \mathbb{R}e$: on peut décomposer $x \in E$ comme $(x = x - (x \mid e)e + (x \mid e)e)$. L'application $S = T|_F$ est linéaire symétrique et on a $dimF = n - 1$. et $S(F) \subset F$. On peut donc utiliser l'hypothèse de récurrence : $\exists(\lambda_1 \dots \lambda_{n-1}) \in \mathbb{R}^n$ et $\exists(\boldsymbol{f}_1 \dots \boldsymbol{f}_{n-1}) \in E^n$ tels que $\forall\, i \in \{1 \dots n-1\}$, $S\boldsymbol{f}_i = T\boldsymbol{f}_i = \lambda_i \boldsymbol{f}_i$, et $\forall i, j \in \{1 \dots n-1\}$, $\boldsymbol{f}_i \cdot \boldsymbol{f}_j = \delta_{i,j}$. Et donc $(\lambda_1 \dots \lambda_n)$ et $(\boldsymbol{f}_1, \dots, \boldsymbol{f}_n)$ conviennent.

■

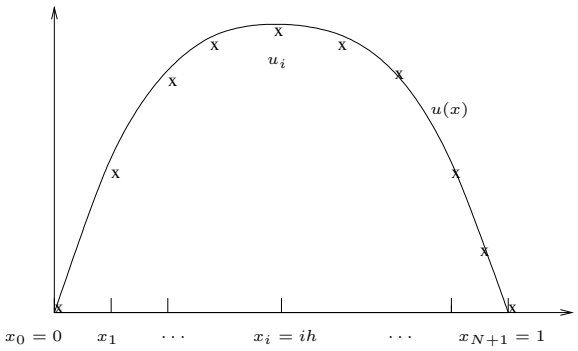

FIGURE 1.1: Solution exacte et approchée de $-u'' = f$

1.2.2 Discrétisation de l'équation de la chaleur

Dans ce paragraphe, nous prenons un exemple très simple pour obtenir un système linéaire à partir de la discrétisation d'un problème continu.

L'équation de la chaleur unidimensionnelle

Discrétisation par différences finies de $-u'' = f$ Soit $f \in C([0,1], \mathbb{R})$. On cherche u tel que

$$-u''(x) = f(x) \tag{1.5a}$$
$$u(0) = u(1) = 0. \tag{1.5b}$$

Remarque 1.7 (Problèmes aux limites, problèmes à conditions initiales)**.** *L'équation différentielle $-u'' = f$ admet une infinité de solutions. Pour avoir existence et unicité, il est nécessaire d'avoir des conditions supplémentaires. Si l'on considère deux conditions en 0 (ou en 1, l'origine importe peu) on a ce qu'on appelle un problème de Cauchy, ou problème à conditions initiales. Le problème* (1.5) *est lui un problème aux limites : il y a une condition pour chaque bord du domaine. En dimension supérieure, le problème $-\Delta u = f$ nécessite une condition sur au moins "un bout" de frontière pour être bien posé : voir le cours d'équations aux dërivées partielles de master pour plus de détails à ce propos.*

On peut montrer (on l'admettra ici) qu'il existe une unique solution $u \in C^2([0,1], \mathbb{R})$. On cherche à calculer u de manière approchée. On va pour cela introduire la méthode de discrétisation dite *par différences finies*. Soit $n \in \mathbb{N}^*$, on définit $h = 1/(n+1)$ le *pas de discrétisation*, c.à.d. la distance entre deux points de discrétisation, et pour $i = 0, \ldots, n+1$ on définit les points de discrétisation $x_i = ih$ (voir Figure 1.1), qui sont les points où l'on va écrire l'équation $-u'' = f$ en vue de se ramener à un système discret, c.à.d. à un système avec un nombre fini d'inconnues $u_1, \ldots, u_n$. Remarquons que $x_0 = 0$ et $x_{n+1} = 1$, et qu'en ces points, u est spécifiée par les conditions limites (1.5b). Soit $u(x_i)$ la valeur exacte de u en x_i. On écrit la première équation de (1.5a) en chaque point x_i, pour $i = 1 \ldots n$.

$$-u''(x_i) = f(x_i) = b_i \, \forall i \in \{1 \ldots n\}. \tag{1.6}$$

Supposons que $u \in C^4([0,1], \mathbb{R})$ (ce qui est vrai si $f \in C^2$). Par développement de Taylor, on a :

$$u(x_{i+1}) = u(x_i) + hu'(x_i) + \frac{h^2}{2}u''(x_i) + \frac{h^3}{6}u'''(x_i) + \frac{h^4}{24}u^{(4)}(\xi_i),$$
$$u(x_{i-1}) = u(x_i) - hu'(x_i) + \frac{h^2}{2}u''(x_i) - \frac{h^3}{6}u'''(x_i) + \frac{h^4}{24}u^{(4)}(\eta_i),$$

avec $\xi_i \in (x_i, x_{i+1})$ et $\eta_i \in (x_i, x_{i+1})$. En sommant ces deux égalités, on en déduit que :

$$u(x_{i+1}) + u(x_{i-1}) = 2u(x_i) + h^2 u''(x_i) + \frac{h^4}{24}u^{(4)}(\xi_i) + \frac{h^4}{24}u^{(4)}(\eta_i).$$

On définit l'erreur de consistance, qui mesure la manière dont on a approché $-u''(x_i)$; l'erreur de consistance R_i au point x_i est définie par

$$R_i = -u''(x_i) - \frac{u(x_{i+1}) + u(x_{i-1}) - 2u(x_i)}{h^2}. \tag{1.7}$$

On a donc :

$$\begin{aligned} |R_i| &= \left| -\frac{u(x_{i+1}) + u(x_{i-1}) - 2u(x_i)}{h^2} + u''(x_i) \right| \\ &\leq \left| \frac{h^4}{24}u^{(4)}(\xi_i) + \frac{h^4}{24}u^{(4)}(\eta_i) \right| \\ &\leq \frac{h^2}{12}\|u^{(4)}\|_\infty. \end{aligned} \tag{1.8}$$

où $\|u^{(4)}\|_\infty = \sup_{x\in]0,1[} |u^{(4)}(x)|$. Cette majoration nous montre que l'erreur de consistance tend vers 0 comme h^2, et on en dit que le schéma est *consistant d'ordre 2*.
On introduit alors les inconnues $(u_i)_{i=1,\dots,n}$ qu'on espère être des valeurs approchées de u aux points x_i et qui sont les composantes de la solution (si elle existe) du système suivant

$$\begin{cases} -\dfrac{u_{i+1} + u_{i-1} - 2u_i}{h^2} = b_i, \quad \forall i; 1 \leq i \leq n, \\ u_0 = u_{n+1} = 0. \end{cases} \tag{1.9}$$

On cherche donc $\boldsymbol{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix} \in \mathbb{R}^n$ solution de (1.9). Ce système peut s'écrire sous forme matricielle :$K_n \boldsymbol{u} = \boldsymbol{b}$

où $\boldsymbol{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}$ et K_n est la matrice carrée d'ordre n de coefficients $(k_{i,j})_{i,j=1,n}$ définis par :

$$\begin{cases} k_{i,i} &= \dfrac{2}{h^2}, \forall\, i = 1, \dots, n, \\ k_{i,j} &= -\dfrac{1}{h^2}, \forall\, i = 1, \dots, n, \ \ j = i \pm 1, \\ k_{i,j} &= 0, \forall\, i = 1, \dots, n, \ |i - j| > 1. \end{cases} \tag{1.10}$$

On remarque immédiatement que K_n est tridiagonale.
On peut montrer que K_n est symétrique définie positive (voir exercice 8 page 19), et elle est donc inversible Le système $K_n \boldsymbol{u} = \boldsymbol{b}$ admet donc une unique solution. C'est bien, mais encore faut il que cette solution soit ce qu'on espérait, c.à.d. que chaque valeur u_i soit une approximation pas trop mauvaise de $u(x_i)$. On appelle erreur de discrétisation en x_i la différence de ces deux valeurs :

$$e_i = u(x_i) - u_i, \ i = 1, \dots, n. \tag{1.11}$$

mSi on appelle $\boldsymbol{e}$ le vecteur de composantes e_i, on déduit de la définition 1.11 de l'erreur de consistance et des équations (exactes) 1.6 que

$$K_n \boldsymbol{e} = R \text{ et donc } \boldsymbol{e} = K_n^{-1} R. \tag{1.12}$$

Le fait que le schéma soit consistant est une bonne chose, mais cela ne suffit pas à montrer que le schéma est convergent, c.à.d. que l'erreur entre $\max_{i=1,\dots,n} e_i$ tend vers 0 lorsque h tend vers 0, parce que A dépend de

h ! Pour cela, il faut de plus que le schéma soit *stable*, au sens où l'on puisse montrer que $\|K_n^{-1}\|$ est borné indépendamment de h, ce qui revient à trouver une estimation sur les valeurs approchées u_i indépendante de h. La stabilité et la convergence font l'objet de l'exercice 43, où l'on montre que le schéma est convergent, et qu'on a l'estimation d'erreur suivante :

$$\max_{i=1\dots n}\{|u_i - u(x_i)|\} \leq \frac{h^2}{96}\|u^{(4)}\|_\infty.$$

Cette inégalité donne la précision de la méthode (c'est une méthode dite d'ordre 2). On remarque en particulier que si on raffine la discrétisation, c'est–à–dire si on augmente le nombre de points n ou, ce qui revient au même, si on diminue le pas de discrétisation h, on augmente la précision avec laquelle on calcule la solution approchée.

L'équation de la chaleur bidimensionnelle

Prenons maintenant le cas d'une discrétisation du Laplacien sur un carré par différences finies. Si u est une fonction de deux variables x et y à valeurs dans $\mathbb{R}$, et si u admet des dérivées partielles d'ordre 2 en x et y, l'opérateur laplacien est défini par $\Delta u = \partial_{xx}u + \partial_{yy}u$. L'équation de la chaleur bidimensionnelle s'écrit avec cet opérateur. On cherche à résoudre le problème :

$$\begin{array}{l} -\Delta u = f \text{ sur } \Omega =]0,1[\times]0,1[, \\ u = 0 \text{ sur } \partial\Omega, \end{array} \tag{1.13}$$

On rappelle que l'opérateur Laplacien est défini pour $u \in C^2(\Omega)$, où Ω est un ouvert de $\mathbb{R}^2$, par

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}.$$

Définissons une discrétisation uniforme du carré par les points (x_i, y_j), pour $i = 1,\dots,M$ et $j = 1,\dots,M$ avec $x_i = ih$, $y_j = jh$ et $h = 1/(M+1)$, representée en figure 1.2 pour $M = 6$. On peut alors approcher les dérivées secondes par des quotients différentiels comme dans le cas unidimensionnel (voir page 11), pour obtenir un système linéaire : $A\boldsymbol{u} = \boldsymbol{b}$ où $A \in \mathcal{M}_n(\mathbb{R})$ et $\boldsymbol{b} \in \mathbb{R}^n$ avec $n = M^2$. Utilisons l'ordre"*lexicographique*" pour numéroter les inconnues, c.à.d. de bas en haut et de gauche à droite : les inconnues sont alors numérotées de 1 à $n = M^2$ et le second membre s'écrit $b = (b_1,\dots,b_n)^t$. Les composantes $b_1,\dots,b_n$ sont définies par :pour $i,j = 1,\dots,M$, on pose $k = j + (i-1)M$ et $b_k = f(x_i, y_j)$.
Les coefficients de $A = (a_{k,\ell})_{k,l=1,n}$ peuvent être calculés de la manière suivante :

$$\left\{\begin{array}{ll} \text{Pour} & i,j = 1,\dots,M, \text{ on pose } k = j + (i-1)M, \\ a_{k,k} & = \dfrac{4}{h^2}, \\ a_{k,k+1} = & \left\{\begin{array}{ll} -\dfrac{1}{h^2} & \text{si } j \neq M, \\ 0 & \text{sinon,} \end{array}\right. \\ a_{k,k-1} = & \left\{\begin{array}{ll} -\dfrac{1}{h^2} & \text{si } j \neq 1, \\ 0 & \text{sinon,} \end{array}\right. \\ a_{k,k+M} = & \left\{\begin{array}{ll} -\dfrac{1}{h^2} & \text{si } i < M, \\ 0 & \text{sinon,} \end{array}\right. \\ a_{k,k-M} = & \left\{\begin{array}{ll} -\dfrac{1}{h^2} & \text{si } i > 1, \\ 0 & \text{sinon,} \end{array}\right. \\ \text{Pour} & k = 1,\dots,n, \text{ et } \ell = 1,\dots,n; \\ a_{k,\ell} & = 0, \forall\, k = 1,\dots,n,\ 1 < |k-\ell| < n \text{ ou } |k-\ell| > n. \end{array}\right.$$

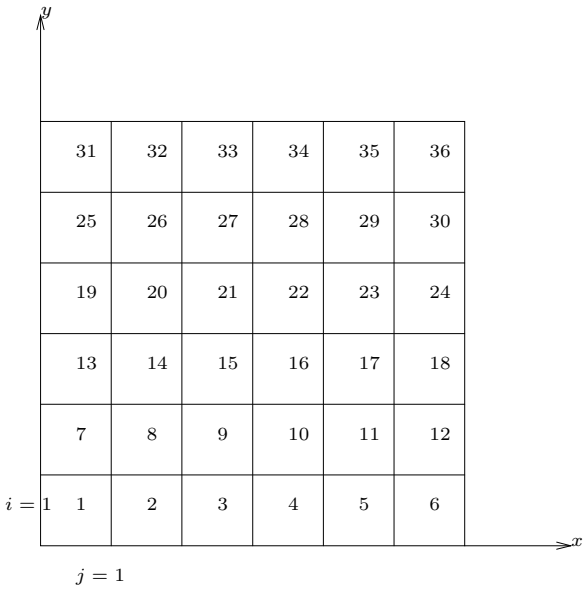

FIGURE 1.2: Ordre lexicographique des inconnues, exemple dans le cas $M = 6$

La matrice est donc tridiagonale par blocs, plus précisément si on note

$$D = \begin{pmatrix} 4 & -1 & 0 & \dots & \dots & 0 \\ -1 & 4 & -1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & & & \\ 0 & & \ddots & \ddots & \ddots & -1 \\ 0 & \dots & & 0 & -1 & 4 \end{pmatrix},$$

les blocs diagonaux (qui sont des matrices de dimension $M \times M$), on a :

$$A = \begin{pmatrix} D & -Id & 0 & \dots & \dots & 0 \\ -Id & D & -Id & 0 & \dots & 0 \\ 0 & -Id & D & -Id & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & & \ddots & -Id & D & -Id \\ 0 & \dots & & 0 & -Id & D \end{pmatrix}, \tag{1.14}$$

où Id désigne la matrice identité d'ordre M.

Matrices monotones, ou à inverse positive Une propriété qui revient souvent dans l'étude des matrices issues de la discrétisation d'équations différentielles est le fait que si leur action sur un vecteur $\boldsymbol{u}$ donne un vecteur positif $\boldsymbol{v}$ (composante par composante) alors le vecteur $\boldsymbol{u}$ de départ doit être positif (composante par composante) ; on dit souvent que la matrice est "monotone", ce qui n'est pas un terme très 'evocateuri. . . Dans ce cours, on lui préf'erera le terme "à inverse positive" ; en effet, on montre à la proposition qu'une matrice A est monotone si et seulement si elle inversible et à inverse positive.

Définition 1.8 (IP-matrice ou matrice monotone)**.** *Si $\boldsymbol{x} \in \mathbb{R}^n$, on dit que $\boldsymbol{x} \geq 0$ [resp. $\boldsymbol{x} > 0$] si toutes les composantes de $\boldsymbol{x}$ sont positives [resp. strictement positives].*
Soit $A \in \mathcal{M}_n(\mathbb{R})$, on dit que A est une matrice monotone si elle vérifie la propriété suivante :

$$\text{Si } \boldsymbol{x} \in \mathbb{R}^n \text{ est tel que } A\boldsymbol{x} \geq 0, \text{ alors } \boldsymbol{x} \geq 0,$$

ce qui peut encore s'écrire : $\{\boldsymbol{x} \in \mathbb{R}^n \text{ t.q. } A\boldsymbol{x} \geq 0\} \subset \{\boldsymbol{x} \in \mathbb{R}^n \text{ t.q. } \boldsymbol{x} \geq 0\}$.

Proposition 1.9 (Caractérisation des matrices monotones)**.** *Une matrice A est monotone si et seulement si elle inversible et à inverse positive (c.à.d. dont tous les coefficients sont positifs).*

La démonstration de ce résultat est l'objet de l'exercice 6. Retenez que toute matrice monotone est inversible et d'inverse positive que cette propriété de monotonie est utilisée pour établir une borne de $\|A^{-1}\|$ pour la matrice de discrétisation du Laplacien, dont on a besoin pour montrer la convergence du schéma. C'est donc une propriété qui est importante au niveau de l'analyse numérique.

1.2.3 Exercices

Exercice 1 (Vrai ou faux ? Motiver les réponses…)**.**

On suppose dans toutes les questions suivantes que $n \geq 2$.

1. Soit $Z \in \mathbb{R}^n$ un vecteur non nul. La matrice ZZ^t est inversible.
2. La matrice inverse d'une matrice triangulaire inférieure est triangulaire supérieure.
3. Les valeurs propres sont les racines du polynôme caractéristique.
4. Toute matrice inversible est diagonalisable dans $\mathbb{R}$.
5. Toute matrice inversible est diagonalisable dans $\mathbb{C}$.
6. Le déterminant d'une matrice A est égal au produit de ses valeurs propres.
7. Soit A une matrice carrée telle que $A\boldsymbol{x} = \mathbf{0} \Longrightarrow \boldsymbol{x} = \mathbf{0}$, alors A est inversible.
8. Soit A une matrice carrée telle que $A\boldsymbol{x} \geq \mathbf{0} \Longrightarrow \boldsymbol{x} \geq \mathbf{0}$, alors A est inversible.
9. Une matrice symétrique est inversible.
10. Une matrice symétrique définie positive est inversible.

Exercice 2 (Sur quelques notions connues)**.**

1. Soit A une matrice carrée d'ordre n et $\boldsymbol{b} \in \mathbb{R}^3$. Peut il exister exactement deux solutions distinctes au système $A\boldsymbol{x} = \boldsymbol{b}$?
2. Soient A, B et C de dimensions telles que AB et BC existent. Montrer que si $AB = Id$ et $BC = Id$, alors $A = C$.
3. Combien y a -til de matrices carrées d'ordre 2 ne comportant que des 1 ou des 0 comme coefficients ? Combien d'entre elles sont inversibles ?
4. Soit $B = \begin{bmatrix} 3 & 2 \\ -5 & -3 \end{bmatrix}$. Montrer que $B^{1024} = Id$.

Exercice 3 (La matrice K_3)**.** *Soit $f \in C([0,1], \mathbb{R})$. On cherche u tel que*

$$-u''(x) = f(x), \ \forall x \in (0,1), \tag{1.15a}$$
$$u(0) = u(1) = 0. \tag{1.15b}$$

1. *Calculer la solution exacte $u(x)$ du problème lorsque f est la fonction identiquement égale à 1 (on admettra que cette solution est unique), et vérifier que $u(x) \geq 0$ pour tout $x \in [0,1]$.*

On discrétise le problème suivant par différences finies, avec un pas $h = \frac{1}{4}$ avec la technique vue en cours.

2. *A l'aide de dévloppements de Taylor, écrire l'approximation de $u''(x_i)$ au deuxième ordre en fonction de $u(x_i)$, $u(x_{i-1})$ et $u(x_{i+1})$. En déduire le schéma aux différences finies pour l'approximation de* (1.15)*, qu'on écrira sous la forme :*

$$K_3 \boldsymbol{u} = \boldsymbol{b}, \tag{1.16}$$

où K_3 est la matrice de discrétisation qu'on explicitera, $\boldsymbol{u} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}$ et $\boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} f(x_1) \\ f(x_2) \\ f(x_3) \end{bmatrix}$.

3. *Résoudre le système linéaire* (1.16) *par la méthode de Gauss. Comparer u_i et $u(x_i)$ pour $i = 1, 2, 3$, et expliquer pourquoi l'erreur de discrétisation $u(x_i) - u_i$ est nulle.*
4. *Reprendre les questions précédentes en remplaçant les conditions limites* (1.15b) *par :*

$$u(0) = 0, \quad u'(1) = 0. \tag{1.17}$$

5. *Soit $c \in \mathbb{R}$. On considère maintenant le problème suivant :*

$$-u''(x) = c, \ \forall x \in (0,1), \tag{1.18a}$$
$$u'(0) = u'(1) = 0, \tag{1.18b}$$

   (a) *Montrer que le problème* (1.18) *admet soit une infinité de solution, soit pas de solution.*
   (b) *Ecrire la discrétisation du problème* (1.18)*, toujours avec $h = \frac{1}{4}$, sous la forme $\widetilde{K}_3 \boldsymbol{u} = \widetilde{\boldsymbol{b}}$ en explicitant $\widetilde{K}_3$ et $\widetilde{\boldsymbol{b}}$.*
   (c) *Montrer que la matrice $\widetilde{K}_3$ n'est pas inversible : on part d'un problème continu mal posé, et on obtient par discrétisation un problème discret mal posé...*

Exercice 4 (Matrices symétriques définies positives)**.** *Suggestions en page 19, corrigé en page 19. On rappelle que toute matrice $A \in \mathcal{M}_n(\mathbb{R})$ symétrique est diagonalisable dans $\mathbb{R}$ (cf. lemme 1.6 page 9). Plus précisément, on a montré en cours que, si $A \in \mathcal{M}_n(\mathbb{R})$ est une matrice symétrique, il existe une base de $\mathbb{R}^n$, notée $\{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n\}$, et il existe $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ t.q. $A\boldsymbol{f}_i = \lambda_i \boldsymbol{f}_i$, pour tout $i \in \{1, \ldots, n\}$, et $\boldsymbol{f}_i \cdot \boldsymbol{f}_j = \delta_{i,j}$ pour tout $i, j \in \{1, \ldots, n\}$ ($x \cdot y$ désigne le produit scalaire de x avec y dans $\mathbb{R}^n$).*

1. *Soit $A \in \mathcal{M}_n(\mathbb{R})$. On suppose que A est symétrique définie positive, montrer que les éléments diagonaux de A sont strictements positifs.*
2. *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique. Montrer que A est symétrique définie positive si et seulement si toutes les valeurs propres de A sont strictement positives.*
3. *Soit $A \in \mathcal{M}_n(\mathbb{R})$. On suppose que A est symétrique définie positive. Montrer qu'on peut définir une unique matrice $B \in \mathcal{M}_n(\mathbb{R})$, symétrique définie positive t.q. $B^2 = A$ (on note $B = A^{\frac{1}{2}}$).*

Exercice 5 (Diagonalisation dans $\mathbb{R}$)**.**

Soit E un espace vectoriel réel de dimension $n \in \mathbb{N}$ muni d'un produit scalaire, noté $(\cdot, \cdot)$. Soient T et S deux applications linéaires symétriques de E dans E (T symétrique signifie $(Tx, y) = (x, Ty)$ pour tous $x, y \in E$). On suppose que T est définie positive (c'est-à-dire $(Tx, x) > 0$ pour tout $x \in E \setminus \{0\}$).

1. Montrer que T est inversible. Pour $x, y \in E$, on pose $(x, y)_T = (Tx, y)$. Montrer que l'application $(x, y) \mapsto (x, y)_T$ définit un nouveau produit scalaire sur E.
2. Montrer que $T^{-1}S$ est symétrique pour le produit scalaire défini à la question précédente. En déduire, avec le lemme 1.6 page 9, qu'il existe une base de E, notée $\{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n\}$ et une famille $\{\lambda_1, \ldots, \lambda_n\} \subset \mathbb{R}$ telles que $T^{-1}S\boldsymbol{f}_i = \lambda_i \boldsymbol{f}_i$ pour tout $i \in \{1, \ldots, n\}$ et t.q. $(T\boldsymbol{f}_i, \boldsymbol{f}_j) = \delta_{i,j}$ pour tout $i, j \in \{1, \ldots, n\}$.

Exercice 6 (IP-matrice)**.** *Corrigé en page 21*

Soit $n \in \mathbb{N}^\star$, on note $\mathcal{M}_n(\mathbb{R})$ l'ensemble des matrices de n lignes et n colonnes et à coefficients réels. Si $x \in \mathbb{R}^n$, on dit que $x \geq 0$ [resp. $x > 0$] si toutes les composantes de x sont positives [resp. strictement positives].
Soit $A \in \mathcal{M}_n(\mathbb{R})$, on dit que A est une IP-matrice si elle vérifie la propriété suivante :

$$\text{Si } x \in \mathbb{R}^n \text{ est tel que } Ax \geq 0, \text{ alors } x \geq 0,$$

ce qui peut encore s'écrire : $\{x \in \mathbb{R}^n \text{ t.q. } Ax \geq 0\} \subset \{x \in \mathbb{R}^n \text{ t.q. } x \geq 0\}$.

1. Soit $A = (a_{i,j})_{i,j=1,\ldots,n} \in \mathcal{M}_n(\mathbb{R})$. Montrer que A est une IP-matrice si et seulement si A est inversible et $A^{-1} \geq 0$ (c'est-à-dire que tous les coefficients de A^{-1} sont positifs).

2. Soit $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ une matrice réelle d'ordre 2. Montrer que A est une IP-matrice si et seulement si :

$$\begin{cases} ad < bc, \\ a < 0,\ d < 0 \\ b \geq 0,\ c \geq 0 \end{cases} \quad \text{ou} \quad \begin{cases} ad > bc, \\ a > 0,\ d > 0, \\ b \leq 0,\ c \leq 0. \end{cases} \tag{1.19}$$

3. Montrer que si $A \in \mathcal{M}_n(\mathbb{R})$ est une IP-matrice alors A^t (la transposée de A) est une IP-matrice.

4. Montrer que si A est telle que

$$\begin{aligned} &a_{i,j} \leq 0, \text{ pour tout } i, j = 1, \ldots, n,\ i \neq j, \\ &a_{i,i} > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{i,j}|, \text{ pour tout } i = 1, \ldots, n, \end{aligned} \tag{1.20}$$

alors A est une IP-matrice.

5. En déduire que si A est telle que

$$\begin{aligned} &a_{i,j} \leq 0, \text{ pour tout } i, j = 1, \ldots, n,\ i \neq j, \\ &a_{i,i} > \sum_{\substack{j=1 \\ j \neq k}}^{n} |a_{j,i}|, \text{ pour tout } i = 1, \ldots, n, \end{aligned} \tag{1.21}$$

alors A est une IP-matrice.

6. Montrer que si $A \in \mathcal{M}_n(\mathbb{R})$ est une IP-matrice et si $x \in \mathbb{R}^n$ alors :

$$Ax > 0 \Rightarrow x > 0.$$

c'est-à-dire que $\{x \in \mathbb{R}^n \text{ t.q. } Ax > 0\} \subset \{x \in \mathbb{R}^n \text{ t.q. } x > 0\}$

7. Montrer, en donnant un exemple, qu'une matrice A de $\mathcal{M}_n(\mathbb{R})$ peut vérifier $\{x \in \mathbb{R}^n \text{ t.q. } Ax > 0\} \subset \{x \in \mathbb{R}^n$ t.q. $x > 0\}$ et ne pas être une IP-matrice.

8. On suppose dans cette question que $A \in \mathcal{M}_n(\mathbb{R})$ est inversible et que $\{x \in \mathbb{R}^n \text{ t.q. } Ax > 0\} \subset \{x \in \mathbb{R}^n \text{ t.q. } x > 0\}$. Montrer que A est une IP-matrice.

9. (Question plus difficile) Soit E l'espace des fonctions continues sur $\mathbb{R}$ et admettant la même limite finie en $+\infty$ et $-\infty$. Soit $\mathcal{L}(E)$ l'ensemble des applications linéaires continues de E dans E. Pour $f \in E$, on dit que $f > 0$ (resp. $f \geq 0$) si $f(x) > 0$ (resp. $f(x) \geq 0$) pour tout $x \in \mathbb{R}$. Montrer qu'il existe $T \in \mathcal{L}(E)$ tel que $Tf \geq 0 \Longrightarrow f \geq 0$, et $g \in E$ tel que $Tg > 0$ et $g \not> 0$ (ceci démontre que le raisonnement utilisé en 2 (b) ne marche pas en dimension infinie).

Exercice 7 (M-matrice)**.**

Dans ce qui suit, toutes les inégalités écrites sur des vecteurs ou des matrices sont à entendre au sens composante par composante.
Soit $A = (a_{i,j})_{i,j=1,\dots,n}$ une matrice carrée d'ordre n. On dit que A est une M-matrice si A est une IP-matrice (A est inversible et $A^{-1} \geq 0$, voir exercice 6) qui vérifie de plus

(a) $a_{i,i} > 0$ pour $i = 1, \dots, n$;

(b) $a_{i,j} \leq 0$ pour $i, j = 1, \dots, n$, $i \neq j$;

1. Soit A une IP-matrice ; montrer que A est une M-matrice si et seulement si la propriété (b) est vérifiée.

2. Soit $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ une matrice réelle d'ordre 2. Montrer que A est une M-matrice si et seulement si :

$$\begin{cases} ad > bc, \\ a > 0,\ d > 0, \\ b \leq 0,\ c \leq 0. \end{cases} \tag{1.22}$$

3. Les matrices $A = \begin{pmatrix} -1 & 2 \\ 2 & -1 \end{pmatrix}$ et $B = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$ sont-elles des IP-matrices ? des M-matrices ?

4. Soit A la matrice carrée d'ordre 3 définie par :

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

Montrer que $A^{-1} \geq 0$ mais que A n'est pas une M–matrice.

5. Soit A une matrice carrée d'ordre $n = m + p$, avec $m, p \in \mathbb{N}$ tels que $m \geq 1$ et $p \geq 1$, vérifiant :

$$\left.\begin{array}{l} a_{i,i} \geq 0, \\ a_{i,j} \leq 0, \text{ pour } j = 1, \dots, n,\ j \neq i, \\ a_{i,i} + \displaystyle\sum_{\substack{j=1 \\ j \neq i}}^{n} a_{i,j} = 0 \end{array}\right\} \text{ pour } i = 1, \dots, m, \tag{1.23}$$

$$\left.\begin{array}{l} a_{i,i} = 1, \\ a_{i,j} = 0, \text{ pour } j = 1, \dots, n,\ j \neq i, \end{array}\right\} \text{ pour } i = m+1, \dots, n. \tag{1.24}$$

$$\forall i \leq m,\ \exists (k_\ell)_{\ell=1,\dots,L_i};\, k_1 = i, k_{L_i} > m, \text{ et } a_{k_\ell, k_{\ell+1}} < 0,\ \forall\, \ell = 1, \dots, L_i. \tag{1.25}$$

Soit $b \in \mathbb{R}^n$ tel que $b_i = 0$ pour $i = 1, \dots, m$. On considère le système linéaire

$$Au = b \tag{1.26}$$

5.1 Montrer que le système (1.26) admet une et une seule solution.

5.2 Montrer que u est tel que $\min_{k=m+1,n} b_k \leq u_i \leq \max_{k=m+1,n} b_k$. (On pourra pour simplifier supposer que les équations sont numérotées de telle sorte que $\min_{k=m+1,n} b_k = b_{m+2} \leq b_2 \leq \ldots \leq b_n = \max_{k=m+1,n} b_k$.)

6. On considère le problème de Dirichlet suivant :

$$-u'' = 0 \text{ sur } [0, 1] \tag{1.27a}$$

$$u(0) = -1 \tag{1.27b}$$

$$u(1) = 1. \tag{1.27c}$$

6.1 Calculer la solution exacte u de ce problème et vérifier qu'elle reste comprise entre -1 et 1.

6.2 Soit $m > 1$ et soient A et b et la matrice et le second membre du système linéaire d'ordre $n = m+2$ obtenu par discrétisation par différences finies avec un pas uniforme $h = \frac{1}{m}$ du problème (1.27) (en écrivant les conditions aux limites dans le système). Montrer que la solution $\boldsymbol{u} = (u_1, \dots, u_n)^t \in \mathbb{R}^n$ du système $A\boldsymbol{u} = \boldsymbol{b}$ vérifie $-1 \leq u_i \leq 1$.

Exercice 8 (Matrice du Laplacien discret 1D)**.** *Corrigé détaillé en page 20.*
Soit $f \in C([0,1])$. Soit $n \in \mathbb{N}^\star$, n impair. On pose $h = 1/(n+1)$. Soit K_n la matrice définie par (1.10) page 12, issue d'une discrétisation par différences finies avec pas constant du problème (1.5a) page 11.
Montrer que K_n est symétrique définie positive.

Exercice 9 (Pas non constant)**.** *Reprendre la discrétisation vue en cours avec un pas $h_i = x_{i+1} - x_i$ non constant, et montrer que dans ce cas,le schéma est consistant d'ordre 1 seulement.*

Exercice 10 (Réaction diffusion 1d.)**.** *Corrigé détaillé en page 20. On s'intéresse à la discrétisation par Différences Finies du problème aux limites suivant :*

$$\begin{array}{l} -u''(x) + u(x) = f(x),\ x \in]0,1[, \\ u(0) = u(1) = 0. \end{array} \tag{1.28}$$

Soit $n \in \mathbb{N}^\star$. On note $U = (u_j)_{j=1,\ldots,n}$ une "valeur approchée" de la solution u du problème (1.28) aux points $\left(\frac{j}{n+1}\right)_{j=1,\ldots,n}$. Donner la discrétisation par différences finies de ce problème sous la forme $AU = b$.

Exercice 11 (Discrétisation)**.**

On considère la discrétisation à pas constant par le schéma aux différences finies symétrique à trois points du problème (1.5a) page 11, avec $f \in C([0,1])$. Soit $n \in \mathbb{N}^\star$, n impair. On pose $h = 1/(n+1)$. On note u est la solution exacte, $x_i = ih$, pour $i = 1, \ldots, n$ les points de discrétisation, et $(u_i)_{i=1,\ldots,n}$ la solution du système discrétisé (1.9).

1. Montrer que si $u \in C^4([0,1]$, alors la propriété (1.7) est vérifiée, c.à.d. :

$$-\frac{u(x_{i+1}) + u(x_{i-1}) - 2u(x_i)}{h^2} = -u''(x_i) + R_i \text{ avec } |R_i| \leq \frac{h^2}{12}\|u^{(4)}\|_\infty.$$

2. Montrer que si f est constante, alors

$$\max_{1\leq i\leq n} |u_i - u(x_i)| = 0.$$

.

3. Soit n fixé, et $\max_{1\leq i\leq n} |u_i - u(x_i)| = 0$. A-t-on forcément que f est constante sur $[0,1]$? (justifier la réponse.)

1.2.4 Suggestions pour les exercices

Exercice 4 page 16 (Matrices symétriques définies positives)

3. Utiliser la diagonalisation sur les opérateurs linéaires associés.

1.2.5 Corrigés des exercices

Exercice 4 page 16 (Matrices symétriques définies positives)

1. Supposons qu'il existe un élément diagonal $a_{i,i}$ négatif. Alors $Ae_i \cdot e_i \leq 0$ ce qui contredit le fait que A est définie positive.

2. Soit $x \in \mathbb{R}^n$, décomposons x sur la base orthonormée $(\boldsymbol{f}_i)_{i=1,n}$: $x = \sum_{i=1}^n x_i \boldsymbol{f}_i$. On a donc :

$$Ax \cdot x = \sum_{i=1}^n \lambda_i x_i^2. \tag{1.29}$$

Montrons d'abord que si les valeurs propres sont strictement positives alors A est définie positive :

Supposons que $\lambda_i \geq 0$, $\forall i = 1, \ldots, n$. Alors pour $\forall x \in \mathbb{R}^n$, d'après (1.29), $Ax \cdot x \geq 0$ et la matrice A est positive. Supposons maintenant que $\lambda_i > 0$, $\forall i = 1, \ldots, n$. Alors pour $\forall x \in \mathbb{R}^n$, toujours d'après (1.29), $(Ax \cdot x = 0) \Rightarrow (x = 0)$, et la matrice A est donc bien définie.

Montrons maintenant la réciproque : si A est définie positive, alors $A\boldsymbol{f}_i \cdot \boldsymbol{f}_i > 0$, $\forall i = 1, \ldots, n$ et donc $\lambda_i > 0$, $\forall i = 1, \ldots, n$.

3. Comme A est s.d.p., toutes ses valeurs propres sont strictement positives, et on peut donc définir l'application linéaire S dans la base orthonormée $(f_i)_{i=1,n}$ par : $S(\boldsymbol{f}_i) = \sqrt{\lambda_i}\boldsymbol{f}_i$, $\forall i = 1, \ldots, n$. On a évidemment $S \circ S = T$, et donc si on désigne par B la matrice représentative de l'application S dans la base canonique, on a bien $B^2 = A$.

Exercice 8 page 19 (Matrice du laplacien discret 1D.)

Il est clair que la matrice A est symétrique.
Pour montrer que A est définie positive (car A est évidemment symétrique), on peut procéder de plusieurs façons :

1. *Par échelonnement* :
2. *Par les valeurs propres* :Les valeurs propres sont calculées à l'exercice 41 ; elles sont de la forme :
$$\lambda_k = \frac{2}{h^2}(1 - \cos k\pi h) = \frac{2}{h^2}(1 - \cos\frac{k\pi}{n+1}), k = 1, \ldots, n,$$
et elles sont donc toutes strictement positives ; de ce fait, la matrice est symétrique définie positive (voir exercice 4).
3. *Par la forme quadratique associée* : on montre que $Ax \cdot x > 0$ si $x \neq 0$ et $Ax \cdot x = 0$ ssi $x = 0$. En effet, on a
$$Ax \cdot x = \frac{1}{h^2}\left[x_1(2x_1 - x_2) + \sum_{i=2}^{n-1} x_i(-x_{i-1} + 2x_i - x_{i+1}) + 2x_n^2 - x_{n-1}x_n\right]$$
On a donc
$$\begin{aligned} h^2 Ax \cdot x &= 2x_1^2 - x_1x_2 - \sum_{i=2}^{n-1}\left(x_ix_{i-1} + 2x_i^2\right) - \sum_{i=3}^{n} x_ix_{i-1} + 2x_n^2 - x_{n-1}x_n \\ &= \sum_{i=1}^{n} x_i^2 + \sum_{i=2}^{n} x_{1-i}^2 + x_n^2 - 2\sum_{i=1}^{n} x_ix_{i-1} \\ &= \sum_{i=2}^{n}(x_i - x_{i-1})^2 + x_1^2 + x_n^2 \geq 0. \end{aligned}$$
De plus, $Ax \cdot x = 0 \Rightarrow x_1^2 = x_n = 0$ et $x_i = x_{i-1}$ pour $i = 2$ à n, donc $x = 0$.

Exercice 10 page 19 (Réaction diffusion 1D.)

La discrétisation du probllème consiste à chercher U comme solution du système linéaire
$$AU = \left(f(\frac{j}{N+1})\right)_{j=1,\ldots,n}$$
où la matrice $A \in \mathcal{M}_n(\mathbb{R})$ est définie par $A = (N+1)^2K_n + Id$, Id désigne la matrice identité et
$$K_n = \begin{pmatrix} 2 & -1 & 0 & \ldots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \ldots & 0 & -1 & 2 \end{pmatrix}$$

Exercice 6 page 17 (IP-matrice)

1. Supposons d'abord que A est inversible et que $A^{-1} \geq 0$; soit $x \in \mathbb{R}^n$ tel que $b = Ax \geq 0$. On a donc $x = A^{-1}b$, et comme tous les coefficients de A^{-1} et de b sont positifs ou nuls, on a bien $x \geq 0$.
   Réciproquement, si A est une IP-matrice, alors $Ax = 0$ entraine $x = 0$ ce qui montre que A est inversible. Soit e_i le i-ème vecteur de la base canonique de $\mathbb{R}^n$, on a : $AA^{-1}e_i = e_i \geq 0$, et donc par la propriété de IP-matrice, $A^{-1}e_i \geq 0$, ce qui montre que tous les coefficients de A^{-1} sont positifs.
2. La matrice inverse de A est $A^{-1} = \frac{1}{\Delta}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$ avec $\Delta = ad - bc$. Les coefficients de A^{-1} sont donc positifs ou nuls si et seulement si
$$\begin{cases} ad < bc, \\ a \leq 0,\ d \leq 0 \\ b \geq 0,\ c \geq 0 \end{cases} \quad \text{ou} \quad \begin{cases} ad > bc, \\ a \geq 0,\ d \geq 0, \\ b \leq 0,\ c \leq 0. \end{cases}$$
   Or on a forcément $ad \neq 0$: en effet sinon on aurait dans le premier cas) $bc < 0$, or $b \leq 0$ et $c \leq 0$, ce qui aboutit à une contradiction. De même dans le deuxième cas, on aurait $bc > 0$, or $b \geq 0$ et $c \geq 0$. Les conditions précédentes sont donc équivalentes aux conditions (1.19).
3. La matrice A^t est une IP-matrice si et seulement A^t est inversible et $(A^t)^{-1} \geq 0$. Or $(A^t)^{-1} = (A^{-1})^t$. D'où l'équivalence.
4. Supposons que A vérifie (1.20), et soit $x \in \mathbb{R}^n$ tel que $Ax \geq 0$. Soit $k \in 1, \ldots, n$ tel que $x_k = \min\{x_i, i = 1, \ldots, n\}$. Alors
$$(Ax)_k = a_{k,k}x_k + \sum_{\substack{j=1 \\ j\neq k}}^{n} a_{k,j}x_j \geq 0.$$
   Par hypothèse, $a_{k,j} \leq 0$ pour $k \neq j$, et donc $a_{k,j} = -|a_{k,j}|$. On peut donc écrire :
$$a_{k,k}x_k - \sum_{\substack{j=1 \\ j\neq k}}^{n} |a_{k,j}|x_j \geq 0,$$
   et donc :
$$(a_{k,k} - \sum_{\substack{j=1 \\ j\neq k}}^{n} |a_{k,j}|)x_k \geq \sum_{\substack{j=1 \\ j\neq k}}^{n} |a_{k,j}|(x_j - x_k).$$
   Comme $x_k = \min\{x_i, i = 1, \ldots, n\}$, on en déduit que le second membre de cette inégalité est positif ou nul, et donc que $x_k \geq 0$. On a donc $x \geq 0$.
5. Si la matrice A vérifie (1.21), alors la matrice A^t vérifie (1.20). On en déduit par les questions précédentes que A^t et A sont des IP-matrices.
6. Soit $\mathbf{1}$ le vecteur de $\mathbb{R}^n$ dont toutes les composantes sont égales à 1. Si $Ax > 0$, comme l'espace $\mathbb{R}^n$ est de dimension finie, il existe $\epsilon > 0$ tel que $Ax \geq \epsilon\mathbf{1}$. Soit $z = \epsilon A^{-1}\mathbf{1} \geq 0$; on a alors $A(x - z) \geq 0$ et donc $x \geq z$, car A est une IP-matrice.
   Montrons maintenant que $z > 0$: tous les coefficients de A^{-1} sont positifs ou nuls et au moins l'un d'entre eux est non nul par ligne (puisque la matrice A^{-1} est inversible). On en déduit que $z_i = \epsilon \sum_{i=1}^{n} (A^{-1})_{i,j} > 0$ pour tout $i = 1, \ldots, n$. On a donc bien $x \geq z > 0$.
7. Soit A la matrice nulle, on a alors $\{x \in \mathbb{R}^n \text{ t.q. } Ax > 0\} = \emptyset$, et donc $\{x \in \mathbb{R}^n \text{ t.q. } Ax > 0\} \subset \{x \in \mathbb{R}^n \text{ t.q. } x > 0\}$. Pourtant A n'est pas inversible, et n'est donc pas une IP-matrice.
8. Soit x tel que $Ax \geq 0$, alors il existe $\varepsilon \geq 0$ tel que $Ax + \varepsilon\mathbf{1} \geq 0$. Soit maintenant $b = A^{-1}\mathbf{1}$; on a $A(x + \varepsilon b) > 0$ et donc $x + \varepsilon b > 0$. En faisant tendre ε vers 0, on en déduit que $x \geq 0$.

9. Soit $T \in \mathcal{L}(E)$ défini par $f \in E \mapsto Tf$, avec $Tf(x) = f(\frac{1}{x})$ si $x \neq 0$ et $f(0) = \ell$, avec $\ell = \lim_{\pm\infty} f$. On vérifie facilement que $Tf \in E$. Si $Tf \geq 0$, alors $f(\frac{1}{x}) \geq 0$ pour tout $x \in \mathbb{R}$; donc $f(x) \geq 0$ pour tout $x \in \mathbb{R} \setminus \{0\}$; on en déduit que $f(0) \geq 0$ par continuité. On a donc bien $f \geq 0$.
Soit maintenant g définie de $\mathbb{R}$ dans $\mathbb{R}$ par $g(x) = |\arctan x|$. On a $g(0) = 0$, donc $g \not> 0$. Or $Tg(0) = \frac{\pi}{2}$ et $Tg(x) = |\arctan \frac{1}{x}| > 0$ si $x > 0$, donc $Tg > 0$.

1.3 Les méthodes directes

1.3.1 Définition

Définition 1.10 (Méthode directe)**.** *On appelle méthode directe de résolution de (1.1) une méthode qui donne exactement x (A et b étant connus) solution de (1.1) après un nombre fini d'opérations élémentaires $(+, -, \times, /)$.*

Parmi les méthodes de résolution du système (1.1), la plus connue est la *méthode de Gauss* (avec pivot), encore appelée *méthode d'échelonnement* ou *méthode LU* dans sa forme matricielle.
Nous rappelons la méthode de Gauss et sa réécriture matricielle qui donne la méthode LU et nous étudierons plus en détails la méthode de Choleski, qui est adaptée aux matrices symétriques.

1.3.2 Méthode de Gauss, méthode LU

Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible, et $b \in \mathbb{R}^n$. On cherche à calculer $x \in \mathbb{R}^n$ tel que $Ax = b$. Le principe de la méthode de Gauss est de se ramener, par des opérations simples (combinaisons linéaires), à un système triangulaire équivalent, qui sera donc facile à inverser.
Commençons par un exemple pour une matrice 3×3. Nous donnerons ensuite la méthode pour une matrice $n \times n$.

Un exemple 3×3

On considère le système $Ax = b$, avec

$$A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 2 & -1 \\ -1 & 1 & -2 \end{bmatrix} \qquad b = \begin{bmatrix} 2 \\ 1 \\ -2 \end{bmatrix}.$$

On écrit la **matrice augmentée**, constituée de la matrice A et du second membre b.

$$\tilde{A} = \begin{bmatrix} A & \boldsymbol{b} \end{bmatrix} = \begin{bmatrix} \mathbf{1} & 0 & 1 & 2 \\ 0 & 2 & -1 & 1 \\ -1 & 1 & -2 & -2 \end{bmatrix}.$$

Gauss et opérations matricielles Allons y pour Gauss :
La première ligne a un 1 en première position (en gras dans la matrice), ce coefficient est non nul, et c'est un **pivot**. On va pouvoir diviser toute la première ligne par ce nombre pour en soustraire un multiple à toutes les lignes d'après, dans le but de faire apparaître des 0 dans tout le bas de la colonne.
La deuxième équation a déjà un 0 dessous, donc on n'a rien besoin de faire. On veut ensuite annuler le premier coefficient de la troisième ligne. On retranche donc (-1) fois la première ligne à la troisième [3] :

$$\begin{bmatrix} \mathbf{1} & 0 & 1 & 2 \\ 0 & 2 & -1 & 1 \\ -1 & 1 & -2 & -2 \end{bmatrix} \xrightarrow{\ell_3 \leftarrow to \ell_3 + \ell_1} \begin{bmatrix} \mathbf{1} & 0 & 1 & 2 \\ 0 & \mathbf{2} & -1 & 1 \\ 0 & 1 & -1 & 0 \end{bmatrix}$$

3. Bien sûr, ceci revient à ajouter la première ligne ! Il est cependant préférable de parler systématiquement de "retrancher" quitte à utiliser un coefficient négatif, car c'est ce qu'on fait conceptuellement : pour l'élimination on enlève un multiple de la ligne du pivot à la ligne courante.

Ceci revient à multiplier $\tilde{A}$ à gauche par la matrice $E_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$.
La deuxième ligne a un terme non nul en deuxième position (2) : c'est un pivot. On va maintenant annuler le deuxième terme de la troisième ligne ; pour cela, on retranche 1/2 fois la ligne 2 à la ligne 3 :

$$\begin{bmatrix} \mathbf{1} & 1 & 1 & 2 \\ 0 & \mathbf{2} & -1 & 1 \\ 0 & 1 & -1 & 0 \end{bmatrix} \overset{\ell_3 \leftarrow \ell_3 - 1/2\ell_2}{\longrightarrow} \begin{bmatrix} \mathbf{1} & 0 & 1 & 2 \\ 0 & \mathbf{2} & -1 & 1 \\ 0 & 0 & -\mathbf{\frac{1}{2}} & -\frac{1}{2} \end{bmatrix}.$$

Ceci revient à multiplier la matrice précédente à gauche par la matrice $E_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -\frac{1}{2} & 1 \end{bmatrix}$. On a ici obtenu une matrice sous forme triangulaire supérieure à trois pivots : on peut donc faire la remontée pour obtenir la solution du système, et on obtient (en notant x_i les composantes de $\boldsymbol{x}$) : $x_3 = 1$ puis $x_2 = 1$ et enfin $x_1 = 1$.
On a ainsi résolu le système linéaire.
On rappelle que le fait de travailler sur la matrice augmentée est extrêmement pratique car il permet de travailler simultanément sur les coefficients du système linéaire et sur le second membre.
Finalement, au moyen des opérations décrites ci-dessus, on a transformé le système linéaire

$$A\boldsymbol{x} = \boldsymbol{b} \text{ en } U\boldsymbol{x} = E_2E_1b, \text{ où } U = E_2E_1A$$

est une matrice triangulaire supérieure.

Factorisation LU Tout va donc très bien pour ce système, mais supposons maintenant qu'on ait à résoudre 3089 systèmes, avec la même matrice A mais 3089 seconds membres b différents[4]. Il serait un peu dommage de recommencer les opérations ci-dessus 3089 fois, alors qu'on peut en éviter une bonne partie. Comment faire ? L'idée est de "factoriser" la matrice A, c.à.d de l'écrire comme un produit $A = LU$, où L est triangulaire inférieure (lower triangular) et U triangulaire supérieure (upper triangular). On reformule alors le système $A\boldsymbol{x} = \boldsymbol{b}$ sous la forme $LU\boldsymbol{x} = \boldsymbol{b}$ et on résout maintenant deux systèmes faciles à résoudre car triangulaires : $L\boldsymbol{y} = \boldsymbol{b}$ et $U\boldsymbol{x} = \boldsymbol{y}$.
La factorisation LU de la matrice découle immédiatement de l'algorithme de Gauss. Voyons comment sur l'exemple précédent.

1/ On remarque que $U = E_2E_1A$ peut aussi s'écrire $A = LU$, avec $L = (E_2E_1)^{-1}$.

2/ On sait que $(E_2E_1)^{-1} = (E_1)^{-1}(E_2)^{-1}$.

3/ Les matrices inverses E_1^{-1} et E_2^{-1} sont faciles à déterminer : comme E_2 consiste à retrancher 1/2 fois la ligne 2 à la ligne 3, l'opération inverse consiste à ajouter 1/2 fois la ligne 2 à la ligne 3, et donc

$$E_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{1}{2} & 1 \end{bmatrix}.$$

Il est facile de voir que $E_1^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}$ et donc $L = E_1^{-1}E_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & \frac{1}{2} & 1 \end{bmatrix}$.
La matrice L est une matrice triangulaire inférieure (et c'est d'ailleurs pour cela qu'on l'appelle L, pour "lower" in English...) dont les coefficients sont particulièrement simples à trouver : les termes diagonaux sont tous égaux à un, et **chaque terme non nul sous-diagonal $\ell_{i,j}$ est égal au coefficient par lequel on a multiplié la ligne pivot i avant de la retrancher à la ligne j.**

4/ On a bien donc $A = LU$ avec L triangulaire inférieure (lower triangular) et U triangulaire supérieure (upper triangular).

4. Ceci est courant dans les applications. Par exemple on peut vouloir calculer la réponse d'une structure de génie civil à 3089 chargements différents.

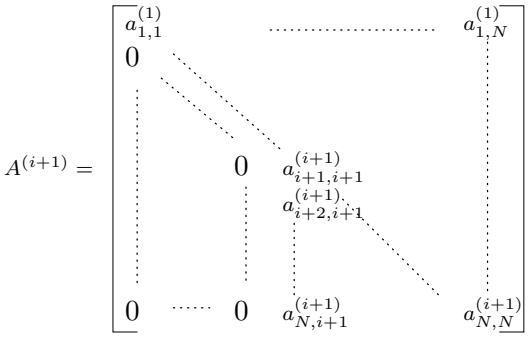

FIGURE 1.3: Allure de la matrice de Gauss à l'étape $i+1$

La procédure qu'on vient d'expliquer s'appelle **méthode LU** pour la résolution des systèmes linéaires, et elle est d'une importance considérable dans les sciences de l'ingénieur, puisqu'elle est utilisée dans les programmes informatiques pour la résolution des systèmes linéaires.
Dans l'exemple que nous avons étudié, tout se passait très bien car nous n'avons pas eu de zéro en position pivotale. Si on a un zéro en position pivotale, la factorisation peut quand même se faire, mais au prix d'une permutation. Le résultat général que l'on peut démontrer est que si la matrice A est inversible, alors il existe une matrice de permutation P, une matrice triangulaire inférieure L et une matrice triangulaire supérieure U telles que $PA = LU$: voir le théorème 1.19.

Le cas général d'une matrice $n \times n$

De manière plus générale, pour une matrice A carrée d'ordre n, la méthode de Gauss s'écrit :
On pose $A^{(1)} = A$ et $b^{(1)} = b$. Pour $i = 1, \ldots, n-1$, on cherche à calculer $A^{(i+1)}$ et $\boldsymbol{b}^{(i+1)}$ tels que les systèmes $A^{(i)}\boldsymbol{x} = \boldsymbol{b}^{(i)}$ et $A^{(i+1)}\boldsymbol{x} = \boldsymbol{b}^{(i+1)}$ soient équivalents, où $A^{(i+1)}$ est une matrice dont les coefficients sous-diagonaux des colonnes 1 à i sont tous nuls, voir figure 1.3 :
Une fois la matrice $A^{(n)}$ (triangulaire supérieure) et le vecteur $\boldsymbol{b}^{(n)}$ calculés, il sera facile de résoudre le système $A^{(n)}\boldsymbol{x} = \boldsymbol{b}^{(n)}$. Le calcul de $A^{(n)}$ est l'étape de "factorisation", le calcul de $\boldsymbol{b}^{(n)}$ l'étape de "descente", et le calcul de $\boldsymbol{x}$ l'étape de "remontée". Donnons les détails de ces trois étapes.

Etape de factorisation et descente Pour passer de la matrice $A^{(i)}$ à la matrice $A^{(i+1)}$, on va effectuer des combinaisons linéaires entre lignes qui permettront d'annuler les coefficients de la i-ème colonne situés en dessous de la ligne i (dans le but de se rapprocher d'une matrice triangulaire supérieure). Evidemment, lorsqu'on fait ceci, il faut également modifier le second membre $\boldsymbol{b}$ en conséquence. L'étape de factorisation et descente s'écrit donc :

1. Pour $k \leq i$ et pour $j = 1, \ldots, n$, on pose $a_{k,j}^{(i+1)} = a_{k,j}^{(i)}$ et $b_k^{(i+1)} = b_k^{(i)}$.
2. Pour $k > i$, **si** $a_{i,i}^{(i)} \neq 0$, on pose :

$$a_{k,j}^{(i+1)} = a_{k,j}^{(i)} - \frac{a_{k,i}^{(i)}}{a_{i,i}^{(i)}} a_{i,j}^{(i)}, \text{ pour } k = j, \ldots, n, \tag{1.30}$$

$$b_k^{(i+1)} = b_k^{(i)} - \frac{a_{k,i}^{(i)}}{a_{i,i}^{(i)}} b_i^{(i)}. \tag{1.31}$$

La matrice $A^{(i+1)}$ est de la forme donnée sur la figure 1.3. Remarquons que le système $A^{(i+1)}x = b^{(i+1)}$ est bien équivalent au système $A^{(i)}x = b^{(i)}$.
Si la condition $a_{i,i}^{(i)} \neq 0$ est vérifiée pour $i = 1$ à n, on obtient par le procédé de calcul ci-dessus un système linéaire $A^{(n)}x = b^{(n)}$ équivalent au système $Ax = b$, avec une matrice $A^{(n)}$ triangulaire supérieure facile à inverser. On

verra un peu plus loin les techniques de pivot qui permettent de régler le cas où la condition $a_{i,i}^{(i)} \neq 0$ n'est pas vérifiée.

Etape de remontée Il reste à résoudre le système $A^{(n)}x = b^{(n)}$. Ceci est une étape facile. Comme $A^{(n)}$ est une matrice inversible, on a $a_{i,i}^{(i)} \neq 0$ pour tout $i = 1, \ldots, n$, et comme $A^{(n)}$ est une matrice triangulaire supérieure, on peut donc calculer les composantes de x en "remontant", c'est–à–dire de la composante x_n à la composante x_1 :

$$x_n = \frac{b_n^{(n)}}{a_{n,n}^{(n)}},$$

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

Il est important de savoir mettre sous forme algorithmique les opérations que nous venons de décrire : c'est l'étape clef avant l'écriture d'un programme informatique qui nous permettra de faire faire le boulot par l'ordinateur !

Algorithme 1.11 (Gauss sans permutation)**.**

1. *(Factorisation et descente)*
   Pour i allant de 1 à n, on effectue les calculs suivants :
   (a) *On ne change pas la i-ème ligne (qui est la ligne du pivot)*
   $u_{i,j} = a_{i,j}$ *pour* $j = i, \ldots, n$,
   $y_i = b_i$
   (b) *On calcule les lignes $i+1$ à n de U et le second membre y en utilisant la ligne i.*
   Pour k allant de $i+1$ à n :
   $\ell_{k,i} = \frac{a_{k,i}}{a_{i,i}}$ *(si $a_{i,i} = 0$, prendre la méthode avec pivot partiel)*
   pour j allant de $i+1$ à n,
   $u_{k,j} = a_{k,j} - \ell_{k,i} u_{i,j}$ *(noter que $a_{k,i} = 0$)*
   Fin pour
   $y_k = b_k - \ell_{k,i} y_i$
   Fin pour
2. *(Remontée) On calcule x :*
   $x_n = \frac{y_n}{u_{n,n}}$
   Pour i allant de $n-1$ à 1,
   $x_i = y_i$
   Pour j allant de $i+1$ à n,
   $x_i = x_i - u_{i,j} x_j$
   Fin pour
   $x_i = \frac{1}{u_{i,i}} x_i$
   Fin pour

Coût de la méthode de Gauss (nombre d'opérations) On peut montrer (on fera le calcul de manière détaillée pour la méthode de Choleski dans la section suivante, le calcul pour Gauss est similaire) que le nombre d'opérations nécessaires n_G pour effectuer les étapes de factorisation, descente et remontée est $\frac{2}{3}n^3 + O(n^2)$; on rappelle qu'une fonction f de $\mathbb{N}$ dans $\mathbb{N}$ est $O(n^2)$ veut dire qu'i existe un réel constant C tel que $f(n) \leq Cn^2$. On a donc $\lim_{n\to+\infty} \frac{n_G}{n^3} = \frac{2}{3}$: lorsque n est grand, le nombre d'opérations se comporte comme n^3.
En ce qui concerne la place mémoire, on peut très bien stocker les itérés $A^{(i)}$ dans la matrice A de départ, ce qu'on n'a pas voulu faire dans le calcul précédent, par souci de clarté.

Décomposition LU Si le système $A\boldsymbol{x} = \boldsymbol{b}$ doit être résolu pour plusieurs second membres $\boldsymbol{b}$, on a déjà dit qu'on a intérêt à ne faire l'étape de factorisation (*i.e.* le calcul de $A^{(n)}$), qu'une seule fois, alors que les étapes de descente et remontée (*i.e.* le calcul de $b^{(n)}$ et x) seront faits pour chaque vecteur $\boldsymbol{b}$. L'étape de factorisation peut se faire en décomposant la matrice A sous la forme LU. Supposons toujours pour l'instant que lors de l'algorithme de Gauss, la condition $a_{i,i}^{(i)} \neq 0$ est vérifiée pour tout $i = 1, \ldots, n$. La matrice L a comme coefficients $\ell_{k,i} = \frac{a_{k,i}^{(i)}}{a_{i,i}^{(i)}}$ pour $k > i$, $\ell_{i,i} = 1$ pour tout $i = 1, \ldots, n$, et $\ell_{i,j} = 0$ pour $j > i$, et la matrice U est égale à la matrice $A^{(n)}$. On peut vérifier que $A = LU$ grâce au fait que le système $A^{(n)}x = b^{(n)}$ est équivalent au système $Ax = b$. En effet, comme $A^{(n)}x = b^{(n)}$ et $b^{(n)} = L^{-1}b$, on en déduit que $LUx = b$, et comme A et LU sont inversibles, on en déduit que $A^{-1}b = (LU)^{-1}b$ pour tout $b \in \mathbb{R}^n$. Ceci démontre que $A = LU$. La méthode LU se déduit donc de la méthode de Gauss en remarquant simplement que, ayant conservé la matrice L, on peut effectuer les calculs sur $\boldsymbol{b}$ après les calculs sur A, ce qui donne :

Algorithme 1.12 (LU simple (sans permutation))**.**

1. *(Factorisation) Pour i allant de 1 à n, on effectue les calculs suivants :*
   (a) *On ne change pas la i-ème ligne*
   $u_{i,j} = a_{i,j}$ *pour* $j = i, \ldots, n$,
   (b) *On calcule les lignes $i+1$ à n de U en utilisant la ligne i (mais pas le second membre).*
   Pour k allant de $i+1$ à n :
   $\ell_{k,i} = \frac{a_{k,i}}{a_{i,i}}$ *(si $a_{i,i} = 0$, prendre la méthode avec pivot partiel)*
   pour j allant de $i+1$ à n,
   $u_{k,j} = a_{k,j} - \ell_{k,i}u_{i,j}$ *(noter que $a_{k,i} = 0$)*
   Fin pour
2. *(Descente) On calcule y (avec $Ly = b$)*
   Pour i allant de 1 à n,
   $y_i = b_i - \sum_{k=1}^{i-1} \ell_{i,k}y_k$ *(on a ainsi implicitement $\ell_{i,i} = 1$)*
3. *(Remontée) On calcule x (avec $Ux = y$)*
   Pour i allant de n à 1,
   $x_i = \frac{1}{u_{i,i}}(y_i - \sum_{j=i+1}^{n} u_{i,j}x_j)$

Remarque 1.13 (Optimisation mémoire)**.** *L'introduction des matrices L et U et des vecteurs $\boldsymbol{y}$ et $\boldsymbol{x}$ n'est pas nécessaire (tout peut s'écrire avec la matrice A et le vecteur $\boldsymbol{b}$, que l'on modifie au cours de l'algorithme). L'introduction de L, U, x et y peut toutefois aider à comprendre la méthode. Le principe retenu est que, dans les algorithmes (Gauss ou LU), on modifie la matrice A et le second membre $\boldsymbol{b}$ (en remplaçant le système à résoudre par un système équivalent) mais on ne modifie jamais L, U, y et x (qui sont définis au cours de l'algorithme).*

Nous allons maintenant donner une condition nécessaire et suffisante (CNS) pour qu'une matrice A admette une décomposition LU avec U inversible et sans permutation. Commençons par un petit lemme technique qui va nous permettre de prouver cette CNS.

Lemme 1.14 (Décomposition LU de la matrice principale d'ordre k)**.** *Soit $n \in \mathbb{N}$, $A \in \mathcal{M}_n(\mathbb{R})$ et $k \in \{1, \ldots, N\}$. On appelle matrice principale d'ordre k de A la matrice $A_k \in \mathcal{M}_k(\mathbb{R})$ définie par $(A_k)_{i,j} = a_{i,j}$ pour $i = 1, \ldots, k$ et $j = 1, \ldots, k$. On suppose qu'il existe une matrice $L_k \in \mathcal{M}_k(\mathbb{R})$ triangulaire inférieure de coefficients diagonaux tous égaux à 1 et une matrice triangulaire supérieure $U_k \in \mathcal{M}_k(\mathbb{R})$ inversible, telles que $A_k = L_kU_k$. Alors A s'écrit sous la forme "par blocs" suivante :*

$$A = \begin{bmatrix} L_k & 0_{k\times(n-k)} \\ C_k & \mathrm{Id}_{n-k} \end{bmatrix} \begin{bmatrix} U_k & B_k \\ 0_{(n-k)\times k} & D_k \end{bmatrix}, \tag{1.32}$$

où $0_{p,q}$ désigne la matrince nulle de dimension $p \times q$, $B_k \in \mathcal{M}_{k,n-k}(\mathbb{R})$ et $C_k \in \mathcal{M}_{n-k,k}(\mathbb{R})$ et $D_k \in \mathcal{M}_{n-k,n-k}(\mathbb{R})$; de plus, la matrice principale d'ordre $k+1$ s'écrit sous la forme

$$A_{k+1} = \begin{bmatrix} L_k & 0_{1\times k} \\ c_k & 1 \end{bmatrix} \begin{bmatrix} U_k & b_k \\ 0_{k\times 1} & d_k \end{bmatrix} \tag{1.33}$$

où $b \in \mathcal{M}_{k,1}(\mathbb{R})$ est la première colonne de la matrice B_k, $c_k \in \mathcal{M}_{1,k}$ est la première ligne de la matrice C_k, et d_k est le coefficient de la ligne 1 et colonne 1 de D_k.

DÉMONSTRATION – On écrit la décomposition par blocs de A :

$$A = \begin{bmatrix} A_k & E_k \\ F_k & G_k \end{bmatrix},$$

avec $A_k \in \mathcal{M}_k(\mathbb{R})$, $E_k \in \mathcal{M}_{k,n-k}(\mathbb{R})$, $F_k \in \mathcal{M}_{n-k,k}(\mathbb{R})$ et $G_k \in \mathcal{M}_{n-k,n-k}(\mathbb{R})$. Par hypothèse, on a $A_k = L_kU_k$. De plus L_k et U_k sont inversibles, et il existe donc une unique matrice $B_k \in \mathcal{M}_k(\mathbb{R})$ (resp. $C_k \in \mathcal{M}_k(\mathbb{R})$) telle que $L_kB_k = E_k$ (resp $C_kU_k = G_k$). En posant $D_k = G_k - B_kC_k$, on obtient (1.32). L'égalité (1.33) en découle immédiatement. ∎

Proposition 1.15 (CNS pour LU sans permutation)**.** *Soit $n \in \mathbb{N}$, $A \in \mathcal{M}_n(\mathbb{R})$. Les deux propriétés suivantes sont équivalentes.*

(P1) Il existe un unique couple (L, U), avec L matrice triangulaire inférieure de coefficients égaux à 1 et U une matrice inversible triangulaire supérieure, tel que $A = LU$.

(P2) Les mineurs principaux[5] *de A sont tous non nuls.*

DÉMONSTRATION – Si $A = LU$ avec L triangulaire inférieure de coefficients égaux à 1 et U inversible triangulaire supérieure, alors $A_k = L_kU_k$ où les matrices L_k et U_k les matrices principales d'ordre k de L et U, qui sont encore respectivement triangulaire inférieure de coefficients égaux à 1 et inversible triangulaire supérieure. On a donc

$$\det(A_k) = \det(L_k)\det(U_k) \neq 0 \text{ pour tout } k = 1, \ldots, n,$$

et donc (P1) ⇒ (P2).

Montrons maintenant la réciproque. On suppose que les mineurs sont non nuls, et on va montrer que $A = LU$. On va en fait montrer que pour tout $k = 1, \ldots n$, on a $A_k = L_kU_k$ où L_k triangulaire inférieure de coefficients égaux à 1 et U_k inversible triangulaire supérieure. Le premier mineur est non nul, donc $a_{11} = 1 \times a_{11}$, et la récurrence est bien initialisée. On la suppose vraie à l'étape k. Par le lemme 1.14, on a donc A_{k+1} qui est de la forme (1.33), et donc une $A_{k+1} = L_{k+1}U_{k+1}$ Comme $\det(A_{k+1}) \neq 0$, la matrice U_{k+1} est inversible, et l'hypothèse de récurrence est vérifiée à l'ordre $k+1$. On a donc bien (P2) ⇒ (P1). ∎

Que faire en cas de pivot nul : la technique de permutation La caractérisation que nous venons de donner pour qu'une matrice admette une décomposition LU sans permutation est intéressante mathématiquement, mais de peu d'intérêt en pratique. On ne va en effet jamais calculer n déterminants pour savoir si non peut ou non permuter. En pratique, on effectue la décomposition LU sans savoir si on a le droit ou non de le faire, avec ou sans permutation. Au cours de l'élimination, si $a_{i,i}^{(i)} = 0$, on va permuter la ligne i avec une des lignes suivantes telle que $a_{k,i}^{(i)} \neq 0$. Notons que si le "pivot" $a_{i,i}^{(i)}$ est très petit, son utilisation peut entraîner des erreurs d'arrondi importantes dans les calculs et on va là encore permuter. En fait, même dans le cas où la CNS donnée par la proposition 1.15 est verifiée, la plupart des fonctions de libraries scientifiques vont permuter.
Plaçons-nous à l'itération i de la méthode de Gauss. Comme la matrice $A^{(i)}$ est forcément non singulière, on a :

$$\det(A^{(i)}) = a_{1,1}^{(i)}a_{2,2}^{(i)}\cdots a_{i-1,i-1}^{(i)}\det\begin{pmatrix} a_{i,i}^{(i)} & \ldots & a_{i,n}^{(i)} \\ \vdots & \ddots & \vdots \\ a_{n,i}^{(i)} & \ldots & a_{n,n}^{(i)} \end{pmatrix} \neq 0.$$

5. On rappelle que le mineur principal d'ordre k est le déterminant de la matrice prinicipale d'ordre k.

On a donc en particulier

$$\det \begin{pmatrix} a_{i,i}^{(i)} & \dots & a_{i,n}^{(i)} \\ \vdots & \ddots & \vdots \\ a_{n,i}^{(i)} & \dots & a_{n,n}^{(i)} \end{pmatrix} \neq 0.$$

On déduit qu'il existe $i_0 \in \{i, \dots, n\}$ tel que $a_{i_0,i}^{(i)} \neq 0$. On choisit alors $i_0 \in \{i, \dots, n\}$ tel que $|a_{i_0,i}^{(i)}| = \max\{|a_{k,i}^{(i)}|, k = i, \dots, n\}$. Le choix de ce max est motivé par le fait qu'on aura ainsi moins d'erreur d'arrondi. On échange alors les lignes i et i_0 (dans la matrice A et le second membre $\boldsymbol{b}$) et on continue la procédure de Gauss décrite plus haut.
L'intérêt de cette stratégie de pivot est qu'on aboutit toujours à la résolution du système (dès que A est inversible).

Remarque 1.16 (Pivot total)**.** *La méthode que nous venons de d'écrire est souvent nommée technique de pivot "partiel". On peut vouloir rendre la norme du pivot encore plus grande en considérant tous les coefficients restants et pas uniquement ceux de la colonne i. A l'etape i, on choisit maintenant i_0 et $j_0 \in \{i, \dots, n\}$ tels que $|a_{i_0,j_0}^{(i)}| = \max\{|a_{k,j}^{(i)}|, k = i, \dots, n, j = i, \dots, n\}$, et on échange alors les lignes i et i_0 (dans la matrice A et le second membre $\boldsymbol{b}$), les colonnes i et j_0 de A et les inconnues x_i et x_{j_0}. La stratégie du pivot total permet une moins grande sensibilité aux erreurs d'arrondi. L'inconvénient majeur est qu'on change la structure de A : si, par exemple la matrice avait tous ses termes non nuls sur quelques diagonales seulement, ceci n'est plus vrai pour la matrice $A^{(n)}$.*

Ecrivons maintenant l'algorithme de la méthode LU avec pivot partiel ; pour ce faire, on va simplement remarquer que l'ordre dans lequel les équations sont prises n'a aucune importance pour l'algorithme. Au départ de l'algorithme, on initialise la bijection t de $\{1, \dots, n\}$ dans $\{1, \dots, n\}$ par l'identité, c.à.d. $t(i) = i$; cette bijection t va être modifiée au cours de l'algorithme pour tenir compte du choix du pivot.

Algorithme 1.17 (LU avec pivot partiel)**.**

1. *(Initialisation de t) Pour i allant de 1 à n, $t(i) = i$. Fin pour*
2. *(Factorisation)*

   Pour i allant de 1 à n, on effectue les calculs suivants :

   (a) Choix du pivot (et de $t(i)$) : on cherche $i^* \in \{i, \dots, n\}$ t.q. $|a_{t(i^*),i}| = \max\{|a_{t(k),i}|, k \in \{i, \dots, n\}\}$ (noter que ce max est forcément non nul car la matrice est inversible).
   On modifie alors t en inversant les valeurs de $t(i)$ et $t(i^*)$.
   $p = t(i^*)$; $t(i^*) = t(i)$; $t(i) = p$.
   On ne change pas la ligne $t(i)$:
   $u_{t(i),j} = a_{t(i),j}$ *pour* $j = i, \dots, n$,

   (b) On modifie les lignes $t(k)$, $k > i$ (et le second membre), en utilisant la ligne $t(i)$.
   Pour $k = i + 1, \dots, n\}$ (noter qu'on a uniquement besoin de connaître l'ensemble , et pas l'ordre) :
   $\ell_{t(k),i} = \dfrac{a_{t(k),i}}{a_{t(i),i}}$
   Pour j allant de $i + 1$ à n,
   $u_{t(k),j} = a_{t(k),j} - \ell_{t(k),i} u_{t(i),j}$ *(noter que $u_{t(k),i} = 0$)*
   Fin pour
   Fin pour

3. *(Descente) On calcule y*
   Pour i allant de 1 à n,
   $y_{t(i)} = b_{t(i)} - \sum_{j=1}^{i-1} \ell_{t(j),k} y_k$
   Fin pour

4. *(Remontée) On calcule x*
   Pour i allant de n à 1,
   $x_{(t(i)} = \frac{1}{u_{t(i),i}}(y_i - \sum_{j=i+1}^{n} u_{t(i),j}x_j)$
   Fin pour

NB : On a changé l'ordre dans lequel les équations sont considérées (le tableau t donne cet ordre, et donc la matrice P). Donc, on a aussi changé l'ordre dans lequel interviennent les composantes du second membre : le système $Ax = b$ est devenu $PAx = Pb$. Par contre, on n'a pas touché à l'ordre dans lequel interviennent les composantes de x et y.

Il reste maintenant à signaler la propriété magnifique de cet algorithme…Il est inutile de connaitre *a priori* la bijection pour cet algorithme. A l'étape i de l'item 1 (et d'ailleurs aussi à l'étape i de l'item 2), il suffit de connaître $t(j)$ pour j allant de 1 à i, les opérations de $1(b)$ se faisant alors sur toutes les autres lignes (dans un ordre quelconque). Il suffit donc de partir d'une bijection arbitraire de $\{1, \ldots, n\}$ dans $\{1, \ldots, n\}$ (par exemple l'identité) et de la modifier à chaque étape. Pour que l'algorithme aboutisse, il suffit que $a_{t(i),i} \neq 0$ (ce qui toujours possible car A est inversible).

Remarque 1.18 (Ordre des équations et des inconnues)**.** *L'algorithme se ramène donc à résoudre $LU\boldsymbol{x} = \boldsymbol{b}$, en résolvant d'abord $L\boldsymbol{y} = b$ puis $Ux = \boldsymbol{y}$. Notons que lors de la résolution du système $L\boldsymbol{y} = \boldsymbol{b}$, les équations sont dans l'ordre $t(1), \ldots, t(k)$ (les composantes de $\boldsymbol{b}$ sont donc aussi prises dans cet ordre), mais le vecteur $\boldsymbol{y}$ est bien le vecteur de composantes $(y_1, \ldots, y_n)$, dans l'ordre initial. Puis, on résout $U\boldsymbol{x} = \boldsymbol{y}$, et les équations sont encore dans l'ordre $t(1), \ldots, t(k)$ mais les vecteurs $\boldsymbol{x}$ et $\boldsymbol{y}$ ont comme composantes respectives $(x_1, \ldots, x_n)$ et $(y_1, \ldots, y_n)$.*

Le théorème d'existence L'algorithme LU avec pivot partiel nous permet de démontrer le théorème d'existence de la décomposition LU pour una matrice inversible.

Théorème 1.19 (Décomposition LU d'une matrice)**.** *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible, il existe une matrice de permutation P telle que, pour cette matrice de permutation, il existe un et un seul couple de matrices (L, U) où L est triangulaire inférieure de termes diagonaux égaux à 1 et U est triangulaire supérieure, vérifiant*

$$PA = LU.$$

DÉMONSTRATION –
1. **L'existence** de la matrice P et des matrices L U peut s'effectuer en s'inspirant de l'algorithme "LU avec pivot partiel" 1.17). Posons $A^{(0)} = A$.
À chaque étape i de l'algorithme 1.17 peut s'écrire comme $A^{(i)} = E^{(i)}P^{(i)}A^{(i-1)}$, où $P^{(i)}$ est la matrice de permutation qui permet le choix du pivot partiel, et $E^{(i)}$ est une matrice d'élimination qui effectue les combinaisons linéaires de lignes permettant de mettre à zéro tous les coefficients de la colonne i situés en dessous de la ligne i. Pour simplifier, raisonnons sur une matrice 4×4 (le raisonnement est le même pour une matrice $n \times n$. On a donc en appliquant l'algorithme de Gauss :

$$E^{(3)}P^{(3)}E^{(2)}P^{(2)}E^{(1)}P^{(1)}A = U$$

Les matrices $P^{(i+1)}$ et $E^{(i)}$ ne permutent en gén'al pas. Prenons par exemple E_2, qui est de la forme

$$E^{(2)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & a & 1 & 0 \\ 0 & b & 0 & 1 \end{bmatrix}$$

Si $P^{(3)}$ est la matrice qui échange les lignes 3 et 4, alors

$$P^{(3)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \text{ et } P^{(3)}E^{(2)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & b & 0 & 1 \\ 0 & a & 1 & 0 \end{bmatrix}, \text{ alors que } E^{(2)}P^{(3)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & a & 0 & 1 \\ 0 & b & 1 & 0 \end{bmatrix}$$

Mais par contre, comme la multiplication *à gauche* par $P^{(i+1)}$ permute *les lignes* $i+1$ et $i+k$, pour un certain $k \geq 1$, et que la multiplication *à droite* permute *les colonnes* $i+1$ et $i+k$, la matrice $\widetilde{E^{(i)}} = P^{(i+1)}E^{(i)}P^{(i+1)}$ est encore une matrice triangulaire inférieure avec la même structure que $E^{(i)}$: on a juste échangé les coefficients extradiagonaux des lignes $i+1$ et $i+k$. On a donc

$$P^{(i+1)}E^{(i)} = \widetilde{E^{(i)}}P^{(i+1)}. \tag{1.34}$$

Dans l'exemple précédent, on effectue le calcul :

$$P^{(3)}E^{(2)}P^{(3)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & b & 1 & 0 \\ 0 & a & 0 & 1 \end{bmatrix} = \widetilde{E^{(2}},$$

qui est une matrice triangulaire inférieure de coefficients tous égaux à 1, et comme $P^{(3)}P^{(3)} = \mathrm{Id}$, on a donc :

$$P^{(3)}E^{(2)} = \widetilde{E^{(2)}}P^{(3)}.$$

Pour revenir à notre exemple $n = 4$, on peut donc écrire :

$$E^{(3)}\widetilde{E^{(2)}}P^{(3)}\widetilde{E^{(1)}}P^{(2)}P^{(1)}A = U$$

Mais par le même raisonnement que précédemment, on a $P^{(3)}\widetilde{E^{(1)}} = \widetilde{\widetilde{E^{(1)}}}P^{(3)}$ où $\widetilde{\widetilde{E^{(1)}}}$ est encore une matrice triangulaire inférieure avec des 1 sur la diagonale. On en déduit que

$$E^{(3)}\widetilde{E^{(2)}}\widetilde{\widetilde{E^{(1)}}}P^{(3)}P^{(2)}P^{(1)}A = U, \text{ soit encore } PA = LU$$

où $P = P^{(3)}P^{(2)}P^{(1)}$ bien une matrice de permutation, et $L = (E^{(3)}\widetilde{E^{(2)}}\widetilde{\widetilde{E^{(1)}}})^{-1}$ est une matrice triangulaire inférieure avec des 1 sur la diagonale.

Le raisonnement que nous venons de faire pour $n = 3$ se généralise facilement à n quelconque. Dans ce cas, l'échelonnement de la matrice s'écrit sous la forme

$$U = E^{(n-1)}P^{(n-1)}\dots E^{(2)}P^{(2)}E^{(1)}P^{(1)}A,$$

et se transforme grâce à (1.34) en

$$U = F^{(n-1)}\dots F^{(2)}F^{(1)}P^{(n-1)}\dots P^{(2)}P^{(1)}A,$$

où les matrices $F^{(i)}$ sont des matrices triangulaires inférieures de coefficients diagonaux tous égaux à 1. Plus précisément, $F^{(n-1)} = E^{(n-1)}$, $F^{(n-2)} = \widetilde{E^{(n-2)}}$, $F^{(n-3)} = \widetilde{\widetilde{E^{(n-3)}}}$, etc... On a ainsi démontré l'existence de la décomposition LU.

2. Pour montrer **l'unicité** du couple (L, U) à P donnée, supposons qu'il existe une matrice P et des matrices L_1, L_2, triangulaires inférieures et U_1, U_2, triangulaires supérieures, telles que

$$PA = L_1U_1 = L_2U_2$$

Dans ce cas, on a donc $L_2^{-1}L_1 = U_2U_1^{-1}$. Or la matrice $L_2^{-1}L_1$ est une matrice triangulaire inférieure dont les coefficients diagonaux sont tout égaux à 1, et la matrice $U_2U_1^{-1}$ est une matrice triangulaire supérieure. On en déduit que $L_2^{-1}L_1 = U_2U_1^{-1} = Id$, et donc que $L_1 = L_2$ et $U_1 = U_2$. ∎

Remarque 1.20 (Décomposition LU pour les matrices non inversibles)**.** *En fait n'importe quelle matrice carrée admet une décomposition de la forme $PA = LU$. Mais si la matrice A n'est pas inversible, son échelonnement va nous donner des lignes de zéros pour les dernières lignes . La décomposition LU n'est dans ce cas pas unique. Cette remarque fait l'objet de l'exercice 20.*

1.3.3 Méthode de Choleski

On va maintenant étudier la méthode de Choleski, qui est une méthode directe adaptée au cas où A est symétrique définie positive. On rappelle qu'une matrice $A \in \mathcal{M}_n(\mathbb{R})$ de coefficients $(a_{i,j})_{i=1,n,j=1,n}$ est symétrique si $A = A^t$, où A^t désigne la transposée de A, définie par les coefficients $(a_{j,i})_{i=1,n,j=1,n}$, et que A est définie positive si $Ax \cdot x > 0$ pour tout $x \in \mathbb{R}^n$ tel que $x \neq 0$. Dans toute la suite, $x \cdot y$ désigne le produit scalaire des deux vecteurs x et y de $\mathbb{R}^n$. On rappelle (exercice) que si A est symétrique définie positive elle est en particulier inversible.

Description de la méthode

Commençons par un exemple. On considère la matrice $A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$, qui est symétrique. Calculons sa décomposition LU. Par échelonnement, on obtient

$$A = LU = \begin{bmatrix} 1 & 0 & 0 \\ -\frac{1}{2} & 1 & 0 \\ 0 & -\frac{2}{3} & 1 \end{bmatrix} \begin{bmatrix} 2 & -1 & 0 \\ 0 & \frac{3}{2} & -1 \\ 0 & 0 & \frac{4}{3} \end{bmatrix}$$

La structure LU ne conserve pas la symétrie de la matrice A. Pour des raisons de coût mémoire, il est important de pouvoir la conserver. Une façon de faire est de décomposer U en sa partie diagonale fois une matrice triangulaire. On obtient

$$U = \begin{bmatrix} 2 & 0 & 0 \\ 0 & \frac{3}{2} & 0 \\ 0 & 0 & \frac{4}{3} \end{bmatrix} \begin{bmatrix} 1 & -\frac{1}{2} & 0 \\ 0 & 1 & -\frac{2}{3} \\ 0 & 0 & 1 \end{bmatrix}$$

On a donc $U = DL^t$, et comme tous les coefficients de D sont positifs, on peut écrire $D = \sqrt{D}\sqrt{D}$, où $\sqrt{D}$ est la matrice diagonale dont les éléments diagonaux sont les racines carrées des éléments diagonaux de A. On a donc $A = L\sqrt{D}\sqrt{D}L^t = \widetilde{L}\widetilde{L}^t$, avec $\widetilde{L} = L\sqrt{D}$. Notons que la matrice $\widetilde{L}$ est toujours triangulaire inférieure, mais ses coefficients diagonaux ne sont plus astreints à être égaux à 1. C'est la décomposition de Choleski de la matrice A.

De fait, la méthode de Choleski consiste donc à trouver une décomposition d'une matrice A symétrique définie positive de la forme $A = LL^t$, où L est triangulaire inférieure de coefficients diagonaux strictement positifs. On résout alors le système $A\boldsymbol{x} = \boldsymbol{b}$ en résolvant d'abord $L\boldsymbol{y} = \boldsymbol{b}$ puis le système $L^t\boldsymbol{x} = \boldsymbol{y}$. Une fois la matrice A "factorisée", c'est–à–dire la décomposition LL^t obtenue (voir paragraphe suivant), on effectue les étapes de "descente" et "remontée" :

1. Etape 1 : "descente" Le système $L\boldsymbol{y} = \boldsymbol{b}$ s'écrit :

$$L\boldsymbol{y} = \begin{bmatrix} \ell_{1,1} & 0 & \\ \vdots & \ddots & \vdots \\ \ell_{n,1} & \dots & \ell_{n,n} \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}.$$

Ce système s'écrit composante par composante en partant de $i = 1$.

$$\begin{aligned} \ell_{1,1}y_1 = b_1, \text{ donc} \qquad & y_1 = \frac{b_1}{\ell_{1,1}} \\ \ell_{2,1}y_1 + \ell_{2,2}y_2 = b_2, \text{ donc} \qquad & y_2 = \frac{1}{\ell_{2,2}}(b_2 - \ell_{2,1}y_1) \\ \vdots \qquad & \vdots \\ \sum_{j=1,i} \ell_{i,j}y_j = b_i, \text{ donc} \qquad & y_i = \frac{1}{\ell_{i,i}}(b_i - \sum_{j=1,i-1} \ell_{i,j}y_j) \\ \vdots \qquad & \vdots \\ \sum_{j=1,n} \ell_{n,j}y_j = b_n, \text{ donc} \qquad & y_n = \frac{1}{\ell_{n,n}}(b_n - \sum_{j=1,n-1} \ell_{n,j}y_j). \end{aligned}$$

On calcule ainsi $y_1, y_2, \dots, y_n$.

2. Etape 2 : “remontée" On calcule maintenant x solution de $L^t x = y$.

$$L^t x = \begin{bmatrix} \ell_{1,1} & \ell_{2,1} & \dots & \ell_{n,1} \\ 0 & \ddots & & \\ \vdots & & & \vdots \\ 0 & \dots & & \ell_{n,n} \end{bmatrix} \begin{bmatrix} x_1 \\ \\ \vdots \\ \\ x_n \end{bmatrix} = \begin{bmatrix} y_1 \\ \\ \vdots \\ \\ y_n \end{bmatrix}.$$

On a donc :

$$\ell_{n,n}\, x_n = y_n \text{ donc } x_n = \frac{y_n}{\ell_{n,n}}$$

$$\ell_{n-1,n-1} x_{n-1} + \ell_{n,n-1} x_n = y_{n-1} \text{ donc } x_{n-1} = \tfrac{y_{n-1} - \ell_{n,n-1} x_n}{\ell_{n-1,n-1}}$$

$$\vdots$$

$$\sum_{j=1,n} \ell_{j,1} x_j = y_1 \text{ donc } x_1 = \frac{y_1 - \sum_{j=2,n} \ell_{j,1} x_j}{\ell_{1,1}}.$$

On calcule ainsi $x_n, x_{n-1}, \dots, x_1$.

Existence et unicité de la décomposition

Soit A une matrice symétrique définie positive. On sait déjà par le théorème 1.19 page 29, qu'il existe une matrice de permutation et L triangulaire inférieure et U triangulaire supérieure telles que $PA = LU$. L'avantage dans le cas où la matrice est symétrique définie positive, est que la décomposition est toujours possible sans permutation. On prouve l'existence et unicité en construisant la décomposition, c.à.d. en construisant la matrice L.
Pour comprendre le principe de la preuve, commençons d'abord par le cas $n = 2$. Dans ce cas on peut écrire $A = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$. On sait que $a > 0$ car A est s.d.p. . L'échelonnement de A donne donc

$$A = LU = \begin{bmatrix} 1 & 0 \\ \frac{b}{a} & 1 \end{bmatrix} \begin{bmatrix} a & b \\ 0 & c - \frac{b^2}{a} \end{bmatrix}$$

En extrayant la diagonale de U, on obtient :

$$A = LU = \begin{bmatrix} 1 & 0 \\ \frac{b}{a} & 1 \end{bmatrix} \begin{bmatrix} a & 0 \\ 0 & c - \frac{b^2}{a} \end{bmatrix} \begin{bmatrix} a & \frac{b}{a} \\ 0 & 1 \end{bmatrix}.$$

Et donc

$$A = \widetilde{L}\widetilde{L}^t \text{ avec } \widetilde{L} = \begin{bmatrix} \sqrt{a} & 0 \\ b\sqrt{\frac{ac-b^2}{a}} & \end{bmatrix}.$$

Théorème 1.21 (Décomposition de Choleski)**.** *Soit $A \in \mathcal{M}_n(\mathbb{R})$ ($n \geq 1$) une matrice symétrique définie positive. Alors il existe une unique matrice $L \in \mathcal{M}_n(\mathbb{R})$, $L = (\ell_{i,j})_{i,j=1}^n$, telle que :*

1. *L est triangulaire inférieure (c'est–à–dire $\ell_{i,j} = 0$ si $j > i$),*
2. *$\ell_{i,i} > 0$, pour tout $i \in \{1, \dots, n\}$,*
3. *$A = LL^t$.*

DÉMONSTRATION –

I- Existence de L : démonstration par récurrence sur n

1. Dans le cas $n = 1$, on a $A = (a_{1,1})$. Comme A est symétrique définie positive, on a $a_{1,1} > 0$. On peut donc définir $L = (\ell_{1,1})$ où $\ell_{1,1} = \sqrt{a_{1,1}}$, et on a bien $A = LL^t$.
2. On suppose que la décomposition de Choleski s'obtient pour $A \in \mathcal{M}_p(\mathbb{R})$ symétrique définie positive, pour $1 \leq p \leq n$ et on va démontrer que la propriété est encore vraie pour $A \in \mathcal{M}_{n+1}(\mathbb{R})$ symétrique définie positive. Soit donc $A \in \mathcal{M}_{n+1}(\mathbb{R})$ symétrique définie positive ; on peut écrire A sous la forme :

$$A = \left[\begin{array}{c|c} B & a \\ \hline a^t & \alpha \end{array}\right] \tag{1.35}$$

où $B \in \mathcal{M}_n(\mathbb{R})$ est symétrique, $a \in \mathbb{R}^n$ et $\alpha \in \mathbb{R}$. Montrons que B est définie positive, c.à.d. que $By \cdot y > 0$, pour tout $y \in \mathbb{R}^n$ tel que $y \neq 0$. Soit donc $y \in \mathbb{R}^n \setminus \{0\}$, et $x = \begin{bmatrix} y \\ 0 \end{bmatrix} \in \mathbb{R}^{n+1}$. Comme A est symétrique définie positive, on a :

$$0 < Ax \cdot x = \left[\begin{array}{c|c} B & a \\ \hline a^t & \alpha \end{array}\right] \left[\begin{array}{c} y \\ \hline 0 \end{array}\right] \cdot \left[\begin{array}{c} y \\ \hline 0 \end{array}\right] = \left[\begin{array}{c} By \\ \hline a^t y \end{array}\right] \cdot \left[\begin{array}{c} y \\ \hline 0 \end{array}\right] = By \cdot y$$

donc B est définie positive. Par hypothèse de récurrence, il existe une matrice $M \in \mathcal{M}_n(\mathbb{R})$ $M = (m_{i,j})_{i,j=1}^n$ telle que :

(a) $m_{i,j} = 0$ si $j > i$

(b) $m_{i,i} > 0$

(c) $B = MM^t$.

On va chercher L sous la forme :

$$L = \left[\begin{array}{c|c} M & 0 \\ \hline b^t & \lambda \end{array}\right] \tag{1.36}$$

avec $b \in \mathbb{R}^n$, $\lambda \in \mathbb{R}_+^*$ tels que $LL^t = A$. Pour déterminer $\boldsymbol{b}$ et λ, calculons LL^t où L est de la forme (1.36) et identifions avec A :

$$LL^t = \left[\begin{array}{c|c} M & 0 \\ \hline b^t & \lambda \end{array}\right] \left[\begin{array}{c|c} M^t & b \\ \hline 0 & \lambda \end{array}\right] = \left[\begin{array}{c|c} MM^t & Mb \\ \hline b^t M^t & b^t b + \lambda^2 \end{array}\right]$$

On cherche $b \in \mathbb{R}^n$ et $\lambda \in \mathbb{R}_+^*$ tels que $LL^t = A$, et on veut donc que les égalités suivantes soient vérifiées :

$$Mb = a \text{ et } b^t b + \lambda^2 = \alpha.$$

Comme M est inversible (en effet, le déterminant de M s'écrit $\det(M) = \prod_{i=1}^n m_{i,i} > 0$), la première égalité ci-dessus donne : $b = M^{-1}a$ et en remplaçant dans la deuxième égalité, on obtient : $(M^{-1}a)^t(M^{-1}a) + \lambda^2 = \alpha$, donc $a^t(M^t)^{-1}M^{-1}a + \lambda^2 = \alpha$ soit encore $a^t(MM^t)^{-1}a + \lambda^2 = \alpha$, c'est–à–dire :

$$a^t B^{-1} a + \lambda^2 = \alpha \tag{1.37}$$

Pour que (1.37) soit vérifiée, il faut que

$$\alpha - a^t B^{-1} a > 0 \tag{1.38}$$

Montrons que la condition (1.38) est effectivement vérifiée : Soit $z = \begin{pmatrix} B^{-1}a \\ -1 \end{pmatrix} \in \mathbb{R}^{n+1}$. On a $z \neq 0$ et donc $Az \cdot z > 0$ car A est symétrique définie positive. Calculons Az :

$$Az = \left[\begin{array}{c|c} B & a \\ \hline a^t & \alpha \end{array}\right] \left[\begin{array}{c} B^{-1}a \\ \hline -1 \end{array}\right] = \begin{bmatrix} 0 \\ \\ a^t B^{-1} a - \alpha \end{bmatrix}.$$

On a donc $Az \cdot z = \alpha - a^t B^{-1} a > 0$ ce qui montre que (1.38) est vérifiée. On peut ainsi choisir $\lambda = \sqrt{\alpha - a^t B^{-1} a}$ (> 0) de telle sorte que (1.37) est vérifiée. Posons :

$$L = \left[\begin{array}{c|c} M & 0 \\ \hline (M^{-1}a)^t & \lambda \end{array}\right].$$

La matrice L est bien triangulaire inférieure et vérifie $\ell_{i,i} > 0$ et $A = LL^t$.

On a terminé ainsi la partie "existence".

II- Unicité et calcul de L. Soit $A \in \mathcal{M}_n(\mathbb{R})$ symétrique définie positive ; on vient de montrer qu'il existe $L \in \mathcal{M}_n(\mathbb{R})$ triangulaire inférieure telle que $\ell_{i,j} = 0$ si $j > i$, $\ell_{i,i} > 0$ et $A = LL^t$. On a donc :

$$a_{i,j} = \sum_{k=1}^{n} \ell_{i,k}\ell_{j,k}, \quad \forall\, (i,j) \in \{1 \dots n\}^2. \tag{1.39}$$

1. Calculons la 1-ère colonne de L ; pour $j = 1$, on a :
$$\begin{aligned} &a_{1,1} = \ell_{1,1}\ell_{1,1} \text{ donc } \ell_{1,1} = \sqrt{a_{1,1}} \quad (a_{1,1} > 0 \text{ car } \ell_{1,1} \text{ existe }),\\ &a_{2,1} = \ell_{2,1}\ell_{1,1} \text{ donc } \ell_{2,1} = \frac{a_{2,1}}{\ell_{1,1}},\\ &a_{i,1} = \ell_{i,1}\ell_{1,1} \text{ donc } \ell_{i,1} = \frac{a_{i,1}}{\ell_{1,1}} \ \forall i \in \{2, \dots, n\}. \end{aligned}$$

2. On suppose avoir calculé les q premières colonnes de L. On calcule la colonne $(q+1)$ en prenant $j = q+1$ dans (1.39)

   Pour $i = q+1$, $a_{q+1,q+1} = \sum_{k=1}^{q+1} \ell_{q+1,k}\ell_{q+1,k}$ donc

$$\ell_{q+1,q+1} = (a_{q+1,q+1} - \sum_{k=1}^{q} \ell_{q+1,k}^2)^{1/2} > 0. \tag{1.40}$$

   Notons que $a_{q+1,q+1} - \sum_{k=1}^{q} \ell_{q+1,k}^2 > 0$ car L existe : il est indispensable d'avoir d'abord montré l'existence de L pour pouvoir exhiber le coefficient $\ell_{q+1,q+1}$.
   On procède de la même manière pour $i = q+2, \dots, n$; on a :

$$a_{i,q+1} = \sum_{k=1}^{q+1} \ell_{i,k}\ell_{q+1,k} = \sum_{k=1}^{q} \ell_{i,k}\ell_{q+1,k} + \ell_{i,q+1}\ell_{q+1,q+1}$$

   et donc

$$\ell_{i,q+1} = \left(a_{i,q+1} - \sum_{k=1}^{q} \ell_{i,k}\ell_{q+1,k}\right) \frac{1}{\ell_{q+1,q+1}}. \tag{1.41}$$

   On calcule ainsi toutes les colonnes de L. On a donc montré que L est unique par un moyen constructif de calcul de L.

■

Remarque 1.22 (Choleski et LU.)**.** *Considérons une matrice A symétrique définie positive. Alors une matrice P de permutation dans le théorème 1.21 possible n'est autre que l'identité. Il suffit pour s'en convaincre de remarquer qu'une fois qu'on s'est donné la bijection $t = Id$ dans l'algorithme 1.17, celle-ci n'est jamais modifiée et donc on a $P = Id$. Les théorèmes d'existence et d'unicité 1.19 et 1.21 nous permettent alors de remarquer que $A = LU = \tilde{L}\tilde{L}^t$ avec $\tilde{L} = L\sqrt{D}$, où D est la matrice diagonale extraite de U, et $\sqrt{D}$ désigne la matrice dont les coefficients sont les racines carrées des coefficients de D (qui sont tous positifs). Voir à ce sujet l'exercice 21 41.*

La décomposition LU permet de caractériser les matrices symétriques définies positives.

Proposition 1.23 (Caractérisation des matrices symétriques définies positives par la décomposition LU)**.** *Soit A une matrice symétrique admettant une décomposition LU sans permutation, c'est-à-dire qu'on suppose qu'il existe L triangulaire inférieure de coefficients diagonaux tous égaux à 1, et U triangulaire supérieure telle que $A = LU$. Alors A est symérique définie positive si et seulement si tous les pivots (c'est-à-dire les coefficients diagonaux de la matrice U) sont strictement positifs.*

DÉMONSTRATION – Soit A une matrice symétrique admettant une décomposition LU sans permutation. Si A est symétrique définie positive, le théorème 1.21 de décomposition de Choleski donne immédiatement le résultat.
Montrons maintenant la réciproque : supposons que $A = LU$ avec tous les pivots strictement positifs. On a donc $A = LU$, et U est inversible car c'est une matrice triangulaire supérieure dont tous les coefficients diagonaux sont strictement positifs. Donc A est aussi inversible, et la décomposition LU est donc unique, par le théorème 1.19 de décomposition LU d'une matrice inversible. On a donc $A = LU = LD\tilde{L}^t$ où D est la matrice diagonale dont la diagonale est celle de U, et $\tilde{L}$ est la matrice triangulaire inférieure de coefficients diagonaux tous égaux à 1 définie par $\tilde{L}^t = D^{-1}U$. On a donc aussi par symétrie de A

$$A^t = \tilde{L}DL^t = A = LU$$

et par unicité de la décomposition LU, on en déduit que $\tilde{L} = L$ et $DL^t = U$, ce qui entraîne que $A = LDL^t = CC^t$ avec $C = L\sqrt{D}$. On a donc pour tout $\boldsymbol{x} \in \mathbb{R}^n$, $A\boldsymbol{x} \cdot \boldsymbol{x} = CC^t\boldsymbol{x} \cdot \boldsymbol{x} = \|C\boldsymbol{x}\|^2$ et donc que A est symétrique définie positive. ∎

Attention : la proposition précédente est fausse si la décomposition est avec permutation, méditer pour s'en convaincre l'exemple $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$.

Remarque 1.24 (Pivot partiel et Choleski.)**.** *Considérons une matrice A symétrique définie positive. On a vu dans le théorème qu'on n'a pas besoin de permutation pour obtenir la décomposition LL^t d'une matrice symétrique définie positive. Par contre, on utilise malgré tout la technique de pivot partiel pour minimiser les erreurs d'arrondi. On peut illustrer cette raison par l'exemple suivant :*

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

À titre d'illustration, pour $n = 12$ en FORTRAN (double précision), on obtient la bonne solution, c.à.d. $(-1, 1)$, avec le programme `gausslupivot` *donné plus haut, alors que le programme sans pivot* `gausslu` *donne comme solution $(0, 1)$.*

Calcul du coût de la méthode de Choleski

Calcul du coût de calcul de la matrice L Dans le procédé de calcul de L exposé ci-dessus, le nombre d'opérations pour calculer la première colonne est n. Calculons, pour $p = 0, \ldots, n-1$, le nombre d'opérations pour calculer la $(p+1)$-ième colonne : pour la colonne $(p+1)$, le nombre d'opérations par ligne est $2p+1$, car le calcul de $\ell_{p+1,p+1}$ par la formule (1.40) nécessite p multiplications, p soustractions et une extraction de racine, soit $2p+1$ opérations ; le calcul de $\ell_{i,p+1}$ par la formule (1.41) nécessite p multiplications, p soustractions et une division, soit encore $2p+1$ opérations. Comme les calculs se font des lignes $p+1$ à n (car $\ell_{i,p+1} = 0$ pour $i \leq p$), le nombre d'opérations pour calculer la $(p+1)$-ième colonne est donc $(2p+1)(n-p)$. On en déduit que le nombre d'opérations N_L nécessaires au calcul de L est :

$$N_L = \sum_{p=0}^{n-1}(2p+1)(n-p) = 2n\sum_{p=0}^{n-1}p - 2\sum_{p=0}^{n-1}p^2 + n\sum_{p=0}^{n-1}1 - \sum_{p=0}^{n-1}p$$

$$= (2n-1)\frac{n(n-1)}{2} + n^2 - 2\sum_{p=0}^{n-1}p^2.$$

(On rappelle que $2\sum_{p=0}^{n-1} p = n(n-1)$.) Il reste à calculer $C_n = \sum_{p=0}^{n} p^2$, en remarquant par exemple que

$$\begin{aligned}\sum_{p=0}^{n}(1+p)^3 &= \sum_{p=0}^{n} 1+p^3+3p^2+3p = \sum_{p=0}^{n} 1 + \sum_{p=0}^{n} p^3 + 3\sum_{p=0}^{n} p^2 + 3\sum_{p=0}^{n} p \\ &= \sum_{p=1}^{n+1} p^3 = \sum_{p=0}^{n} p^3 + (n+1)^3.\end{aligned}$$

On a donc $3C_n + 3\frac{n(n+1)}{2} + n + 1 = (n+1)^3$, d'où on déduit que

$$C_n = \frac{n(n+1)(2n+1)}{6}.$$

On a donc :

$$\begin{aligned}N_L &= (2n-1)\frac{n(n-1)}{2} - 2C_{n-1} + n^2 \\ &= n\left(\frac{2n^2+3n+1}{6}\right) = \frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6} = \frac{n^3}{3} + 0(n^2).\end{aligned}$$

Coût de la résolution d'un système linéaire par la méthode LL^t Nous pouvons maintenant calculer le coût (en termes de nombre d'opérations élémentaires) nécessaire à la résolution de (1.1) par la méthode de Choleski pour $A \in \mathcal{M}_n(\mathbb{R})$ symétrique définie positive. On a besoin de N_L opérations pour le calcul de L, auquel il faut rajouter le nombre d'opérations nécessaires pour les étapes de descente et remontée. Le calcul de y solution de $Ly = b$ s'effectue en résolvant le système :

$$\begin{bmatrix} \ell_{1,1} & & 0 \\ \vdots & \ddots & \vdots \\ \ell_{n,1} & \dots & \ell_{n,1} \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}$$

Pour la ligne 1, le calcul $y_1 = \dfrac{b_1}{\ell_{1,1}}$ s'effectue en une opération.

Pour les lignes $p = 2$ à n, le calcul $y_p = \left(b_p - \sum_{i=1}^{p-1} \ell_{i,p} y_i\right) / \ell_{p,p}$ s'effectue en $(p-1)$ (multiplications) $+(p-2)$ (additions) $+1$ soustraction $+1$ (division) $= 2p-1$ opérations. Le calcul de y (descente) s'effectue donc en $N_1 = \sum_{p=1}^{n}(2p-1) = n(n+1) - n = n^2$. On peut calculer de manière similaire le nombre d'opérations nécessaires pour l'étape de remontée $N_2 = n^2$. Le nombre total d'opérations pour calculer x solution de (1.1) par la méthode de Choleski est $N_C = N_L + N_1 + N_2 = \frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6} + 2n^2 = \frac{n^3}{3} + \frac{5n^2}{2} + \frac{n}{6}$. L'étape la plus coûteuse est donc la factorisation de A.

Remarque 1.25 (Décomposition LDL^t)**.** *Dans les programmes informatiques, on préfère implanter la variante suivante de la décomposition de Choleski : $A = \tilde{L}D\tilde{L}^t$ où D est la matrice diagonale définie par $d_{i,i} = \ell_{i,i}^2$, $\tilde{L}_{i,i} = L\tilde{D}^{-1}$, où $\tilde{D}$ est la matrice diagonale définie par $d_{i,i} = \ell_{i,i}$. Cette décomposition a l'avantage de ne pas faire intervenir le calcul de racines carrées, qui est une opération plus compliquée que les opérations "élémentaires" ($\times$, $+$, $-$).*

1.3.4 Quelques propriétés

Comparaison Gauss/Choleski

Soit $A \in \mathcal{M}_n(\mathbb{R})$ inversible, la résolution de (1.1) par la méthode de Gauss demande $2n^3/3 + 0(n^2)$ opérations (exercice). Dans le cas d'une matrice symétrique définie positive, la méthode de Choleski est donc environ deux fois moins chère.

Et la méthode de Cramer ?

Soit $A \in \mathcal{M}_n(\mathbb{R})$ inversible. On rappelle que la méthode de Cramer pour la résolution de (1.1) consiste à calculer les composantes de x par les formules :

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

où A_i est la matrice carrée d'ordre n obtenue à partir de A en remplaçant la i-ème colonne de A par le vecteur $\boldsymbol{b}$, et $\det(A)$ désigne le déterminant de A.
Chaque calcul de déterminant d'une matrice carrée d'ordre n nécessite au moins $n!$ opérations (voir cours L1-L2, ou livres d'algèbre linéaire proposés en avant-propos). Par exemple, pour $n = 10$, la méthode de Gauss nécessite environ 700 opérations, la méthode de Cholesky environ 350 et la méthode de Cramer plus de 4 000 000. . . . Cette dernière méthode est donc à proscrire.

Conservation du profil de A

Dans de nombreuses applications, par exemple lors de la résolution de systèmes linéaires issus de la discrétisation [6] de problèmes réels, la matrice $A \in \mathcal{M}_n(\mathbb{R})$ est "creuse", au sens où un grand nombre de ses coefficients sont nuls. Il est intéressant dans ce cas pour des raisons d'économie de mémoire de connaître le "profil" de la matrice, donné dans le cas où la matrice est symétrique, par les indices $j_i = \min\{j \in \{1, \ldots, n\}$ tels que $a_{i,j} \neq 0\}$. Le profil de la matrice est donc déterminé par les diagonales contenant des coefficients non nuls qui sont les plus éloignées de la diagonale principale. Dans le cas d'une matrice creuse, il est avantageux de faire un stockage "profil" de A, en stockant, pour chaque ligne i la valeur de j_i et des coefficients $a_{i,k}$, pour $k = i - j_i, \ldots, i$, ce qui peut permettre un large gain de place mémoire, comme on peut s'en rendre compte sur la figure **??**.

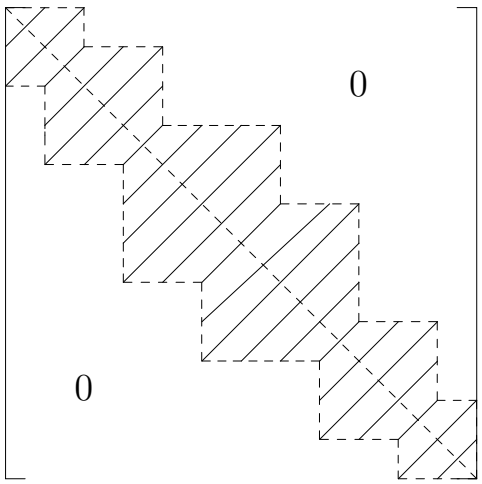

FIGURE 1.4: Exemple de profil d'une matrice symétrique

Une propriété intéressante de la méthode de Cholesky est de conserver le profil. On peut montrer (en reprenant les calculs effectués dans la deuxième partie de la démonstration du théorème 1.21) que $\ell_{i,j} = 0$ si $j < j_i$. Donc si on a adopté un stockage "profil" de A, on peut utiliser le même stockage pour L.

Matrices non symétriques

Soit $A \in \mathcal{M}_n(\mathbb{R})$ inversible. On ne suppose plus ici que A est symétrique. On cherche à calculer $x \in \mathbb{R}^n$ solution de (1.1) par la méthode de Cholesky. Ceci est possible en remarquant que : $Ax = b \Leftrightarrow A^tAx = A^tb$ car

6. On appelle discrétisation le fait de se ramener d'un problème où l'inconnue est une fonction en un problème ayant un nombre fini d'inconnues.

$\det(A) = \det(A^t) \neq 0$. Il ne reste alors plus qu'à vérifier que A^tA est symétrique définie positive. Remarquons d'abord que pour toute matrice $A \in \mathcal{M}_n(\mathbb{R})$, la matrice AA^t est symétrique. Pour cela on utilise le fait que si $B \in \mathcal{M}_n(\mathbb{R})$, alors B est symétrique si et seulement si $Bx \cdot y = x \cdot By$ et $Bx \cdot y = x \cdot B^t y$ pour tout $(x, y) \in (\mathbb{R}^n)^2$. En prenant $B = A^tA$, on en déduit que A^tA est symétrique. De plus, comme A est inversible, $A^tAx \cdot x = Ax \cdot Ax = |Ax|^2 > 0$ si $x \neq 0$. La matrice A^tA est donc bien symétrique définie positive.
La méthode de Choleski dans le cas d'une matrice non symétrique consiste donc à calculer A^tA et A^tb, puis à résoudre le système linéaire$A^tA \cdot x = A^tb$ par la méthode de Choleski "symétrique".
Cette manière de faire est plutôt moins efficace que la décomposition LU puisque le coût de la décomposition LU est de $2n^3/3$ alors que la méthode de Choleski dans le cas d'une matrice non symétrique nécessite au moins $4n^3/3$ opérations (voir exercice 22).

Systèmes linéaires non carrés

On considère ici des matrices qui ne sont plus carrées. On désigne par $\mathcal{M}_{M,n}(\mathbb{R})$ l'ensemble des matrices réelles à M lignes et n colonnes. Pour $A \in \mathcal{M}_{M,n}(\mathbb{R})$, $M > n$ et $b \in \mathbb{R}^M$, on cherche $x \in \mathbb{R}^n$ tel que

$$Ax = b. \tag{1.42}$$

Ce système contient plus d'équations que d'inconnues et n'admet donc en général pas de solution. On cherche $x \in \mathbb{R}^n$ qui vérifie le sytème (1.42) "au mieux". On introduit pour cela une fonction f définie de $\mathbb{R}^n$ dans $\mathbb{R}$ par :

$$f(x) = |Ax - b|^2,$$

où $|x| = \sqrt{x \cdot x}$ désigne la norme euclidienne sur $\mathbb{R}^n$. La fonction f ainsi définie est évidemment positive, et s'il existe x qui annule f, alors x est solution du système (1.42). Comme on l'a dit, un tel x n'existe pas forcément, et on cherche alors un vecteur x qui vérifie (1.42) "au mieux", au sens où $f(x)$ soit le plus proche de 0. On cherche donc $x \in \mathbb{R}^n$ satisfaisant (1.42) en minimisant f, c.à.d. en cherchant $x \in \mathbb{R}^n$ solution du problème d'optimisation :

$$f(x) \leq f(y) \quad \forall y \in \mathbb{R}^n \tag{1.43}$$

On peut réécrire f sous la forme : $f(x) = A^tAx \cdot x - 2b \cdot Ax + b \cdot b$. On montrera au chapitre III que s'il existe une solution au problème (1.43), elle est donnée par la résolution du système linéaire suivant :

$$AA^tx = A^tb \ \in \mathbb{R}^n, \tag{1.44}$$

qu'on appelle équations normales du problème de minimisation. La résolution approchée du problème (1.42) par cette procédure est appelée méthode des moindres carrés . La matrice AA^t étant symétrique, on peut alors employer la méthode de Choleski pour la résolution du système (1.44).

1.3.5 Exercices

Exercice 12 (Vrai ou faux ?). *Corrigé en page 42*

Les propositions suivantes sont-elles vraies ou fausses ?

1. La matrice $B = \begin{pmatrix} 1 & -2 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & 3 \end{pmatrix}$ est symétrique définie positive.
2. La matrice B ci-dessus admet une décomposition LU.
3. La matrice $\begin{bmatrix} 1 & -1 \\ 1 & 3 \end{bmatrix}$ s'écrit C^tC.
4. La matrice $A = \begin{pmatrix} 1 & 1 \\ 1 & 5 \end{pmatrix}$ admet une décomposition de Choleski $A = C^tC$ avec $C = \begin{pmatrix} -1 & -1 \\ 0 & -2 \end{pmatrix}$.

Exercice 13 (LU). *Corrigé en page 42*

1. Donner la décomposition LU de la matrice $A = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 2 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 2 & 1 & 0 \end{pmatrix}$.
2. Montrer que la matrice $A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$ vérifie $PA = LU$ avec P une matrice de permutation, L triangulaire inférieure et U triangulaire supérieure à déterminer.
3. Calculer la décomposition LU de la matrice $\begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}$

Exercice 14 (Echelonnement et factorisation LU et LDU). *Corrigé en page 43.*
Echelonner les matrices suivantes (c.à.d. appliquer l'algorithme de Gauss), et lorsqu'elle existe, donner leur décomposition LU et LDU

$$A = \begin{bmatrix} 2 & -1 & 4 & 0 \\ 4 & -1 & 5 & 1 \\ -2 & 2 & -2 & 3 \\ 0 & 3 & -9 & 4 \end{bmatrix} ; \qquad B = \begin{bmatrix} 1. & 2. & 1. & 2. \\ -1. & -1. & 0. & -2. \\ 1. & 2. & 2. & 3. \\ -1. & -1. & 1. & 0. \end{bmatrix}.$$

Exercice 15 (Décomposition LU d'une matrice à paramètres). *Corrigé en page 44.*
Soient a, b, c et d des nombres réels. On considère la matrice suivante :

$$A = \begin{bmatrix} a & a & a & a \\ a & b & b & b \\ a & b & c & c \\ a & b & c & d \end{bmatrix}.$$

Appliquer l'algorithme d'élimination de Gauss à A pour obtenir sa décomposition LU.
Donner les conditions sur a, b, c et d pour que la matrice A soit inversible.

Exercice 16 (Décomposition de Choleski d'une matrice particulière). *Soit $n \in \mathbb{N} \setminus \{0\}$. On considère la matrice A_n carrée d'ordre n dont les coefficients sont donnés par $(A_n)_{i,j} : \min(i, j)$, et qui s'écrit donc :*

$$A_n = \begin{bmatrix} 1 & 1 & \cdots & \cdots & 1 \\ 1 & 2 & \cdots & \cdots & 2 \\ \vdots & \vdots & & & \\ \vdots & \vdots & & n-1 & n-1 \\ 1 & 2 & & n-1 & n \end{bmatrix}$$

1. *Écrire et échelonner les matrices A_2 et A_3. Montrer que A_2 et A_3 sont des matrices symétriques définies positives et donner leur décomposition de Choleski.*
2. *En déduire la décomposition de Choleski de la matrice A_n.*

Exercice 17 (Décomposition LU et mineurs principaux). *Montrer que la matrice A de coefficients*

$$a_{ij} = \begin{cases} -1 \text{ si } i > j \\ 1 \text{ si } i = j \text{ et } j = n \\ 0 \text{ sinon} \end{cases}$$

admet une décomposition LU (sans permutation préalable). Calculer les coefficients diagonaux de la matrice U.

Exercice 18 (Conservation du profil)**.** *On considère des matrices A et $B \in \mathcal{M}_4(\mathbb{R})$ de la forme suivante, où x en position (i,j) de la matrice signifie que le coefficient $a_{i,j}$ est non nul) et 0 en position (i,j) de la matrice signifie que $a_{i,j} = 0$.)*

$$A = \begin{bmatrix} x & x & x & x \\ x & x & x & 0 \\ 0 & x & x & 0 \\ 0 & 0. & x. & x \end{bmatrix} \text{ et } B = \begin{bmatrix} x & x & x & 0 \\ x & x & 0 & x \\ 0 & x & x & x \\ 0 & x. & x. & x \end{bmatrix}.$$

Quels sont les coefficients nuls (notés 0 dans les matrices) qui resteront nuls dans les matrices L et U de la factorisation LU (si elle existe) ?

Exercice 19 (Matrices définies positives et décomposition LU)**.** *On rappelle que les mineurs principaux d'une matrice $A \in \mathcal{M}_n(\mathbb{R})$, sont les déterminants Δ_p des matrices $A_p = A(1:p, 1:p)$ extraites de la matrice A.*

1. *Montrer qu'une matrice symétrique définie positive a tous ses mineurs pricipaux strictement positifs.*
2. *En déduire que toute matrice symétrique définie positive admet une décomposition LU.*

Exercice 20 (Matrices non inversibles et décomposition LU)**.**

1. *Matrices 2×2*
   (a) *Soit $A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$ On suppose que $a_{11} \neq 0$.*
      i. *Echelonner la matrice A et en déduire qu'il existe une matrice $\widetilde{L}$ triangulaire inférieure dont les coefficients diagonaux sont égaux à 1, et une matrice $\widetilde{U}$ triangulaire supérieure telles que $A = \widetilde{L}\widetilde{U}$.*
      ii. *Montrer que $\widetilde{L}$ et $\widetilde{U}$ sont uniques.*
      iii. *Donner une condition nécessaire et suffisante sur les coefficients de A pour que la matrice $\widetilde{U}$ soit inversible.*
   (b) *On pose maintenant $A = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$. Trouver deux matrices $\widetilde{L}_1$ et $\widetilde{L}_2$ distinctes, toutes deux triangulaires inférieures et dont les coefficients diagonaux sont égaux à 1, et des matrices $\widetilde{U}_1$ et $\widetilde{U}_2$ triangulaires supérieures avec $A = \widetilde{L}_1\widetilde{U}_1 = \widetilde{L}_2\widetilde{U}_2$.*
2. *Matrices 3×3*
   (a) *Echelonner la matrice $A = \begin{bmatrix} 1. & 2. & 3. \\ 5. & 7. & 9. \\ 12. & 15. & 18. \end{bmatrix}$. et en déduire que la matrice A peut se décomposer en $A = \widetilde{L}\widetilde{U}$ où $\widetilde{L}$ est une matrice triangulaire inférieure dont les coefficients diagonaux sont égaux à 1, et $\widetilde{U}$ est une matrice triangulaire supérieure.*
   (b) *Soit $A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$. Montrer que si $a_{11} \neq 0$ et que la matrice $\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$ est inversible, alors il existe un unique couple de matrices $(\widetilde{L}, \widetilde{U})$ tel que $A = \widetilde{L}\widetilde{U}$, où $\widetilde{L}$ est triangulaire inférieure dont les coefficients diagonaux sont égaux à 1, et une matrice $\widetilde{U}$ triangulaire supérieure.*
3. *Matrices $n \times n$.*
   (a) *Généraliser le résultat de la question précédente à une matrice de dimension n : donner le résultat espéré sous forme de théorème et le démontrer.*
   (b) *Soit maintenant A une matrice de dimensions $n \times n$. . Montrer qu'il existe une matrice de permutation P et des matrices $\widetilde{L}$ triangulaire inférieure et de coefficients diagonaux égaux à 1, et $\widetilde{U}$ triangulaire supérieure, telles que $PA = LU$. (On pourra commencer par le cas où est de rang égal à $n-1$.)*

Exercice 21 (Décomposition LL^t "pratique")**.** *Corrigé en page 45.*
1. Soit A une matrice symétrique définie positive. Montrer que la décomposition de Choleski $\tilde{L}\tilde{L}^t$ de la matrice A est obtenue à partir de sa décomposition LU en posant $\tilde{L} = L\sqrt{D}$ où D est la matrice diagonale extraite de U. (Voir remarque 1.22.)

En déduire la décomposition LL^t de la matrice particulière $A = \begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix}$.

2. Que deviennent les coefficients nuls dans la décomposition LL^t ci-dessus ? Quelle est la propriété vue en cours qui est ainsi vérifiée ?

Exercice 22 (Sur la méthode LL^t)**.** *Corrigé détaillé en page 47.*
Soit A une matrice carrée d'ordre n, symétrique définie positive et pleine. On cherche à résoudre le système $A^2x = b$.
On propose deux méthodes de résolution de ce système :

1. *Calculer A^2, effectuer la décomposition LL^t de A^2, résoudre le système $LL^tx = b$.*
2. *Calculer la décomposition LL^t de A, résoudre les systèmes $LL^ty = b$ et $LL^tx = y$.*

Calculer le nombre d'opérations élémentaires nécessaires pour chacune des deux méthodes et comparer.

Exercice 23 (Décomposition LDL^t et LL^t)**.** *Corrigé en page 47*
1. Soit $A = \begin{pmatrix} 2 & 1 \\ 1 & 0 \end{pmatrix}$.
Calculer la décomposition LDL^t de A. Existe-t-il une décomposition LL^t de A ?

2. Montrer que toute matrice de $\mathcal{M}_n(\mathbb{R})$ symétrique définie positive admet une décomposition LDL^t.

3. Ecrire l'algorithme de décomposition LDL^t. La matrice $A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ *admet-elle une décomposition LDL^t ?*

Exercice 24 (Décomposition LL^t d'une matrice tridiagonale symétrique)**.** *Soit $A \in \mathcal{M}_n(\mathbb{R})$ symétrique définie positive et tridiagonale (i.e. $a_{i,j} = 0$ si $i - j > 1$).*

1. *Montrer que A admet une décomposition LL^t, où L est de la forme*

$$L = \begin{pmatrix} \alpha_1 & 0 & \dots & & 0 \\ \beta_2 & \alpha_2 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \dots & 0 \\ \vdots & \ddots & \ddots & \dots & \vdots \\ 0 & \cdots & 0 & \beta_n & \alpha_n \end{pmatrix}.$$

2. *Donner un algorithme de calcul des coefficients α_i et β_i, en fonction des coefficients $a_{i,j}$, et calculer le nombre d'opérations élementaires nécessaires dans ce cas.*
3. *En déduire la décomposition LL^t de la matrice :*

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

4. *L'inverse d'une matrice inversible tridiagonale est elle tridiagonale ?*

Exercice 25 (Choleski pour matrice bande)**.** *Suggestions en page 42, corrigé en page 49*
Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique définie positive.

1. *On suppose ici que A est tridiagonale. Estimer le nombre d'opérations de la factorisation LL^t dans ce cas.*
2. *Même question si A est une matrice bande (c'est-à-dire p diagonales non nulles).*
3. *En déduire une estimation du nombre d'opérations nécessaires pour la discrétisation de l'équation $-u'' = f$ vue page 11. Même question pour la discrétisation de l'équation $-\Delta u = f$ présentée page 13.*

1.3.6 Suggestions

Exercice 25 page 42

2. Soit q le nombre de sur- ou sous-diagonales ($p = 2q + 1$). Compter le nombre c_q d'opérations nécessaires pour le calcul des colonnes 1 à q et $n - q + 1$ à n, puis le nombre d_n d'opérations nécessaires pour le calcul des colonnes $n = q + 1$ $n - q$. En déduire l'estimation sur le nombre d'opérations nécessaires pour le calcul de toutes les colonnes, $Z_p(n)$, par :

$$2c_q \leq Z_p(n)2c_q + \sum_{n=q+1}^{n-q} c_n.$$

1.3.7 Corrigés

Exercice 12 page 38 (Vrai ou faux ?)

1. La matrice B n'est pas symétrique.
2. L'élimination de Gauss donne $A = LU$ avec

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \text{ et } U = \begin{bmatrix} 1 & -2 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix}.$$

   La matrice B ci-dessus admet une décomposition LU.
3. Non car elle n'est pas symétrique.
4. La matrice $A = \begin{bmatrix} 1 & 1 \\ 1 & 5 \end{bmatrix}$ admet une décomposition de Choleski $A = C^tC$ avec $C = \begin{bmatrix} -1 & -1 \\ 0 & -2 \end{bmatrix}$. Non la décomposition de Choleski fait apparaître des termes positifs sur la diagonale. Elle s'écrit

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix}.$$

Exercice 13 page 39 (Décomposition LU)

1. L'échelonnement donne

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0. & 0. \\ 0 & 0 & 1 & 0. \\ 1 & 1 & 1 & 1. \end{bmatrix} \text{ et } U = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 2 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & -3 \end{bmatrix}$$

2. La matrice A est une matrice de permutation (des lignes 2 et 3). Donc on a $P = A$ et $PA = \mathrm{Id} = LU$ avec $L = U = \mathrm{Id}$.

3. Calculer la décomposition LU de la matrice $\begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}$ L'échelonnement donne

$$L = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ 0 & \frac{2}{3} & 1 \end{bmatrix} \text{ et } U = \begin{bmatrix} 2 & 1 & 0 \\ 0 & \frac{3}{2} & 1 \\ 0 & 0 & \frac{4}{3} \end{bmatrix}$$

Exercice 14 page 39 (Décomposition LU)

Pour la première matrice, on donne le détail de l'élimination de Gauss sur cette matrice, et on montre ainsi qu'on peut stocker les multiplicateurs qu'on utilise au fur et à mesure dans la matrice L pour chaque étape k.
Etape $k = 1$

$$A = A^{(1)} = \begin{bmatrix} 2 & -1 & 4 & 0 \\ 4 & -1 & 5 & 1 \\ -2 & 2 & -2 & 3 \\ 0 & 3 & -9 & 4 \end{bmatrix} \overset{\lambda_2 \leftarrow \lambda_2 - 2\lambda_1}{\underset{\lambda_3 \leftarrow \lambda_3 + +\lambda_1}{\longrightarrow}} \begin{bmatrix} 2 & -1 & 4 & 0 \\ 0 & 1 & -3 & 1 \\ 0 & 1 & 2 & 3 \\ 0 & 3 & -9 & 4 \end{bmatrix} = A^{(2)}$$

où $\lambda_i \leftarrow \lambda_i - \alpha\lambda_j$ veut dire qu'on a soustrait α fois la ligne j à la ligne i. On a donc, sous forme matricielle,

$$A^{(2)} = E^{(1)}A^{(1)} \text{ avec } E^{(1)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Notons que $A = A^{(1)} = (E^{(1)})^{-1}A^{(2)}$ avec $(E^{(1)})^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ et donc $L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 1 & \mathrm{x} & 1 & 0 \\ \mathrm{x} & \mathrm{x} & \mathrm{x} & 1 \end{bmatrix}$

Etape $k = 2$

$$A^{(2)} = \begin{bmatrix} 2 & -1 & 4 & 0 \\ 0 & 1 & -3 & 1 \\ 0 & 1 & 2 & 3 \\ 0 & 3 & -9 & 4 \end{bmatrix} \overset{\lambda_3 \leftarrow \lambda_3 - \lambda_2}{\underset{\lambda_4 \leftarrow \lambda_4 - 3\lambda_2}{\longrightarrow}} = \begin{bmatrix} 2 & -1 & 4 & 0 \\ 0 & 1 & -3 & 1 \\ 0 & 0 & 5 & 2 \\ 0 & 0 & 0 & 1 \end{bmatrix} = A^{(3)} = E^{(2)}A^{(2)} \text{ avec } E^{(2)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & -3 & 0 & 1 \end{bmatrix}.$$

Notons que $A^{(2)} = (E^{(2)})^{-1}A^{(3)}$ avec $(E^{(2)})^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 3 & 0 & 1 \end{bmatrix}$ et donc $L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 3 & 0 & 1 \end{bmatrix}$.

Et la vie est belle... car $A^{(3)}$ est déjà triangulaire supérieure, avec tous les coefficients diagonaux non nuls (ce qui prouve A est inversible). On n'a donc pas besoin d'étape 4 :

$$U = A^{(3)} = \begin{bmatrix} 2 & -1 & 4 & 0 \\ 0 & 1 & -3 & 1 \\ 0 & 0 & 5 & 2 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

On a également $U = A^{(3)} = E^{(2)}E^{(1)}A$, soit encore $A = (E^{(1)})^{-1}(E^{(2)})^{-1}U = LU$ avec

$$L = (E^{(1)})^{-1}(E^{(2)})^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ -1 & 1 & 1 & 0 \\ 0 & 3 & 0 & 1 \end{bmatrix}$$

On peut vérifier par le calcul qu'on a bien $A = LU$. Une fois que le mécanisme d'élimination est bien compris, il est inutile de calculer les matrices $E^{k)}$: on peut directement stocker les multiplicateurs de lélimination de Gauss dans la matrice L. C'est ce quon va faire pour la prochaine matrice.

Pour la troisième matrice, le même type de raisonnement donne donc : $L = \begin{bmatrix} 1. & 0. & 0. & 0. \\ -1. & 1. & 0. & 0. \\ 1. & 0. & 1. & 0. \\ -1. & 1. & 1. & 1. \end{bmatrix}, U = \begin{bmatrix} 1. & 2. & 1. & 2. \\ 0. & 1. & 1. & 0. \\ 0. & 0. & 1. & 1. \\ 0. & 0. & 0. & 1. \end{bmatrix}$

Exercice 15 page 39 (Sur la méthode LL^t)

Appliquons l'algorithme de Gauss ; la première étape de l'élimination consiste à retrancher la première ligne à toutes les autres, c.à.d. à multiplier A à gauche par E_1, avec

$$E_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix}.$$

On obtient :

$$E_1 A = \begin{bmatrix} a & a & a & a \\ 0 & b-a & b-a & b-a \\ 0 & b-a & c-a & c-a \\ 0 & b-a & c-a & d-a \end{bmatrix}.$$

La deuxième étape consiste à multiplier A à gauche par E_2, avec

$$E_2 == \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix}.$$

On obtient :

$$E_2 E_1 A = \begin{bmatrix} a & a & a & a \\ 0 & b-a & b-a & b-a \\ 0 & 0 & c-b & c-b \\ 0 & 0 & c-b & d-b \end{bmatrix}.$$

Enfin, la troisième étape consiste à multiplier A à gauche par E_3, avec

$$E_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}.$$

On obtient :

$$E_3 E_2 E_1 A = \begin{bmatrix} a & a & a & a \\ 0 & b-a & b-a & b-a \\ 0 & 0 & c-b & c-b \\ 0 & 0 & 0 & d-c \end{bmatrix}.$$

On $A = LU$ avec $L = (E_3 E_2 E_1)^{-1} = (E_1)^{-1}(E_2)^{-1}(E_3)^{-1}$; les matrices $(E_1)^{-1}$, $(E_2)^{-1}$ et $(E_3)^{-1}$ sont faciles à calculer : la multiplication à gauche par $(E_1)^{-1}$ consiste à ajouter la première ligne à toutes les suivantes ; on calcule de la même façon $(E_2)^{-1}$ et $(E_3)^{-1}$. On obtient :

$$(E_1)^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}, \quad (E_2)^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}, \quad (E_3)^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix},$$

$$\text{et donc } L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix} \text{ et } U = \begin{bmatrix} a & a & a & a \\ 0 & b-a & b-a & b-a \\ 0 & 0 & c-b & c-b \\ 0 & 0 & 0 & d-c \end{bmatrix}.$$

La matrice L est inversible car produit de matrices élémentaires, et la matrice A est donc inversible si et seulement si la matrice U l'est. Or U est une matrice triangulaire qui est inversible si et seulement si ses éléments diagonaux sont non nuls, c.à.d. $a \neq 0$, $b \neq c$ et $c \neq d$.

Exercice 21 page 41 (Décomposition LL^t "pratique")

1. Ecrivons l'élimination de Gauss sur cette matrice, en stockant les multiplicateurs qu'on utilise au fur et à mesure dans la matrice $E^{(k)}$ pour chaque étape k.
Etape $k = 1$

$$A = A^{(1)} = \begin{bmatrix} 2 & -1 & 4 & 0 \\ 4 & -1 & 5 & 1 \\ -2 & 2 & -2 & 3 \\ 0 & 3 & -9 & 4 \end{bmatrix} \underset{\lambda_3 \leftarrow \lambda_3 + +\lambda_1}{\overset{\lambda_2 \leftarrow \lambda_2 - 2\lambda_1}{\longrightarrow}} \begin{bmatrix} 2 & -1 & 4 & 0 \\ 0 & 1 & -3 & 1 \\ 0 & 1 & 2 & 3 \\ 0 & 3 & -9 & 4 \end{bmatrix} = A^{(2)}$$

où $\lambda_i \leftarrow \lambda_i - \alpha\lambda_j$ veut dire qu'on a soustrait α fois la ligne j à la ligne i. On a donc, sous forme matricielle,

$$A^{(2)} = E^{(1)}A^{(1)} \text{ avec } E^{(1)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Notons que $A = A^{(1)} = (E^{(1)})^{-1}A^{(2)}$ avec $(E^{(1)})^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$

Etape $k = 2$

$$A^{(2)} = \begin{bmatrix} 2 & -1 & 4 & 0 \\ 0 & 1 & -3 & 1 \\ 0 & 1 & 2 & 3 \\ 0 & 3 & -9 & 4 \end{bmatrix} \underset{\lambda_4 \leftarrow \lambda_4 - 3\lambda_2}{\overset{\lambda_3 \leftarrow \lambda_3 - \lambda_2}{\longrightarrow}} = \begin{bmatrix} 2 & -1 & 4 & 0 \\ 0 & 1 & -3 & 1 \\ 0 & 0 & 5 & 2 \\ 0 & 0 & 0 & 1 \end{bmatrix} = A^{(3)} = E^{(2)}A^{(2)} \text{ avec } E^{(2)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & -3 & 0 & 1 \end{bmatrix}.$$

Notons que $A^{(2)} = (E^{(2)})^{-1}A^{(3)}$ avec $(E^{(2)})^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 3 & 0 & 1 \end{bmatrix}$.

Et la vie est belle... car $A^{(3)}$ est déjà triangulaire supérieure, avec tous les coefficients diagonaux non nuls (ce qui prouve A est inversible). On n'a donc pas besoin d'étape 4 :

$$U = A^{(3)} = \begin{bmatrix} 2 & -1 & 4 & 0 \\ 0 & 1 & -3 & 1 \\ 0 & 0 & 5 & 2 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

On a également $U = A^{(3)} = E^{(2)}E^{(1)}A$, soit encore $A = (E^{(1)})^{-1}(E^{(2)})^{-1}U = LU$ avec

$$L = (E^{(1)})^{-1}(E^{(2)})^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ -1 & 1 & 1 & 0 \\ 0 & 3 & 0 & 1 \end{bmatrix}$$

2. Si A est une matrice symétrique définie positive, on sait par le théorème 1.19 et la remarque 1.22 qu'il existe une unique décomposition LU : $A = LU$. Le théorème 1.21 nous donne l'existence (et l'unicité) de la décomposition $A = \tilde{L}\tilde{L}^t$. Soit $\tilde{D}$ la matrice diagonale extraite de $\tilde{L}$, qui est strictement positive par construction de $\tilde{L}$; on pose $\bar{L} = \tilde{L}\tilde{D}^{-1}$. On a donc $A = \bar{L}\tilde{D}\tilde{D}\bar{L}^t = \bar{L}\bar{U}$, avec $\bar{U} = \tilde{D}^2\bar{L}^t$. La matrice $\bar{D} = \tilde{D}^2$ est donc la diagonale de la matrice $\bar{U}$. Par unicité de la décomposition LU, on a $\bar{L} = L$, $\bar{U} = U$ et $\bar{D} = D$, et donc $\tilde{L} = L\sqrt{D}$.

Montrons maintenant que $A = \begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix}$ est s.d.p (symétrique définite positive). Elle est évidemment symétrique. Soit $x = (a, b, c, d) \in \mathbb{R}^4$. Calculons $Ax \cdot x$:

$$Ax \cdot x = \begin{bmatrix} 2a - b \\ -a + 2b - c \\ -b + 2c - d \\ -c + 2d \end{bmatrix} \cdot \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix}$$

Donc $Ax \cdot x = 2a^2 - ab - ab + 2b^2 - bc - bc + 2c^2 - cd - cd + 2d^2 = a^2 + (a-b)^2 + (b-c)^2 + (c-d)^2 + d^2 \geq 0$. De plus $Ax \cdot x = 0$ ssi $a = b = c = d = 0$. Donc A est sdp.

On peut soit appliquer ici l'algorithme de construction de la matrice donné dans la partie unicité de la preuve du théorème 1.21 d'existence et d'unicité de la décomposition de Choleski, soit procéder comme en 1, calculer la décomposition LU habituelle, puis calculer la décomposition de $A = LU$, écrire $A = \tilde{L}\tilde{L}^t$ avec $\tilde{L} = L\sqrt{D}$, où $\sqrt{D}$et D la matrice diagonale extraite de U, comme décrit plus haut. Nous allons procéder selon le deuxième choix, qui est un peu plus rapide à écrire. (on utilise ici la notation $\tilde{L}$ parce que les matrices L dans les décompositions LU et LL^t ne sont pas les mêmes...)

Etape $k = 1$

$$A = A^{(1)} = \begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix} \underset{\lambda_2 \leftarrow \lambda_2 + \frac{1}{2}\lambda_1}{\longrightarrow} \begin{bmatrix} 2 & -1 & 0 & 0 \\ 0 & \frac{3}{2} & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix} = A^{(2)}$$

Etape $k = 2$

$$A^{(2)} = \begin{bmatrix} 2 & -1 & 0 & 0 \\ 0 & \frac{3}{2} & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix} \underset{\lambda_3 \leftarrow \lambda_3 + \frac{2}{3}\lambda_2}{\longrightarrow} \begin{bmatrix} 2 & -1 & 0 & 0 \\ 0 & \frac{3}{2} & -1 & 0 \\ 0 & 0 & \frac{4}{3} & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix} = A^{(3)}$$

Etape $k = 3$

$$A^{(3)} = \begin{bmatrix} 2 & -1 & 0 & 0 \\ 0 & \frac{3}{2} & -1 & 0 \\ 0 & 0 & \frac{4}{3} & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix} \underset{\lambda_4 \leftarrow \lambda_4 + \frac{3}{4}\lambda_3}{\longrightarrow} \begin{bmatrix} 2 & -1 & 0 & 0 \\ 0 & \frac{3}{2} & -1 & 0 \\ 0 & 0 & \frac{4}{3} & -1 \\ 0 & 0 & 0 & \frac{5}{4} \end{bmatrix} = A^{(4)}$$

On vérifie alors qu'on a bien $U = A^{(4)} = DL^t$ où L est la matrice inverse du produit des matrices élémentaires utilisées pour transformer A en une matrice élémentaire (même raisonnement qu'en 1), c.à.d.

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -\frac{1}{2} & 1 & 0 & 0 \\ 0 & -\frac{2}{3} & 1 & 0 \\ 0 & 0 & -\frac{3}{4} & 1 \end{bmatrix}$$

On en déduit la décomposition $A = \tilde{L}\tilde{L}^t$ avec

$$\tilde{L} = \begin{bmatrix} \sqrt{2} & 0 & 0 & 0 \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{6}}{2} & 0 & 0 \\ 0 & -\frac{\sqrt{6}}{3} & \frac{2\sqrt{3}}{3} & 0 \\ 0 & 0 & -\frac{\sqrt{3}}{2} & \frac{\sqrt{5}}{2} \end{bmatrix}$$

3. Que deviennent les coefficients nuls dans la décomposition LL^t ci-dessus ? Quelle est la propriété vue en cours qui est ainsi vérifiée ?

Ils restent nuls : le profil est préservé, comme expliqué dans le cours page 17.

Exercice 22 page 41 (Sur la méthode LL^t)

Calculons le nombre d'opérations élémentaires nécessaires pour chacune des méthodes :

1. Le calcul de chaque coefficient nécessite n multiplications et $n-1$ additions, et la matrice comporte n^2 coefficients. Comme la matrice est symétrique, seuls $n(n+1)/2$ coefficients doivent être calculés. Le calcul de A^2 nécessite nécessite donc $\frac{(2n-1)n(n+1)}{2}$ opérations élémentaires.
   Le nombre d'opérations élémentaires pour effectuer la décomposition LL^t de A^2 nécessite $\frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6}$ (cours).
   La résolution du système $A^2x = b$ nécessite $2n^2$ opérations (n^2 pour la descente, n^2 pour la remontée, voir cours).
   Le nombre total d'opérations pour le calcul de la solution du système $A^2x = b$ par la première méthode est donc $\frac{(2n-1)n(n+1)}{2} + \frac{n^3}{3} + \frac{3n^2}{2} + \frac{n}{6} = \frac{4n^3}{3} + O(n^2)$ opérations.
2. La décomposition LL^t de A nécessite $\frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6}$, et la résolution des systèmes $LL^ty = b$ et $LL^tx = y$ nécessite $4n^2$ opérations. Le nombre total d'opérations pour le calcul de la solution du système $A^2x = b$ par la deuxième méthode est donc $\frac{n^3}{3} + \frac{9n^2}{2} + \frac{n}{6} = \frac{n^3}{3} + O(n^2)$ opérations.

Pour les valeurs de n assez grandes, il est donc avantageux de choisir la deuxième méthode.

Exercice 23 page 41 (Décompositions LL^t et LDL^t)

1. On pose $L = \begin{pmatrix} 1 & 0 \\ \gamma & 1 \end{pmatrix}$ et $D = \begin{pmatrix} \alpha & 0 \\ 0 & \beta \end{pmatrix}$.
Par identification, on obtient $\alpha = 2$, $\beta = -\frac{1}{2}$ et $\gamma = \frac{1}{2}$.

Si maintenant on essaye d'écrire $A = LL^t$ avec $L = \begin{pmatrix} a & 0 \\ b & c \end{pmatrix}$, on obtient $c^2 = -\frac{1}{2}$ ce qui est impossible dans $\mathbb{R}$.

En fait, on peut remarquer qu'il est normal que A n'admette pas de décomposition LL^t, car elle n'est pas définie positive. En effet, soit $x = (x_1, x_2)^t \in \mathbb{R}^2$,, alors $Ax \cdot x = 2x_1(x_1 + x_2)$, et en prenant $x = (1, -2)^t$, on a $Ax \cdot x < 0$.

2. 2. Reprenons en l'adaptant la démonstration du théorème 1.3. On raisonne donc par récurrence sur la dimension.

1. Dans le cas $n = 1$, on a $A = (a_{1,1})$. On peut donc définir $L = (\ell_{1,1})$ où $\ell_{1,1} = 1$, $D = (a_{1,1})$, $d_{1,1} \neq 0$, et on a bien $A = LDL^t$.
2. On suppose que, pour $1 \leq p \leq n$, la décomposition $A = LDL^t$ s'obtient pour $A \in \mathcal{M}_p(\mathbb{R})$ symétrique définie positive ou négative, avec $d_{i,i} \neq 0$ pour $1 \leq i \leq n$ et on va démontrer que la propriété est encore vraie pour $A \in \mathcal{M}_{n+1}(\mathbb{R})$ symétrique définie positive ou négative. Soit donc $A \in \mathcal{M}_{n+1}(\mathbb{R})$ symétrique définie positive ou négative ; on peut écrire A sous la forme :

$$A = \left[\begin{array}{c|c} B & a \\ \hline a^t & \alpha \end{array}\right] \tag{1.45}$$

où $B \in \mathcal{M}_n(\mathbb{R})$ est symétrique définie positive ou négative (calculer $Ax\cdot x$ avec $x=(y,0)^t$, avec $y\in\mathbb{R}^n$ pour le vérifier), $a\in\mathbb{R}^n$ et $\alpha\in\mathbb{R}$.

Par hypothèse de récurrence, il existe une matrice $M\in\mathcal{M}_n(\mathbb{R})$ $M=(m_{i,j})_{i,j=1}^n$ et une matrice diagonale $\tilde{D}=\mathrm{diag}(d_{1,1},d_{2,2},\dots,d_{n,n})$ dont les coefficients sont tous non nuls, telles que :

(a) $m_{i,j}=0$ si $j>i$

(b) $m_{i,i}=1$

(c) $B=M\tilde{D}M^t$.

On va chercher L et D sous la forme :

$$L=\left[\begin{array}{c|c} M & 0 \\ \hline b^t & 1 \end{array}\right],\ D=\left[\begin{array}{c|c} \tilde{D} & 0 \\ \hline 0 & \lambda \end{array}\right], \tag{1.46}$$

avec $b\in\mathbb{R}^n$, $\lambda\in\mathbb{R}$ tels que $LDL^t=A$. Pour déterminer b et λ, calculons LDL^t avec L et D de la forme (1.46) et identifions avec A :

$$LDL^t=\left[\begin{array}{c|c} M & 0 \\ \hline b^t & 1 \end{array}\right]\left[\begin{array}{c|c} \tilde{D} & 0 \\ \hline 0 & \lambda \end{array}\right]\left[\begin{array}{c|c} M^t & b \\ \hline 0 & 1 \end{array}\right]=\left[\begin{array}{c|c} M\tilde{D}M^t & M\tilde{D}b \\ \hline b^t\tilde{D}M^t & b^t\tilde{D}b+\lambda \end{array}\right]$$

On cherche $b\in\mathbb{R}^n$ et $\lambda\in\mathbb{R}$ tels que $LDL^t=A$, et on veut donc que les égalités suivantes soient vérifiées :

$$M\tilde{D}b=a \text{ et } b^t\tilde{D}b+\lambda=\alpha.$$

La matrice M est inversible (en effet, le déterminant de M s'écrit $det(M)=\prod_{i=1}^n 1=1$). Par hypothèse de récurrence, la matrice $\tilde{D}$ est aussi inversible. La première égalité ci-dessus donne : $b=\tilde{D}^{-1}M^{-1}a$. On calcule alors $\lambda=\alpha-b^tM^{-1}a$. Remarquons qu'on a forcément $\lambda\neq 0$, car si $\lambda=0$,

$$A=LDL^t=\left[\begin{array}{c|c} M\tilde{D}M^t & M\tilde{D}b \\ \hline b^t\tilde{D}M^t & b^t\tilde{D}b \end{array}\right]$$

qui n'est pas inversible. En effet, si on cherche $(x,y)\in\mathbb{R}^n\times\mathbb{R}$ solution de

$$\left[\begin{array}{c|c} M\tilde{D}M^t & M\tilde{D}b \\ \hline b^t\tilde{D}M^t & b^t\tilde{D}b \end{array}\right]\left[\begin{array}{c} x \\ \hline y \end{array}\right]=\left[\begin{array}{c} 0 \\ \hline 0 \end{array}\right],$$

on se rend compte facilement que tous les couples de la forme $(-M^{-t}by,y)^t$, $y\in\mathbb{R}$, sont solutions. Le noyau de la matrice n'est donc pas réduit à $\{0\}$ et la matrice n'est donc pas inversible. On a ainsi montré que $d_{n+1,n+1}\neq 0$ ce qui termine la récurrence.

3. On reprend l'algorithme de décomposition LL^t :

Soit $A\in\mathcal{M}_n(\mathbb{R})$ symétrique définie positive ou négative ; on vient de montrer qu'il existe une matrice $L\in\mathcal{M}_n(\mathbb{R})$ triangulaire inférieure telle que $\ell_{i,j}=0$ si $j>i$, $\ell_{i,i}=1$, et une matrice $D\in\mathcal{M}_n(\mathbb{R})$ diagonale inversible, telles que et $A=LDL^t$. On a donc :

$$a_{i,j}=\sum_{k=1}^{n}\ell_{i,k}d_{k,k}\ell_{j,k},\quad \forall\,(i,j)\in\{1,\dots,n\}^2. \tag{1.47}$$

1. Calculons la 1ère colonne de L ; pour $j = 1$, on a :

$$\begin{aligned} &a_{1,1} = d_{1,1} \text{ donc } d_{1,1} = a_{1,1}, \\ &a_{2,1} = \ell_{2,1} d_{1,1} \text{ donc } \ell_{2,1} = \frac{a_{2,1}}{d_{1,1}}, \\ &a_{i,1} = \ell_{i,1}\ell_{1,1} \text{ donc } \ell_{i,1} = \frac{a_{i,1}}{d_{1,1}} \ \forall i \in \{2, \ldots, n\}. \end{aligned}$$

2. On suppose avoir calculé les n premières colonnes de L. On calcule la colonne $(k+1)$ en prenant $j = n+1$ dans (1.39).

   Pour $i = n+1$, $a_{n+1,n+1} = \sum_{k=1}^{n} \ell_{n+1,k}^2 d_{k,k} + d_{n+1,n+1}$ donc

$$d_{n+1,n+1} = a_{n+1,n+1} - \sum_{k=1}^{n} \ell_{n+1,k}^2 d_{k,k}. \tag{1.48}$$

   On procède de la même manière pour $i = n+2, \ldots, n$; on a :

$$a_{i,n+1} = \sum_{k=1}^{n+1} \ell_{i,k} d_{k,k} \ell_{n+1,k} = \sum_{k=1}^{n} \ell_{i,k} d_{k,k} \ell_{n+1,k} + \ell_{i,n+1} d_{n+1,n+1} \ell_{n+1,n+1},$$

   et donc, comme on a montré dans la question 2 que les coefficients $d_{k,k}$ sont tous non nuls, on peut écrire :

$$\ell_{i,n+1} = \left(a_{i,n+1} - \sum_{k=1}^{n} \ell_{i,k} d_{k,k} \ell_{n+1,k} \right) \frac{1}{d_{n+1,n+1}}. \tag{1.49}$$

Exercice 25 page 42 (Décomposition LL^t d'une matrice bande)

On utilise le résultat de conservation du profil de la matrice énoncé dans le cours. Comme A est symétrique, le nombre p de diagonales de la matrice A est forcément impair si A ; notons $q = \frac{p-1}{2}$ le nombre de sous- et sur-diagonales non nulles de la matrice A, alors la matrice L aura également q sous-diagonales non nulles.

1. Cas d'une matrice tridiagonale. Si on reprend l'algorithme de construction de la matrice L vu en cours, on remarque que pour le calcul de la colonne $n+1$, avec $1 \leq n < n-1$, on a le nombre d'opérations suivant :

— Calcul de $\ell_{n+1,n+1} = (a_{n+1,n+1} - \sum_{k=1}^{n} \ell_{n+1,k}\ell_{n+1,k})^{1/2} > 0$:
une multiplication, une soustraction, une extraction de racine, soit 3 opérations élémentaires.

— Calcul de $\ell_{n+2,n+1} = \left(a_{n+2,n+1} - \sum_{k=1}^{n} \ell_{n+2,k}\ell_{n+1,k} \right) \frac{1}{\ell_{n+1,n+1}}$:
une division seulement car $\ell_{n+2,k} = 0$.

On en déduit que le nombre d'opérations élémentaires pour le calcul de la colonne $n+1$, avec $1 \leq n < n-1$, est de 4.
Or le nombre d'opérations pour la première et dernière colonnes est inférieur à 4 (2 opérations pour la première colonne, une seule pour la dernière). Le nombre $Z_1(n)$ d'opérations élémentaires pour la décomposition LL^t de A peut donc être estimé par : $4(n-2) \leq Z_1(n) \leq 4n$, ce qui donne que $Z_1(n)$ est de l'ordre de $4n$ (le calcul exact du nombre d'opérations, inutile ici car on demande une estimation, est $4n-3$.)

2. Cas d'une matrice à p diagonales.
On cherche une estimation du nombre d'opérations $Z_p(n)$ pour une matrice à p diagonales non nulles (ou q sous-diagonales non nulles) en fonction de n.

On remarque que le nombre d'opérations nécessaires au calcul de

$$\ell_{n+1,n+1} = (a_{n+1,n+1} - \sum_{k=1}^{n} \ell_{n+1,k}\ell_{n+1,k})^{1/2} > 0, \tag{1.50}$$

$$\text{et } \ell_{i,n+1} = \left(a_{i,n+1} - \sum_{k=1}^{n} \ell_{i,k}\ell_{n+1,k} \right) \frac{1}{\ell_{n+1,n+1}}, \tag{1.51}$$

est toujours inférieur à $2q+1$, car la somme $\sum_{k=1}^{n}$ fait intervenir au plus q termes non nuls.
De plus, pour chaque colonne $n+1$, il y a au plus $q+1$ coefficients $\ell_{i,n+1}$ non nuls, donc au plus $q+1$ coefficients à calculer. Donc le nombre d'opérations pour chaque colonne peut être majoré par $(2q+1)(q+1)$.
On peut donc majorer le nombre d'opérations z_q pour les q premières colonnes et les q dernières par $2q(2q+1)(q+1)$, qui est indépendant de n (on rappelle qu'on cherche une estimation en fonction de n, et donc le nombre z_q est $O(1)$ par rapport à n.)
Calculons maintenant le nombre d'opérations x_n nécessaires une colonne $n = q+1$ à $n-q-1$. Dans (1.50) et (1.51), les termes non nuls de la somme sont pour $k = i-q, \ldots, n$, et donc on a $(n-i+q+1)$ multiplications et additions, une division ou extraction de racine. On a donc

$$\begin{aligned} x_n &= \sum_{i=n+1}^{n+q+1} \big(2(n-i+q+1)+1\big) \\ &= \sum_{j=1}^{q+1} \big(2(-j+q+1)+1\big) \\ &= (q+1)(2q+3) - 2\sum_{j=1}^{q+1} j \\ &= (q+1)^2. \end{aligned}$$

Le nombre z_i d'opérations nécessaires pour les colonnes $n = q+1$ à $n-q-1$ est donc

$$z_i = (q+1)^2(n-2q).$$

Un encadrement du nombre d'opérations nécessaires pour la décomposition LL^t d'une matrice à p diagonales est donc donnée par :

$$(q+1)^2(n-2q) \leq Z_p(n) \leq (q+1)^2(n-2q) + 2q(2q+1)(q+1), \tag{1.52}$$

et que, à q constant, $Z_p(n) = O((q+1)^2 n))$. Remarquons qu'on retrouve bien l'estimation obtenue pour $q = 1$.

3. Dans le cas de la discrétisation de l'équation $-u'' = f$ (voir page 11), on a $q = 1$ et la méthode de Choleski nécessite de l'ordre de $4n$ opérations élémentaires, alors que dans le cas de la discrétisation de l'équation $-\Delta u = f$ (voir page 13), on a $q = \sqrt{n}$ et la méthode de Choleski nécessite de l'ordre de n^2 opérations élémentaires (dans les deux cas n est le nombre d'inconnues).

On peut noter que l'encadrement (1.52) est intéressant dès que q est d'ordre inférieur à n^α, $\alpha < 1$.

1.4 Normes et conditionnement d'une matrice

Dans ce paragraphe, nous allons définir la notion de conditionnement d'une matrice, qui peut servir à établir une majoration des erreurs d'arrondi dues aux erreurs sur les données. Malheureusement, nous verrons également que cette majoration n'est pas forcément très utile dans des cas pratiques, et nous nous efforcerons d'y remédier. La notion de conditionnement est également utilisée dans l'étude des méthodes itératives que nous verrons plus loin. Pour l'étude du conditionnement comme pour l'étude des erreurs, nous avons tout d'abord besoin de la notion de norme et de rayon spectral, que nous rappelons maintenant.

1.4.1 Normes, rayon spectral

Définition 1.26 (Norme matricielle, norme induite). *On note $\mathcal{M}_n(\mathbb{R})$ l'espace vectoriel (sur $\mathbb{R}$) des matrices carrées d'ordre n.*

1. *On appelle norme matricielle sur $\mathcal{M}_n(\mathbb{R})$ une norme $\|\cdot\|$ sur $\mathcal{M}_n(\mathbb{R})$ t.q.*

$$\|AB\| \leq \|A\|\|B\|, \ \forall A, B \in \mathcal{M}_n(\mathbb{R}) \tag{1.53}$$

2. *On considère $\mathbb{R}^n$ muni d'une norme $\|\cdot\|$. On appelle norme matricielle induite (ou norme induite) sur $\mathcal{M}_n(\mathbb{R})$ par la norme $\|\cdot\|$, encore notée $\|\cdot\|$, la norme sur $\mathcal{M}_n(\mathbb{R})$ définie par :*

$$\|A\| = \sup\{\|A\boldsymbol{x}\|; \ \boldsymbol{x} \in \mathbb{R}^n, \|\boldsymbol{x}\| = 1\}, \ \forall A \in \mathcal{M}_n(\mathbb{R}) \tag{1.54}$$

Proposition 1.27 (Propriétés des normes induites). *Soit $\mathcal{M}_n(\mathbb{R})$ muni d'une norme induite $\|\cdot\|$. Alors pour toute matrice $A \in \mathcal{M}_n(\mathbb{R})$, on a :*

1. $\|A\boldsymbol{x}\| \leq \|A\| \ \|\boldsymbol{x}\|, \quad \forall \boldsymbol{x} \in \mathbb{R}^n$,
2. $\|A\| = \max\{\|A\boldsymbol{x}\| \ ; \ \|\boldsymbol{x}\| = 1, \boldsymbol{x} \in \mathbb{R}^n\}$,
3. $\|A\| = \max\left\{\dfrac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \ ; \ \boldsymbol{x} \in \mathbb{R}^n \setminus \{0\}\right\}$.
4. $\|\cdot\|$ *est une norme matricielle.*

DÉMONSTRATION

1. Soit $\boldsymbol{x} \in \mathbb{R}^n \setminus \{0\}$, posons $\boldsymbol{y} = \dfrac{\boldsymbol{x}}{\|\boldsymbol{x}\|}$, alors $\|\boldsymbol{y}\| = 1$ donc $\|A\boldsymbol{y}\| \leq \|A\|$. On en déduit que $\dfrac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leq \|A\|$ et donc que $\|A\boldsymbol{x}\| \leq \|A\| \ \|\boldsymbol{x}\|$. Si maintenant $\boldsymbol{x} = 0$, alors $A\boldsymbol{x} = 0$, et donc $\|\boldsymbol{x}\| = 0$ et $\|A\boldsymbol{x}\| = 0$; l'inégalité $\|A\boldsymbol{x}\| \leq \|A\| \ \|\boldsymbol{x}\|$ est encore vérifiée.
2. L'application φ définie de $\mathbb{R}^n$ dans $\mathbb{R}$ par : $\varphi(\boldsymbol{x}) = \|A\boldsymbol{x}\|$ est continue sur la sphère unité $S_1 = \{\boldsymbol{x} \in \mathbb{R}^n \,|\, \|\boldsymbol{x}\| = 1\}$ qui est un compact de $\mathbb{R}^n$. Donc φ est bornée et atteint ses bornes : il existe $\boldsymbol{x}_0 \in \mathbb{R}^n$ tel que $\|A\| = \|A\boldsymbol{x}_0\|$.
3. Cette égalité résulte du fait que
$$\frac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \|A\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}\| \text{ et } \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} \in S_1 \text{ et } \boldsymbol{x} \neq 0.$$
4. Soient A et $B \in \mathcal{M}_n(\mathbb{R})$, on a $\|AB\| = \max\{\|AB\boldsymbol{x}\| \ ; \ \|\boldsymbol{x}\| = 1, \boldsymbol{x} \in \mathbb{R}^n\}$. Or
$$\|AB\boldsymbol{x}\| \leq \|A\|\|B\boldsymbol{x}\| \leq \|A\|\|B\|\|\boldsymbol{x}\| \leq \|A\|\|B\|.$$
On en déduit que $\|\cdot\|$ est une norme matricielle.

■

Définition 1.28 (Rayon spectral). *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible. On appelle rayon spectral de A la quantité $\rho(A) = \max\{|\lambda|; \lambda \in \mathbb{C}, \lambda$ valeur propre de $A\}$.*

La proposition suivante caractérise les principales normes matricielles induites.

Proposition 1.29 (Caractérisation de normes induites). *Soit $A = (a_{i,j})_{i,j\in\{1,\ldots,n\}} \in \mathcal{M}_n(\mathbb{R})$.*

1. *On munit $\mathbb{R}^n$ de la norme $\|\cdot\|_\infty$ et $\mathcal{M}_n(\mathbb{R})$ de la norme induite correspondante, notée aussi $\|\cdot\|_\infty$. Alors*

$$\|A\|_\infty = \max_{i\in\{1,\ldots,n\}} \sum_{j=1}^{n} |a_{i,j}|. \tag{1.55}$$

2. *On munit $\mathbb{R}^n$ de la norme $\|\cdot\|_1$ et $\mathcal{M}_n(\mathbb{R})$ de la norme induite correspondante, notée aussi $\|\cdot\|_1$. Alors*

$$\|A\|_1 = \max_{j\in\{1,\dots,n\}} \sum_{i=1}^{n} |a_{i,j}| \tag{1.56}$$

3. *On munit $\mathbb{R}^n$ de la norme $\|\cdot\|_2$ et $\mathcal{M}_n(\mathbb{R})$ de la norme induite correspondante, notée aussi $\|\cdot\|_2$.*

$$\|A\|_2 = (\rho(A^tA))^{\frac{1}{2}}. \tag{1.57}$$

En particulier, si A est symétrique, $\|A\|_2 = \rho(A)$.

DÉMONSTRATION – La démonstration des points 1 et 2 fait l'objet de l'exercice 27 page 61. On démontre ici uniquement le point 3.
Par définition de la norme 2, on a :

$$\|A\|_2^2 = \sup_{\substack{\boldsymbol{x}\in\mathbb{R}^n\\ \|\boldsymbol{x}\|_2=1}} A\boldsymbol{x}\cdot A\boldsymbol{x} = \sup_{\substack{\boldsymbol{x}\in\mathbb{R}^n\\ \|\boldsymbol{x}\|_2=1}} A^tA\boldsymbol{x}\cdot\boldsymbol{x}.$$

Comme A^tA est une matrice symétrique positive (car $A^tA\boldsymbol{x}\cdot\boldsymbol{x} = A\boldsymbol{x}\cdot A\boldsymbol{x} \geq 0$), il existe une base orthonormée $(\boldsymbol{f}_i)_{i=1,\dots,n}$ et des valeurs propres $(\mu_i)_{i=1,\dots,n}$, avec $0 \leq \mu_1 \leq \mu_2 \leq \dots \leq \mu_n$ tels que $A\boldsymbol{f}_i = \mu_i\boldsymbol{f}_i$ pour tout $i \in \{1,\dots,n\}$. Soit $\boldsymbol{x} = \sum_{i=1,\dots,n}\alpha_i\boldsymbol{f}_i \in \mathbb{R}^n$. On a donc :

$$A^tA\boldsymbol{x}\cdot\boldsymbol{x} = \Big(\sum_{i=1,\dots,n}\mu_i\alpha_i\boldsymbol{f}_i\Big)\cdot\Big(\sum_{i=1,\dots,n}\alpha_i\boldsymbol{f}_i\Big) = \sum_{i=1,\dots,n}\alpha_i^2\mu_i \leq \mu_n\|\boldsymbol{x}\|_2^2.$$

On en déduit que $\|A\|_2^2 \leq \rho(A^tA)$.
Pour montrer qu'on a égalité, il suffit de considérer le vecteur $\boldsymbol{x} = \boldsymbol{f}_n$; on a en effet $\|\boldsymbol{f}_n\|_2 = 1$, et $\|A\boldsymbol{f}_n\|_2^2 = A^tA\boldsymbol{f}_n\cdot\boldsymbol{f}_n = \mu_n = \rho(A^tA)$.

■

Nous allons maintenant comparer le rayon spectral d'une matrice avec des normes. Rappelons d'abord le théorème de triangularisation (ou trigonalisation) des matrices complexes.

Théorème 1.30 (Décomposition de Schur, triangularisation d'une matrice)**.** *Soit $A \in \mathcal{M}_n(\mathbb{R})$ ou $\mathcal{M}_n(\mathbb{C})$ une matrice carrée quelconque, réelle ou complexe ; alors il existe une matrice complexe Q inversible et une matrice complexe triangulaire supérieure T telles que $A = QTQ^{-1}$.*
Ce résultat s'énonce de manière équivalente de la manière suivante : Soit ψ une application linéaire de E dans E, où E est un espace vectoriel normé de dimension finie n sur $\mathbb{C}$. Alors il existe une base $(\boldsymbol{f}_1,\dots,\boldsymbol{f}_n)$ de $\mathbb{C}$ et une famille de complexes $(t_{i,j})_{i=1,\dots,n,j=1,\dots,n,j\geq i}$ telles que $\psi(\boldsymbol{f}_i) = t_{i,i}\boldsymbol{f}_i + \sum_{k<i}t_{k,i}\boldsymbol{f}_k$. De plus $t_{i,i}$ est valeur propre de ψ et de A pour tout $i \in \{1,\dots,n\}$.
Les deux énoncés sont équivalents au sens où la matrice A de l'application linéaire ψ s'écrit $A = QTQ^{-}1$, où T est la matrice triangulaire supérieure de coefficients $(t_{i,j})_{i,j=1,\dots,n,j\geq i}$ et Q la matrice inversible dont la colonne j est le vecteur $\boldsymbol{f}_j$).

DÉMONSTRATION – On démontre cette propriété par récurrence sur n. Elle est évidemment vraie pour $n = 1$. Soit $n \geq 1$, on suppose la propriété vraie pour n et on la démontre pour $n+1$. Soint donc E un espace vectoriel sur $\mathbb{C}$ de dimension $n+1$ et ψ une application linéaire de E dans $\mathbb{C}$. On sait qu'il existe $\lambda \in \mathbb{C}$ (qui résulte du caractère algébriquement clos de $\mathbb{C}$) et $\boldsymbol{f}_1 \in E$ tels que $\psi(\boldsymbol{f}_1) = \lambda\boldsymbol{f}_1$ et $\|\boldsymbol{f}_1\| = 1$; on pose $t_{1,1} = \lambda$ et on note F le sous espace vectoriel de E supplémentaire orthogonal de $\mathbb{C}\boldsymbol{f}_1$. Soit $\boldsymbol{u} \in F$, il existe un unique couple $(\mu,\boldsymbol{v}) \in \mathbb{C}\times F$ tel que $\psi(\boldsymbol{u}) = \mu\boldsymbol{f}_1 + \boldsymbol{v}$. On note $\tilde{\psi}$ l'application qui à $\boldsymbol{u}$ associe $\boldsymbol{v}$. On peut appliquer l'hypothèse de récurrence à $\tilde{\psi}$ (car $\tilde{\psi}$ est une application linéaire de F dans F, et F est de dimension n). Il existe donc une base orthonormée $\boldsymbol{f}_2,\dots,\boldsymbol{f}_{n+1}$ de F et $(t_{i,j})_{j\geq i\geq 2}$ tels que

$$\tilde{\psi}(\boldsymbol{f}_i) = \sum_{2\leq j\leq i} t_{j,i}\boldsymbol{f}_j,\ i = 2,\dots,n+1.$$

On en déduit que

$$\psi(\boldsymbol{f}_i) = \sum_{1\le j\le i\le n} t_{j,i}\boldsymbol{f}_j,\ i=1,\dots,n+1.$$

■

Dans la proposition suivante, nous montrons qu'on peut toujours trouver une norme (qui dépend de la matrice) pour approcher son rayon spectral d'aussi près que l'on veut par valeurs supérieures.

Théorème 1.31 (Approximation du rayon spectral par une norme induite)**.**
1. Soit $\|\cdot\|$ une norme induite. Alors

$$\rho(A) \le \|A\|, \text{ pour tout } A\in\mathcal{M}_n(\mathbb{R}).$$

2. Soient maintenant $A\in\mathcal{M}_n(\mathbb{R})$ et $\varepsilon>0$, alors il existe une norme sur $\mathbb{R}^n$ (qui dépend de A et ε) telle que la norme induite sur $\mathcal{M}_n(\mathbb{R})$, notée $\|\cdot\|_{A,\varepsilon}$, vérifie $\|A\|_{A,\varepsilon}\le\rho(A)+\varepsilon$.

DÉMONSTRATION – 1. Soit $\lambda\in\mathcal{C}$ valeur propre de A et $\boldsymbol{x}$ un vecteur propre associé, alors $A\boldsymbol{x}=\lambda\boldsymbol{x}$, et comme $\|\cdot\|$ est une norme induite, on a :

$$\|\lambda\boldsymbol{x}\| = |\lambda|\|\boldsymbol{x}\| = \|A\boldsymbol{x}\| \le \|A\|\|\boldsymbol{x}\|.$$

On en déduit que toute valeur propre λ vérifie $\lambda\le\|A\|$ et donc $\rho(A)\le\|A\|$.

2. Soit $A\in\mathcal{M}_n(\mathbb{R})$, alors par le théorème de triangularisation de Schur (théorème 1.30 ppécédent), il existe une base $(\boldsymbol{f}_1,\dots,\boldsymbol{f}_n)$ de $\mathbb{C}^n$ et une famille de complexes $(t_{i,j})_{i,j=1,\dots,n,j\ge i}$ telles que $A\boldsymbol{f}_i=\sum_{j\le i}t_{j,i}\boldsymbol{f}_j$. Soit $\eta\in]0,1[$, qu'on choisira plus précisément plus tard. Pour $i=1,\dots,n$, on définit $\boldsymbol{e}_i=\eta^{i-1}\boldsymbol{f}_i$. La famille $(\boldsymbol{e}_i)_{i=1,\dots,n}$ forme une base de $\mathbb{C}^n$. On définit alors une norme sur $\mathbb{R}^n$ par $\|\boldsymbol{x}\|=(\sum_{i=1}^n\alpha_i\overline{\alpha_i})^{1/2}$, où les α_i sont les composantes de $\boldsymbol{x}$ dans la base $(\boldsymbol{e}_i)_{i=1,\dots,n}$. Notons que cette norme dépend de A et de η. Soit $\varepsilon>0$; montrons que pour η bien choisi, on a $\|A\|\le\rho(A)+\varepsilon$. Remarquons d'abord que

$$A\boldsymbol{e}_i = A(\eta^{i-1}\boldsymbol{f}_i) = \eta^{i-1}A\boldsymbol{f}_i\eta^{i-1}\sum_{j\le i}t_{k,i}\boldsymbol{f}_j = \eta^{i-1}\sum_{j\le i}t_{j,i}\eta^{1-j}\boldsymbol{e}_j = \sum_{1\le j\le i}\eta^{i-j}t_{j,i}\boldsymbol{e}_j,$$

Soit maintenant $\boldsymbol{x}=\sum_{i=1,\dots,n}\alpha_i\boldsymbol{e}_i$. On a

$$A\boldsymbol{x} = \sum_{i=1}^n\alpha_iA\boldsymbol{e}_i = \sum_{i=1}^n\sum_{1\le j\le i}\eta^{i-j}t_{j,i}\alpha_i\boldsymbol{e}_j = \sum_{j=1}^n\big(\sum_{i=j}^n\eta^{i-j}\lambda_{i,j}\alpha_i\big)\boldsymbol{e}_j.$$

On en déduit que

$$\begin{aligned}\|A\boldsymbol{x}\|^2 &= \sum_{j=1}^n\big(\sum_{i=j}^n\eta^{i-j}t_{j,i}\alpha_i\big)\big(\sum_{i=j}^n\eta^{i-j}\overline{t}_{j,i}\overline{\alpha}_i\big),\\ &= \sum_{j=1}^n t_{j,j}\overline{t_{j,j}}\alpha_j\overline{\alpha_j} + \sum_{j=1}^n\sum_{\substack{k,\ell\ge j\\(k,\ell)\ne(j,j)}}\eta^{k+\ell-2j}t_{j,k}\overline{t_{j,\ell}}\alpha_k\overline{\alpha_\ell}\\ &\le \rho(A)^2\|\boldsymbol{x}\|^2 + \max_{k=1,\dots,n}|\alpha_k|^2\sum_{j=1}^n\sum_{\substack{k,\ell\ge j\\(k,\ell)\ne(j,j)}}\eta^{k+\ell-2j}t_{j,k}\overline{t_{j,\ell}}.\end{aligned}$$

Comme $\eta\in[0,1]$ et $k+\ell-2j\ge1$ dans la dernière sommation, on a

$$\sum_{j=1}^n\sum_{\substack{k,\ell\ge j\\(k,\ell)\ne(j,j)}}\eta^{k+\ell-2j}t_{j,k}\overline{t_{j,\ell}} \le \eta C_Tn^3,$$

où $C_T=\max_{j,k,\ell=1,\dots,n}|t_{j,k}||t_{j,k}|$ ne dépend que de la matrice T, qui elle même ne dépend que de A. Comme

$$\max_{k=1,\dots,n}|\alpha_k|^2 \le \sum_{k=1,\dots,n}|\alpha_k|^2 = \|\boldsymbol{x}\|^2,$$

on a donc

$$\frac{\|A\boldsymbol{x}\|^2}{\|\boldsymbol{x}\|^2} \leq \rho(A)^2 + \eta C_T n^3.$$

On en conclut que :

$$\|A\| \leq \rho(A)\left(1 + \frac{\eta C_T n^3}{\rho(A)^2}\right) \leq \rho(A) + \frac{\eta C_T n^3}{2\rho(A)},$$

D'où le résultat, en prenant $\|\cdot\|_{A,\varepsilon} = \|\cdot\|$ et η tel que $\eta = \min\left(1, \frac{2\rho(A)\varepsilon}{C_T n^3}\right)$..

■

Corollaire 1.32 (Convergence et rayon spectral)**.** *Soit $A \in \mathcal{M}_n(\mathbb{R})$. Alors :*

$$\rho(A) < 1 \textit{ si et seulement si } A^k \to 0 \textit{ quand } k \to \infty.$$

DÉMONSTRATION – Si $\rho(A) < 1$, grâce au résultat d'approximation du rayon spectral de la proposition précédente, il existe $\varepsilon > 0$ tel que $\rho(A) < 1 - 2\varepsilon$ et une norme induite $\|.\|_{A,\varepsilon}$ tels que $\|A\|_{A,\varepsilon} = \mu \leq \rho(A) + \varepsilon = 1 - \varepsilon < 1$. Comme $\|.\|_{A,\varepsilon}$ est une norme matricielle, on a $\|A^k\|_{A,\varepsilon} \leq \mu^k \to 0$ lorsque $k \to \infty$. Comme l'espace $\mathcal{M}_n(\mathbb{R})$ est de dimension finie, toutes les normes sont équivalentes, et on a donc $\|A^k\| \to 0$ lorsque $k \to \infty$.

Montrons maintenant la réciproque : supposons que $A^k \to 0$ lorsque $k \to \infty$, et montrons que $\rho(A) < 1$. Soient λ une valeur propre de A et $\boldsymbol{x}$ un vecteur propre associé. Alors $A^k\boldsymbol{x} = \lambda^k\boldsymbol{x}$, et si $A^k \to 0$, alors $A^k\boldsymbol{x} \to 0$, et donc $\lambda^k\boldsymbol{x} \to 0$, ce qui n'est possible que si $|\lambda| < 1$.

■

Remarque 1.33 (Convergence des suites)**.** *Une conséquence immédiate du corollaire précédent est que la suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ définie par $\boldsymbol{x}^{(k+1)} = A\boldsymbol{x}^{(k)}$ converge vers $\mathbf{0}$ (le vecteur nul) pour tout $\boldsymbol{x}^{(0)}$ donné si et seulement si $\rho(A) < 1$.*

Proposition 1.34 (Convergence et rayon spectral)**.** *On munit $\mathcal{M}_n(\mathbb{R})$ d'une norme, notée $\|\cdot\|$. Soit $A \in \mathcal{M}_n(\mathbb{R})$. Alors*

$$\rho(A) = \lim_{k\to\infty} \|A^k\|^{\frac{1}{k}}. \tag{1.58}$$

DÉMONSTRATION – La démonstration se fait par des arguments d'homogénéité, en trois étapes. Rappelons tout d'abord que

$$\limsup_{k\to+\infty} u_k = \lim_{k\to+\infty} \sup_{n\geq k} u_n,$$
$$\liminf_{k\to+\infty} u_k = \lim_{k\to+\infty} \inf_{n\geq k} u_n,$$

et que si $\limsup_{k\to+\infty} u_k \leq \liminf_{k\to+\infty} u_k$, alors la suite $(u_k)_{k\in\mathbb{N}}$ converge vers $\lim_{k\to+\infty} u_k = \liminf_{k\to+\infty} u_k = \limsup_{k\to+\infty} u_k$.

Etape 1. On montre que

$$\rho(A) < 1 \Rightarrow \limsup_{k\to\infty} \|A^k\|^{\frac{1}{k}} \leq 1. \tag{1.59}$$

En effet, si i $\rho(A) < 1$, d'après le corollaire 1.32 on a : $\|A^k\| \to 0$ donc il existe $K \in \mathbb{N}$ tel que pour $k \geq K$, $\|A^k\| < 1$. On en déduit que pour $k \geq K$, $\|A^k\|^{1/k} < 1$, et donc en passant à la limite sup sur k, on obtient bien que

$$\limsup_{k\to+\infty} \|A^k\|^{\frac{1}{k}} \leq 1.$$

Etape 2. On montre maintenant que

$$\liminf_{k\to\infty} \|A^k\|^{\frac{1}{k}} < 1 \Rightarrow \rho(A) < 1.. \tag{1.60}$$

Pour démontrer cette assertion, rappelons que pour toute suite $(u_k)_{k\in\mathbb{N}}$ d'éléments de $\mathbb{R}$ ou $\mathbb{R}^n$, la limite inférieure $\liminf_{k\to+\infty} u_k$ est une valeur d'adhérence de la suite $(u_k)_{k\in\mathbb{N}}$, donc qu'il existe une suite extraite $(u_{k_n})_{n\in\mathbb{N}}$ telle que $u_{k_n} \to \liminf_{k\to+\infty} u_k$ lorsque $k \to +\infty$. Or $\liminf_{k\to+\infty} \|A^k\|^{1/k} < 1$; donc il existe une sous-suite $(k_n)_{n\in\mathbb{N}} \subset \mathbb{N}$ telle que $\|A^{k_n}\|^{1/k_n} \to \ell < 1$ lorsque $n \to +\infty$, et donc il existe n tel que pour $n \geq n$, $\|A^{k_n}\|^{1/k_n} \leq \eta$, avec $\eta \in]0,1[$. On en déduit que pour $n \geq n$, $\|A^{k_n}\| \leq \eta^{k_n}$, et donc que $A^{k_n} \to 0$ lorsque $n \to +\infty$. Soient λ une valeur propre de A et x un vecteur propre associé, on a : $A^{k_n}x = \lambda^{k_n}x$; on en déduit que $|\lambda| < 1$, et donc que $\rho(A) < 1$.

Etape 3. On montre que $\rho(A) = \lim_{k\to\infty} \|A^k\|^{\frac{1}{k}}$.
Soit $\alpha \in \mathbb{R}_+$ tel que $\rho(A) < \alpha$. Alors $\rho(\frac{1}{\alpha}A) < 1$, et donc grâce à (1.59),

$$\limsup_{k\to+\infty} \|A^k\|^{\frac{1}{k}} < \alpha,\ \forall \alpha > \rho(A).$$

En faisant tendre α vers $\rho(A)$, on obtient donc :

$$\limsup_{k\to+\infty} \|A^k\|^{\frac{1}{k}} \leq \rho(A). \tag{1.61}$$

Soit maintenant $\beta \in \mathbb{R}_+$ tel que $\liminf_{k\to+\infty} \|A^k\|^{\frac{1}{k}} < \beta$. On a alors $\liminf_{k\to+\infty} \|(\frac{1}{\beta}A)^k\|^{\frac{1}{k}} < 1$ et donc en vertu de (1.60), $\rho(\frac{1}{\beta}A) < 1$, donc $\rho(A) < \beta$ pour tout $\beta \in \mathbb{R}_+$ tel que $\liminf_{k\to+\infty} \|A^k\|^{\frac{1}{k}} < \beta$. En faisant tendre β vers $\liminf_{k\to+\infty} \|A^k\|^{\frac{1}{k}}$, on obtient donc

$$\rho(A) \leq \liminf_{k\to+\infty} \|A^k\|^{\frac{1}{k}}. \tag{1.62}$$

De (1.61) et (1.62), on déduit que

$$\limsup_{k\to+\infty} \|A^k\|^{\frac{1}{k}} = \liminf_{k\to+\infty} \|A^k\|^{\frac{1}{k}} = \lim_{k\to+\infty} \|A^k\|^{\frac{1}{k}} = \rho(A). \tag{1.63}$$

■

Un corollaire important de la proposition 1.34 est le suivant.

Corollaire 1.35 (Comparaison rayon spectral et norme)**.** *On munit $\mathcal{M}_n(\mathbb{R})$ d'une norme **matricielle**, notée $\|\cdot\|$. Soit $A \in \mathcal{M}_n(\mathbb{R})$. Alors :*

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

Par conséquent, si $M \in \mathcal{M}_n(\mathbb{R})$ et $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$, pour montrer que la suite $\boldsymbol{x}^{(k)}$ définie par $\boldsymbol{x}^{(k)} = M^k\boldsymbol{x}^{(0)}$ converge vers $\mathbf{0}$ dans $\mathbb{R}^n$, il suffit de trouver une norme matricielle $\|\cdot\|$ telle que $\|M\| < 1$.

DÉMONSTRATION – Si $\|.\|$ est une norme matricielle, alors $\|A^k\| \leq \|A\|^k$ et donc par la caractérisation (1.58) du rayon spectral donnée dans la proposition précédente, on obtient que $\rho(A) \leq \|A\|$. ■

Ce dernier résultat est évidemment bien utile pour montrer la convergence de la suite A^k, ou de suites de la forme $A^k\boldsymbol{x}^{(0)}$ avec $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$. Une fois qu'on a trouvé une norme matricielle pour laquelle A est de norme strictement inférieure à 1, on a gagné. Attention cependant au piège suivant : pour toute matrice A, on peut toujours trouver une norme pour laquelle $\|A\| < 1$, alors que la série de terme général A^k peut ne pas être convergente.
Prenons un exemple dans $\mathbb{R}$, $\|x\| = \frac{1}{4}|x|$. Pour $x = 2$ on a $\|x\| = \frac{1}{2} < 1$. Et pourtant la série de terme général x^k n'est pas convergente ; le problème ici est que la norme choisie n'est pas une norme matricielle (on n'a pas $\|xy\| \leq \|x\|\|y\|$).
De même, on peut trouver une matrice et une norme telles que $\|A\| \geq 1$, alors que la série de terme général A^k converge...

Nous donnons maintenant un théorème qui nous sera utile dans l'étude du conditionnement, ainsi que plus tard dans l'étude des méthodes itératives.

Théorème 1.36 (Matrices de la forme $Id + A$)**.**

1. *Soit une norme matricielle induite, Id la matrice identité de $\mathcal{M}_n(\mathbb{R})$ et $A \in \mathcal{M}_n(\mathbb{R})$ telle que $\|A\| < 1$. Alors la matrice $Id + A$ est inversible et*
$$\|(Id + A)^{-1}\| \leq \frac{1}{1 - \|A\|}.$$
2. *Si une matrice de la forme $Id + A \in \mathcal{M}_n(\mathbb{R})$ est singulière, alors $\|A\| \geq 1$ pour toute norme matricielle $\|\cdot\|$.*

DÉMONSTRATION –

1. La démonstration du point 1 fait l'objet de l'exercice 32 page 62.
2. Si la matrice $Id + A \in \mathcal{M}_n(\mathbb{R})$ est singulière, alors $\lambda = -1$ est valeur propre, et donc $\rho(A) \geq 1$. En utilisant le corollaire 1.35, on obtient que $\|A\| \geq \rho(A) \geq 1$.

■

1.4.2 Le problème des erreurs d'arrondis

Soient $A \in \mathcal{M}_n(\mathbb{R})$ inversible et $\boldsymbol{b} \in \mathbb{R}^n$; supposons que les données A et $\boldsymbol{b}$ ne soient connues qu'à une erreur près. Ceci est souvent le cas dans les applications pratiques. Considérons par exemple le problème de la conduction thermique dans une tige métallique de longueur 1, modélisée par l'intervalle $[0, 1]$. Supposons que la température u de la tige soit imposée aux extrémités, $u(0) = u_0$ et $u(1) = u_1$. On suppose que la température dans la tige satisfait à l'équation de conduction de la chaleur, qui s'écrit $(k(x)u'(x))' = 0$, où k est la conductivité thermique. Cette équation différentielle du second ordre peut se discrétiser par exemple par différences finies (on verra une description de la méthode page 11), et donne lieu à un système linéaire de matrice A. Si la conductivité k n'est connue qu'avec une certaine précision, alors la matrice A sera également connue à une erreur près, notée δ_A. On aimerait que l'erreur commise sur les données du modèle (ici la conductivité thermique k) n'ait pas une conséquence trop grave sur le calcul de la solution du modèle (ici la température u). Si par exemple 1% d'erreur sur k entraîne 100% d'erreur sur u, le modèle ne sera pas d'une utilité redoutable…
L'objectif est donc d'estimer les erreurs commises sur $\boldsymbol{x}$ solution de (1.1) à partir des erreurs commises sur $\boldsymbol{b}$ et A. Notons $\boldsymbol{\delta_b} \in \mathbb{R}^n$ l'erreur commise sur $\boldsymbol{b}$ et $\delta_A \in \mathcal{M}_n(\mathbb{R})$ l'erreur commise sur A. On cherche alors à évaluer $\boldsymbol{\delta_x}$ où $\boldsymbol{x} + \boldsymbol{\delta_x}$ est solution (si elle existe) du système :
$$\begin{cases} \boldsymbol{x} + \boldsymbol{\delta_x} \in \mathbb{R}^n \\ (A + \delta_A)(\boldsymbol{x} + \boldsymbol{\delta_x}) = \boldsymbol{b} + \boldsymbol{\delta_b}. \end{cases} \tag{1.64}$$
On va montrer que si δ_A "n'est pas trop grand", alors la matrice $A + \delta_A$ est inversible, et qu'on peut estimer $\boldsymbol{\delta_x}$ en fonction de δ_A et $\boldsymbol{\delta_b}$.

1.4.3 Conditionnement et majoration de l'erreur d'arrondi

Définition 1.37 (Conditionnement)**.** *Soit $\mathbb{R}^n$ muni d'une norme $\|\cdot\|$ et $\mathcal{M}_n(\mathbb{R})$ muni de la norme induite. Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible. On appelle conditionnement de A par rapport à la norme $\|\cdot\|$ le nombre réel positif $\mathrm{cond}(A)$ défini par :*
$$\mathrm{cond}(A) = \|A\| \, \|A^{-1}\|.$$

Proposition 1.38 (Propriétés générales du conditionnement)**.** *Soit $\mathbb{R}^n$ muni d'une norme $\|\cdot\|$ et $\mathcal{M}_n(\mathbb{R})$ muni de la norme induite.*

1. *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible, alors* $\mathrm{cond}(A) \geq 1$.
2. *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible et $\alpha \in \mathbb{R}^*$, alors* $\mathrm{cond}(\alpha A) = \mathrm{cond}(A)$.
3. *Soient A et $B \in \mathcal{M}_n(\mathbb{R})$ des matrices inversibles, alors* $\mathrm{cond}(AB) \leq \mathrm{cond}(A)\mathrm{cond}(B)$.

DÉMONSTRATION – 1. Comme $\|\cdot\|$ est une norme induite, c'est donc une norme matricielle. On a donc pour toute matrice $A \in \mathcal{M}_n(\mathbb{R})$,

$$\|\mathrm{Id}\| \leq \|A\| \, \|A^{-1}\|$$

ce qui prouve que $\mathrm{cond}(A) \geq 1$.

2. Par définition,

$$\begin{aligned} \mathrm{cond}(\alpha A) &= \|\alpha A\| \, \|(\alpha A)^{-1}\| \\ &= |\alpha| \, \|A\| \frac{1}{|\alpha|} \|A^{-1}\| = \mathrm{cond}(A) \end{aligned}$$

3. Soient A et B des matrices inversibles, alors AB est une matrice inversible et comme $\|\cdot\|$ est une norme matricielle,

$$\begin{aligned} \mathrm{cond}(AB) &= \|AB\| \, \|(AB)^{-1}\| \\ &= \|AB\| \, \|B^{-1}A^{-1}\| \\ &\leq \|A\| \, \|B\| \, \|B^{-1}\| \, \|A^{-1}\|. \end{aligned}$$

Donc $\mathrm{cond}(AB) \leq \mathrm{cond}(A)\mathrm{cond}(B)$. ∎

Proposition 1.39 (Caractérisation du conditionnement pour la norme 2)**.** *Soit $\mathbb{R}^n$ muni de la norme euclidienne $\|\cdot\|_2$ et $\mathcal{M}_n(\mathbb{R})$ muni de la norme induite. Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible. On note $cond_2(A)$ le conditionnement associé à la norme induite par la norme euclidienne sur $\mathbb{R}^n$.*

1. *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible. On note σ_n [resp. σ_1] la plus grande [resp. petite] valeur propre de A^tA (noter que A^tA est une matrice symétrique définie positive). Alors*

$$\mathrm{cond}_2(A) = \sqrt{\frac{\sigma_n}{\sigma_1}}.$$

2. *Si de plus A une matrice symétrique définie positive, alors*

$$\mathrm{cond}_2(A) = \frac{\lambda_n}{\lambda_1},$$

où λ_n [resp. λ_1] est la plus grande [resp. petite] valeur propre de A.

DÉMONSTRATION – On rappelle que si A a comme valeurs propres $\lambda_1, \ldots, \lambda_n$, alors A^{-1} a comme valeurs propres $\lambda_1^{-1}, \ldots, \lambda_n^{-1}$ et A^t a comme valeurs propres $\lambda_1, \ldots, \lambda_n$.

1. Par définition, on a $\mathrm{cond}_2(A) = \|A\|_2 \|A^{-1}\|_2$. Or par le point 3. de la proposition 1.29 que $\|A\|_2 = (\rho(A^tA))^{1/2} = \sqrt{\sigma_n}$. On a donc

$$\|A^{-1}\|_2 = (\rho((A^{-1})^tA^{-1}))^{1/2} = \left(\rho(AA^t)^{-1}\right)^{1/2} \text{ ; or } \rho((AA^t)^{-1}) = \frac{1}{\tilde{\sigma}_1},$$

où $\tilde{\sigma}_1$ est la plus petite valeur propre de la matrice AA^t. Mais les valeurs propres de AA^t sont les valeurs propres de A^tA : en effet, si λ est valeur propre de AA^t associée au vecteur propre x alors λ est valeur propre de AA^t associée au vecteur propre A^tx. On a donc

$$cond_2(A) = \sqrt{\frac{\sigma_n}{\sigma_1}}.$$

2. Si A est s.d.p., alors $A^tA = A^2$ et $\sigma_i = \lambda_i^2$ où λ_i est valeur propre de la matrice A. On a dans ce cas $cond_2(A) = \dfrac{\lambda_n}{\lambda_1}$. ∎

Les propriétés suivantes sont moins fondamentales, mais cependant intéressantes !

Proposition 1.40 (Propriétés du conditionnement pour la norme 2). *Soit $\mathbb{R}^n$ muni de la norme euclidienne $\|\cdot\|_2$ et $\mathcal{M}_n(\mathbb{R})$ muni de la norme induite. Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible. On note $cond_2(A)$ le conditionnement associé à la norme induite par la norme euclidienne sur $\mathbb{R}^n$.*

1. *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible. Alors $\mathrm{cond}_2(A) = 1$ si et seulement si $A = \alpha Q$ où $\alpha \in \mathbb{R}^\star$ et Q est une matrice orthogonale (c'est-à-dire $Q^t = Q^{-1}$).*
2. *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible. On suppose que $A = QR$ où Q est une matrice orthogonale. Alors $\mathrm{cond}_2(A) = \mathrm{cond}_2(R)$.*
3. *Si A et B sont deux matrices symétriques définies positives, alors $\mathrm{cond}_2(A+B) \leq \max(\mathrm{cond}_2(A), \mathrm{cond}_2(B))$.*

La démonstration de la proposition 1.40 fait l'objet de l'exercice 35 page 63.
On va maintenant majorer l'erreur relative commise sur $\boldsymbol{x}$ solution de $A\boldsymbol{x} = \boldsymbol{b}$ lorsque l'on commet une erreur $\delta_{\boldsymbol{b}}$ sur le second membre $\boldsymbol{b}$.

Proposition 1.41 (Majoration de l'erreur relative pour une erreur sur le second membre). *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible, et $\boldsymbol{b} \in \mathbb{R}^n$, $\boldsymbol{b} \neq 0$. On munit $\mathbb{R}^n$ d'une norme $\|\cdot\|$ et $\mathcal{M}_n(\mathbb{R})$ de la norme induite. Soit $\boldsymbol{\delta_b} \in \mathbb{R}^n$. Si $\boldsymbol{x}$ est solution de* (1.1) *et $\boldsymbol{x} + \boldsymbol{\delta_x}$ est solution de*

$$A(\boldsymbol{x} + \boldsymbol{\delta_x}) = \boldsymbol{b} + \boldsymbol{\delta_b}, \tag{1.65}$$

alors

$$\frac{\|\boldsymbol{\delta_x}\|}{\|\boldsymbol{x}\|} \leq \mathrm{cond}(A)\frac{\|\boldsymbol{\delta_b}\|}{\|\boldsymbol{b}\|} \tag{1.66}$$

DÉMONSTRATION – En retranchant (1.1) à (1.65), on obtient :

$$A\boldsymbol{\delta_x} = \boldsymbol{\delta_b}$$

et donc

$$\|\boldsymbol{\delta_x}\| \leq \|A^{-1}\|\|\boldsymbol{\delta_b}\|. \tag{1.67}$$

Cette première estimation n'est pas satisfaisante car elle porte sur l'erreur globale ; or la notion intéressante est celle d'erreur relative. On obtient l'estimation sur l'erreur relative en remarquant que $\boldsymbol{b} = A\boldsymbol{x}$, ce qui entraîne que $\|\boldsymbol{b}\| \leq \|A\|\|\boldsymbol{x}\|$. On en déduit que

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

En multipliant membre à membre cette dernière inégalité et (1.67), on obtient le résultat souhaité. ∎

Remarquons que l'estimation (1.66) est optimale. En effet, on va démontrer qu'on peut avoir égalité dans (1.66). Pour cela, il faut choisir convenablement $\boldsymbol{b}$ et $\boldsymbol{\delta_b}$. On sait déjà que si $\boldsymbol{x}$ est solution de (1.1) et $\boldsymbol{x} + \boldsymbol{\delta_x}$ est solution de (1.64), alors

$$\boldsymbol{\delta_x} = A^{-1}\boldsymbol{\delta_b}, \text{ et donc } \|\boldsymbol{\delta_x}\| = \|A^{-1}\boldsymbol{\delta_b}\|.$$

Soit $\boldsymbol{x} \in \mathbb{R}^n$ tel que $\|\boldsymbol{x}\| = 1$ et $\|A\boldsymbol{x}\| = \|A\|$. Notons qu'un tel $\boldsymbol{x}$ existe parce que

$$\|A\| = \sup\{\|A\boldsymbol{x}\|; \|\boldsymbol{x}\| = 1\} = \max\{\|A\boldsymbol{x}\|; \|\boldsymbol{x}\| = 1\}$$

(voir proposition 1.27 page 51). On a donc

$$\frac{\|\boldsymbol{\delta_x}\|}{\|\boldsymbol{x}\|} = \|A^{-1}\boldsymbol{\delta_b}\|\frac{\|A\|}{\|A\boldsymbol{x}\|}.$$

Posons $\boldsymbol{b} = A\boldsymbol{x}$; on a donc $\|\boldsymbol{b}\| = \|A\|$, et donc

$$\frac{\|\boldsymbol{\delta_x}\|}{\|\boldsymbol{x}\|} = \|A^{-1}\boldsymbol{\delta_b}\|\frac{\|A\|}{\|b\|}.$$

De même, grâce à la proposition 1.27, il existe $\boldsymbol{y} \in \mathbb{R}^n$ tel que $\|\boldsymbol{y}\| = 1$, et $\|A^{-1}\boldsymbol{y}\| = \|A^{-1}\|$. On choisit alors $\boldsymbol{\delta_b}$ tel que $\delta_b = \boldsymbol{y}$. Comme $A(\boldsymbol{x} + \boldsymbol{\delta_x}) = \boldsymbol{b} + \boldsymbol{\delta_b}$, on a $\boldsymbol{\delta_x} = A^{-1}\boldsymbol{\delta_b}$ et donc :

$$\|\boldsymbol{\delta_x}\| = \|A^{-1}\boldsymbol{\delta_b}\| = \varepsilon\|A^{-1}\boldsymbol{y}\| = \varepsilon\|A^{-1}\| = \|\boldsymbol{\delta_b}\|\ \|A^{-1}\|.$$

On en déduit que

$$\frac{\|\boldsymbol{\delta_x}\|}{\|\boldsymbol{x}\|} = \|\boldsymbol{\delta_x}\| = \|\boldsymbol{\delta_b}\|\ \|A^{-1}\|\ \ \frac{\|A\|}{\|\boldsymbol{b}\|} \text{ car } \|\boldsymbol{b}\| = \|A\| \text{ et } \|\boldsymbol{x}\| = 1.$$

Par ce choix de $\boldsymbol{b}$ et $\boldsymbol{\delta_b}$ on a bien égalité dans (1.66) qui est donc optimale.
Majorons maintenant l'erreur relative commise sur $\boldsymbol{x}$ solution de $A\boldsymbol{x} = \boldsymbol{b}$ lorsque l'on commet une erreur δ_A sur la matrice A.

Proposition 1.42 (Majoration de l'erreur relative pour une erreur sur la matrice)**.** *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible, et $\boldsymbol{b} \in \mathbb{R}^n$, $\boldsymbol{b} \neq 0$. On munit $\mathbb{R}^n$ d'une norme $\|\cdot\|$, et $\mathcal{M}_n(\mathbb{R})$ de la norme induite. Soit $\delta_A \in \mathbb{R}^n$; on suppose que $A + \delta_A$ est une matrice inversible. Si $\boldsymbol{x}$ est solution de* (1.1) *et $\boldsymbol{x} + \delta\boldsymbol{x}$ est solution de*

$$(A + \delta_A)(\boldsymbol{x} + \boldsymbol{\delta_x}) = \boldsymbol{b} \tag{1.68}$$

alors

$$\frac{\|\boldsymbol{\delta_x}\|}{\|\boldsymbol{x} + \boldsymbol{\delta_x}\|} \leq \mathrm{cond}(A)\frac{\|\delta_A\|}{\|A\|} \tag{1.69}$$

DÉMONSTRATION – En retranchant (1.1) à (1.68), on obtient :

$$A\delta(\boldsymbol{\delta_x} = -\delta_A(\boldsymbol{x} + \boldsymbol{\delta_x})$$

et donc

$$\boldsymbol{\delta_x} = -A^{-1}\delta_A(\boldsymbol{x} + \boldsymbol{\delta_x}).$$

On en déduit que $\|\boldsymbol{\delta_x}\| \leq \|A^{-1}\|\|\delta_A\|\|\boldsymbol{x} + \boldsymbol{\delta_x}\|$, d'où on déduit le résultat souhaité. ∎

On peut en fait majorer l'erreur relative dans le cas où l'on commet à la fois une erreur sur A et une erreur sur $\boldsymbol{b}$. On donne le théorème à cet effet ; la démonstration est toutefois nettement plus compliquée.

Théorème 1.43 (Majoration de l'erreur relative pour une erreur sur matrice et second membre)**.** *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible, et $\boldsymbol{b} \in \mathbb{R}^n$, $bfb \neq \mathbf{0}$. On munit $\mathbb{R}^n$ d'une norme $\|\cdot\|$, et $\mathcal{M}_n(\mathbb{R})$ de la norme induite. Soient $\delta_A \in \mathcal{M}_n(\mathbb{R})$ et $\boldsymbol{\delta_b} \in \mathbb{R}^n$. On suppose que $\|\delta_A\| < \dfrac{1}{\|A^{-1}\|}$. Alors la matrice $(A + \delta_A)$ est inversible et si $\boldsymbol{x}$ est solution de* (1.1) *et $\boldsymbol{x} + \boldsymbol{\delta_x}$ est solution de* (1.64)*, alors*

$$\frac{\|\boldsymbol{\delta_x}\|}{\|\boldsymbol{x}\|} \leq \frac{\mathrm{cond}(A)}{1 - \|A^{-1}\|\ \|\delta_A\|}\left(\frac{\boldsymbol{\delta_b}\|}{\|\boldsymbol{b}\|} + \frac{\|\delta_A\|}{\|A\|}\right). \tag{1.70}$$

DÉMONSTRATION – On peut écrire $A + \delta_A = A(\mathrm{Id} + B)$ avec $B = A^{-1}\delta_A$. Or le rayon spectral de B, $\rho(B)$, vérifie $\rho(B) \leq \|B\| \leq \|\delta_A\|\ \|A^{-1}\| < 1$, et donc (voir le théorème 1.36 page 55 et l'exercice 32 page 62) $(\mathrm{Id} + B)$ est inversible et $(\mathrm{Id} + B)^{-1} = \sum_{n=0}^{\infty}(-1)^n B^n$. On a aussi $\|(\mathrm{Id} + B)^{-1}\| \leq \sum_{n=0}^{\infty}\|B\|^n = \dfrac{1}{1 - \|B\|} \leq \dfrac{1}{1 - \|A^{-1}\|\ \|\delta_A\|}$. On en déduit que $A + \delta_A$ est inversible, car $A + \delta_A = A(Id + B)$ et comme A est inversible, $(A + \delta_A)^{-1} = (\mathrm{Id} + B)^{-1}A^{-1}$.

Comme A et $A+\delta_A$ sont inversibles, il existe un unique $\boldsymbol{x} \in \mathbb{R}^n$ tel que $A\boldsymbol{x} = b$ et il existe un unique $\boldsymbol{\delta_x} \in \mathbb{R}^n$ tel que $(A+\delta_A)(\boldsymbol{x}+\boldsymbol{\delta_x}) = \boldsymbol{b}+\boldsymbol{\delta_b}$. Comme $A\boldsymbol{x} = \boldsymbol{b}$, on a $(A+\delta_A)\boldsymbol{\delta_x} + \delta_A\boldsymbol{x} = \boldsymbol{\delta_b}$ et donc $\boldsymbol{\delta_x} = (A+\delta_A)^{-1}\boldsymbol{\delta_b} - \delta_A\boldsymbol{x})$. Or $(A+\delta_A)^{-1} = (Id+B)^{-1}A^{-1}$, on en déduit :

$$\begin{aligned}\|(A+\delta_A)^{-1}\| &\leq \|(\mathrm{Id}+B)^{-1}\| \ \|A^{-1}\| \\ &\leq \frac{\|A^{-1}\|}{1-\|A^{-1}\| \ \|\delta_A\|}.\end{aligned}$$

On peut donc écrire la majoration suivante :

$$\frac{\|\boldsymbol{\delta_x}\|}{\|x\|} \leq \frac{\|A^{-1}\| \ \|A\|}{1-\|A^{-1}\| \ \|\delta_A\|}\left(\frac{\|\boldsymbol{\delta_b}\|}{\|A\| \ \|\boldsymbol{x}\|} + \frac{\|\delta_A\|}{\|A\|}\right).$$

En utilisant le fait que $\boldsymbol{b} = A\boldsymbol{x}$ et que par suite $\|\boldsymbol{b}\| \leq \|A\| \ \|\boldsymbol{x}\|$, on obtient :

$$\frac{\|\boldsymbol{\delta_x}\|}{\|\boldsymbol{x}\|} \leq \frac{\|A^{-1}\| \ \|A\|}{1-\|A^{-1}\| \ \|\delta_A\|}\left(\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta_A\|}{\|A\|}\right),$$

ce qui termine la démonstration. ∎

1.4.4 Discrétisation d'équations différentielles, conditionnement "efficace"

On suppose encore ici que $\delta_A = 0$. On suppose que la matrice A du système linéaire à résoudre provient de la discrétisation par différences finies du problème de la chaleur unidimensionnel (1.5a). On peut alors montrer (voir exercice 41 page 65) que le conditionnement de A est d'ordre n^2, où n est le nombre de points de discrétisation. Pour $n = 10$, on a donc $\mathrm{cond}(A) \simeq 100$ et l'estimation (1.66) donne :

$$\frac{\|\delta_x\|}{\|x\|} \leq 100 \ \frac{\|\delta_b\|}{\|b\|}.$$

Une erreur de 1% sur $\boldsymbol{b}$ peut donc entraîner une erreur de 100% sur x. Autant dire que dans ce cas, il est inutile de rechercher la solution de l'équation discrétisée... Heureusement, on peut montrer que l'estimation (1.66) n'est pas significative pour l'étude de la propagation des erreurs lors de la résolution des systèmes linéraires provenant de la discrétisation d'une équation différentielle ou d'une équation aux dérivées partielles[7]. Pour illustrer notre propos, reprenons l'étude du système linéaire obtenu à partir de la discrétisation de l'équation de la chaleur (1.5a) qu'on écrit : $A\boldsymbol{u} = \boldsymbol{b}$ avec $\boldsymbol{b} = (b_1, \ldots, b_n)$ et A la matrice carrée d'ordre n de coefficients $(a_{i,j})_{i,j=1,n}$ définis par (1.10). On rappelle que A est symétrique définie positive (voir exercice 8 page 19), et que

$$\max_{i=1\ldots n}\{|u_i - u(x_i)|\} \leq \frac{h^2}{96}\|u^{(4)}\|_\infty.$$

En effet, si on note $\overline{u}$ le vecteur de $\mathbb{R}^n$ de composantes $u(x_i)$, $i = 1, \ldots, n$, et R le vecteur de $\mathbb{R}^n$ de composantes R_i, $i = 1, \ldots, n$, on a par définition de R (formule (1.7)) $A(u-\overline{u}) = R$, et donc $\|u-\overline{u}\|_\infty \leq \|A^{-1}\|_\infty\|R\|_\infty$. Or on peut montrer (voir exercice 41 page 65) que $\mathrm{cond}(A) \simeq n^2$. Donc si on augmente le nombre de points, le conditionnement de A augmente aussi. Par exemple si $n = 10^4$, alors $\|\delta_x\|/\|x\| = 10^8\|\delta_b\|/\|b\|$. Or sur un ordinateur en simple précision, on a $\|\delta_b\|/\|b\| \geq 10^{-7}$, donc l'estimation (1.66) donne une estimation de l'erreur relative $\|\delta_x\|/\|x\|$ de 1000%, ce qui laisse à désirer pour un calcul qu'on espère précis.
En fait, l'estimation (1.66) ne sert à rien pour ce genre de problème, il faut faire une analyse un peu plus poussée, comme c'est fait dans l'exercice 43 page 65. On se rend compte alors que pour f donnée il existe $C \in \mathbb{R}_+$ ne dépendant que de f (mais pas de n) tel que

$$\frac{\|\boldsymbol{\delta_u}\|}{\|\boldsymbol{u}\|} \leq C\frac{\|\boldsymbol{\delta_b}\|}{\|\boldsymbol{b}\|} \text{ avec } \boldsymbol{b} = \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix}. \tag{1.71}$$

7. On appelle équation aux dérivées partielles une équation qui fait intervenir les dérivées partielles de la fonction inconnue, par exemple $\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$, où u est une fonction de $\mathbb{R}^2$ dans $\mathbb{R}$

L'estimation (1.71) est évidemment bien meilleure que l'estimation (1.66) puisqu'elle montre que l'erreur relative commise sur u est du même ordre que celle commise sur $\boldsymbol{b}$. En particulier, elle n'augmente pas avec le nombre de points de discrétisation. En conclusion, l'estimation (1.66) est peut-être optimale dans le cas d'une matrice quelconque, (on a montré ci-dessus qu'il peut y avoir égalité dans (1.66)) mais elle n'est pas toujours significative pour l'étude des systèmes linéaires issus de la discrétisation des équations aux dérivées partielles.

1.4.5 Exercices

Exercice 26 (Normes de l'Identité)**.**

Soit Id la matrice "Identité" de $\mathcal{M}_n(\mathbb{R})$. Montrer que pour toute norme induite on a $\|Id\| = 1$ et que pour toute norme matricielle on a $\|Id\| \geq 1$.

Exercice 27 (Normes induites particulières)**.** *Suggestions en page 66, corrigé détaillé en page 67.*

Soit $A = (a_{i,j})_{i,j\in\{1,\dots,n\}} \in \mathcal{M}_n(\mathbb{R})$.

1. On munit $\mathbb{R}^n$ de la norme $\|\cdot\|_\infty$ et $\mathcal{M}_n(\mathbb{R})$ de la norme induite correspondante, notée aussi $\|\cdot\|_\infty$. Montrer que
$$\|A\|_\infty = \max_{i\in\{1,\dots,n\}} \sum_{j=1}^{n} |a_{i,j}|.$$
2. On munit $\mathbb{R}^n$ de la norme $\|\cdot\|_1$ et $\mathcal{M}_n(\mathbb{R})$ de la norme induite correspondante, notée aussi $\|\cdot\|_1$. Montrer que
$$\|A\|_1 = \max_{j\in\{1,\dots,n\}} \sum_{i=1}^{n} |a_{i,j}|.$$

Exercice 28 (Norme non induite)**.**

Pour $A = (a_{i,j})_{i,j\in\{1,\dots,n\}} \in \mathcal{M}_n(\mathbb{R})$, on pose $\|A\|_s = (\sum_{i,j=1}^n a_{i,j}^2)^{\frac{1}{2}}$.

1. Montrer que $\|\cdot\|_s$ est une norme matricielle mais n'est pas une norme induite (pour $n > 1$).
2. Montrer que $\|A\|_s^2 = tr(A^tA)$. En déduire que $\|A\|_2 \leq \|A\|_s \leq \sqrt{n}\|A\|_2$ et que $\|Ax\|_2 \leq \|A\|_s\|x\|_2$, pour tout $A \in \mathcal{M}_n(\mathbb{R})$ et tout $x \in \mathbb{R}^n$.
3. Chercher un exemple de norme non matricielle.

Exercice 29 (Valeurs propres d'un produit de matrices)**.**

Soient p et n des entiers naturels non nuls tels que $n \leq p$, et soient $A \in \mathcal{M}_{n,p}(\mathbb{R})$ et $B \in \mathcal{M}_{p,n}(\mathbb{R})$. (On rappelle que $\mathcal{M}_{n,p}(\mathbb{R})$ désigne l'ensemble des matrices à n lignes et p colonnes.)

1. Montrer que λ est valeur propre non nulle de AB si et seulement si λ est valeur propre non nulle de BA.
2. Montrer que si $\lambda = 0$ est valeur propre de AB alors λ est valeur propre nulle de BA. *(Il est conseillé de distinguer les cas $Bx \neq 0$ et $Bx = 0$, où x est un vecteur propre associé à la $\lambda = 0$ valeur propre de AB. Pour le deuxième cas, on pourra distinguer selon que $ImA = \mathbb{R}^n$ ou non.)*
3. Montrer en donnant un exemple que λ peut être une valeur propre nulle de BA sans être valeur propre de AB. *(Prendre par exemple $n = 1$, $p = 2$.)*
4. On suppose maintenant que $n = p$, déduire des questions 1 et 2 que l'ensemble des valeurs propres de AB est égal à l'ensemble des valeurs propres de la matrice BA.

Exercice 30 (Rayon spectral)**.** *Corrigé en page 68.*

Soit $A \in \mathcal{M}_n(\mathbb{R})$. Montrer que si A est diagonalisable, il existe une norme induite sur $\mathcal{M}_n(\mathbb{R})$ telle que $\rho(A) = \|A\|$. Montrer par un contre exemple que ceci peut être faux si A n'est pas diagonalisable.

Exercice 31 (Sur le rayon spectral)**.**

On définit les matrices carrées d'ordre 2 suivantes :

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix},\ B = \begin{pmatrix} -1 & 0 \\ -1 & -1 \end{pmatrix},\ C = A + B.$$

Calculer le rayon spectral de chacune des matrices A, B et C et en déduire que le rayon spectral ne peut être ni une norme, ni même une semi-norme sur l'espace vectoriel des matrices.

Exercice 32 (Série de Neumann)**.** *Suggestions en page 67, corrigé détaillé en page 68.*

Soient $A \in \mathcal{M}_n(\mathbb{R})$.

1. Montrer que si $\rho(A) < 1$, les matrices $Id - A$ et $Id + A$ sont inversibles.

2. Montrer que la série de terme général A^k converge (vers $(Id - A)^{-1}$) si et seulement si $\rho(A) < 1$.

3. Montrer que si $\rho(A) < 1$, et si $\|\cdot\|$ une norme matricielle telle que $\|A\| < 1$, alors $\|(Id - A)^{-1}\| \leq \frac{1}{1-\|A\|}$ et $\|(Id + A)^{-1}\| \leq \frac{1}{1-\|A\|}$.

Exercice 33 (Normes induites)**.**

Soit $\|.\|$ une norme induite sur $\mathcal{M}_n(\mathbb{R})$ par une norme quelconque sur $\mathbb{R}^n$, et soit $A \in \mathcal{M}_n(\mathbb{R})$ telle que $\rho(A) < 1$ (on rappelle qu'on note $\rho(A)$ le rayon spectral de la matrice A). Pour $x \in \mathbb{R}^n$, on définit $\|x\|_*$ par :

$$\|x\|_* = \sum_{j=0}^{\infty} \|A^j x\|.$$

1. Montrer que l'application définie de $\mathbb{R}^n$ dans $\mathbb{R}$ par $x \mapsto \|x\|_*$ est une norme.
2. Soit $x \in \mathbb{R}^n$ tel que $\|x\|_* = 1$. Calculer $\|Ax\|_*$ en fonction de $\|x\|$, et en déduire que $\|A\|_* < 1$.
3. On ne suppose plus que $\rho(A) < 1$. Soit $\varepsilon > 0$ donné. Construire à partir de la norme $\|.\|$ une norme induite $\|.\|_{**}$ telle que $\|A\|_{**} \leq \rho(A) + \varepsilon$.

Exercice 34 (Un système par blocs)**.**

Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice carrée d'ordre N inversible, $b, c, f \in \mathbb{R}^n$. Soient α et $\gamma \in \mathbb{R}$. On cherche à résoudre le système suivant (avec $x \in \mathbb{R}^n$, $\lambda \in \mathbb{R}$) :

$$\begin{aligned} Ax + b\lambda &= f, \\ c \cdot x + \alpha\lambda &= \gamma. \end{aligned} \tag{1.72}$$

1. Ecrire le système (1.72) sous la forme : $My = g$, où M est une matrice carrée d'ordre $n+1$, $y \in \mathbb{R}^{n+1}$, $g \in \mathbb{R}^{n+1}$. Donner l'expression de M, y et g.
2. Donner une relation entre A, b, c et α, qui soit une condition nécessaire et suffisante pour que le système (1.72) soit inversible. Dans toute la suite, on supposera que cette relation est vérifiée.
3. On propose la méthode suivante pour la résolution du système (1.72) :
   (a) Soient z solution de $Az = b$, et h solution de $Ah = f$.
   (b) $x = h - \dfrac{\gamma - c \cdot h}{\alpha - c \cdot z} z, \qquad \lambda = \dfrac{\gamma - c \cdot h}{\alpha - c \cdot z}.$

   Montrer que $x \in \mathbb{R}^n$ et $\lambda \in \mathbb{R}$ ainsi calculés sont bien solutions du système (1.72).
4. On suppose dans cette question que A est une matrice bande, dont la largeur de bande est p.
   (a) Calculer le coût de la méthode de résolution proposée ci-dessus en utilisant la méthode LU pour la résolution des systèmes linéaires.

(b) Calculer le coût de la résolution du système $My = g$ par la méthode LU (en profitant ici encore de la structure creuse de la matrice A).

(c) Comparer et conclure.
Dans les deux cas, le terme d'ordre supérieur est $2nq^2$, et les coûts sont donc comparables.

Exercice 35 (Propriétés générales du conditionnement)**.** *Corrigé détaillé en page 69.*
On suppose que $\mathbb{R}^n$ est muni de la norme euclidienne usuelle $\|\cdot\| = \|\cdot\|_2$ et $\mathcal{M}_n(\mathbb{R})$ de la norme induite (notée aussi $\|\cdot\|_2$. On note alors $\mathrm{cond}_2(A)$ le conditionnement d'une matrice A inversible.

1. *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible. Montrer que $\mathrm{cond}_2(A) = 1$ si et seulement si $A = \alpha Q$ où $\alpha \in \mathbb{R}^\star$ et Q est une matrice orthogonale (c'est-à-dire $Q^t = Q^{-1}$).*
2. *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible. On suppose que $A = QR$ où Q est une matrice orthogonale. Montrer que $\mathrm{cond}_2(A) = \mathrm{cond}_2(R)$.*
3. *Soit A, $B \in \mathcal{M}_n(\mathbb{R})$ deux matrices symétriques définies positives. Montrer que*

$$\mathrm{cond}_2(A+B) \leq \max\{\mathrm{cond}_2(A), \mathrm{cond}_2(B)\}.$$

Exercice 36 (Conditionnement de la matrice transposée)**.** *On suppose que $A \in M_n(\mathbb{R})$ est inversible.*

1. *Montrer que si $B \in \mathcal{M}_n(\mathbb{R})$, on a pour tout $\lambda \in \mathcal{C}$,*

$$\det(AB - \lambda Id) = \det(BA - \lambda Id).$$

2. *En déduire que les rayons spectraux des deux matrices AB et BA sont identiques.*
3. *Montrer que $\|A^t\|_2 = \|A\|_2$.*
4. *En déduire que $cond_2(A) = cond_2(A^t)$.*
5. *Montrer que la norme induite associée à la norme 1 de $\mathbb{R}^n$ est donnée par :*

$$\|A\|_1 = \max_{j\in\{1,\ldots,nt\}} \sum_{i=1}^{n} |a_{ij}|$$

A-t-on $\|A^t\|_1 = \|A\|_1$?

6. *Montrer que dans le cas $n = 2$, on a toujours $cond_1(A) = cond_1(A^t), \forall A \in M_2(\mathbb{R})$.*
7. *Calculer $cond_1(A)$ pour $A = \begin{pmatrix} 2 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}$ et conclure.*

Exercice 37 (Majoration du conditionnement)**.**

Soit $\|\,.\,\|$ une norme induite sur $\mathcal{M}_n(\mathbb{R})$ et soit $A \in \mathcal{M}_n(\mathbb{R})$ telle que $\det(A) \neq 0$.

1. Montrer que si $\|A - B\| < \frac{1}{\|A^{-1}\|}$, alors B est inversible.
2. Montrer que cond $(A) \geq \sup_{\substack{B\in\mathcal{M}_n(\mathbb{R}) \\ \det B=0}} \frac{\|A\|}{\|A-B\|}$

Exercice 38 (Minoration du conditionnement)**.** *Corrigé détaillé en page 70.*

On note $\|\cdot\|$ une norme matricielle sur $\mathcal{M}_n(\mathbb{R})$. Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice carrée inversible, $\mathrm{cond}(A) = \|A\|\|A^{-1}\|$ le conditionnement de A, et soit $\delta_A \in \mathcal{M}_n(\mathbb{R})$.

1. Montrer que si $A + \delta_A$ est singulière, alors

$$\mathrm{cond}(A) \geq \frac{\|A\|}{\|\delta_A\|}. \tag{1.73}$$

2. On suppose dans cette question que la norme $\|\cdot\|$ est la norme induite par la norme euclidienne sur $\mathbb{R}^n$. Montrer que la minoration (1.73) est optimale, c'est-à-dire qu'il existe $\delta_A \in \mathcal{M}_n(\mathbb{R})$ telle que $A+\delta_A$ soit singulière et telle que l'égalité soit vérifiée dans (1.73).
   [*On pourra chercher δ_A de la forme*
   $$\delta_A = -\frac{y\,x^t}{x^t\,x},$$
   avec $y \in \mathbb{R}^n$ convenablement choisi et $x = A^{-1}y$.]
3. On suppose ici que la norme $\|\cdot\|$ est la norme induite par la norme infinie sur $\mathbb{R}^n$. Soit $\alpha \in]0,1[$. Utiliser l'inégalité (1.73) pour trouver un minorant, qui tend vers $+\infty$ lorsque α tend vers 0, de $\mathrm{cond}(A)$ pour la matrice
   $$A = \begin{pmatrix} 1 & -1 & 1 \\ -1 & \alpha & -\alpha \\ 1 & \alpha & \alpha \end{pmatrix}.$$

Exercice 39 (Conditionnement du carré)**.**

Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice telle que $\det A \neq 0$.

1. Quelle relation existe-t-il en général entre $\mathrm{cond}\,(A^2)$ et $(\mathrm{cond}\,A)^2$?
2. On suppose que A symétrique. Montrer que $\mathrm{cond}_2\,(A^2) = (\mathrm{cond}_2\,A)^2$.
3. On suppose que $\mathrm{cond}_2\,(A^2) = (\mathrm{cond}_2\,A)^2$. Peut-on conclure que A est symétrique ? (justifier la réponse.)

Exercice 40 (Calcul de l'inverse d'une matrice et conditionnement)**.** *Corrigé détaillé en page 71.*

On note $\|\cdot\|$ une norme matricielle sur $\mathcal{M}_n(\mathbb{R})$. Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice carrée inversible. On cherche ici des moyens d'évaluer la précision de calcul de l'inverse de A.

1. On suppose qu'on a calculé B, approximation (en raison par exemple d'erreurs d'arrondi) de la matrice A^{-1}. On pose :
   $$\left\{ \begin{array}{lllll} e_1 &=& \dfrac{\|B-A^{-1}\|}{\|A^{-1}\|}, & e_2 &= \dfrac{\|B^{-1}-A\|}{\|A\|} \\ \\ e_3 &=& \|AB-Id\|, & e_4 &= \|BA-Id\| \end{array} \right. \tag{1.74}$$
   (a) Expliquer en quoi les quantités e_1, e_2, e_3 et e_4 mesurent la qualité de l'approximation de A^{-1}.
   (b) On suppose ici que $B = A^{-1}+E$, où $\|E\| \leq \varepsilon\|A^{-1}\|$, et que
   $$\varepsilon \mathrm{cond}(A) < 1.$$
   Montrer que dans ce cas,
   $$e_1 \leq \varepsilon,\ e_2 \leq \frac{\varepsilon\mathrm{cond}(A)}{1-\varepsilon\mathrm{cond}(A)},\ e_3 \leq \varepsilon\mathrm{cond}(A) \text{ et } e_4 \leq \varepsilon\mathrm{cond}(A).$$
   (c) On suppose maintenant que $AB - Id = E'$ avec $\|E'\| \leq \varepsilon < 1$. Montrer que dans ce cas :
   $$e_1 \leq \varepsilon,\ e_2 \leq \frac{\varepsilon}{1-\varepsilon},\ e_3 \leq \varepsilon \text{ et } e_4 \leq \varepsilon\mathrm{cond}(A).$$
2. On suppose maintenant que la matrice A n'est connue qu'à une certaine matrice d'erreurs près, qu'on note δ_A.
   (a) Montrer que la matrice $A+\delta_A$ est inversible si $\|\delta_A\| < \dfrac{1}{\|A^{-1}\|}$.
   (b) Montrer que si la matrice $A+\delta_A$ est inversible, alors
   $$\frac{\|(A+\delta_A)^{-1}-A^{-1}\|}{\|(A+\delta_A)^{-1}\|} \leq \mathrm{cond}(A)\frac{\|\delta_A\|}{\|A\|}.$$

Exercice 41 (Conditionnement du Laplacien discret 1D)**.** *Suggestions en page 67, corrigé détaillé en page 72. Soit $f \in C([0,1])$. Soit $n \in \mathbb{N}^\star$, n impair. On pose $h = 1/(n+1)$. Soit A la matrice définie par (1.10) page 12, issue d'une discrétisation par différences finies (vue en cours) du problème (1.5a) page 11.*
Calculer les valeurs propres et les vecteurs propres de A. [On pourra commencer par chercher $\lambda \in \mathbb{R}$ et $\varphi \in C^2(\mathbb{R},\mathbb{R})$ (φ non identiquement nulle) t.q. $-\varphi''(x) = \lambda\varphi(x)$ pour tout $x \in]0,1[$ et $\varphi(0) = \varphi(1) = 0$].
Calculer $\mathrm{cond}_2(A)$ et montrer que $h^2\mathrm{cond}_2(A) \to \frac{4}{\pi^2}$ lorsque $h \to 0$.

Exercice 42 (Conditionnement, réaction diffusion 1d.)**.**

On s'intéresse au conditionnement pour la norme euclidienne de la matrice issue d'une discrétisation par Différences Finies du problème (1.28) étudié à l'exercice 10, qu'on rappelle :

$$\begin{array}{l} -u''(x) + u(x) = f(x),\ x \in]0,1[, \\ u(0) = u(1) = 0. \end{array} \tag{1.75}$$

Soit $n \in \mathbb{N}^\star$. On note $U = (u_j)_{j=1,\ldots,n}$ une "valeur approchée" de la solution u du problème (1.28) aux points $\left(\frac{j}{n+1}\right)_{j=1,\ldots,n}$. On rappelle que la discrétisation par différences finies de ce problème consiste à chercher U comme solution du système linéaire $AU = \left(f(\frac{j}{N+1})\right)_{j=1,\ldots,n}$ où la matrice $A \in \mathcal{M}_n(\mathbb{R})$ est définie par $A = (N+1)^2B + Id$, Id désigne la matrice identité et

$$B = \begin{pmatrix} 2 & -1 & 0 & \ldots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \ldots & 0 & -1 & 2 \end{pmatrix}$$

1. (Valeurs propres de la matrice B.)
   On rappelle que le problème aux valeurs propres

   $$\begin{array}{l} -u''(x) = \lambda u(x),\ x \in]0,1[, \\ u(0) = u(1) = 0. \end{array} \tag{1.76}$$

   admet la famille $(\lambda_k, u_k)_{k\in\mathbb{N}^*}$, $\lambda_k = (k\pi)^2$ et $u_k(x) = \sin(k\pi x)$ comme solution. Montrer que les vecteurs $U_k = \left(u_k(\frac{j}{n+1})\right)_{j=1,\ldots,n}$ sont des vecteurs propres de la matrice B. En déduire toutes les valeurs propres de la matrice B.
2. En déduire les valeurs propres de la matrice A.
3. En déduire le conditionnement pour la norme euclidienne de la matrice A.

Exercice 43 (Conditionnement "efficace")**.** *Suggestions en page 67.*

Soit $f \in C([0,1])$. Soit $n \in \mathbb{N}^\star$, n impair. On pose $h = 1/(n+1)$. Soit A la matrice définie par (1.10) page 12, issue d'une discrétisation par différences finies (vue en cours) du problème (1.5a) page 11.

Pour $u \in \mathbb{R}^n$, on note $u_1, \ldots, u_n$ les composantes de u. Pour $u \in \mathbb{R}^n$, on dit que $u \geq 0$ si $u_i \geq 0$ pour tout $i \in \{1,\ldots,n\}$. Pour u, $v \in \mathbb{R}^n$, on note $u \cdot v = \sum_{i=1}^n u_i v_i$.

On munit $\mathbb{R}^n$ de la norme suivante : pour $u \in \mathbb{R}^n$, $\|u\| = \max\{|u_i|,\ i \in \{1,\ldots,n\}\}$. On munit alors $\mathcal{M}_n(\mathbb{R})$ de la norme induite, également notée $\|\cdot\|$, c'est-à-dire $\|B\| = \max\{\|Bu\|,\ u \in \mathbb{R}^n \text{ t.q. } \|u\| = 1\}$, pour tout $B \in \mathcal{M}_n(\mathbb{R})$.

Partie I Conditionnement de la matrice et borne sur l'erreur relative

1. (Existence et positivité de A^{-1}) Soient $b \in \mathbb{R}^n$ et $u \in \mathbb{R}^n$ t.q. $Au = b$. Remarquer que $Au = b$ peut s'écrire :

$$\begin{cases} \frac{1}{h^2}(u_i - u_{i-1}) + \frac{1}{h^2}(u_i - u_{i+1}) = b_i, \ \forall i \in \{1, \dots, n\}, \\ u_0 = u_{n+1} = 0. \end{cases} \tag{1.77}$$

   Montrer que $b \geq 0 \Rightarrow u \geq 0$. [On pourra considérer $p \in \{0, \dots, n+1\}$ t.q. $u_p = \min\{u_j, \ j \in \{0, \dots, n+1\}$.]

   En déduire que A est inversible.

2. (Préliminaire) On considère la fonction $\varphi \in C([0,1], \mathbb{R})$ définie par $\varphi(x) = (1/2)x(1-x)$ pour tout $x \in [0,1]$. On définit alors $\boldsymbol{\phi} = (\phi_1, \dots \phi_n) \in \mathbb{R}^n$ par $\phi_i = \varphi(ih)$ pour tout $i \in \{1, \dots, n\}$. Montrer que $(A\boldsymbol{\phi})_i = 1$ pour tout $i \in \{1, \dots, n\}$.

3. (Calcul de $\|A^{-1}\|$) Soient $\boldsymbol{b} \in \mathbb{R}^n$ et $\boldsymbol{u} \in \mathbb{R}^n$ t.q. $A\boldsymbol{u} = \boldsymbol{b}$. Montrer que $\|\boldsymbol{u}\| \leq (1/8)\|\boldsymbol{b}\|$ [Calculer $A(\boldsymbol{u} \pm \|\boldsymbol{b}\|\boldsymbol{\phi})$ avec
   $bfphi$ défini à la question 2 et utiliser la question 1]. En déduire que $\|A^{-1}\| \leq 1/8$ puis montrer que $\|A^{-1}\| = 1/8$.

4. (Calcul de $\|A\|$) Montrer que $\|A\| = \frac{4}{h^2}$.

5. (Conditionnement pour la norme $\|\cdot\|$). Calculer $\|A^{-1}\|\|A\|$. Soient $\boldsymbol{b}$, $\delta_b \in \mathbb{R}^n$ et soient u, $\delta_u \in \mathbb{R}^n$ t.q. $Au = b$ et $A(u + \delta_u) = b + \delta_b$. Montrer que $\dfrac{\|\delta_u\|}{\|u\|} \leq \|A^{-1}\|\|A\| \dfrac{\|\delta_b\|}{\|b\|}$.

   Montrer qu'un choix convenable de $\boldsymbol{b}$ et δ_b donne l'égalité dans l'inégalité précédente.

Partie II Borne réaliste sur l'erreur relative : Conditionnement "efficace"

On se donne maintenant $f \in C([0,1], \mathbb{R})$ et on suppose (pour simplifier...) que $f(x) > 0$ pour tout $x \in]0,1[$. On prend alors, dans cette partie, $b_i = f(ih)$ pour tout $i \in \{1, \dots, n\}$. On considère aussi le vecteur ϕ défini à la question 2 de la partie I.

1. Montrer que

$$h \sum_{i=1}^{n} b_i \phi_i \to \int_0^1 f(x)\varphi(x)dx \text{ quand } n \to \infty$$

   et que

$$\sum_{i=1}^{n} b_i \phi_i > 0 \text{ pour tout } n \in \mathbb{N}^*.$$

   En déduire qu'il existe $\alpha > 0$, ne dépendant que de f, t.q. $h \sum_{i=1}^n b_i \phi_i \geq \alpha$ pour tout $n \in \mathbb{N}^*$.

2. Soit $u \in \mathbb{R}^n$ t.q. $Au = b$. Montrer que $n\|u\| \geq \sum_{i=1}^n u_i = u \cdot A\phi \geq \frac{\alpha}{h}$ (avec α donné à la question 1).

   Soit $\delta_b \in \mathbb{R}^n$ et $\delta_u \in \mathbb{R}^n$ t.q. $A(u + \delta_u) = b + \delta_b$. Montrer que $\dfrac{\|\delta_u\|}{\|u\|} \leq \dfrac{\|f\|_{L^\infty(]0,1[)}}{8\alpha} \dfrac{\|\delta_b\|}{\|b\|}$.

3. Comparer $\|A^{-1}\|\|A\|$ (question I.5) et $\dfrac{\|f\|_{L^\infty(]0,1[)}}{8\alpha}$ (question II.2) quand n est "grand" (ou quand $n \to \infty$).

1.4.6 Suggestions pour les exercices

Exercice 27 page 61 (Normes induites particulières)

1. Pour montrer l'égalité, prendre x tel que $x_j = \text{sign}(a_{i_0,j})$ où i_0 est tel que $\sum_{j=1,\dots,n} |a_{i_0,j}| \geq \sum_{j=1,\dots,n} |a_{i,j}|$, $\forall i = 1, \dots, n$, et $\text{sign}(s)$ désigne le signe de s.

2. Pour montrer l'égalité, prendre x tel que $x_{j_0} = 1$ et $x_j = 0$ si $j \neq j_0$, où j_0 est tel que $\sum_{i=1,\dots,n} |a_{i,j_0}| = \max_{j=1,\dots,n} \sum_{i=1,\dots,n} |a_{i,j}|$.

Exercice 32 page 62 (Série de Neumann)

1. Montrer que si $\rho(A) < 1$, alors 0 n'est pas valeur propre de $Id + A$ et $Id - A$.

2. Utiliser le corollaire 1.32.

Exercice 35 page 63 (Propriétés générales du conditionnement)

3. Soient $0 < \lambda_1 \leq \lambda_2 \ldots \leq \lambda_n$ et $0 < \mu_1 \leq \mu_2 \ldots \leq \mu_n$ les valeurs propres de A et B (qui sont s.d.p.). Montrer d'abord que :

$$cond_2(A+B) \leq \frac{\lambda_n + \mu_n}{\lambda_1 + \mu_1}.$$

Montrer ensuite que

$$\frac{a+b}{c+d} \leq \max(\frac{a}{c}, \frac{b}{d}),\ \forall (a,b,c,d) \in (\mathbb{R}_+^*)^4.$$

et conclure

Exercice 41 page 65(Conditionnement du Laplacien discret 1D)

2. Chercher les vecteurs propres $\Phi \in \mathbb{R}^n$ de A sous la forme $\Phi_j = \varphi(x_j)$, $j = 1, \ldots, n$ où φ est introduite dans les indications de l'énoncé. Montrer que les valeurs propres associées à ces vecteurs propres sont de la forme :

$$\lambda_k = \frac{2}{h^2}(1 - \cos k\pi h) = \frac{2}{h^2}(1 - \cos \frac{k\pi}{n+1}).$$

Exercice 43 page 65 (Conditionnement efficace)

Partie 1
1. Pour montrer que A est inversible, utiliser le théorème du rang.

2. Utiliser le fait que φ est un polynôme de degré 2.

3. Pour montrer que $\|A^{-1}\| = \dfrac{1}{8}$, remarquer que le maximum de φ est atteint en $x = .5$, qui correspond à un point de discrétisation car n est impair.

Partie 2 Conditionnement efficace
1. Utiliser la convergence uniforme des fonctions constantes par morceaux φ_h et f_h définies par

$$\begin{array}{ll} \varphi_h(x) & = \left\{ \begin{array}{l} \varphi(ih) = \phi_i \text{ si } x \in]x_i - \frac{h}{2},\ x_i + \frac{h}{2}[,\ i = 1, \ldots, n, \\ 0 \text{ si } x \in [0, \frac{h}{2}] \text{ ou } x \in]1 - \frac{h}{2}, 1]. \end{array} \right. \quad \text{et} \\ f_h(x) & = \left\{ \begin{array}{l} f(ih) = b_i \text{ si } x \in]x_i - \frac{h}{2},\ x_i + \frac{h}{2}[, \\ f(ih) = 0 \text{ si } x \in [0, \frac{h}{2}] \text{ ou } x \in]1 - \frac{h}{2}, 1]. \end{array} \right. \end{array}$$

2. Utiliser le fait que $A\phi = (1 \ldots 1)^t$.

1.4.7 Corrigés

Exercice 27 page 61 (Normes induites particulières)

1. Par définition, $\|A\|_\infty = \sup_{\substack{\boldsymbol{x} \in \mathbb{R}^n \\ \|\boldsymbol{x}\|_\infty = 1}} \|A\boldsymbol{x}\|_\infty$, et

$$\|A\boldsymbol{x}\|_\infty = \max_{i=1,\ldots,n} |\sum_{j=1,\ldots,n} a_{i,j} x_j| \leq \max_{i=1,\ldots,n} \sum_{j=1,\ldots,n} |a_{i,j}||x_j|.$$

Or $\|\boldsymbol{x}\|_\infty = 1$ donc $|x_j| \leq 1$ et

$$\|A\boldsymbol{x}\|_\infty \leq \max_{i=1,\ldots,n} \sum_{j=1,\ldots,n} |a_{i,j}|.$$

Montrons maintenant que la valeur $\alpha = \max_{i=1,\ldots,n} |\sum_{j=1,\ldots,n} |a_{i,j}|$ est atteinte, c'est-à-dire qu'il existe $\boldsymbol{x} \in \mathbb{R}^n$, $\|\boldsymbol{x}\|_\infty = 1$, tel que $\|A\boldsymbol{x}\|_\infty = \alpha$. Pour $s \in \mathbb{R}$, on note $\mathrm{sign}(s)$ le signe de s, c'est–à–dire

$$\mathrm{sign}(s) = |\begin{cases} s/|s \text{ si } s \neq 0, \\ 0 \text{ si } s = 0. \end{cases}$$

Choisissons $\boldsymbol{x} \in \mathbb{R}^n$ défini par $x_j = \mathrm{sign}(a_{i_0,j})$ où i_0 est tel que $\sum_{j=1,\ldots,n} |a_{i_0,j}| \geq \sum_{j=1,\ldots,n} |a_{i,j}|$, $\forall i = 1,\ldots,n$. On a bien $\|\boldsymbol{x}\|_\infty = 1$, et

$$\|A\boldsymbol{x}\|_\infty = \max_{i=1,\ldots,n} |\sum_{j=1}^{n} a_{i,j} sgn(a_{i_0,j})|.$$

Or, par choix de $\boldsymbol{x}$, on a

$$\sum_{j=1,\ldots,n} |a_{i_0,j}| = max_{i=1,\ldots,n} \sum_{j=1,\ldots,n} |a_{i,j}|.$$

On en déduit que pour ce choix de $\boldsymbol{x}$, on a bien $\|A\boldsymbol{x}\| = \max_{i=1,\ldots,n} |\sum_{j=1,\ldots,n} |a_{i,j}|$.

2. Par définition, $\|A\|_1 = \sup_{\substack{\boldsymbol{x}\in\mathbb{R}^n \\ \|\boldsymbol{x}\|_1=1}} \|A\boldsymbol{x}\|_1$, et

$$\|A\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |\sum_{j=1}^{n} a_{i,j}x_j| \leq \sum_{j=1}^{n} |x_j| \left(\sum_{i=1}^{n} |a_{i,j}|\right) \leq \max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{i,j}| \sum_{j=1,\ldots,n} |x_j|.$$

Et comme $\sum_{j=1}^{n} |x_j| = 1$, on a bien que $\|A\|_1 \leq \max_{j=1,\ldots,n} \sum_{i=1,\ldots,n} |a_{i,j}|$.
Montrons maintenant qu'il existe $\boldsymbol{x} \in \mathbb{R}^n$, $\|\boldsymbol{x}\|_1 = 1$, tel que $\|A\boldsymbol{x}\|_1 = \sum_{i=1,\ldots,n} |a_{i,j}|$. Il suffit de considérer pour cela le vecteur $\boldsymbol{x} \in \mathbb{R}^n$ défini par $x_{j_0} = 1$ et $x_j = 0$ si $j \neq j_0$, où j_0 est tel que $\sum_{i=1,\ldots,n} |a_{i,j_0}| = \max_{j=1,\ldots,n} \sum_{i=1,\ldots,n} |a_{i,j}|$. On vérifie alors facilement qu'on a bien $\|A\boldsymbol{x}\|_1 = \max_{j=1,\ldots,n} \sum_{i=1,\ldots,n} |a_{i,j}|$.

Exercice 30 page 61 (Rayon spectral)

Il suffit de prendre comme norme la norme définie par : $\|x\|^2 = \sum_{i=1}^{n} \alpha_i^2$ où les $(\alpha_i)_{i=1,n}$ sont les composantes de x dans la base des vecteurs propres associés à A. Pour montrer que ceci est faux dans le cas où A n'est pas diagonalisable, il suffit de prendre $A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, on a alors $\rho(A) = 0$, et comme A est non nulle, $\|A\| \neq 0$.

Exercice 32 page 62 (Série de Neumann)

1. Si $\rho(A) < 1$, les valeurs propres de A sont toutes différentes de 1 et -1. Donc 0 n'est pas valeur propre des matrices $Id - A$ et $Id + A$, qui sont donc inversibles.

2. Supposons que $\rho(A) < 1$. Il est facile de remarquer que

$$(\sum_{k=0}^{n} A^k)(Id - A) = Id - A^{n+1}. \tag{1.78}$$

Si $\rho(A) < 1$, d'après le corollaire 1.32, on a $A^k \to 0$ lorsque $k \to 0$. De plus, $Id - A$ est inversible. En passant à la limite dans (1.78) et on a donc

$$(Id - A)^{-1} = \sum_{k=0}^{+\infty} A^k. \tag{1.79}$$

Réciproquement, si $\rho(A) \geq 1$, la série ne peut pas converger en raison du corollaire 1.32.
3. On a démontré plus haut que si $\rho(A) < 1$, la série de terme général A^k est absolument convergente et qu'elle vérifie (1.79). On en déduit que si $\|A\| < 1$,

$$\|(Id - A)^{-1}\| \leq \sum_{k=0}^{+\infty} \|A^k\| = \frac{1}{1 - \|A\|}.$$

On a de même

$$(Id + A)^{-1} = \sum_{k=0}^{+\infty} (-1)^k A^k,$$

d'où on déduit de manière similaire que

$$\|(Id + A)^{-1}\| \leq \sum_{k=0}^{+\infty} \|A^k\| = \frac{1}{1 - \|A\|}.$$

Exercice 35 page 63 (propriétés générales du conditionnement)

1. Si $cond_2(A) = 1$, alors $\sqrt{\frac{\sigma_n}{\sigma_1}} = 1$ et donc toutes les valeurs propres de A^tA sont égales. Comme A^tA est symétrique définie positive (car A est inversible), il existe une base orthonormée $(\boldsymbol{f}_1 \dots \boldsymbol{f}_n)$ telle que $A^tA\boldsymbol{f}_i = \sigma \boldsymbol{f}_i, \ \forall i$ et $\sigma > 0$ (car A^tA est s.d.p.). On a donc $A^tA = \sigma Id \ A^t = \alpha^2 A^{-1}$ avec $\alpha = \sqrt{\sigma}$. En posant $Q = \frac{1}{\alpha}A$, on a donc $Q^t = \frac{1}{\alpha}A^t = \alpha A^{-1} = Q^{-1}$.
Réciproquement, si $A = \alpha Q$, alors $A^tA = \alpha^2 Id$, $\frac{\sigma_n}{\sigma_1} = 1$, et donc $cond_2(A) = 1$.

2. $A \in \mathcal{M}_n(\mathbb{R})$ est une matrice inversible. On suppose que $A = QR$ où Q est une matrice orthogonale. On a donc $cond_2(A) = \sqrt{\frac{\sigma_n}{\sigma_1}}$ où $\sigma_1 \leq \ldots \leq \sigma_n$ sont les valeurs propres de A^tA. Or $A^tA = (QR)^t(QR) = R^tQ^{-1}QR = R^tR$. Donc $cond_2(A) = cond_2(R)$.

3. Soient $0 < \lambda_1 \leq \lambda_2 \dots \leq \lambda_n$ et $0 < \mu_1 \leq \mu_2 \dots \leq \mu_n$ les valeurs propres de A et B (qui sont s.d.p.). Alors $cond_2(A + B) = \frac{\nu_n}{\nu_1}$, où $0 < \nu_1 \leq \ldots \leq \nu_n$ sont les valeurs propres de $A + B$.

a) On va d'abord montrer que

$$cond_2(A + B) \leq \frac{\lambda_n + \mu_n}{\lambda_1 + \mu_1}.$$

Remarquons en premier lieu que si A est s.d.p., alors

$$cond_2(A) = \frac{\sup_{\|x\|=1} Ax \cdot x}{\inf_{\|x\|_2=1} Ax \cdot x}$$

En effet, si A est s.d.p., alors $\sup_{\|x\|_2=1} Ax \cdot x = \lambda_n$; il suffit pour s'en rendre compte de décomposer x sur la base $(\boldsymbol{f}_i)_{i=1\dots n}$. Soit $x = \sum_{i=1}^{n} \alpha_i \boldsymbol{f}_i$. Alors : $Ax \cdot x = \sum_{i=1}^{n} \alpha_i^2 \lambda_i \leq \lambda_n \Sigma \alpha_i^2 = \lambda_n$. Et $A\boldsymbol{f}_n \cdot \boldsymbol{f}_n = \lambda_n$.
De même, $Ax \cdot x \geq \lambda_1 \Sigma \alpha_i^2 = \lambda_1$ et $Ax \cdot x = \lambda_1$ si $x = f_1$. Donc $\inf_{\|x\|=1} Ax \cdot x = \lambda_1$.

On en déduit que si A est s.d.p., $cond_2(A) = \frac{\sup_{\|x\|=1} Ax \cdot x}{\inf_{\|x\|=1} Ax \cdot x}$

Donc $cond_2(A + B) = \frac{\sup_{\|x\|=1}(A+B)x \cdot x}{\inf_{\|x\|=1}(A+B)x \cdot x}$

$$\text{Or } \sup_{\|x\|=1}(Ax\cdot x+Bx\cdot x)\leq \sup_{\|x\|=1}Ax\cdot x+\sup_{\|x\|=1}Bx\cdot x=\lambda_n+\mu_n$$
$$\text{et } \inf_{\|x\|=1}(Ax\cdot x+Bx\cdot x)\geq \inf_{\|x\|=1}Ax\cdot x+\inf_{\|x\|=1}Bx\cdot x=\lambda_1+\mu_1$$

donc

$$cond_2(A+B)\leq \frac{\lambda_n+\mu_n}{\lambda_1+\mu_1}.$$

b) On va montrer que

$$\frac{a+b}{c+d}\leq \max(\frac{a}{c},\frac{b}{d}),\ \forall (a,b,c,d)\in(\mathbb{R}_+^*)^4.$$

Supposons que $\frac{a+b}{c+d}\geq\frac{a}{c}$ alors $(a+b)c\geq(c+d)a$ c'est–à–dire $bc\geq da$ donc $bc+bd\geq da+db$ soit $b(c+d)\geq d(a+b)$; donc $\frac{a+b}{c+d}\leq\frac{b}{d}$. On en déduit que $cond_2(A+B)\leq\max(cond_2(A),cond_2(B))$.

Exercice 38 page 63 (Minoration du conditionnement)

1. Comme A est inversible, $A+\delta_A=A(Id+A^{-1}\delta_A)$, et donc si $A+\delta_A$ est singulière, alors $Id+A^{-1}\delta_A$ est singulière. Or on a vu en cours que toute matrice de la forme $Id+B$ est inversible si et seulement si $\rho(B)<1$. On en déduit que $\rho(A^{-1}\delta_A)\geq 1$, et comme

$$\rho(A^{-1}\delta_A)\leq\|A^{-1}\delta_A\|\leq\|A^{-1}\|\|\delta_A\|,$$

on obtient

$$\|A^{-1}\|\|\delta_A\|\geq 1,\ \text{soit encore } \mathrm{cond}(A)\geq\frac{\|A\|}{\|\delta_A\|}.$$

2. Soit $y\in\mathbb{R}^n$ tel que $\|y\|=1$ et $\|A^{-1}y\|=\|A^{-1}\|$. Soit $x=A^{-1}y$, et $\delta_A=\frac{-y\ x^t}{x^t\ x}$,on a donc

$$(A+\delta_A)x=Ax-\frac{-y\ x^t}{x^t\ x}x=y-\frac{-y\ x^tx}{x^t\ x}=0.$$

La matrice $A+\delta_A$ est donc singulière. De plus,

$$\|\delta_A\|=\frac{1}{\|x\|^2}\|y\ y^t\ A^{-t}\|.$$

Or par définition de x et y, on a $\|x\|^2=\|A^{-1}\|^2$. D'autre part, comme il s'agit ici de la norme L^2, on a $\|A^{-t}\|=\|A^{-1}\|$. On en déduit que

$$\|\delta_A\|=\frac{1}{\|A^{-1}\|^2}\|y\|^2\|A^{-1}\|=\frac{1}{\|A^{-1}\|}.$$

On a donc dans ce cas égalité dans (1.73).

3. Remarquons tout d'abord que la matrice A est inversible. En effet, $\det A=2\alpha^2>0$. Soit $\delta_A=\begin{pmatrix}0&0&0\\0&-\alpha&\alpha\\0&-\alpha&-\alpha\end{pmatrix}$.

Comme $\det(A+\delta_A)=0$, la matrice $A+\delta_A$ est singulière, et donc

$$\mathrm{cond}(A)\geq\frac{\|A\|}{\|\delta_A\|}. \tag{1.80}$$

Or $\|\delta_A\|=2\alpha$ et $\|A\|=\max(3,1+2\alpha)=3$, car $\alpha\in]0,1[$. Donc $\mathrm{cond}(A)\geq\frac{3}{2\alpha}$.

Exercice 40 page 64 (Calcul de l'inverse d'une matrice et conditionnement)

1. (a) L'inverse de la matrice A vérifie les quatre équations suivantes :

$$\begin{cases} X - A^{-1} = 0, & X^{-1} - A = 0, \\ AX - Id = 0, & XA - Id = 0. \end{cases}$$

Les quantités e_1, e_2, e_3 et e_4 sont les erreurs relatives commises sur ces quatre équations lorsqu'on remplace X par B ; en ce sens, elles mesurent la qualité de l'approximation de A^{-1}.

(b) On remarque d'abord que comme la norme est matricielle, on a $\|MP\| \leq \|M\|\|P\|$ pour toutes matrices M et P de $\mathcal{M}_n(\mathbb{R})$. On va se servir de cette propriété plusieurs fois par la suite.

(α) Comme $B = A^{-1} + E$, on a

$$e_1 = \frac{\|E\|}{\|A^{-1}\|} \leq \varepsilon \frac{\|A^{-1}\|}{\|A^{-1}\|} = \varepsilon.$$

(β) Par définition,

$$e_2 = \frac{\|B^{-1} - A\|}{\|A\|} = \frac{\|(A^{-1} + E)^{-1} - A\|}{\|A\|}.$$

Or

$$\begin{aligned} (A^{-1} + E)^{-1} - A &= (A^{-1}(Id + AE))^{-1} - A \\ &= (Id + AE)^{-1}A - A \\ &= (Id + AE)^{-1}(Id - (Id + AE))A \\ &= -(Id + AE)^{-1}AEA. \end{aligned}$$

On a donc

$$e_2 \leq \|(Id + AE)^{-1}\|\|A\|\|E\|.$$

Or par hypothèse, $\|AE\| \leq \|A\|\|E\| \leq \mathrm{cond}(A)\varepsilon < 1$; on en déduit, en utilisant le théorème 1.11, que :

$$\|(Id + AE))^{-1}\| \leq \frac{1}{1 - \|AE\|}, \text{ et donc } e_2 \leq \frac{\varepsilon \mathrm{cond}(A)}{1 - \varepsilon \mathrm{cond}(A)}.$$

(γ) Par définition, $e_3 = \|AB - Id\| = \|A(A^{-1} + E) - Id\| = \|AE\| \leq \|A\|\|E\| \leq \|A\|\varepsilon\|A^{-1}\| = \varepsilon \mathrm{cond}(A)$.

(δ) Enfin, $e_4 = \|BA - Id\| = \|(A^{-1} + E)A - Id\| \leq \|EA\| \leq \|E\|\|A\| \leq \varepsilon \mathrm{cond}(A)$.

(c) (α) Comme $B = A^{-1}(Id + E')$, on a

$$e_1 = \frac{\|A^{-1}(Id + E') - A^{-1}\|}{\|A^{-1}\|} \leq \|Id + E' - Id\| \leq \varepsilon.$$

(β) Par définition,

$$\begin{aligned} e_2 &= \frac{\|(Id+E')^{-1}A - A\|}{\|A\|} \\ &= \frac{\|(Id+E')^{-1}(A - (Id+E')A)\|}{\|A\|} \\ &\leq \|(Id + E')^{-1}\|\|Id - (Id + E')\| \leq \frac{\varepsilon}{1-\varepsilon} \end{aligned}$$

car $\varepsilon < 1$ (théorème 1.1).

(γ) Par définition, $e_3 = \|AB - Id\| = \|AA^{-1}(Id + E') - Id\| = \|E'\| \leq \varepsilon$.

(δ) Enfin, $e_4 = \|BA - Id\| = \|A^{-1}(Id + E')A - Id\| = \|A^{-1}(A + E'A - A)\| \leq \|A^{-1}\|\|AE'\| \leq \varepsilon \mathrm{cond}(A)$.

2. (a) On peut écrire $A+\delta_A = A(Id + A^{-1}\delta_A)$. On a vu en cours (théorème 1.11) que si $\|A^{-1}\delta_A\| < 1$, alors la matrice $Id + A^{-1}\delta_A$ est inversible. Or $\|A^{-1}\delta_A\| \leq \|A^{-1}\|\|\delta_A\|$, et donc la matrice $A+\delta_A$ est inversible si $\|\delta_A\| < \dfrac{1}{\|A^{-1}\|}$.

   (b) On peut écrire $\|(A+\delta_A)^{-1} - A^{-1}\| = \|(A+\delta_A)^{-1}(Id - (A+\delta_A)A^{-1}\| \leq \|(A+\delta_A)^{-1}\|\|Id - Id - \delta_A A^{-1}\| \leq \|(A+\delta_A)^{-1}\|\delta_A\|\|A^{-1}\|$. On en déduit le résultat.

Exercice 41 page 65 (Conditionnement du Laplacien discret 1D)

Pour chercher les valeurs propres et vecteurs propres de A, on s'inspire des valeurs propres et vecteurs propres du problème continu, c'est–à–dire des valeurs λ et fonctions φ telles que

$$\begin{cases} -\varphi''(x) = \lambda\varphi(x) & x \in]0,1[\\ \varphi(0) = \varphi(1) = 0 \end{cases} \tag{1.81}$$

(Notons que ce "truc" ne marche pas dans n'importe quel cas.)
L'ensemble des solutions de l'équation différentielle $-\varphi'' = \lambda\varphi$ est un espace vectoriel d'ordre 2. donc φ est de la forme $\varphi(x) = \alpha\cos\sqrt{\lambda}x + \beta\sin\sqrt{\lambda}x$ ($\lambda \geq 0$) et α et β dont déterminés par les conditions aux limites $\varphi(0) = \alpha = 0$ et $\varphi(1) = \alpha\cos\sqrt{\lambda} + \beta\sin\sqrt{\lambda} = 0$; on veut $\beta \neq 0$ car on cherche $\varphi \neq 0$ et donc on obtient $\lambda = k^2\pi^2$. Les couples (λ, φ) vérifiant (1.81) sont donc de la forme $(k^2\pi^2, \sin k\pi x)$.
2. Pour $k = 1$ à n, posons $\Phi_i^{(k)} = \sin k\pi x_i$, où $x_i = ih$, pour $i = 1$ à n, et calculons $A\Phi^{(k)}$:

$$(A\Phi^{(k)})_i = -sink\pi(i-1)h + 2\sin k\pi(ih) - \sin k\pi(i+1)h.$$

En utilisant le fait que $\sin(a+b) = \sin a\cos b + \cos a\sin b$ pour développer $\sin k\pi(1-i)h$ et $\sin k\pi(i+1)h$, on obtient (après calculs) :

$$(A\Phi^{(k)})_i = \lambda_k\Phi_i^{(k)}, \quad i = 1, \ldots, n,$$

avec

$$\lambda_k = \frac{2}{h^2}(1 - \cos k\pi h) = \frac{2}{h^2}(1 - \cos\frac{k\pi}{n+1}) \tag{1.82}$$

On a donc trouvé n valeurs propres $\lambda_1, \ldots, \lambda_n$ associées aux vecteurs propres $\Phi^{(1)}, \ldots, \Phi^{(n)}$ de $\mathbb{R}^n$ définis par $\Phi_i^{(k)} = \sin\dfrac{k\pi i}{n+1}$, $i = 1 \ldots n$.
Remarque : Lorsque $n \to +\infty$ (ou $h \to 0$), on a

$$\lambda_k^{(h)} = \frac{2}{h^2}\left(1 - 1 + \frac{k^2\pi^2h^2}{2} + O(h^4)\right) = k^2\phi^2 + O(h^2)$$

Donc

$$\lambda_k^{(h)} \to k^2\pi^2 = \lambda_k \text{ lorsque } h \to 0.$$

Calculons maintenant $\mathrm{cond}_2(A)$. Comme A est s.d.p., on a

$$cond_2(A) = \frac{\lambda_n}{\lambda_1} = \frac{1 - \cos\frac{n\pi}{n+1}}{1 - \cos\frac{\pi}{n+1}}$$

On a : $h^2\lambda_n = 2(1 - \cos\frac{n\pi}{n+1}) \to 4$ et $\lambda_1 \to \pi^2$ lorsque $h \to 0$. Donc

$$h^2 cond_2(A) \to \frac{4}{\pi^2} \text{ lorsque } h \to 0.$$

Exercice 43 page 65 (Conditionnement "efficace")

Partie I

1. Soit $u = (u_1, \dots, u_n)^t$. On a

$$Au = b \Leftrightarrow \left\{ \begin{array}{l} \frac{1}{h^2}(u_i - u_{i-1}) + \frac{1}{h^2}(u_i - u_{i+1}) = b_i, \ \forall i = 1, \dots, n, \\ u_0 = u_{n+1} = 0. \end{array} \right.$$

Supposons $b_i \geq 0, \forall i = 1, \dots, n$, et soit

$$p = \min\{k \in \{0, \dots, n+1\}; u_k = \min\{u_i, \ i = 0, \dots, n+1\}\}.$$

Remarquons que p ne peut pas être égal à $n+1$ car $u_0 = u_{n+1} = 0$. Si $p = 0$, alors $u_i \geq 0 \ \forall i = 0, n+1$ et donc $u \geq 0$.
Si $p \in \{1, \dots, n\}$, alors

$$\frac{1}{h^2}(u_p - u_{p-1}) + \frac{1}{h^2}(u_p - u_{p+1}) \geq 0;$$

mais par définition de p, on a $u_p - u_{p-1} < 0$ et $u_p - u_{p+1} \leq 0$, et on aboutit donc à une contradiction.

Montrons maintenant que A est inversible. On vient de montrer que si $Au \geq 0$ alors $u \geq 0$. On en déduit par linéarité que si $Au \leq 0$ alors $u \leq 0$, et donc que si $Au = 0$ alors $u = 0$. Ceci démontre que l'application linéaire représentée par la matrice A est injective donc bijective (car on est en dimension finie).

2. Soit $\varphi \in C([0,1], \mathbb{R})$ tel que $\varphi(x) = \frac{1}{2}x(1-x)$ et $\phi_i = \varphi(x_i), \ i = 1, n$, où $x_i = ih$.
On remarque que $(A\phi)_i$ est le développement de Taylor à l'ordre 2 de $\varphi(x_i)$. En effet, φ est un polynôme de degré 2, sa dérivée troisième est nulle ; de plus on a $\varphi'(x) = \frac{1}{2} - x$ et $\varphi''(x) = 1$. On a donc :

$$\phi_{i+1} = \phi_i + h\varphi'(x_i) - \frac{h^2}{2}$$
$$\phi_{i-1} = \phi_i - h\varphi'(x_i) - \frac{h^2}{2}$$

On en déduit que $\frac{1}{h^2}(2\phi_i - \phi_{i+1} - \phi_{i+1}) = 1$, et donc que $(A\phi)_i = 1$.

3. Soient $b \in \mathbb{R}^n$ et $u \in \mathbb{R}^n$ tels que $Au = b$. On a :

$$(A(u \pm \|b\|\varphi))_i = (Au)_i \pm \|b\|(A\phi)_i = b_i \pm \|b\|.$$

Prenons d'abord $\tilde{b}_i = b_i + \|b\| \geq 0$, alors par la question (1),

$$u_i + \|b\|\phi_i \geq 0 \ \ \forall i = 1 \dots n.$$

Si maintenant on prend $\bar{b}_i = b_i - \|b\| \leq 0$, alors

$$u_i - \|b\|\phi_i \leq 0 \ \ \forall i = 1, \dots, n.$$

On a donc $-\|b\|\phi_i \leq u_i \leq \|b\|\phi_i$.
On en déduit que $\|u\| \leq \|b\| \ \|\phi\|$; or $\|\phi\| = \frac{1}{8}$. D'où $\|u\| \leq \frac{1}{8}\|b\|$.
On peut alors écrire que pour tout $b \in \mathbb{R}^n$,

$$\|A^{-1}b\| \leq \frac{1}{8}\|b\|, \text{ donc } \frac{\|A^{-1}b\|}{\|b\|} \leq \frac{1}{8}, \text{ d'où } \|A^{-1}\| \leq \frac{1}{8}.$$

On montre que $\|A^{-1}\| = \dfrac{1}{8}$ en prenant le vecteur $\boldsymbol{b}$ défini par $b(x_i) = 1, \forall i = 1, \ldots, n$. On a en effet $A^{-1}b = \phi$, et comme n est impair, $\exists i \in \{1, \ldots, n\}$ tel que $x_i = \frac{1}{2}$;or $\|\varphi\| = \varphi(\frac{1}{2}) = \frac{1}{8}$.

4. Par définition, on a $\|A\| = \sup_{\|x\|=1} \|Ax\|$, et donc $\|A\| = \max_{i=1,n} \sum_{j=1,n} |a_{i,j}|$, d'où le résultat.

5. Grâce aux questions 3 et 4, on a, par définition du conditionnement pour la norme $\|\cdot\|$, $\mathrm{cond}(A) = \|A\|\|A^{-1}\| = \frac{1}{2h^2}$.
Comme $A\delta_u = \delta_b$, on a :

$$\|\delta_u\| \leq \|A^{-1}\|\|\delta_b\|\frac{\|b\|}{\|b\|} \leq \|A^{-1}\|\|\delta_b\|\frac{\|A\|\|u\|}{\|b\|},$$

d'où le résultat.
Pour obtenir l'égalité, il suffit de prendre $b = Au$ où u est tel que $\|u\| = 1$ et $\|Au\| = \|A\|$, et δ_b tel que $\|\delta_b\| = 1$ et $\|A^{-1}\delta_b\| = \|A^{-1}\|$. On obtient alors

$$\frac{\|\delta_b\|}{\|b\|} = \frac{1}{\|A\|} \text{ et } \frac{\|\delta u\|}{\|u\|} = \|A^{-1}\|.$$

D'où l'égalité.

Partie 2 Conditionnement "efficace"

1. Soient φ_h et f_h les fonctions constantes par morceaux définies par

$$\begin{array}{rcl} \varphi_h(x) & = & \left\{ \begin{array}{l} \varphi(ih) = \phi_i \text{ si } x \in]x_i - \frac{h}{2},\ x_i + \frac{h}{2}[,\ i = 1, \ldots, n, \\ 0 \text{ si } x \in [0, \frac{h}{2}] \text{ ou } x \in]1 - \frac{h}{2}, 1]. \end{array} \right. \quad \text{et} \\ f_h(x) & = & \left\{ \begin{array}{l} f(ih) = b_i \text{ si } x \in]x_i - \frac{h}{2},\ x_i + \frac{h}{2}[, \\ f(ih) = 0 \text{ si } x \in [0, \frac{h}{2}] \text{ ou } x \in]1 - \frac{h}{2}, 1]. \end{array} \right. \end{array}$$

Comme $f \in C([0,1], \mathbb{R})$ et $\varphi \in C^2([0,1], \mathbb{R})$, la fonction f_h (resp. φ_h) converge uniformément vers f (resp. φ) lorsque $h \to 0$. En effet,

$$\begin{aligned} \|f - f_h\|_\infty &= \sup_{x\in[0,1]} |f(x) - f_h(x)| \\ &= \max_{i=0,\ldots,n} \sup_{x\in[x_i,x_{i+1}]} |f(x) - f_h(x)| \\ &= \max_{i=0,\ldots,n} \sup_{x\in[x_i,x_{i+1}]} |f(x) - f(x_i)| \end{aligned}$$

Comme f est continue, elle est uniformément continue sur $[0,1]$ et donc pour tout $\varepsilon > 0$, il existe $h_\varepsilon > 0$ tel que si $|s - t| \leq h_\varepsilon$, alors $|f(s) - f(t)|$. On en conclut que si l'on prend $h \leq h_\varepsilon$, on a $\|f - f_h\| \leq \varepsilon$. Le raisonnement est le même pour φ_h, et donc $f_h\varphi_h$ converge uniformément vers $f\varphi$. On peut donc passer à la limite sous l'intégrale et écrire que :

$$h\sum_{i=1}^{n} b_i\varphi_i = \int_0^1 f_h(x)\varphi_h(x)dx \to \int_0^1 f(x)\varphi(x)dx \text{ lorsque } h \to 0.$$

Comme $b_i > 0$ et $\phi_i > 0\ \forall i = 1, \ldots, n$, on a évidemment

$$S_n = \sum_{i=1}^{n} b_i\varphi_i > 0 \text{ et } S_n \to \int_0^1 f(x)\varphi(x)dx = \beta > 0 \text{ lorsque } h \to 0.$$

Donc il existe $n_0 \in \mathbb{N}$ tel que si $n \geq n_0$, $S_n \geq \dfrac{\beta}{2}$, et donc $S_n \geq \alpha = \min(S_0, S_1 \ldots S_{n_0}, \dfrac{\beta}{2}) > 0$.

2. On a $n\|u\| = n \sup_{i=1,n} |u_i| \geq \sum_{i=1}^n u_i$. D'autre part, $A\varphi = (1 \dots 1)^t$ donc $u \cdot A\varphi = \sum_{i=1}^n u_i$; or $u \cdot A\varphi = A^t u \cdot \varphi = Au \cdot \varphi$ car A est symétrique. Donc $u \cdot A\varphi = \sum_{i=1}^n b_i \varphi_i \geq \frac{\alpha}{h}$ d'après la question 1. Comme $\delta_u = A^{-1}\delta_b$, on a donc $\|\delta_u\| \leq \|A^{-1}\| \ \|\delta_b\|$; et comme $n\|u\| \geq \frac{\alpha}{h}$, on obtient : $\frac{\|\delta_u\|}{\|u\|} \leq \frac{1}{8}\frac{hn}{\alpha}\|\delta_b\|\frac{\|f\|}{\|b\|}$. Or $hn \leq 1$ et on a donc bien :

$$\frac{\|\delta_u\|}{\|u\|} \leq \frac{\|f\|}{8\alpha}\frac{\|\delta_b\|}{\|b\|}.$$

3. Le conditionnement $\mathrm{cond}(A)$ calculé dans la partie 1 est d'ordre $1/h^2$, et donc tend vers l'infini lorsque le pas de discrétisation tend vers 0, alors qu'on vient de montrer dans la partie 2 que la variation relative $\frac{\|\delta_u\|}{\|u\|}$ est inférieure à une constante multipliée par la variation relative de $\frac{\|\delta_b\|}{\|b\|}$. Cette dernière information est nettement plus utile et réjouissante pour la résolution effective du système linéaire.

1.5 Méthodes itératives

Les méthodes directes sont très efficaces : elles donnent la solution exacte (aux erreurs d'arrondi près) du système linéaire considéré. Elles ont l'inconvénient de nécessiter une assez grande place mémoire car elles nécessitent le stockage de toute la matrice en mémoire vive. Si la matrice est pleine, c.à.d. si la plupart des coefficients de la matrice sont non nuls et qu'elle est trop grosse pour la mémoire vive de l'ordinateur dont on dispose, il ne reste plus qu'à gérer habilement le "swapping" c'est–à–dire l'échange de données entre mémoire disque et mémoire vive pour pouvoir résoudre le système.
Cependant, si le système a été obtenu à partir de la discrétisation d'équations aux dérivés partielles, il est en général "creux", c.à. d. qu'un grand nombre des coefficients de la matrice du système sont nuls ; de plus la matrice a souvent une structure "bande", *i.e.* les éléments non nuls de la matrice sont localisés sur certaines diagonales. On a vu au chapitre précédent que dans ce cas, la méthode de Choleski "conserve le profil" (voir à ce propos page 37). Si on utilise une méthode directe genre Choleski, on aura donc besoin d'une place mémoire de $n \times M = M^3$. (Notons que pour une matrice pleine on a besoin de M^4.)
Lorsqu'on a affaire à de très gros systèmes issus par exemple de l'ingénierie (calcul des structures, mécanique des fluides, ...), où n peut être de l'ordre de plusieurs milliers, on cherche à utiliser des méthodes nécessitant le moins de mémoire possible. On a intérêt dans ce cas à utiliser des méthodes itératives. Ces méthodes ne font appel qu'à des produits matrice vecteur, et ne nécessitent donc pas le stockage du profil de la matrice mais uniquement des termes non nuls. Dans l'exemple précédent, on a 5 diagonales non nulles, donc la place mémoire nécessaire pour un produit matrice vecteur est $5n = 5M^2$. Ainsi pour les gros systèmes, il est souvent avantageux d'utiliser des méthodes itératives qui ne donnent pas toujours la solution exacte du système en un nombre fini d'itérations, mais qui donnent une solution approchée à coût moindre qu'une méthode directe, car elles ne font appel qu'à des produits matrice vecteur.

Remarque 1.44 (Sur la méthode du gradient conjugué)**.**
Il existe une méthode itérative "miraculeuse" de résolution des systèmes linéaires lorsque la matrice A est symétrique définie positive : c'est la méthode du gradient conjugué. Elle est miraculeuse en ce sens qu'elle donne la solution exacte du système $Ax = b$ en un nombre fini d'opérations (en ce sens c'est une méthode directe) : moins de n itérations où n est l'ordre de la matrice A, bien qu'elle ne nécessite que des produits matrice vecteur ou des produits scalaires. La méthode du gradient conjugué est en fait une méthode d'optimisation pour la recherche du minimum dans $\mathbb{R}^n$ de la fonction de $\mathbb{R}^n$ dans $\mathbb{R}$ définie par : $f(x) = \frac{1}{2}Ax \cdot x - b \cdot x$. Or on peut montrer que lorsque A est symétrique définie positive, la recherche de x minimisant f dans $\mathbb{R}^n$ est équivalent à la résolution du système $Ax = b$. (Voir paragraphe **??** *page* **??**.*) En fait, la méthode du gradient conjugué n'est pas si miraculeuse que cela en pratique : en effet, le nombre n est en général très grand et on ne peut en géneral pas envisager d'effectuer un tel nombre d'itérations pour résoudre le système. De plus, si on utilise la méthode du gradient conjugué brutalement, non seulement elle ne donne pas la solution en n itérations en raison de l'accumulation des erreurs d'arrondi, mais plus la taille du système croît et plus le nombre d'itérations nécessaires devient élevé. On a alors recours aux techniques de "préconditionnement". Nous reviendrons sur ce point au chapitre 3.*
La méthode itérative du gradient à pas fixe, qui est elle aussi obtenue comme méthode de minimisation de la fonction f ci-dessus, fait l'objet de l'exercice 45 page 88 et du théorème **??** *page* **??**.

1.5.1 Définition et propriétés

Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible et $\boldsymbol{b} \in \mathbb{R}^n$, on cherche toujours ici à résoudre le système linéaire (1.1) c'est–à–dire à trouver $\boldsymbol{x} \in \mathbb{R}^n$ tel que $A\boldsymbol{x} = \boldsymbol{b}$, mais de façon itérative, c.à.d. par la construction d'une suite.

Définition 1.45 (Méthode itérative)**.** *On appelle méthode itérative de résolution du système linéaire* (1.1) *une méthode qui construit une suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ (où l'itéré $\boldsymbol{x}^{(k)}$ est calculé à partir des itérés $\boldsymbol{x}^{(0)} \dots \boldsymbol{x}^{(k-1)}$ censée converger vers x solution de* (1.1)*).*

Bien sûr, on souhaite que cette suite converge vers la solution $\boldsymbol{x}$ du système.

Définition 1.46 (Méthode itérative convergente)**.** *On dit qu'une méthode itérative est convergente si pour tout choix initial $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$, on a :*

$$\boldsymbol{x}^{(k)} \longrightarrow \boldsymbol{x} \text{ quand } k \to +\infty$$

Enfin, on veut que cette suite soit simple à calculer. Une idée naturelle est de travailler avec une matrice P inversible qui soit "proche" de A, mais plus facile que A à inverser. On appelle matrice de préconditionnement cette matrice P. On écrit alors $A = P - (P - A) = P - N$ (avec $N = P - A$), et on réécrit le système linéaire $A\boldsymbol{x} = \boldsymbol{b}$ sous la forme

$$P\boldsymbol{x} = (P - A)\boldsymbol{x} + \boldsymbol{b} = N\boldsymbol{x} + \boldsymbol{b}. \tag{1.83}$$

Cette forme suggère la construction de la suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ à partir d'un choix initial $\boldsymbol{x}^{(0)}$ donné, par la formule suivante :

$$\begin{aligned} P\boldsymbol{x}^{(k+1)} &= (P - A)\boldsymbol{x}^{(k)} + \boldsymbol{b} \\ &= N\boldsymbol{x}^{(k)} + \boldsymbol{b}, \end{aligned} \tag{1.84}$$

ce qui peut également s'écrire :.

$$\boldsymbol{x}^{(k+1)} = B\boldsymbol{x}^{(k)} + \boldsymbol{c}, \text{ avec } B = P^{-1}(P - A) = \mathrm{Id} - P^{-1}A = P^{-1}N \text{ et } \boldsymbol{c} = P^{-1}\boldsymbol{b}. \tag{1.85}$$

Remarque 1.47 (Convergence vers $A^{-1}\boldsymbol{b}$)**.** *Si $P\boldsymbol{x}^{(k+1)} = (P - A)\boldsymbol{x}^{(k)} + b$ pour tout $k \in \mathbb{N}$ et $\boldsymbol{x}^{(k)} \longrightarrow \bar{\boldsymbol{x}}$ quand $k \longrightarrow +\infty$ alors $P\bar{\boldsymbol{x}} = (P - A)\bar{\boldsymbol{x}} + b$, et donc $A\bar{\boldsymbol{x}} = \boldsymbol{b}$, c.à.d. $\bar{\boldsymbol{x}} = \boldsymbol{x}$. En conclusion, si la suite converge, alors elle converge bien vers la solution du système linéaire.*

On introduit l'erreur d'approximation $\boldsymbol{e}^{(k)}$ à l'itération k, définie par

$$\boldsymbol{e}^{(k)} = \boldsymbol{x}^{(k)} - \boldsymbol{x}, \ k \in \mathbb{N} \tag{1.86}$$

où $\boldsymbol{x}^{(k)}$ est construit par (1.85) et $\boldsymbol{x} = A^{-1}\boldsymbol{b}$. Il est facile de vérifier que $\boldsymbol{x}^{(k)} \to \boldsymbol{x} = A^{-1}\boldsymbol{b}$ lorsque $k \to +\infty$ si et seulement si $\boldsymbol{e}^{(k)} \to \boldsymbol{0}$ lorsque $k \to +\infty$

Lemme 1.48. *La suite $(\boldsymbol{e}^{(k)})_{k\in\mathbb{N}}$ définie par* (1.86) *est également définie par*

$$\begin{aligned} \boldsymbol{e}^{(0)} &= \boldsymbol{x}^{(0)} - \boldsymbol{x} \\ \boldsymbol{e}^{(k)} &= B^k\boldsymbol{e}^{(0)} \end{aligned} \tag{1.87}$$

DÉMONSTRATION – Comme $\boldsymbol{c} = P^{-1}\boldsymbol{b} = P^{-1}A\boldsymbol{x}$, on a

$$\boldsymbol{e}^{(k+1)} = \boldsymbol{x}^{(k+1)} - \boldsymbol{x} = B\boldsymbol{x}^{(k)} - \boldsymbol{x} + P^{-1}A\boldsymbol{x} \tag{1.88}$$

$$= B(\boldsymbol{x}^{(k)} - \boldsymbol{x}). \tag{1.89}$$

Par récurrence sur k,

$$\boldsymbol{e}^{(k)} = B^k(\boldsymbol{x}^{(0)} - \boldsymbol{x}), \quad \forall k \in \mathbb{N}. \tag{1.90}$$

■

Théorème 1.49 (Convergence de la suite)**.** *Soit A et $P \in \mathcal{M}_n(\mathbb{R})$ des matrices inversibles. Soit $\boldsymbol{x}^{(0)}$ donné et soit $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ la suite définie par* (1.85)*.*

1. *La suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ converge vers $\boldsymbol{x} = A^{-1}\boldsymbol{b}$ si et seulement si $\rho(B) < 1$.*
2. *La suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ converge si et seulement si il existe une norme induite notée $\|\cdot\|$ telle que $\|B\| < 1$.*

DÉMONSTRATION –

1. On a vu que la suite $(\boldsymbol{x})^{(k)})_{k\in\mathbb{N}}$ définie par (1.85) converge vers $\boldsymbol{x} = A^{-1}\boldsymbol{b}$ si et seulement si la suite $\boldsymbol{e}^{(k)}$ définie par (1.87) tend vers $\mathbf{0}$. On en déduit par le lemme 1.32 que la suite $(\boldsymbol{x})^{(k)})_{k\in\mathbb{N}}$ converge (vers $\boldsymbol{x}$) si et seulement si $\rho(B) < 1$.
2. Si il existe une norme induite notée $\|\cdot\|$ telle que $\|B\| < 1$, alors en vertu du corollaire 1.32, $\rho(B) < 1$ et donc la méthode converge ce qui précède.
   Réciproquement, si la méthode converge alors $\rho(B) < 1$, et donc il existe $\eta > 0$ tel que $\rho(B) = 1 - \eta$. Prenons maintenant $\varepsilon = \frac{\eta}{2}$ et appliquons la proposition 1.31 : il existe une norme induite $\|\cdot\|$ telle que $\|B\| \leq \rho(B) + \varepsilon < 1$, ce qui démontre le résultat.

■

Pour trouver des méthodes itératives de résolution du système (1.1), on cherche donc une décomposition de la matrice A de la forme : $A = P - (P - A) = P - N$, où P est inversible et telle que le système $Py = d$ soit un système facile à résoudre (par exemple P diagonale ou triangulaire).

Estimation de la vitesse de convergence Soit $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$ donné et soit $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ la suite définie par (1.85). On a vu que si $\rho(B) < 1$), $\boldsymbol{x}^{(k)} \to x$ quand $k \to \infty$, où $\boldsymbol{x}$ est la solution du système $A\boldsymbol{x} = b$. On montre à l'exercice 59 page 113 que (sauf cas particuliers)

$$\frac{\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}\|}{\|\boldsymbol{x}^{(k)} - \boldsymbol{x}\|} \longrightarrow \rho(B) \quad \text{lorsque } k \to +\infty,$$

indépendamment de la norme sur $\mathbb{R}^n$. Le rayon spectral $\rho(B)$ de la matrice B est donc une bonne estimation de la vitesse de convergence. Pour estimer cette vitesse de convergence lorsqu'on ne connaît pas $\boldsymbol{x}$, on peut utiliser le fait (voir encore l'exercice 59 page 113) qu'on a aussi

$$\frac{\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\|}{\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}\|} \longrightarrow \rho(B) \ : \ \text{lorsque } k \to +\infty,$$

ce qui permet d'évaluer la vitesse de convergence de la méthode par le calcul des itérés courants.

1.5.2 Quelques exemples de méthodes itératives

Une méthode simpliste

Le choix le plus simple pour le système $Px = (P - A)x + b$ soit facile à résoudre (on rappelle que c'est un objectif dans la construction d'une méthode itérative) est de prendre pour P la matrice identité (qui est très facile à inverser !). Voyons ce que cela donne sur la matrice

$$A = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}. \tag{1.91}$$

On a alors $B = P - A = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}$. Les valeurs propres de B sont 0 et -2 et on a donc $\rho(B) = 2 > 1$. La suite $(\boldsymbol{e}^{(k)})_{n\in\mathbb{N}}$ définie par $\boldsymbol{e}^{(k)} = B^k\boldsymbol{e}^{(0)}$ n'est donc en général pas convergente. En effet, si $\boldsymbol{e}^{(0)} = a\boldsymbol{u}_1 + b\boldsymbol{u}_2$, où $\boldsymbol{u}_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$ est vecteur propre de A associé à la valeur propre $\lambda = -2$, on a $\boldsymbol{e}^{(k)} = (-2)^k a$ et donc $|\boldsymbol{e}^{(k)}| \to +\infty$ lorsque $k \to \infty$ dès que $a \neq 0$. Cette première idée n'est donc pas si bonne…

La méthode de Richardson

Affinons un peu et prenons maintenant $P = \beta\mathrm{Id}$, avec $\beta \in \mathbb{R}$. On a dans ce cas $P - A = \beta\mathrm{Id} - A$ et $B = \mathrm{Id} - \frac{1}{\beta}A = \mathrm{Id} - \alpha A$ avec $\alpha = \frac{1}{\beta}$. Les valeurs propres de B sont de la forme $1 - \alpha\lambda$, où λ est valeur propre de A. Pour la matrice A définie par (1.91), les valeurs propres de A sont 1 et 3, et les valeurs propres de

$$B = \begin{bmatrix} 1-2\alpha & \alpha \\ \alpha & 1-2\alpha, \end{bmatrix}$$

sont $1-\alpha$ et $1-3\alpha$. Le rayon spectral de la matrice B, qui dépend de α est donc $\rho(B) = \max(|1-\alpha|, |1-3\alpha|)$, qu'on représente sur la figure ci-dessous. La méthode itérative s'écrit

$$\begin{aligned} &\boldsymbol{x}^{(0)} \in \mathbb{R}^n \text{ donné}, \\ &\boldsymbol{x}^{(k+1)} = B\boldsymbol{x}^{(k)} + \boldsymbol{c}, \text{ avec } \boldsymbol{c} = \alpha\boldsymbol{b}. \end{aligned} \tag{1.92}$$

Pour que la méthode converge, il faut et il suffit que $\rho(B) < 1$, c.à.d. $3\alpha - 1 < 1$, donc $\alpha < \frac{2}{3}$. On voit que le choix $\alpha = 1$ qu'on avait fait au départ n'était pas bon. Mais on peut aussi calculer le meilleur coefficient α pour avoir la meilleure convergence possible : c'est la valeur de α qui minimise le rayon spectral ρ ; il est atteint pour $1 - \alpha = 3 - \alpha$, ce qui donne $\alpha = \frac{1}{2}$. Cette méthode est connue sous le nom de *méthode de Richardson*[8]. Elle est souvent écrite sous la forme :

$$\begin{aligned} &\boldsymbol{x}^{(0)} \in \mathbb{R}^n \text{ donné}, \\ &\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha\boldsymbol{r}^{(k)}, \end{aligned}$$

où $\boldsymbol{r}^{(k)} = \boldsymbol{b} - A\boldsymbol{x}^{(k)}$ est le résidu. On vérifie facilement que cette forme est équivalente à la forme (1.92) qu'on vient d'étudier.

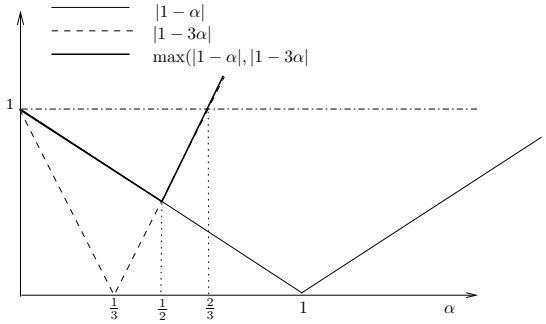

FIGURE 1.5: Rayon spectral de la matrice B de Richardson en fonction du coefficient α.

La méthode de Jacobi

Dans le cas de l'exemple de la matrice A donné par (1.91), la méthode de Richardson avec le coefficient optimal $\alpha = \frac{1}{2}$ revient à prendre comme décomposition de $A = P + A - P$ avec comme matrice $P = D$, où D est la

8. Lewis Fry Richardson, (1881-1953) est un mathématician, physicien, météorologue et psychologue qui a introduit les méthodes mathématiques pour les prévisions météorologiques. Il est également connu pour ses travaux sur les fractals. C'était un pacifiste qui a abandonné ses travaux de météorologie en raison de leur utillisation par l'armée de l'air, pour se tourner vers l'étude des raisons de guerres et de leur prévention.

matrice diagonale dont les coefficients sont les coefficients situés sur la diagonale de A. La *méthode de Jacobi*[9] consiste justement à prendre $P = D$, et ce même si la diagonale de A n'est pas constante.
Elle n'est équivalente à la méthode de Richardson avec coefficient optimal que dans le cas où la diagonale est constante ; c'est le cas de l'exemple (1.91), et donc dans ce cas la méthode de Jacobi s'écrit

$$\begin{aligned} &\boldsymbol{x}^{(0)} = \begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \end{bmatrix} \in \mathbb{R}^2 \text{ donné ,} \\ &\boldsymbol{x}^{(k+1)} = \begin{bmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \end{bmatrix} = B_J \boldsymbol{x}^{(k)} + \boldsymbol{c}, \text{ avec } B_J = \begin{bmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0, \end{bmatrix} \text{ et } \boldsymbol{c} = \frac{1}{2}\boldsymbol{b}. \end{aligned} \tag{1.93}$$

Dans le cas d'une matrice A générale, on décompose A sous la forme $A = D - E - F$, où D représente la diagonale de la matrice A, $(-E)$ la partie triangulaire inférieure et $(-F)$ la partie triangulaire supérieure :

$$D = \begin{bmatrix} a_{1,1} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & & 0 & a_{n,n} \end{bmatrix}, \; -E = \begin{bmatrix} 0 & 0 & \dots & 0 \\ a_{2,1} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & 0 \\ a_{n,1} & \dots & a_{n-1,n} & 0 \end{bmatrix} \text{ et } -F = \begin{bmatrix} 0 & a_{1,2} & \dots & a_{1,n} \\ \vdots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & a_{n,n-1} \\ 0 & \dots & 0 & -0 \end{bmatrix}. \tag{1.94}$$

La méthode de Jacobi s'écrit donc :

$$\begin{cases} \boldsymbol{x}^{(0)} \in \mathbb{R}^n \\ D\boldsymbol{x}^{(k+1)} = (E+F)\boldsymbol{x}^{(k)} + \boldsymbol{b}. \end{cases} \tag{1.95}$$

Lorsqu'on écrit la méthode de Jacobi comme sous la forme (1.85) on a $B = D^{-1}(E+F)$; on notera B_J cette matrice :

$$B_J = \begin{bmatrix} 0 & -\frac{a_{1,2}}{a_{1,1}} & \dots & -\frac{-a_{1,n}}{a_{1,1}} \\ -\frac{a_{2,1}}{a_{2,2}} & \ddots & & -\frac{-a_{2,n}}{a_{2,2}} \\ \vdots & \ddots & \ddots & \vdots \\ -\frac{a_{n,1}}{a_{n,n}} & \dots & -\frac{-a_{n-1,n}}{a_{n,n}} & 0 \end{bmatrix}.$$

La méthode de Jacobi s'écrit aussi :

$$\begin{cases} \boldsymbol{x}^{(0)} \in \mathbb{R}^n \\ a_{i,i} x_i^{(k+1)} = -\sum_{j<i} a_{i,j} x_j^{(k)} - \sum_{j>i} a_{i,j} x_j^{(k)} + b_i \;\; i = 1, \dots, n. \end{cases} \tag{1.96}$$

La méthode de Gauss-Seidel

Dans l'écriture (1.96) de la méthode de Jacobi, on pourrait remplacer les composantes $x_j^{(k)}$ dans la somme pour $j < i$ par les composantes $x_j^{(k+1)}$, puisqu'elles sont déjà calculées au moment où l'on calcule $x_i^{(k+1)}$. C'est l"idée de la méthode de Gauss-Seidel[10] qui consiste à utiliser le calcul des composantes de l'itéré $(k+1)$ dès qu'il est effectué. Par exemple, pour calculer la deuxième composante $x_2^{(k+1)}$ du vecteur $x^{(k+1)}$, on pourrait employer la

9. Carl G. J. Jacobi, (1804 - 1851), mathématicien allemand. Issu d'une famille juive, il étudie à l'Université de Berlin, où il obtient son doctorat à 21 ans. Sa thèse est une discussion analytique de la théorie des fractions. En 1829, il devient professeur de mathématique à l'Université de Königsberg, et ce jusqu'en 1842. Il fait une dépression, et voyage en Italie en 1843. À son retour, il déménage à Berlin où il sera pensionnaire royal jusqu'à sa mort. Sa lettre du 2 juillet 1830 adressée à Legendre est restée célèbre pour la phrase suivante, qui a fait couler beaucoup d'encre : "M. Fourier avait l'opinion que le but principal des mathématiques était l'utilité publique et l'explication des phénomènes naturels ; mais un philosophe comme lui aurait dû savoir que le but unique de la science, c'est l'honneur de l'esprit humain, et que sous ce titre, une question de nombres vaut autant qu'une question du système du monde." C'est une question toujours en discussion. . . .

10. Philipp Ludwig von Seidel (Zweibrücken, Allemagne 1821 ? Munich, 13 August 1896) mathématicien allemand dont il est dit qu'il a découvert en 1847 le concept crucial de la convergence uniforme en étudiant une démonstration incorrecte de Cauchy.

"nouvelle" valeur $x_1^{(k+1)}$ qu'on vient de calculer plutôt que la valeur $x_1^{(k)}$ comme dans (1.96) ; de même, dans le calcul de $x_3^{(k+1)}$, on pourrait employer les "nouvelles" valeurs $x_1^{(k+1)}$ et $x_2^{(k+1)}$ plutôt que les valeurs $x_1^{(k)}$ et $x_2^{(k)}$. Cette idée nous suggère de remplacer dans (1.96) $x_j^{(k)}$ par $x_j^{(k+1)}$ si $j < i$. On obtient donc l'algorithme suivant :

$$\left\{ \begin{array}{l} \boldsymbol{x}^{(0)} \in \mathbb{R}^n \\ a_{i,i} x_i^{(k+1)} = -\sum_{j<i} a_{i,j} x_j^{(k+1)} - \sum_{i<j} a_{i,j} x_j^{(k)} + b_i, \quad i = 1, \ldots, n. \end{array} \right. \tag{1.97}$$

La méthode de Gauss–Seidel s'écrit donc sous la forme $P\boldsymbol{x}^{(k+1)} = (P - A)\boldsymbol{x}^{(k)} + \boldsymbol{b}$, avec $P = D - E$ et $P - A = F$:

$$\left\{ \begin{array}{l} \boldsymbol{x}_0 \in \mathbb{R}^n \\ (D - E)\boldsymbol{x}^{(k+1)} = F\boldsymbol{x}^{(k)} + b. \end{array} \right. \tag{1.98}$$

Si l'on écrit la méthode de Gauss–Seidel sous la forme $\boldsymbol{x}^{(k+1)} = B\boldsymbol{x}^{(k)} + \boldsymbol{c}$, on voit assez vite que $B = (D - E)^{-1}F$; on notera B_{GS} cette matrice, dite matrice de Gauss-Seidel.
Ecrivons la méthode de Gauss-Seidel dans le cas de la matrice A donnée par (1.91) : on a dans ce cas $P = D - E = \begin{bmatrix} 2 & 0 \\ -1 & 2 \end{bmatrix}$, $F = \begin{bmatrix} 0 & 1 \\ 0 & 0, \end{bmatrix}$. L'algorithme de Gauss-Seidel s'écrit donc :

$$\begin{array}{l} \boldsymbol{x}^{(0)} = \begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \end{bmatrix} \in \mathbb{R}^2 \text{ donné} , \\ \boldsymbol{x}^{(k+1)} = \begin{bmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \end{bmatrix} = B_{GS}\boldsymbol{x}^{(k)} + \boldsymbol{c}, \text{ avec } B_{GS} = \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{4} \end{bmatrix} \text{ et } \boldsymbol{c} = \begin{bmatrix} \frac{1}{2} & 0 \\ \frac{1}{4} & \frac{1}{2} \end{bmatrix} \boldsymbol{b}. \end{array} \tag{1.99}$$

On a donc $\rho(B_{GS}) = \frac{1}{4}$. Sur cet exemple la méthode de Gauss-Seidel converge donc beaucoup plus vite que la méthode de Jacobi : Asymptotiquement, l'erreur est divisée par 4 au lieu de 2 pour la méthode de Jacobi. On peut montrer que c'est le cas pour toutes les matrices tridiagonales, comme c'est énoncé dans le théorème suivant :

Théorème 1.50 (Comparaison de Jacobi et Gauss-Seidel pour les matrices tridiagonales)**.** *On considère une matrice $A \in \mathcal{M}_n(\mathbb{R})$ tridiagonale, c.à.d. telle que $a_{i,j} = 0$ si $|i - j| > 1$; soient B_{GS} et B_{J} les matrices d'itération respectives des méthodes de Gauss-Seidel et Jacobi, alors :*

$$\rho(B_{\mathrm{GS}}) = (\rho(B_{\mathrm{J}}))^2.$$

Pour les matrices tridiagonales, la méthode de Gauss–Seidel converge (ou diverge) donc plus vite que celle de Jacobi.

La démonstration de ce résultat se fait en montrant que dans le cas tridiagonal, λ est valeur propre de la matrice d'itération de Jacobi si et seulement si λ^2 est valeur propre de la matrice d'itération de Gauss-Seidel. Elle est laissée à titre d'exercice.

Méthodes SOR et SSOR

L'idée de la méthode de sur-relaxation (SOR = Successive Over Relaxation) est d'utiliser la méthode de Gauss-Seidel pour calculer un itéré intermédiaire $\tilde{x}^{(k+1)}$ qu'on "relaxe" ensuite pour améliorer la vitesse de convergence de la méthode. On se donne $0 < \omega < 2$, et on modifie l'algorithme de Gauss–Seidel de la manière suivante :

$$\left\{ \begin{array}{l} x_0 \in \mathbb{R}^n \\ a_{i,i} \tilde{x}_i^{(k+1)} = -\displaystyle\sum_{j<i} a_{i,j} x_j^{(k+1)} - \sum_{i<j} a_{i,j} x_j^{(k)} + b_i \\ x_i^{(k+1)} = \omega \tilde{x}_i^{(k+1)} + (1 - \omega) x_i^{(k)}, \quad i = 1, \ldots, n. \end{array} \right. \tag{1.100}$$

(Pour $\omega = 1$ on retrouve la méthode de Gauss–Seidel.)
L'algorithme ci-dessus peut aussi s'écrire (en multipliant par $a_{i,i}$ la ligne 3 de l'algorithme (1.109)) :

$$\left\{ \begin{array}{l} x^{(0)} \in \mathbb{R}^n \\ a_{i,i}x_i^{(k+1)} = \omega \left[-\sum_{j<i} a_{i,j}x_j^{(k+1)} - \sum_{j>i} a_{i,j}x_j^{(k)} + b_i \right] \\ \qquad\qquad +(1-\omega)a_{i,i}x_i^{(k)}. \end{array} \right. \tag{1.101}$$

On obtient donc

$$(D - \omega E)x^{(k+1)} = \omega F x^{(k)} + \omega b + (1-\omega)Dx^{(k)}.$$

La matrice d'itération de l'algorithme SOR est donc

$$B_\omega = \left(\frac{D}{\omega} - E\right)^{-1} (F + \left(\frac{1-\omega}{\omega}\right) D) = P^{-1}N, \text{ avec } P = \frac{D}{\omega} - E \text{ et } N = F + \left(\frac{1-\omega}{\omega}\right) D.$$

Il est facile de vérifier que $A = P - N$.

Proposition 1.51 (Condition nécessaire de convergence de la méthode SOR)**.**
Soit $A \in \mathcal{M}_n(\mathbb{R})$ et soiet D, E et F les matrices définies par (1.94)*; on a donc $A = D - E - F$. Soit B_ω la matrice d'itération de la méthode SOR (et de la méthode de Gauss–Seidel pour $\omega = 1$) définie par :*

$$B_\omega = \left(\frac{D}{\omega} - E\right)^{-1} \left(F + \frac{1-\omega}{\omega} D\right), \quad \omega \neq 0.$$

Si $\rho(B_\omega) < 1$ alors $0 < \omega < 2$.

DÉMONSTRATION – Calculons $\det(B_\omega)$. Par définition,

$$B_\omega = P^{-1}N, \text{ avec } P = \frac{1}{\omega}D - E \text{ et } N = F + \frac{1-\omega}{\omega}D.$$

Donc $\det(B_\omega) = (\det(P))^{-1}\det(N)$. Comme P et N sont des matrices triangulaires, leurs déterminants sont les produits ceofficients diagonaux (voir la remarque 1.58 page 85). On a donc :

$$\det(B_\omega) = \frac{(\frac{1-\omega}{\omega})^n \det(D)}{(\frac{1}{\omega})^n \det(D)} = (1-\omega)^n.$$

Or le déterminant d'une matrice est aussi le produit des valeurs propres de cette matrice (comptées avec leur multiplicités algébriques), dont les valeurs absolues sont toutes, par définition, inférieures au rayon spectral. On a donc : $|\det(B_\omega)| = |(1-\omega)^n| \leq (\rho(B_\omega))^n$, d'où le résultat. ∎

On a un résultat de convergence de la méthode SOR (et donc également de Gauss-Seidel) dans le cas où A est symétrique définie positive, grâce au lemme suivant :

Lemme 1.52 (Condition suffisante de convergence pour la suite définie par (1.85))**.** *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique définie positive, et soient P et $N \in \mathcal{M}_n(\mathbb{R})$ telles que $A = P - N$ et P est inversible. Si la matrice $P^t + N$ est symétrique définie positive alors $\rho(P^{-1}N) = \rho(B) < 1$, et donc la suite définie par* (1.85) *converge.*

DÉMONSTRATION – On rappelle (voir le corollaire (1.35) page 55) que si $B \in \mathcal{M}_n(\mathbb{R})$, et si $\|\cdot\|$ est une norme induite sur $\mathcal{M}_n(\mathbb{R})$ par une norme sur $\mathbb{R}^n$, on a toujours $\rho(B) \leq \|B\|$. On va donc chercher une norme sur $\mathbb{R}^n$, notée $\|\cdot\|_*$ telle que

$$\|P^{-1}N\|_* = \max\{\|P^{-1}N\boldsymbol{x}\|_*,\ \boldsymbol{x} \in \mathbb{R}^n,\ \|\boldsymbol{x}\|_* = 1\} < 1,$$

(où on désigne encore par $\|.\|_*$ la norme induite sur $\mathcal{M}_n(\mathbb{R})$) ou encore :

$$\|P^{-1}N\boldsymbol{x}\|_* < \|x\|_*,\ \forall \boldsymbol{x} \in \mathbb{R}^n,\ \boldsymbol{x} \neq 0. \tag{1.102}$$

On définit la norme $\|\cdot\|_*$ par $\|\boldsymbol{x}\|_* = \sqrt{A\boldsymbol{x}\cdot\boldsymbol{x}}$, pour tout $\boldsymbol{x} \in \mathbb{R}^n$. Comme A est symétrique définie positive, $\|\cdot\|_*$ est bien une norme sur $\mathbb{R}^n$, induite par le produit scalaire $(\boldsymbol{x}|\boldsymbol{y})_A = A\boldsymbol{x}\cdot\boldsymbol{y}$. On va montrer que la propriété (1.102) est vérifiée par cette norme. Soit $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{x} \neq 0$, on a : $\|P^{-1}N\boldsymbol{x}\|_*^2 = AP^{-1}N\boldsymbol{x}\cdot Pt^{-1}N\boldsymbol{x}$. Or $N = P - A$, et donc : $\|P^{-1}N\boldsymbol{x}\|_*^2 = A(\mathrm{Id} - Pt^{-1}A)\boldsymbol{x}\cdot(Id - P^{-1}A)\boldsymbol{x}$. Soit $\boldsymbol{y} = P^{-1}A\boldsymbol{x}$; remarquons que $\boldsymbol{y} \neq 0$ car $\boldsymbol{x} \neq 0$ et $P^{-1}A$ est inversible. Exprimons $\|P^{-1}N\boldsymbol{x}\|_*^2$ à l'aide de $\boldsymbol{y}$.

$$\|P^{-1}N\boldsymbol{x}\|_*^2 = A(\boldsymbol{x}-\boldsymbol{y})\cdot(\boldsymbol{x}-\boldsymbol{y}) = A\boldsymbol{x}\cdot\boldsymbol{x} - 2A\boldsymbol{x}\cdot\boldsymbol{y} + A\boldsymbol{y}\cdot\boldsymbol{y} = \|\boldsymbol{x}\|_*^2 - 2A\boldsymbol{x}\cdot\boldsymbol{y} + A\boldsymbol{y}\cdot\boldsymbol{y}.$$

Pour que $\|P^{-1}N\boldsymbol{x}\|_*^2 < \|\boldsymbol{x}\|_*^2$ (et par suite $\rho(Pt^{-1}N) < 1$), il suffit donc de montrer que $-2A\boldsymbol{x}\cdot\boldsymbol{y} + A\boldsymbol{y}\cdot\boldsymbol{y} < 0$. Or, comme $P\boldsymbol{y} = A\boldsymbol{x}$, on a : $-2A\boldsymbol{x}\cdot\boldsymbol{y} + A\boldsymbol{y}\cdot\boldsymbol{y} = -2P\boldsymbol{y}\cdot\boldsymbol{y} + A\boldsymbol{y}\cdot\boldsymbol{y}$. En écrivant : $P\boldsymbol{y}\cdot\boldsymbol{y} = y\cdot P^t y = P^t y\cdot y$, on obtient donc que : $-2A\boldsymbol{x}\cdot\boldsymbol{y} + A\boldsymbol{y}\cdot\boldsymbol{y} = (-P - P^t + A)\boldsymbol{y}\cdot\boldsymbol{y}$, et comme $A = P - N$ on obtient $-2A\boldsymbol{x}\cdot\boldsymbol{y} + A\boldsymbol{y}\cdot\boldsymbol{y} = -(P^t + N)\boldsymbol{y}\cdot\boldsymbol{y}$. Comme $P^t + N$ est symétrique définie positive par hypothèse et que $\boldsymbol{y} \neq \mathbf{0}$, on en déduit que $-2A\boldsymbol{x}\cdot\boldsymbol{y} + A\boldsymbol{y}\cdot\boldsymbol{y} < 0$, ce qui termine la démonstration. ∎

Théorème 1.53 (CNS de convergence de la méthode SOR pour les matrices s.d.p.)**.**
Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique définie positive, et soient D, E et F les matrices définies par (1.94)*; on a donc $A = D - E - F$. Soit B_ω la matrice d'itération de la méthode SOR (et de la méthode de Gauss–Seidel pour $\omega = 1$) définie par :*

$$B_\omega = \left(\frac{D}{\omega} - E\right)^{-1}\left(F + \frac{1-\omega}{\omega}D\right),\quad \omega \neq 0.$$

Alors :

$$\rho(B_\omega) < 1 \text{ si et seulement si } 0 < \omega < 2.$$

En particulier, si A est une matrice symétrique définie positive, la méthode de Gauss–Seidel converge.

DÉMONSTRATION – On sait par la proposition 1.51 que si $\rho(B_\omega) < 1$ alors $0 < \omega < 2$. Supposons maintenant que A est une matrice symétrique définie positive, que $0 < \omega < 2$et montrons que $\rho(B_\omega) < 1$. Par le lemme 1.52 page 82, il suffit pour cela de montrer que $P^t + N$ est une matrice symétrique définie positive. Or,

$$P^t = \left(\frac{D}{\omega} - E\right)^t = \frac{D}{\omega} - F,$$

$$P^t + N = \frac{D}{\omega} - F + F + \frac{1-\omega}{\omega}D = \frac{2-\omega}{\omega}D.$$

La matrice $P^t + N$ est donc bien symétrique définie positive. ∎

Remarque 1.54 (Comparaison Gauss–Seidel/Jacobi)**.** *On a vu (théorème 1.53) que si A est une matrice symétrique définie positive, la méthode de Gauss–Seidel converge. Par contre, même dans le cas où A est symétrique définie positive, il existe des cas où la méthode de Jacobi ne converge pas, voir à ce sujet l'exercice 46 page 88.*

Remarquons que le résultat de convergence des méthodes itératives donné par le théorème précédent n'est que partiel, puisqu'il ne concerne que les matrices symétriques définies positives et que les méthodes Gauss-Seidel et SOR. On a aussi un résultat de convergence de la méthode de Jacobi pour les matrices à diagonale dominante stricte, voir exercice 49 page 89, et un résultat de comparaison des méthodes pour les matrices tridiagonales par blocs, voir le théorème 1.55 donné ci-après. Dans la pratique, il faudra souvent compter sur sa bonne étoile…

Estimation du coefficient de relaxation optimal de SOR La question est ici d'estimer le coefficient de relaxation ω optimal dans la méthode SOR, c.à.d. le coefficient $\omega_0 \in]0,2[$ (condition nécessaire pour que la méthode SOR converge, voir théorème 1.53) tel que

$$\rho(\mathcal{L}_{\omega_0}) < \rho(B_\omega), \forall \omega \in]0,2[.$$

Ce coefficient ω_0 donnera la meilleure convergence possible pour SOR. On sait le faire dans le cas assez restrictif des matrices tridiagonales (ou tridiagonales par blocs, voir paragraphe suivant). On ne fait ici qu'énoncer le résultat dont la démonstration est donnée dans le livre de Ph.Ciarlet conseillé en début de cours.

Théorème 1.55 (Coefficient optimal, matrice tridiagonale)**.** *On considère une matrice $A \in \mathcal{M}_n(\mathbb{R})$ qui admet une décomposition par blocs définie dans la définition 1.103 page 85 ; on suppose que la matrice A est tridiagonale par blocs, c.à.d. $A_{i,j} = 0$ si $|i-j| > 1$; soient B_{GS} et B_{J} les matrices d'itération respectives des méthodes de Gauss-Seidel et Jacobi, alors : On suppose de plus que toutes les valeurs propres de la matrice d'itération J de la méthode de Jacobi sont réelles ; alors le paramètre de relaxation optimal, c.à.d. le paramètre ω_0 tel que $\rho(B_{\omega_0}) = \min\{\rho(B_\omega), \omega \in]0,2[\}$, s'exprime en fonction du rayon spectral $\rho(B_{\mathrm{J}})$ de la matrice J par la formule :*

$$\omega_0 = \frac{2}{1+\sqrt{1-\rho(B_{\mathrm{J}})^2}} > 1,$$

et on a : $\rho(B_{\omega_0}) = \omega_0 - 1$.

La démonstration de ce résultat repose sur la comparaison des valeurs propres des matrices d'itération. On montre que λ est valeur propre de B_ω si et seulement si

$$(\lambda + \omega - 1)^2 = \lambda \omega \mu^2,$$

où μ est valeur propre de B_{J} (voir [Ciarlet] pour plus de détails).

Remarque 1.56 (Méthode de Jacobi relaxée)**.** *On peut aussi appliquer une procédure de relaxation avec comme méthode iérative "de base" la méthode de Jacobi, voir à ce sujet l'exercice 52 page 90). Cette méthode est toutefois beaucoup moins employée en pratique (car moins efficace) que la méthode SOR.*

Méthode SSOR En "symétrisant" le procédé de la méthode SOR, c.à.d. en effectuant les calculs SOR sur les blocs dans l'ordre 1 à n puis dans l'ordre n à 1, on obtient la méthode de sur-relaxation symétrisée (SSOR = Symmetric Successive Over Relaxation) qui s'écrit dans le formalisme de la méthode I avec

$$B_{SSOR} = \underbrace{\left(\frac{D}{\omega} - F\right)^{-1}\left(E + \frac{1-\omega}{\omega}D\right)}_{\text{calcul dans l'ordre } \; n \ldots 1}\underbrace{\left(\frac{D}{\omega} - E\right)^{-1}\left(F + \frac{1-\omega}{\omega}D\right)}_{\text{calcul dans l'ordre } \; 1 \ldots n}.$$

1.5.3 Les méthodes par blocs

Décomposition par blocs d'une matrice

Dans de nombreux cas pratiques, les matrices des systèmes linéaires à résoudre ont une structure "par blocs", et on se sert de cette structure lors de la résolution par une méthode itérative.

Définition 1.57. *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible. Une décomposition par blocs de A est définie par un entier $S \leq n$, des entiers $(n_i)_{i=1,\dots,S}$ tels que $\sum_{i=1}^{S} n_i = n$, et S^2 matrices $A_{i,j} \in \mathcal{M}_{n_i,n_j}(\mathbb{R})$ (ensemble des matrices rectangulaires à n_i lignes et n_j colonnes, telles que les matrices $A_{i,i}$ soient inversibles pour $i = 1, \dots, S$ et*

$$A = \begin{bmatrix} A_{1,1} & A_{1,2} & \dots & & \dots & A_{1,S} \\ A_{2,1} & \ddots & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & A_{S-1,S} \\ A_{S,1} & \dots & & \dots & A_{S,S-1} & A_{S,S} \end{bmatrix} \tag{1.103}$$

Remarque 1.58.

1. *Si $S = n$ et $n_i = 1$ $\forall i \in \{1, \dots, S\}$, chaque bloc est constitué d'un seul coefficient, et on retrouve la structure habituelle d'une matrice. Donc toutes les méthodes que nous allons décrire pour les matrices structurées par blocs s'appliquent évidemment de la même manière aux matrices "habituelles".*
2. *Si A est symétrique définie positive, la condition $A_{i,i}$ inversible dans la définition 1.57 est inutile car $A_{i,i}$ est nécessairement symétrique définie positive donc inversible. Prenons par exemple $i = 1$; soit $y \in \mathbb{R}^{n_1}$, $y \neq 0$ et $x = (y, 0 \dots, 0)^t \in \mathbb{R}^n$. Alors $A_{1,1}y \cdot y = Ax \cdot x > 0$ donc $A_{1,1}$ est symétrique définie positive.*
3. *Si A est une matrice triangulaire par blocs, c.à.d. de la forme* (1.103) *avec $A_{i,j} = 0$ si $j > i$, alors*

$$\det(A) = \prod_{i=1}^{S} \det(A_{i,i}).$$

Par contre si A est décomposée en 2×2 blocs carrés (i.e. tels que $n_i = m_j$, $\forall (i,j) \in \{1,2\}$), on a en général : $\det(A) \neq \det(A_{1,1})\det(A_{2,2}) - \det(A_{1,2})\det(A_{2,1})$.

Méthode de Jacobi

On peut remarquer que le choix le plus simple pour le système $Px = (P - A)x + b$ soit facile à résoudre (on rappelle que c'est un objectif dans la construction d'une méthode itérative) est de prendre pour P une matrice diagonale. La méthode de Jacobi consiste à prendre pour P la matrice diagonale D formée par les blocs diagonaux de A :

$$D = \begin{bmatrix} A_{1,1} & 0 & \dots & & \dots & 0 \\ 0 & \ddots & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & 0 \\ 0 & \dots & & \dots & 0 & A_{S,S} \end{bmatrix}.$$

Dans la matrice ci-dessus, 0 désigne un bloc nul.

On a alors $N = P - A = E + F$, où E et F sont constitués des blocs triangulaires inférieurs et supérieurs de la matrice A :

$$E = \begin{bmatrix} 0 & 0 & \dots & & \dots & 0 \\ -A_{2,1} & \ddots & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & 0 \\ -A_{S,1} & \dots & & \dots & -A_{S,S-1} & 0 \end{bmatrix}$$

et

$$F = \begin{bmatrix} 0 & -A_{1,2} & \dots & & \dots & -A_{1,S} \\ 0 & \ddots & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & -A_{S-1,S} \\ 0 & \dots & & \dots & 0 & 0 \end{bmatrix}.$$

On a bien $A = P - N$ et avec D, E et F définies comme ci-dessus, la méthode de Jacobi s'écrit :

$$\begin{cases} x^{(0)} \in \mathbb{R}^n \\ Dx^{(k+1)} = (E+F)x^{(k)} + b. \end{cases} \tag{1.104}$$

Lorsqu'on écrit la méthode de Jacobi comme sous la forme (1.85) on a $B = D^{-1}(E+F)$; on notera J cette matrice.
En introduisant la décomposition par blocs de x, solution recherchée de (1.1), c.à.d. : $x = [x_1, \dots, x_S]^t$, où $x_i \in \mathbb{R}^{n_i}$, on peut aussi écrire la méthode de Jacobi sous la forme :

$$\begin{cases} x_0 \in \mathbb{R}^n \\ A_{i,i}x_i^{(k+1)} = -\sum_{j<i} A_{i,j}x_j^{(k)} - \sum_{j>i} A_{i,j}x_j^{(k)} + b_i \;\; i = 1, \dots, S. \end{cases} \tag{1.105}$$

Si $S = n$ et $n_i = 1 \; \forall i \in \{1, \dots, S\}$, chaque bloc est constitué d'un seul coefficient, et on obtient la méthode de Jacobi par points (aussi appelée méthode de Jacobi), qui s'écrit donc :

$$\begin{cases} x_0 \in \mathbb{R}^n \\ a_{i,i}x_i^{(k+1)} = -\sum_{j<i} a_{i,j}x_j^{(k)} - \sum_{j>i} a_{i,j}x_j^{(k)} + b_i \;\; i = 1, \dots, n. \end{cases} \tag{1.106}$$

Méthode de Gauss-Seidel

La même procédure que dans le cas $S = n$ et $n_i = 1$ donne :

$$\begin{cases} x^{(0)} \in \mathbb{R}^n \\ A_{i,i}x_i^{(k+1)} = -\sum_{j<i} A_{i,j}x_j^{(k+1)} - \sum_{i<j} A_{i,j}x_j^{(k)} + b_i, \quad i = 1, \dots, S. \end{cases} \tag{1.107}$$

La méthode de Gauss–Seidel s'écrit donc sous forme la forme $Px^{(k+1)} = (P-A)x^{(k)} + b$, $P = D - E$ et $P - A = F$:

$$\begin{cases} x_0 \in \mathbb{R}^n \\ (D-E)x^{(k+1)} = Fx^{(k)} + b. \end{cases} \tag{1.108}$$

Si l'on écrit la méthode de Gauss–Seidel sous la forme $x^{(k+1)} = Bx^{(k)} + c$, on voit assez vite que $B = (D - E)^{-1}F$; on notera B_{GS} cette matrice, dite matrice de Gauss-Seidel.

Méthodes SOR et SSOR

La méthode SOR s'écrit aussi par blocs :On se donne $0 < \omega < 2$, et on modifie l'algorithme de Gauss–Seidel de la manière suivante :

$$\left\{ \begin{array}{l} x_0 \in \mathbb{R}^n \\ A_{i,i}\tilde{x}_i^{(k+1)} = -\sum_{j<i} A_{i,j}x_j^{(k+1)} - \sum_{i<j} A_{i,j}x_j^{(k)} + b_i \\ x_i^{(k+1)} = \omega\tilde{x}_i^{(k+1)} + (1-\omega)x_i^{(k)}, \quad i = 1, \ldots, S. \end{array} \right. \tag{1.109}$$

(Pour $\omega = 1$ on retrouve la méthode de Gauss–Seidel.)
L'algorithme ci-dessus peut aussi s'écrire (en multipliant par $A_{i,i}$ la ligne 3 de l'algorithme (1.109)) :

$$\left\{ \begin{array}{lcl} x^{(0)} \in \mathbb{R}^n & & \\ A_{i,i}x_i^{(k+1)} & = & \omega\left[-\sum_{j<i} A_{i,j}x_j^{(k+1)} - \sum_{j>i} A_{i,j}x_j^{(k)} + b_i\right] \\ & & +(1-\omega)A_{i,i}x_i^{(k)}. \end{array} \right. \tag{1.110}$$

On obtient donc

$$(D - \omega E)x^{(k+1)} = \omega F x^{(k)} + \omega b + (1-\omega)Dx^{(k)}.$$

L'algorithme SOR s'écrit donc comme une méthode II avec

$$P = \frac{D}{\omega} - E \text{ et } N = F + \left(\frac{1-\omega}{\omega}\right) D.$$

Il est facile de vérifier que $A = P - N$.

L'algorithme SOR s'écrit aussi comme une méthode I avec

$$B = \left(\frac{D}{\omega} - E\right)^{-1} (F + \left(\frac{1-\omega}{\omega}\right) D).$$

On notera $\mathcal{L}_\omega$ cette matrice.

Remarque 1.59 (Méthode de Jacobi relaxée)**.** *On peut aussi appliquer une procédure de relaxation avec comme méthode iérative "de base" la méthode de Jacobi, voir à ce sujet l'exercice 52 page 90). Cette méthode est toutefois beaucoup moins employée en pratique (car moins efficace) que la méthode SOR.*

En "symétrisant" le procédé de la méthode SOR, c.à.d. en effectuant les calculs SOR sur les blocs dans l'ordre 1 à n puis dans l'ordre n à 1, on obtient la méthode de sur-relaxation symétrisée (SSOR = Symmetric Successive Over Relaxation) qui s'écrit dans le formalisme de la méthode I avec

$$B = \underbrace{\left(\frac{D}{\omega} - F\right)^{-1}\left(E + \frac{1-\omega}{\omega}D\right)}_{\text{calcul dans l'ordre } S\ldots 1}\underbrace{\left(\frac{D}{\omega} - E\right)^{-1}\left(F + \frac{1-\omega}{\omega}D\right)}_{\text{calcul dans l'ordre } 1\ldots S}.$$

1.5.4 Exercices, énoncés

Exercice 44 (Convergence de suites)**.** *Corrigé en page 96*

Etudier la convergence de la suite $(\boldsymbol{x}^{(k)})_{n\in\mathbb{N}} \subset \mathbb{R}^n$ définie par $\boldsymbol{x}^{(0)}$ donné, $\boldsymbol{x}^{(k)} = B\boldsymbol{x}^{(k)} + \boldsymbol{c}$ dans les cas suivants :

$$\text{(a)} \quad B = \begin{bmatrix} \frac{2}{3} & 1 \\ 0 & \frac{2}{3} \end{bmatrix}, \ \boldsymbol{c} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \qquad \text{(b)} \quad B = \begin{bmatrix} \frac{2}{3} & 1 \\ 0 & 2 \end{bmatrix}, \ \boldsymbol{c} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Exercice 45 (Méthode de Richardson)**.** *Suggestions en page 94, corrigé en page 96*

Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique définie positive, $b \in \mathbb{R}^n$ et $\alpha \in \mathbb{R}$. Pour trouver la solution de $Ax = b$, on considère la méthode itérative suivante :

— Initialisation : $x^{(0)} \in \mathbb{R}^n$,

— Iterations : $x^{(k+1)} = x^{(k)} + \alpha(b - Ax^{(k)})$.

1. Pour quelles valeurs de α (en fonction des valeurs propres de A) la méthode est-elle convergente ?
2. Calculer α_0 (en fonction des valeurs propres de A) t.q. $\rho(Id - \alpha_0 A) = \min\{\rho(Id - \alpha A),\, \alpha \in \mathbb{R}\}$.

Commentaire sur la méthode de Richardson : On peut la voir comme une méthode de gradient à pas fixe pour la minimisation de la fonction f définie de $\mathbb{R}^N$ dans $\mathbb{R}$ par : $\boldsymbol{x} \mapsto f(\boldsymbol{x}) = \frac{1}{2} A\boldsymbol{x}\cdot\boldsymbol{x} - \boldsymbol{b}\cdot\boldsymbol{x}$, qui sera étudiée au chapitre **??** page **??**. On verra en effet que grâce qu caractère symétrique définie positif de A, la fonction $\boldsymbol{f}$ admet un unique minimum, caractérisé par l'annulation du gradient de f en ce point. Or $\nabla f(\boldsymbol{x}) = A\boldsymbol{x} - \boldsymbol{b}$ (voir le paragraphe **??** dans le chapitre optimisation), et annuler le gradient consiste à résoudre le système linéaire $A\boldsymbol{x} = \boldsymbol{b}$.

Exercice 46 (Non convergence de la méthode de Jacobi)**.** *Suggestions en page 95. Corrigé en page 97.*

Soit $a \in \mathbb{R}$ et

$$A = \begin{pmatrix} 1 & a & a \\ a & 1 & a \\ a & a & 1 \end{pmatrix}$$

Montrer que 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 47 (Jacobi et Gauss–Seidel : cas des matrices symétriques)**.** *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice carrée d'ordre n inversible et tridiagonale ; on note $a_{i,j}$ le coefficient de la ligne i et la ligne j de la matrice A. On décompose en $A = D - E - F$, où D représente la diagonale de la matrice A, $(-E)$ la partie triangulaire inférieure stricte et $(-F)$ la partie triangulaire supérieure stricte.*
On note B_J et B_{GS} les matrices d'itération des méthodes de Jacobi et Gauss-Seidel pour la résolution d'un système linéaire de matrice A.

1. *Calculer les matrices B_J et B_{GS} pour la matrice particulère $A = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$ et calculer leurs rayon spectraux.Montrer que les méthodes convergent, et citez les résultat du cours qui s'appliquent pour cette matrice.*
2. *Montrer que λ est valeur propre de B_J si et seulement s'il existe un vecteur complexe $\boldsymbol{x} = (x_1, \dots x_n) \in \mathbb{C}^n$, $\boldsymbol{x} \neq \boldsymbol{0}$, tel que*
$$-a_{p,p-1}x_{p-1} - a_{p,p+1}x_{p+1} = \lambda a_{p,p}x_p, \quad p = 1, \dots, n.$$
avec $x_0 = x_{n+1} = 0$.
3. *Soit $\boldsymbol{y} = (y_1, \dots y_n) \in \mathbb{C}^n$ défini par $y_p = \lambda^{-p}x_p$, où λ est une valeur propre non nulle de B_J et $\boldsymbol{x} = (x_1, \dots, x_n)$ un vecteur propre associé. On pose $y_0 = y_{n+1} = 0$. Montrer que*
$$-a_{p,p-1}y_{p-1} - \lambda^2 a_{p,p+1}y_{p+1} = \lambda^2 a_{p,p}y_p, \quad p = 1, \dots, n.$$
4. *Montrer que μ est valeur propre de B_{GS} associée à un vecteur propre $\boldsymbol{z} \neq \boldsymbol{0}$ si et seulement si*
$$(F - \mu(D - E))\,\boldsymbol{z} = \boldsymbol{0}.$$
5. *Montrer que si λ est valeur propre non nulle de B_J si et seulement si λ^2 est valeur propre de B_{GS}, et en déduire que $\rho(B_{GS}) = \rho(B_J)^2$.*
6. *On considère la matrice :*
$$A = \begin{bmatrix} 1 & \frac{3}{4} & \frac{3}{4} \\ \frac{3}{4} & 1 & \frac{3}{4} \\ \frac{3}{4} & \frac{3}{4} & 1 \end{bmatrix}$$
Montrer que cette matrice est symétrique définie positive. Montrer que $\rho(B_{GS}) \neq \rho(B_J)$. Quelle est l'hypothèse mise en défaut ici ?

Exercice 48 (Une matrice cyclique)**.** *Suggestions en page 95*

Soit $\alpha \in \mathbb{R}$ et soit $A \in \mathcal{M}_4(\mathbb{R})$ la matrice définie par

$$A = \begin{pmatrix} \alpha & -1 & 0 & -1 \\ -1 & \alpha & -1 & 0 \\ 0 & -1 & \alpha & -1 \\ -1 & 0 & -1 & \alpha \end{pmatrix}$$

Cette matrice est dite cyclique : chaque ligne de la matrice peut être déduite de la précédente en décalant chaque coefficient d'une position.

1. Déterminer les valeurs propres de A.
2. Pour quelles valeurs de α la matrice A est-elle symétrique définie positive ? singulière ?
3. On suppose ici que $\alpha \neq 0$. Soit $b = (b_1, b_2, b_3, b_4)^t \in \mathbb{R}^4$ donné. On considère la méthode de Jacobi pour la résolution du système $Ax = b$. Soit $(x^{(k)})_{n\in\mathbb{N}}$ la suite de vecteurs donnés par l'algorithme. On note $x_i^{(k)}$ pour $i = 1, \ldots, 4$ les composantes de $x^{(k)}$. Donner l'expression de $x_i^{(k+1)}$, $i = 1, \ldots, 4$, en fonction de $x_i^{(k)}$ et $b_i^{(k)}$, $i = 1, \ldots, 4$. Pour quelles valeurs de α la méthode de Jacobi converge-t-elle ?
4. On suppose maintenant que A est symétrique définie positive. Reprendre la question précédente pour la méthode de Gauss-Seidel.

Exercice 49 (Jacobi pour les matrices à diagonale dominante stricte)**.** *Suggestions en page 95, corrigé en page 97*

Soit $A = (a_{i,j})_{i,j=1,\ldots,n} \in \mathcal{M}_n(\mathbb{R})$ une matrice à diagonale dominante stricte (c'est-à-dire $|a_{i,i}| > \sum_{j\neq i} |a_{i,j}|$ pour tout $i = 1, \ldots, n$). Montrer que A est inversible et que la méthode de Jacobi (pour calculer la solution de $Ax = b$) converge.

Exercice 50 (Jacobi pour pour un problème de diffusion)**.**

Soit $f \in C([0, 1])$; on considère le système linéaire $Ax = b$ issu de la discrétisation par différences finies de pas uniforme égal à $h = \frac{1}{n+1}$ du problème suivant :

$$\begin{cases} -u''(x) + \alpha u(x) = f(x), \ x \in [0, 1], \\ u(0) = 0, u(1) = 1, \end{cases} \tag{1.111}$$

où $\alpha \geq 0$.

1. Donner l'expression de A et $\boldsymbol{b}$.
2. Montrer que la méthode de Jacobi appliquée à la résolution de ce système converge (distinguer les cas $\alpha > 0$ et $\alpha = 0$).

Exercice 51 (Jacobi et diagonale dominance forte)**.**

1. Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique définie positive.
   (a) Montrer que tous les coefficients diagonaux de A sont strictement positifs.
   (b) En déduire que la méthode de Jacobi pour la résolution du système linéaire $Ax = b$, avec $b \in \mathbb{R}^n$, est bien définie.

Soit $M \in \mathcal{M}_n(\mathbb{R})$ une matrice carrée d'ordre n, avec $n > 1$. On dit que la matrice M est irréductible si :

$$\text{pour tous ensembles d'indices } I \subset \{1, \ldots, n\}, I \neq \emptyset, \text{ et } J = \{1 \ldots, n\} \setminus I, J \neq \emptyset, \ \exists i \in I, \exists j \in J; a_{i,j} \neq 0. \tag{1.112}$$

2 (a) Montrer qu'une matrice diagonale n'est pas irréductible. En déduire qu'une matrice inversible n'est pas forcément irréductible.

2 (b) Soit $M \in \mathcal{M}_n(\mathbb{R})$ une matrice carrée d'ordre n, qui s'écrit sous la forme :

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

où A et C sont des matrices carrées d'ordre p et q, avec $p + q = n$, et $B \in \mathcal{M}_{q,p}(\mathbb{R})$. La matrice M peut-elle être irréductible ?

3. Soit $A \in \mathcal{M}_n(\mathbb{R})$, $n > 1$ une matrice irréductible qui vérifie de plus la propriété suivante :

$$\forall i = 1, \ldots, n, a_{i,i} \geq \sum_{j \neq i} |a_{i,j}| \tag{1.113}$$

(On dit que la matrice est à diagonale dominante). Montrer que la méthode de Jacobi pour la résolution du système linéaire $Ax = b$, avec $b \in \mathbb{R}^n$, est bien définie.

4. Soit $A \in \mathcal{M}_n(\mathbb{R})$, $n > 1$ une matrice irréductible qui vérifie la propriété (1.113). On note B_{J} la matrice d'itération de la méthode de Jacobi pour la résolution du système linéaire $Ax = b$, avec $b \in \mathbb{R}^n$, et $\rho(B_{\mathrm{J}})$ son rayon spectral. On suppose que A vérifie la propriété supplémentaire suivante :

$$\exists i_0; a_{i_0,i_0} > \sum_{j \neq i_0} |a_{i,j}|. \tag{1.114}$$

(a) Montrer que $\rho(B_{\mathrm{J}}) \leq 1$.

(b) Montrer que si $Jx = \lambda x$ avec $|\lambda| = 1$, alors $|x_i| = \|x\|_\infty$, $\forall i = 1, \ldots, n$, où $\|x\|_\infty = \max_{k=1,\ldots,N} |x_k|$. En déduire que $x = 0$ et que la méthode de Jacobi converge.

(c) Retrouver ainsi le résultat de la question 2 de l'exercice 50.

5. En déduire que si A est une matrice qui vérifie les propriétés (1.112), (1.113) et (1.114), alors A est inversible.

6. Montrer que la matrice A suivante est symétrique définie positive et vérifie les propriétés (1.113) et (1.114).

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 1 & 1 \\ 0 & 1 & 2 & 1 \\ 0 & 1 & 1 & 2 \end{bmatrix}$$

La méthode de Jacobi converge-t-elle pour la résolution d'un système linéaire dont la matrice est A ?

Exercice 52 (Méthode de Jacobi et relaxation)**.** *Suggestions en page 95, corrigé en page 98*

Soit $n \geq 1$. Soit $A = (a_{i,j})_{i,j=1,\ldots,n} \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique. On note D la partie diagonale de A, $-E$ la partie triangulaire inférieure de A et $-F$ la partie triangulaire supérieure de A, c'est-à-dire :

$$\begin{array}{l} D = (d_{i,j})_{i,j=1,\ldots,n},\ d_{i,j} = 0 \text{ si } i \neq j,\ d_{i,i} = a_{i,i}, \\ E = (e_{i,j})_{i,j=1,\ldots,n},\ e_{i,j} = 0 \text{ si } i \leq j,\ e_{i,j} = -a_{i,j} \text{ si } i > j, \\ F = (f_{i,j})_{i,j=1,\ldots,n},\ f_{i,j} = 0 \text{ si } i \geq j,\ f_{i,j} = -a_{i,j} \text{ si } i < j. \end{array}$$

Noter que $A = D - E - F$. Soit $b \in \mathbb{R}^n$. On cherche à calculer $x \in \mathbb{R}^n$ t.q. $Ax = b$. On suppose que D est définie positive (noter que A n'est pas forcément inversible). On s'intéresse ici à la méthode de Jacobi (par points), c'est–à–dire à la méthode itérative suivante :

Initialisation. $x^{(0)} \in \mathbb{R}^n$
Itérations. Pour $n \in \mathbb{N}$, $Dx^{(k+1)} = (E + F)x^{(k)} + b$.

On pose $J = D^{-1}(E + F)$.

1. Montrer, en donnant un exemple avec $n = 2$, que J peut ne pas être symétrique.

2. Montrer que J est diagonalisable dans $\mathbb{R}$ et, plus précisement, qu'il existe une base de $\mathbb{R}^n$, notée $\{f_1, \ldots, f_n\}$, et il existe $\{\mu_1, \ldots, \mu_n\} \subset \mathbb{R}$ t.q. $J\boldsymbol{f}_i = \mu_i \boldsymbol{f}_i$ pour tout $i \in \{1, \ldots, n\}$ et t.q. $D\boldsymbol{f}_i \cdot \boldsymbol{f}_j = \delta_{i,j}$ pour tout $i, j \in \{1, \ldots, n\}$.

   En ordonnant les valeurs propres de J, on a donc $\mu_1 \leq \ldots \leq \mu_n$, on conserve cette notation dans la suite.

3. Montrer que la trace de J est nulle et en déduire que $\mu_1 \leq 0$ et $\mu_n \geq 0$.

   On suppose maintenant que A et $2D - A$ sont symétriques définies positives et on pose $x = A^{-1}b$.

4. Montrer que la méthode de Jacobi (par points) converge (c'est-à-dire $x^{(k)} \to x$ quand $n \to \infty$). [Utiliser un théorème du cours.]

   On se propose maintenant d'améliorer la convergence de la méthode par une technique de relaxation. Soit $\omega > 0$, on considère la méthode suivante :
   Initialisation. $x^{(0)} \in \mathbb{R}^n$
   Itérations. Pour $n \in \mathbb{N}$, $D\tilde{x}^{(k+1)} = (E + F)x^{(k)} + b$, $x^{(k+1)} = \omega\tilde{x}^{(k+1)} + (1 - \omega)x^{(k)}$.

5. Calculer les matrices M_ω (inversible) et N_ω telles que $M_\omega x^{(k+1)} = N_\omega x^{(k)} + b$ pour tout $n \in \mathbb{N}$, en fonction de ω, D et A. On note, dans la suite $J_\omega = (M_\omega)^{-1}N_\omega$.

6. On suppose dans cette question que $(2/\omega)D - A$ est symétrique définie positive. Montrer que la méthode converge (c'est-à-dire que $x^{(k)} \to x$ quand $n \to \infty$.)

7. Montrer que $(2/\omega)D - A$ est symétrique définie positive si et seulement si $\omega < 2/(1 - \mu_1)$.

8. Calculer les valeurs propres de J_ω en fonction de celles de J. En déduire, en fonction des μ_i, la valeur "optimale" de ω, c'est-à-dire la valeur de ω minimisant le rayon spectral de J_ω.

Exercice 53 (Méthodes de Jacobi et Gauss Seidel pour une matrice 3×3)**.**

On considère la matrice $A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$ et le vecteur $b = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$. Soit $x^{(0)}$ un vecteur de $\mathbb{R}^3$ donné.

1. *Méthode de Jacobi*

1.a Ecrire la méthode de Jacobi pour la résolution du système $A\boldsymbol{x} = \boldsymbol{b}$, sous la forme $\boldsymbol{x}^{(k+1)} = B_J\boldsymbol{x}^{(k)} + \boldsymbol{c}_J$.

1.b Déterminer le noyau de B_J et en donner une base.

1.c Calculer le rayon spectral de B_J et en déduire que la méthode de Jacobi converge.

1.d Calculer $x^{(1)}$ et $x^{(2)}$ pour les choix suivants de $x^{(0)}$:

$$(i)\, x^{(0)} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \qquad (ii) x^{(0)} = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}.$$

2. *Méthode de Gauss-Seidel.*

2.a Ecrire la méthode de Gauss-Seidel pour la résolution du système $A\boldsymbol{x} = b$, sous la forme $\boldsymbol{x}^{(k+1)} = B_{GS}\boldsymbol{x}^{(k)} + \boldsymbol{c}_{GS}$.

2.b Déterminer le noyau de B_{GS}.

2.c Calculer le rayon spectral de B_{GS} et en déduire que la méthode de Gauss-Seidel converge.

2.d Comparer les rayons spectraux de B_{GS} et B_J et vérifier ainsi un résultat du cours.

2.d Calculer $\boldsymbol{x}^{(1)}$ et $\boldsymbol{x}^{(2)}$ pour les choix suivants de $\boldsymbol{x}^{(0)}$:

$$(i)\ \ \boldsymbol{x}^{(0)} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \qquad (ii)\ \ \boldsymbol{x}^{(0)} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}.$$

3. *Convergence en un nombre fini d'itérations.*

3.1 Soit α et β des réels. Soit $u^{(0)} \in R$ et $(u^{(k)})_{k\in\mathbb{N}}$ la suite réelle définie par $u^{(k+1)} = \alpha u^{(k)} + \beta$.

3.1.a Donner les valeurs de α et β pour lesquelles la suite $(u^{(k)})_{k\in\mathbb{N}}$ converge.

3.1.b On suppose que $\alpha \neq 0$, et que la suite $(u^{(k)})_{k\in\mathbb{N}}$ converge vers une limite qu'on note $\overline{u}$. Montrer que s'il existe $K \in N$ tel que $u_K = \overline{u}$, alors $u^{(k)} = \overline{u}$ pour tout $k \in \mathbb{N}$.

3.2 Soit $N > 1$, B une matrice réelle carrée d'ordre N et $b \in \mathbb{R}^n$. Soit $u^{(0)} \in \mathbb{R}^N$ et $(u^{(k)})_{k\in\mathbb{N}}$ la suite définie par $u^{(k+1)} = Bu^{(k)} + c$.

3.2.a Donner les conditions sur B et c pour que la suite $(u^{(k)})_{k\in\mathbb{N}}$ converge vers une limite indépendante du choix initial $u_0 \in \mathbb{R}^N$.

3.2.b On suppose que la suite $(u^{(k)})_{k\in\mathbb{N}}$ converge vers une limite qu'on note $\overline{u}$. Montrer qu'on peut avoir $u^{(1)} = \overline{u}$ avec $u^{(0)} \neq \overline{u}$.

Exercice 54 (Jacobi et Gauss-Seidel pour une matrice tridiagonale)**.** *Corrigé en page 102*

1. Soit $A = \begin{bmatrix} 3 & -1 \\ -1 & 3 \end{bmatrix}$. Ecrire les méthodes de Jacobi et Gauss Seidel pour la résolution de $Ax = b$. Soient $B_{\rm J}$ et $B_{\rm GS}$ les matrices d'itération respectives. Calculer $\rho(B_{\rm J})$ et $\rho(B_{\rm GS})$ et vérifier que $\rho(B_{\rm GS}) = \rho(B_{\rm J})^2$
Soit $A = (a_{i,j})_{i,j=1,\dots,n} \in \mathcal{M}_n(\mathbb{R})$ une matrice carrée d'ordre n tridiagonale, c'est-à-dire telle que $a_{i,j} = 0$ si $|i-j| > 1$, et telle que la matrice diagonale $D = diag(a_{i,i})_{i=1,\dots,n}$ soit inversible. On note $A = D - E - F$ où $-E$ (resp. $-F$) est la partie triangulaire inférieure (resp. supérieure) de A, et on note J et G les matrices d'itération des méthodes de Jacobi et Gauss-Seidel associées à la matrice A.

2.a. Pour $\mu \in \mathbb{C}$, $\lambda \neq 0$ et $x \in \mathbb{C}^n$, on note

$$x_\mu = (x_1, \mu x_2, \dots, \mu^{k-1}x_k, \mu^{n-1}x_n)^t.$$

Montrer que si λ est valeur propre de J associée au vecteur propre x, alors x_μ vérifie $(\mu E + \frac{1}{\mu}F)x_\mu = \lambda D x_\mu$. En déduire que si $\lambda \neq 0$ est valeur propre de J alors λ^2 est valeur propre de G.

2.b Montrer que si λ^2 est valeur propre non nulle de G, alors λ est valeur propre de J .

3. Montrer que $\rho(B_{\rm GS}) = \rho(B_{\rm J})^2$. En déduire que lorsqu'elle converge, la méthode de Gauss-Seidel pour la résolution du système $Ax = b$ converge plus rapidement que la méthode de Jacobi.

4. Soit B_ω la matrice d'itération de la méthode SOR associée à A. Montrer que λ est valeur propre de J si et seulement si ν_ω est valeur propre de B_ω, où $\nu_\omega = \mu_\omega^2$ et μ_ω vérifie $\mu_\omega^2 - \lambda\omega\mu_\omega + \omega - 1 = 0$.
En déduire que

$$\rho(B_\omega) = \max_{\lambda \text{ valeur propre de } J} \{|\mu_\omega|; \mu_\omega^2 - \lambda\omega\mu_\omega + \omega - 1 = 0\}.$$

Exercice 55 (Méthode de Jacobi pour des matrices particulières)**.** *Suggestions en page 95, corrigé en page 104*

On note $\mathcal{M}_n(\mathbb{R})$ l'ensemble des matrices carrées d'ordre n à coefficients réels, et Id la matrice identité dans $\mathcal{M}_n(\mathbb{R})$. Soit $A = [a_{i,j}]_{i,j=1,\dots,n} \in \mathcal{M}_n(\mathbb{R})$. On suppose que :

$$a_{i,j} \leq 0, \forall i, j = 1, \dots, n, \ i \neq j, \tag{1.115}$$

$$a_{i,i} > 0, \forall i = 1, \dots, n. \tag{1.116}$$

$$\sum_{i=1}^{n} a_{i,j} = 0, \forall j = 1, \dots, n. \tag{1.117}$$

Soit $\lambda \in \mathbb{R}_+^*$.

1. Pour $x \in \mathbb{R}^n$, on définit

$$\|x\|_A = \sum_{i=1}^{n} a_{i,i}|x_i|.$$

Montrer que $\|\cdot\|_A$ est une norme sur $\mathbb{R}^n$.

2. Montrer que la matrice $\lambda Id + A$ est inversible.
3. On considère le système linéaire suivant :

$$(\lambda Id + A)u = b \tag{1.118}$$

Montrer que la méthode de Jacobi pour la recherche de la solution de ce système définit une suite $(u^{(k)})_{k\in n} \subset \mathbb{R}^n$.

4. Montrer que la suite $(u^{(k)})_{k\in\mathbb{N}}$ vérifie :

$$\|u^{(k+1)} - u^{(k)}\|_A \leq (\frac{1}{1+\alpha})^k \|u^{(1)} - u^{(0)}\|_A,$$

où $\alpha = \min_{i=1,\dots,n} a_{i,i}$.

5. Montrer que la suite $(u^{(k)})_{k\in\mathbb{N}}$ est de Cauchy, et en déduire qu'elle converge vers la solution du système (1.118).

Exercice 56 (Une méthode itérative particulière)**.**

Soient $\alpha_1, \dots, \alpha_n$ des réels strictement positifs, et A la matrice $n \times n$ de coefficients $a_{i,j}$ définis par :

$$\begin{cases} a_{i,i} = 2 + \alpha_i \\ a_{i,i+1} = a_{i,i-1} = -1 \\ a_{i,j} = 0 \text{ pour tous les autres cas.} \end{cases}$$

Pour $\beta > 0$ on considère la méthode itérative $Mx^{(k+1)} = Nx^{(k)} + b$ avec $A = M - N$ et $N = \text{diag}(\beta - \alpha_i)$ (c.à.d $\beta - \alpha_i$ pour les coefficients diagonaux, et 0 pour tous les autres).

1. Soit $\lambda \in \mathbb{C}$ une valeur propre de la matrice $M^{-1}N$; montrer qu'il existe un vecteur $x \in \mathbb{C}^n$ non nul tel que $Nx \cdot \overline{x} = \lambda Mx \cdot \overline{x}$ (où $\overline{x}$ désigne le conjugué de x). En déduire que toutes les valeurs propres de la matrice $M^{-1}N$ sont réelles.

2. Montrer que le rayon spectral $\rho(M^{-1}N)$ de la matrice vérifie : $\rho(M^{-1}N) \leq \max_{i=1,n} \dfrac{|\beta - \alpha_i|}{\beta}$

3. Déduire de la question 1. que si $\beta > \dfrac{\overline{\alpha}}{2}$, où $\overline{\alpha} = \max_{i=1,n} \alpha_i$, alors $\rho(M^{-1}N) < 1$, et donc que la méthode itérative converge.

4. Trouver le paramètre β minimisant $\max_{i=1,n} \dfrac{|\beta - \alpha_i|}{\beta}$.
(On pourra d'abord montrer que pour tout $\beta > 0$, $|\beta - \alpha_i| \leq \max(\beta - \underline{\alpha}, \bar{\alpha} - \beta)$ pour tout $i = 1, \dots, n$, avec $\underline{\alpha} = \min_{i=1,\dots,n} \alpha_i$ et $\bar{\alpha} = \max_{i=1,\dots,n} \alpha_i$ et en déduire que $\max_{i=1,n} |\beta - \alpha_i| = \max(\beta - \underline{\alpha}, \bar{\alpha} - \beta)$) .

Exercice 57 (Une matrice 3×3)**.** *Suggestions en page 95, corrigé en page 105*

Soit $A \in M_3(\mathbb{R})$ définie par $A = Id - E - F$ avec

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

1. Montrer que A est inversible.
2. Soit $0 < \omega < 2$. Montrer que pour $(\frac{1}{\omega}Id - E)$ est inversible si et seulement si $\omega \neq \sqrt{2}/2$.

Pour $0 < \omega < 2$, $\omega \neq \sqrt{2}/2$, on considère la méthode itérative (pour trouver la solution de $Ax = b$) suivante :

$$(\frac{1}{\omega}Id - E)x^{n+1} = (F + \frac{1-\omega}{\omega}Id)x^n + b.$$

Il s'agit donc de la "méthode I" du cours avec $B = \mathcal{L}_\omega = (\frac{1}{\omega}Id - E)^{-1}(F + \frac{1-\omega}{\omega}Id)$.

3. Calculer, en fonction de ω, les valeurs propres de $\mathcal{L}_\omega$ et son rayon spectral.
4. Pour quelles valeurs de ω la méthode est-elle convergente ? Déterminer $\omega_0 \in]0,2[$ t.q. $\rho(\mathcal{L}_{\omega_0}) = \min\{\rho(\mathcal{L}_\omega), \omega \in]0,2[, \omega \neq \sqrt{2}/2\}$.

Exercice 58 (Méthode des directions alternées)**.**

Soit $n \in \mathbb{N}$, $n \geq 1$ et oit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice carrée d'ordre n symétrique inversible et $b \in \mathbb{R}^n$. On cherche à calculer $u \in \mathbb{R}^n$, solution du système linéaire suivant :

$$Au = b, \tag{1.119}$$

On suppose connues des matrices X et $Y \in \mathcal{M}_n(\mathbb{R})$, symétriques. Soit $\alpha \in \mathbb{R}_+^*$, choisi tel que $X + \alpha\mathrm{Id}$ et $Y + \alpha\mathrm{Id}$ soient définies positives (où Id désigne la matrice identité d'ordre n) et $X + Y + \alpha\mathrm{Id} = A$.
Soit $u^{(0)} \in \mathbb{R}^n$, on propose la méthode itérative suivante pour résoudre (1.119) :

$$(X + \alpha Id)u^{(k+1/2)} = -Yu^{(k)} + b, \tag{1.120a}$$

$$(Y + \alpha Id)u^{(k+1)} = -Xu^{(k+1/2)} + b. \tag{1.120b}$$

1. Montrer que la méthode itérative (1.120) définit bien une suite $(u^{(k)})_{k\in\mathbb{N}}$ et que cette suite converge vers la solution u de (1.1) si et seulement si

$$\rho\big((Y + \alpha Id)^{-1}X(X + \alpha Id)^{-1}Y\big) < 1.$$

(On rappelle que pour toute matrice carrée d'ordre n, $\rho(M)$ désigne le rayon spectral de la matrice M.)

2. Montrer que si les matrices $(X + \frac{\alpha}{2}Id)$ et $(Y + \frac{\alpha}{2}Id)$ sont définies positives alors la méthode (1.120) converge. On pourra pour cela (mais ce n'est pas obligatoire) suivre la démarche suivante :
   (a) Montrer que

$$\rho\big((Y + \alpha Id)^{-1}X(X + \alpha Id)^{-1}Y\big) = \rho\big(X(X + \alpha Id)^{-1}Y(Y + \alpha Id)^{-1}\big).$$

   (On pourra utiliser l'exercice 29 page 61).
   (b) Montrer que

$$\rho\big(X(X + \alpha Id)^{-1}Y(Y + \alpha Id)^{-1}\big) \leq \rho\big(X(X + \alpha Id)^{-1}\big)\rho\big(Y(Y + \alpha Id)^{-1}\big).$$

   (c) Montrer que $\rho\big(X(X + \alpha Id)^{-1}\big) < 1$ si et seulement si la matrice $(X + \frac{\alpha}{2}Id)$ est définie positive.
   (d) Conclure.

3. Soit $f \in C([0,1] \times [0,1])$ et soit A la matrice carrée d'ordre $n = M \times M$ obtenue par discrétisation de l'équation $-\Delta u = f$ sur le carré $[0,1] \times [0,1]$ avec conditions aux limites de Dirichlet homogènes $u = 0$ sur $\partial\Omega$, par différences finies avec un pas uniforme $h = \frac{1}{M}$, et $\boldsymbol{b}$ le second membre associé.
   (a) Donner l'expression de A et $\boldsymbol{b}$.
   (b) Proposer des choix de X, Y et α pour lesquelles la méthode itérative (1.120) converge dans ce cas et qui justifient l'appellation "méthode des directions alternées" qui lui est donnée.

1.5.5 Exercices, suggestions

Exercice 45 page 88 (Méthode itérative du "gradient à pas fixe".)

1. Calculer le rayon spectral $\rho(B)$ de la matrice d'itération $B = Id - \alpha A$. Calculer les valeurs de α pour lesquelles $\rho(B) < 1$ et en déduire que la méthode itérative du gradient à pas fixe converge si $0 < \alpha < \frac{2}{\rho(A)}$.
2. Remarquer que $\rho(Id - \alpha A) = \max(|1 - \alpha\lambda_1|, |1 - \alpha\lambda_n - 1|$, où $\lambda_1, \ldots, \lambda_n$ sont les valeurs propres de A ordonnées dans le sens croissant. En traçant les graphes des valeurs prises par $|1-\alpha\lambda_1|$ et $|1-\alpha\lambda_n-1|$ en fonction de α, en déduire que le min est atteint pour $\alpha = \frac{2}{\lambda_1+\lambda_n}$.

Exercice 46 page 88 (Non convergence de la méthode de Jacobi)

Considérer d'abord le cas $a = 0$.
Si $a \neq 0$, pour chercher les valeurs de a pour lesquelles A est symétrique définie positive, calculer les valeurs propres de A en cherchant les racines du polynôme caractéristique. Introduire la variable μ telle que $a\mu = 1 - \lambda$.
Pour chercher les valeurs de a pour lesquelles la méthode de Jacobi converge, calculer les valeurs propres de la matrice d'itération J définie en cours.

Exercice 48 page 89 (Une matrice cyclique)

1. On peut trouver les trois valeurs propres (dont une double) sans calcul en remarquant que pour $\alpha = 0$ il y a 2 fois 2 lignes identiques, que la somme des colonnes est un vecteur constant et par le calcul de la trace.
2. Une matrice A est symétrique définie positive si et seulement si elle est diagonalisable et toutes ses valeurs propres sont strictement positives.
3. Appliquer le cours.

Exercice 49 page 89 (Jacobi et diagonale dominante stricte.)

Pour montrer que A est inversible, montrer que $Ax = 0$ si et seulement si $x = 0$. Pour montrer que la méthode de Jacobi converge, montrer que toutes les valeurs propres de la matrice A sont strictement inférieures à 1 en valeur absolue.

Exercice 52 page 90 (Méthode de Jacobi et relaxation.)

1. Prendre pour A une matrice (2,2) symétrique dont les éléments diagonaux sont différents l'un de l'autre.
2. Appliquer l'exercice 5 page 16 en prenant pour T l'application linéaire dont la matrice est D et pour S l'application linéaire dont la matrice est $E + F$.
4. Remarquer que $\rho(B_J) = \max(-\mu_1, \mu_n)$, et montrer que :
si $\mu_1 \leq -1$, alors $2D - A$ n'est pas définie positive,
si $\mu_n \geq 1$, alors A n'est pas définie positive.
6. Reprendre le même raisonnement qu'à la question 2 à 4 avec les matrices M_ω et N_ω au lieu de D et $E + F$.
7. Chercher une condition qui donne que toutes les valeurs propres sont strictement positives en utilisant la base de vecteurs propres ad hoc. (Utiliser la base de $\mathbb{R}^n$, notée $\{f_1, \ldots, f_n\}$, trouvée à la question 2.)
8. Remarquer que les f_i de la question 2 sont aussi vecteurs propres de J_ω et en déduire que les valeurs propres $\mu_i^{(\omega)}$ de J_ω sont de la forme $\mu_i^{(\omega)} = \omega(\mu_i - 1 - 1/\omega)$. Pour trouver le paramètre optimal ω_0, tracer les graphes des fonctions de $\mathbb{R}_+$ dans $\mathbb{R}$ définies par $\omega \mapsto |\mu_1^{(\omega)}|$ et $\omega \mapsto |\mu_n^{(\omega)}|$, et en conclure que le minimum de $\max(|\mu_1^{(\omega)}|, |\mu_n^{(\omega)}|)$ est atteint pour $\omega = \frac{2}{2-\mu_1-\mu_n}$.

Exercice 55 page 92 (Méthode de Jacobi et relaxation.)

2. Utiliser l'exercice 49 page 89

Exercice 57 page 93 (Convergence de SOR.)

1. Calculer le déterminant de A.
2. Calculer le déterminant de $\frac{I}{d}\omega - E$.
3. Remarquer que les valeurs propres de $\mathcal{L}_\omega$ annulent $\det(\frac{1-\omega}{\omega} Id + F - \lambda(\frac{I}{d}\omega - E))$. Après calcul de ce déterminant, on trouve $\lambda_1 = 1 - \omega$, $\lambda_2 = \frac{1-\omega}{1+\sqrt{2}\omega}$, $\lambda_3 = \frac{1-\omega}{1-\sqrt{2}\omega}$.
Montrer que si $\omega < \sqrt{2}$, $\rho(\mathcal{L}_\omega) = |\lambda_3|$ et que $\rho(\mathcal{L}_\omega) = |\lambda_1|$ si $\omega \geq \sqrt{2}$.
4. Utiliser l'expression des valeurs propres pour montrer que la méthode converge si $\omega > \frac{2}{1+\sqrt{2}}$ et que le paramètre de relaxation optimal est $\omega_0 = 1$.

1.5.6 Exercices, corrigés

Exercice 44 page 87

(a) La valeur propre double est $\frac{2}{3}$ et donc le rayon spectral est $\frac{2}{3}$ qui est strictement inférieur à 1, donc la suite converge vers $\overline{\boldsymbol{x}} = (Id - B)^{-1}\boldsymbol{c} = \begin{bmatrix} 3 \\ 9 \end{bmatrix}$

(b) Les valeurs propres sont $\frac{2}{3}$ et 3 et donc le rayon spectral est 2 qui est strictement inférieur à 1, donc la suite diverge

Exercice 45 page 88 (Méthode itérative de Richardson)

1. On peut réécrire l'itération sous la forme : $x_{k+1} = (Id - \alpha A)x_k + \alpha b$. La matrice d'itération est donc $B = Id - \alpha A$. La méthode converge si et seulement si $\rho(B) < 1$; or les valeurs propres de B sont de la forme $1 - \alpha\lambda_i$ où λ_i est v.p. de A. On veut donc :

$$-1 < 1 - \alpha\lambda_i < 1, \ \ \forall i = 1, \ldots, n.$$

c'est–à–dire $-2 < -\alpha\lambda_i$ et $-\alpha\lambda_i < 0, \forall i = 1, \ldots, n$.
Comme A est symétrique définie positive, $\lambda_i > 0, \forall i = 1, \ldots, n$, donc il faut $\alpha > 0$.
De plus, on a :

$$(-2 < -\alpha\lambda_i \ \ \forall i = 1, \ldots, n) \iff (\alpha < \frac{2}{\lambda_i} \ \ \forall i, 1, \ldots, n) \iff (\alpha < \frac{2}{\lambda_n}).$$

La méthode converge donc si et seulement si $0 < \alpha < \dfrac{2}{\rho(A)}$.

2. On a : $\rho(Id - \alpha A) = \sup_i |1 - \alpha\lambda_i| = \max(|1 - \alpha\lambda_1|, |1 - \alpha\lambda_n|)$. Le minimum de $\rho(Id - \alpha A)$ est donc obtenu pour α_0 tel que $1 - \alpha_0\lambda_1 = \alpha_0\lambda_n - 1$, c'est–à–dire (voir Figure (1.6) $\alpha_0 = \dfrac{2}{\lambda_1 + \lambda_n}$.

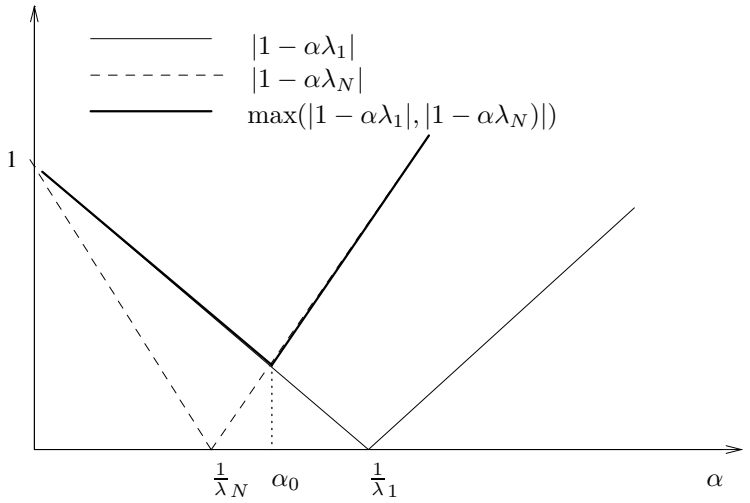

FIGURE 1.6: Graphes de $|1 - \alpha\lambda_1|$ et $|1 - \alpha\lambda_n|$ en fonction de α.

Exercice 46 page 88 (Non convergence de la méthode de Jacobi)

— Si $a = 0$, alors $A = Id$, donc A est s.d.p. et la méthode de Jacobi converge.
— Si $a \neq 0$, posons $a\mu = (1 - \lambda)$, et calculons le polynôme caractéristique de la matrice A en fonction de la variable μ.

$$P(\mu) = det \begin{vmatrix} a\mu & a & a \\ a & a\mu & a \\ a & a & a\mu \end{vmatrix} = a^3 \det \begin{vmatrix} \mu & 1 & 1 \\ 1 & \mu & 1 \\ 1 & 1 & \mu \end{vmatrix} = a^3(\mu^3 - 3\mu + 2).$$

On a donc $P(\mu) = a^3(\mu - 1)^2(\mu + 2)$. Les valeurs propres de la matrice A sont donc obtenues pour $\mu = 1$ et $\mu = 2$, c'est–à–dire : $\lambda_1 = 1 - a$ et $\lambda_2 = 1 + 2a$.
La matrice A est définie positive si $\lambda_1 > 0$ et $\lambda_2 > 0$, c'est–à–dire si $-\frac{1}{2} < a < 1$.
La méthode de Jacobi s'écrit :

$$X^{(k+1)} = D^{-1}(D - A)X^{(k)},$$

avec $D = Id$ dans le cas présent ; donc la méthode converge si et seulement si $\rho(D - A) < 1$.
Les valeurs propres de $D - A$ sont de la forme $\nu = 1 - \lambda$ où λ est valeur propre de A. Les valeurs propres de $D - A$ sont donc $\nu_1 == -a$ (valeur propre double) et $\nu_2 = 2a$.. On en conclut que la méthode de Jacobi converge si et seulement si $-1 < -a < 1$ et $-1 < 2a < 1$, *i.e.* $\frac{1}{2} < a < \frac{1}{2}$.
La méthode de Jacobi ne converge donc que sur l'intervalle $]-\frac{1}{2}, \frac{1}{2}[$ qui est strictement inclus dans l'intervalle $]-\frac{1}{2}, 1[$ des valeurs de a pour lesquelles la matrice A est s.d.p..

Exercice 49 page 89 (Jacobi pour les matrices à diagonale dominante stricte)

Pour montrer que A est inversible, supposons qu'il existe $x \in \mathbb{R}^n$ tel que $Ax = 0$; on a donc

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

Pour $i \in \{1, \ldots, n\}$, on a donc

$$|a_{i,i}||x_i| = |a_{i,i}x_i| = |\sum_{j;i\neq j} a_{i,j}x_j| \leq \sum_{j;i\neq j} |a_{i,j}|\|x\|_\infty, \quad \forall i = 1, \ldots, n.$$

Si $x \neq 0$, on a donc

$$|x_i| \leq \frac{\sum_{j;i\neq j} a_{i,j}x_j|}{|a_{i,i}|}\|x\|_\infty < \|x\|_\infty, \quad \forall i = 1, \ldots, n,$$

ce qui est impossible pour i tel que

$$|x_i| = \|x\|_\infty.$$

Montrons maintenant que la méthode de Jacobi converge : Si on écrit la méthode ous la forme $Px^{(k+1)} = (P - A)x^{(k)} + b$ avec , on a

$$P = D = \begin{bmatrix} a_{1,1} & & 0 \\ & \ddots & \\ 0 & & a_{n,n} \end{bmatrix}.$$

La matrice d'itération est

$$B_{\mathrm{J}} = P^{-1}(P - A) = D^{-1}(E + F) = \begin{bmatrix} a_{1,1}^{-1} & & 0 \\ & \ddots & \\ 0 & & a_{n,n}^{-1} \end{bmatrix} \begin{bmatrix} 0 & & -a_{i,j} \\ & \ddots & \\ -a_{i,j} & & 0 \end{bmatrix}$$
$$= \begin{bmatrix} 0 & -\frac{a_{1,2}}{a_{1,1}} & \cdots \\ & \ddots & \\ -\frac{a_{1,1}}{a_{n,n}} & \cdots & 0 \end{bmatrix}.$$

Cherchons le rayon spectral de B_J : soient $x \in \mathbb{R}^n$ et $\lambda \in \mathbb{R}$ tels que $B_J x = \lambda x$, alors

$$\sum_{j;i\neq j} -\frac{a_{i,j}}{a_{i,i}} x_j = \lambda x_i, \text{ et donc } |\lambda||x_i| \leq \sum_{j;i\neq j} |a_{i,j}| \frac{\|x\|_\infty}{|a_{i,i}|}.$$

Soit i tel que $|x_i| = \|x\|_\infty$ et $x \neq 0$, on déduit de l'inégalité précédente que

$$|\lambda| \leq \frac{\sum_{j;i\neq j} |a_{i,j}|}{|a_{ii}|} < 1 \text{ pour toute valeur propre } \lambda.$$

On a donc $\rho(B_J) < 1$ ce qui prouve que la méthode de Jacobi converge.

Exercice 52 page 90 (Méthode de Jacobi et relaxation)

1. $B_J = D^{-1}(E+F)$ peut ne pas être symétrique, même si A est symétrique :
   En effet, prenons $A = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$.
   Alors
   $$B_J = D^{-1}(E+F) = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{2} \\ 1 & 0 \end{pmatrix} \neq \begin{pmatrix} 0 & 1 \\ \frac{1}{2} & 0 \end{pmatrix}.$$
   donc B_J n'est pas symétrique.
2. On applique l'exercice précédent pour l'application linéaire T de matrice D, qui est, par hypothèse, définie positive (et évidemment symétrique puisque diagonale) et $S = E + F$, symétrique car A est symétrique. Il existe donc $(f_1 \dots f_n)$ base de E et $(\mu_1 \dots \mu_n) \in \mathbb{R}^n$ tels que
   $$B_J f_i = D^{-1}(E+F) f_i = \mu_i f_i, \quad \forall i = 1, \dots, n, \text{ et } (Df_i, f_j) = \delta_{ij}.$$
3. Par définition de B_J, tous les éléments diagonaux de B_J sont nuls et donc sa trace également. Or $TrB_J = \sum_{i=1}^{n} \mu_i$. Si $\mu_i > 0 \ \forall i = 1, \dots, n$, alors $TrB_J > 0$, donc $\exists i_0;\ \mu_i \leq 0$ et comme $\mu_1 \leq \mu_{i_0}$, on a $\mu_1 \leq 0$. Un raisonnement similaire montre que $\mu_n \geq 0$.
4. La méthode de Jacobi converge si et seulement si $\rho(B_J) < 1$ (théorème 1.49 page 77). Or, par la question précédente, $\rho(A) = \max(-\mu_1, \mu_n)$. Supposons que $\mu_1 \leq -1$, alors $\mu_1 = -\alpha$, avec $\alpha \geq 1$. On a alors $D^{-1}(E+F)f_1 = -\alpha f_1$ ou encore $(E+F)f_1 = -\alpha D f_1$, ce qui s'écrit aussi $(D+E+F)f_1 = D(1-\alpha)f_1$ c'est–à–dire $(2D-A)f_1 = \beta D f_1$ avec $\beta \leq 0$. On en déduit que $((2D-A)f_1, f_1) = \beta \leq 0$, ce qui contredit le fait que $2D - A$ est définie positive. En conséquence, on a bien $\mu_1 \geq -1$.
   Supposons maintenant que $\mu_n = \alpha \geq 1$. On a alors $D^{-1}(E+F)f_1 = -\alpha f_n$, soit encore $(E+F)f_n = -\alpha D f_n$. On en déduit que $Af_n = (D-E-F)f_n = D(1-\alpha)f_n = D\beta f_n$ avec $\beta \leq 0$. On a alors$(Af_n, f_n) \leq 0$, ce qui contredit le fait que A est définie positive.
5. Par définition, on a :$D\tilde{x}^{(k+1)}s = (E+F)x^{(k)} + b$ et $x^{(k+1)} = \omega\tilde{x}^{(k+1)} + (1-\omega)x^{(k)}$. On a donc $x^{(k+1)} = \omega[D^{-1}(E+F)x^{(k)} + D^{-1}b] + (1-\omega)x^{(k)}$ c'est–à–dire $x^{(k+1)} = [Id - \omega(Id - D^{-1}(E+F))]x^{(k)} + \omega D^{-1}b$,, soit encore $\frac{1}{\omega}Dx^{(k+1)} = [\frac{1}{\omega}D - (D - (E+F))]x^{(k)} + b$. On en déduit que $M_\omega x^{(k+1)} = N_\omega x^{(k)} + b$ avec $M_\omega = \frac{1}{\omega}D$ et $N_\omega = \frac{1}{\omega}D - A$.
6. La matrice d'itération est donc maintenant $J_\omega = M_\omega^{-1}N_\omega$ qui est symétrique pour le produit scalaire $(\cdot,\cdot)_{M_\omega}$ donc en reprenant le raisonnement de la question 2, il existe une base $(\tilde{f}_1, \dots, \tilde{f}_n) \in (\mathbb{R}^n)^n$ et $(\tilde{\mu}_1, \dots \tilde{\mu}_n) \subset \mathbb{R}^n$ tels que
   $$J_\omega \tilde{f}_i = {M_\omega}^{-1} N_\omega \tilde{f}_i = \omega D^{-1}\left(\frac{1}{\omega}D - A\right)\tilde{f}_i = \tilde{\mu}_i \tilde{f}_i, \quad \forall i = 1, \dots, n,$$
   $$\text{et } \frac{1}{\omega} D\tilde{f}_i \cdot \tilde{f}_j = \delta_{ij}, \quad \forall i, = 1, \dots, n, \forall j, = 1, \dots, n.$$

Supposons $\tilde{\mu}_1 \leq -1$, alors $\tilde{\mu}_1 = -\alpha$, avec $\alpha \geq 1$ et $\omega D^{-1}(\frac{1}{\omega}D - A)\tilde{f}_1 = -\alpha\tilde{f}_1$, ou encore $\frac{1}{\omega}D - A\tilde{f}_1 = -\alpha\frac{1}{\omega}D\tilde{f}_1$. On a donc $\frac{2}{\omega}D - A\tilde{f}_1 = (1-\alpha)\frac{1}{\omega}D\tilde{f}_1$, ce qui entraîne $(\frac{2}{\omega}D - A)\tilde{f}_1 \cdot \tilde{f}_1 \leq 0$. Ceci contredit l'hypothèse $\dfrac{2}{\omega}D - A$ définie positive.

De même, si $\tilde{\mu}_n \geq 1$, alors $\tilde{\mu}_n = \alpha$ avec $\alpha \geq 1$. On a alors

$$(\frac{1}{\omega}D - A)\tilde{f}_n = \alpha\frac{1}{\omega}D\tilde{f}_n,$$

et donc $A\tilde{f}_n = (1-\alpha)\frac{1}{\omega}D\tilde{f}_n$ ce qui entraîne en particulier que $A\tilde{f}_n \cdot \tilde{f}_n \leq 0$; or ceci contredit l'hypothèse A définie positive.

7. On cherche une condition nécessaire et suffisante pour que

$$\left(\frac{2}{\omega}D - A\right) x \cdot x > 0, \quad \forall x \neq 0, \tag{1.121}$$

ce qui est équivalent à

$$\left(\frac{2}{\omega}D - A\right) f_i \cdot f_i > 0, \quad \forall i = 1, \ldots, n, \tag{1.122}$$

où les $(f_i)_{i=1,n}$ sont les vecteurs propres de $D^{-1}(E+F)$. En effet, la famille $(f_i)_{i=1,\ldots,n}$ est une base de $\mathbb{R}^n$, et

$$\begin{aligned} \left(\frac{2}{\omega}D - A\right) f_i &= \left(\frac{2}{\omega}D - D + (E+F)\right) f_i \\ &= \left(\frac{2}{\omega} - 1\right) Df_i + \mu_i Df_i \\ &= \left(\frac{2}{\omega} - 1 + \mu_i\right) Df_i. \end{aligned} \tag{1.123}$$

On a donc en particulier $\left(\frac{2}{\omega}D - A\right) f_i \cdot f_j = 0$ is $i \neq j$, ce qui prouve que (1.121) est équivalent à (1.122). De (1.122), on déduit, grâce au fait que $(Df_i, f_i) = 1$,

$$\left(\left(\frac{2}{\omega}D - A\right) f_i, f_i\right) = \left(\frac{2}{\omega} - 1 + \mu_i\right).$$

On veut donc que $\frac{2}{\omega} - 1 + \mu_1 > 0$ car $\mu_1 = \inf \mu_i$, c'est–à–dire : $-\frac{2}{\omega} < \mu_1 - 1$, ce qui est équivalent à : $\omega < \dfrac{2}{1-\mu_1}$.

8. La matrice d'itération J_ω s'écrit :

$$J_\omega = \left(\frac{1}{\omega}D\right)^{-1}\left(\frac{1}{\omega}D - A\right) = \omega I_\omega, \text{ avec } I_\omega = D^{-1}(\frac{1}{\omega}D - A).$$

Soit λ une valeur propre de I_ω associée à un vecteur propre u ; alors :

$$D^{-1}\left(\frac{1}{\omega}D - A\right) u = \lambda u, \text{ } \textit{i.e.} \left(\frac{1}{\omega}D - A\right) u = \lambda Du.$$

On en déduit que

$$(D-A)u + \left(\frac{1}{\omega} - 1\right) Du = \lambda Du, \text{ soit encore}$$

$$D^{-1}(E+F)u = \left(1 - \frac{1}{\omega} + \lambda\right) u.$$

Or f_i est vecteur propre de $D^{-1}(E+F)$ associée à la valeur propre μ_i (question 2). On a donc :

$$D^{-1}(E+F)f_i = \mu_i f_i = \left(1 - \frac{1}{\omega} + \lambda\right) f_i,$$

ce qui est vrai si $\mu_i = 1 - \frac{1}{\omega} + \lambda$, c'est–à–dire $\lambda = \mu_i - 1 - \frac{1}{\omega}$. Donc $\mu_i^{(\omega)} = \omega\left(\mu_i - 1 - \frac{1}{\omega}\right)$ est valeur propre de J_ω associée au vecteur propre f_i.
On cherche maintenant à minimiser le rayon spectral

$$\rho(J_\omega) = \sup_i |\omega(\mu_i - 1 - \frac{1}{\omega})|$$

On a

$$\omega(\mu_1 - 1 - \frac{1}{\omega}) \leq \omega(\mu_i - 1 - \frac{1}{\omega}) \leq \omega(\mu_n - 1 - \frac{1}{\omega}),$$

et

$$-\omega(\mu_n - 1 - \frac{1}{\omega}) \leq -\omega(\mu_1 - 1 - \frac{1}{\omega}) \leq -\omega(\mu_i - 1 - \frac{1}{\omega}),$$

donc

$$\rho(J_\omega) = \max\left(|\omega(\mu_n - 1 - \frac{1}{\omega})|, |-\omega(\mu_1 - 1 - \frac{1}{\omega}|)\right)$$

dont le minimum est atteint (voir Figure **??**) pour

$$\omega(1-\mu_1) - 1 = 1 - \omega(1-\mu_n) \text{ c'est–à–dire } \omega = \frac{2}{2 - \mu_1 - \mu_n}.$$

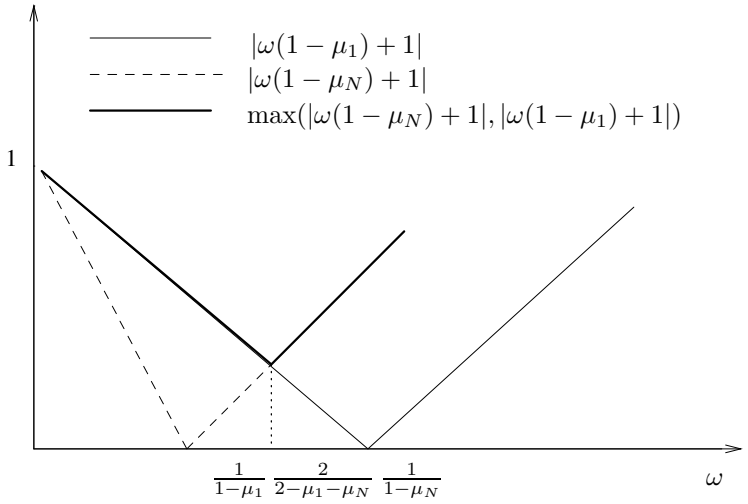

FIGURE 1.7: Détermination de la valeur de ω réalisant le minimum du rayon spectral.

Exercice 53 page 91 (Jacobi et Gauss-Seidel pour une matrice tridiagonale)

On considère la matrice $A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}$ **et le vecteur** $b = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$**. Soit** $x^{(0)}$ **un vecteur de** $\mathbb{R}^3$ **donné.**

1. ***Méthode de Jacobi***

1.a Ecrire la méthode de Jacobi pour la résolution du système $Ax = b$**, sous la forme** $\boldsymbol{x}^{(k+1)} = B_{\mathrm{J}}\boldsymbol{x}^{(k)} + \boldsymbol{c}_J$**.**

La méthode de Jacobi $D\boldsymbol{x}^{(k+1)} = (E+F)\boldsymbol{x}^{(k)} + \boldsymbol{b}$ avec

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

La méthode de Jacobi s'écrit donc $x^{(k+1)} = B_{\mathrm{J}}x^{(k)} + c_J$ avec $B_{\mathrm{J}} = D^{-1}(E+F) = \begin{bmatrix} 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 \end{bmatrix}$ et $c_J = \begin{bmatrix} \frac{1}{2} \\ 0 \\ \frac{1}{2} \end{bmatrix}$.

1.b Déterminer le noyau de B_{J} et en donner une base.

On remarque que $x \in \mathrm{Ker}(B_{\mathrm{J}})$ si $x_2 = 0$ et $x_1 + x_3 = 0$. Donc $\mathrm{Ker}B_{\mathrm{J}} = \{t \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}, t \in \mathbb{R}\}$.

1.c Calculer le rayon spectral de B_{J} et en déduire que la méthode de Jacobi converge.

Le polynôme caractéristique de B_{J} est $P_J(\lambda) = \det(B_{\mathrm{J}} - \lambda Id) = (-\lambda(-\lambda^2 + \frac{1}{2})$ et donc $\rho(B_{\mathrm{J}}) = \frac{\sqrt{2}}{2} < 1$. On en déduit que la méthode de Jacobi converge.

1.d Calculer $x^{(1)}$ et $x^{(2)}$ pour les choix suivants de $x^{(0)}$:

$$(i)\ x^{(0)} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \qquad (ii)\ x^{(0)} = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}.$$

Choix (i) : $x^{(1)} = \begin{bmatrix} \frac{1}{2} \\ 0 \\ \frac{1}{2} \end{bmatrix}, x^{(2)} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}$.

Choix (ii) : $x^{(1)} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, x^{(2)} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$.

2. *Méthode de Gauss-Seidel.*

2.a Ecrire la méthode de Gauss-Seidel pour la résolution du système $Ax = b$, sous la forme $x^{(k+1)} = B_{GS}x^{(k)} + c_{GS}$.

La méthode de Gauss-Seidel s'écrit $(D-E)x^{(k+1)} = Fx^{(k)} + b$, où D, E et F ont été définies à la question 1.a. La méthode s'écrit donc $x^{(k+1)} = B_{GS}x^{(k)} + c_{GS}$ avec $B_{GS} = (D-E)^{-1}F$ et $c_{GS} = (D-E)^{-1}b$. Calculons $(D-E))^{-1}F$ et $(D-E))^{-1}b$ par échelonnement.

$$\begin{bmatrix} 2 & 0 & 0 & 0 & 1 & 0 & 1 \\ -1 & 2 & 0 & 0 & 0 & 1 & 0 \\ 0 & -1 & 2 & 0 & 0 & 0 & 1 \end{bmatrix} \rightsquigarrow \begin{bmatrix} 2 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 2 & 0 & 0 & \frac{1}{2} & 1 & \frac{1}{2} \\ 0 & -1 & 2 & 0 & 0 & 0 & 1 \end{bmatrix} \rightsquigarrow \begin{bmatrix} 2 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 2 & 0 & 0 & \frac{1}{2} & 1 & \frac{1}{2} \\ 0 & 0 & 2 & 0 & \frac{1}{4} & \frac{1}{2} & \frac{5}{4} \end{bmatrix}$$

On a donc

$$B_{GS} = \begin{bmatrix} 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{4} & \frac{1}{2} \\ 0 & \frac{1}{8} & \frac{1}{4} \end{bmatrix} \text{ et } c_{GS} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{4} \\ \frac{5}{8} \end{bmatrix}.$$

2.b Déterminer le noyau de B_{GS}. Il est facile de voir que $x \in \mathrm{Ker}(B_{GS})$ si et seulement si $x_2 = x_3 = 0$. Donc $\mathrm{Ker}B_{GS} = \{t \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, t \in \mathbb{R}\}$.

2.c Calculer le rayon spectral de B_{GS} et en déduire que la méthode de Gauss-Seidel converge.

Le polynôme caractéristique de B_{GS} est $P_{GS}(\lambda) = \det(B_{GS} - \lambda Id)$. On a donc

$$P_{GS}(\lambda) = \begin{vmatrix} -\lambda & \frac{1}{2} & 0 \\ 0 & \frac{1}{4} - \lambda & \frac{1}{2} \\ 0 & \frac{1}{8} & \frac{1}{4} - \lambda \end{vmatrix} = -\lambda\left((\frac{1}{4} - \lambda)^2 - \frac{1}{16}\right) = \lambda^2(\frac{1}{2} - \lambda)$$

et donc $\rho(B_{GS}) = \frac{1}{2} < 1$. On en déduit que la méthode de Gauss-Seidel converge.

2.d Comparer les rayons spectraux de $B_{\rm GS}$ et $B_{\rm J}$ et vérifier ainsi un résultat du cours.
On a bien $\rho(B_{\rm GS}) = \frac{1}{2} = \rho(B_{\rm J})^2$, ce qui est conforme au théorème 1.36 du cours.

2.d Calculer $x^{(1)}$ et $x^{(2)}$ pour les choix suivants de $x^{(0)}$:

$$(i)\ \ x^{(0)} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \qquad\qquad (ii)\ \ x^{(0)} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}.$$

Choix (i) : $x^{(1)} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{4} \\ \frac{5}{8} \end{bmatrix}$, $x^{(2)} = \begin{bmatrix} \frac{5}{8} \\ \frac{5}{8} \\ \frac{13}{16}s \end{bmatrix}$.

Choix (ii) : $x^{(1)} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$, $x^{(2)} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$.

3. *Convergence en un nombre fini d'itérations.*
3.1 Soit α et β des réels. Soit $u^{(0)} \in R$ et $(u^{(k)})_{k\in\mathbb{N}}$ la suite réelle définie par $u^{(k+1)} = \alpha u^{(k)} + \beta$.

3.1.a Donner les valeurs de α et β pour lesquelles la suite $(u^{(k)})_{k\in\mathbb{N}}$ converge.
La suite $(u^{(k)})_{k\in\mathbb{N}}$ converge pour les valeurs suivantes de (α, β) :

1. $|\alpha| < 1$, $\beta \in \mathbb{R}$, auquel cas la suite converge vers $\overline{u} = \frac{\beta}{1-\alpha}$,
2. $\alpha = 1$, $\beta = 0$, auquel cas la suite est constante et égale à u_0.

3.1.b On suppose que $\alpha \neq 0$, et que la suite $(u^{(k)})_{k\in\mathbb{N}}$ converge vers une limite qu'on note $\overline{u}$. Montrer que s'il existe $K \in N$ tel que $u_K = \overline{u}$, alors $u^{(k)} = \overline{u}$ pour tout $k \in \mathbb{N}$.
Si $u_K = \overline{u}$, comme $u_K = \alpha u_{K-1} + \beta$ et que $\overline{u} = \alpha\overline{u} + \beta$, on en déduit que $0 = u_K - \overline{u} = \alpha(u_{K-1} - \overline{u})$, et comme $\alpha \neq 0$, ceci entraîne $u_{K-1} = \overline{u}$. On en déduit par récurrence que $u_k = \overline{u}$ pour tout $k \leq K-1$. On remarque ensuite que si $u_K = \overline{u}$ alors $u_{K+1} = \alpha u_K + \beta = \alpha\overline{u} + \beta = \overline{u}$. On en déduit par récurrence que $u_k = \overline{u}$ pour tout $k \in \mathbb{N}$.

3.2 Soit $N > 1$, B une matrice réelle carrée d'ordre N et $b \in \mathbb{R}^n$. Soit $u^{(0)} \in \mathbb{R}^N$ et $(u^{(k)})_{k\in\mathbb{N}}$ la suite définie par $u^{(k+1)} = Bu^{(k)} + c$.

3.2.a Donner les conditions sur B et c pour que la suite $(u^{(k)})_{k\in\mathbb{N}}$ converge vers une limite indépendante du choix initial $u_0 \in \mathbb{R}^N$.
Pour que la suite $(u^{(k)})_{k\in\mathbb{N}}$ converge, il faut qu'il existe $\overline{u}$ tel que $\overline{u} = B\overline{u} + c$, ou encore $c = (Id - B)\overline{u}$.
Supposons que c'est le cas. On a alors $u^{(k+1)} - \overline{u} = B(u^{(k)} - \overline{u})$, et donc $u^{(k)} - \overline{u} = B^k(u^{(0)} - \overline{u})$. On veut donc que $B^k(u^{(0)} - \overline{u})$ tende vers 0 pour tout choix initial $u^{(0)}$, c.à.d. que B^k tende vers 0. On en déduit, par le lemme 1.5 que la suite $(u^{(k)})_{k\in\mathbb{N}}$ converge si et seulement si $\rho(B) < 1$.

3.2.b On suppose que la suite $(u^{(k)})_{k\in\mathbb{N}}$ converge vers une limite qu'on note $\overline{u}$. Montrer qu'on peut avoir $u^{(1)} = \overline{u}$ avec $u^{(0)} \neq \overline{u}$.
En prenant $B = B_{\rm J}$, $c = c_J$ et $u^{(0)} = (0, 1, 2)^t$, on sait par la question 1.2 qu'on ar $u^{(1)} = (1, 1, 1)^t = \overline{u}$ avec $u^{(0)} \neq \overline{u}$.

Exercice 54 page 92 (Jacobi et Gauss-Seidel pour une matrice tridiagonale)

1.a. Soit $\mu \in \mathcal{C}, \lambda \neq 0$ et $x \in \mathbb{C}^n$, soit $x_\mu = (x_1, \mu x_2, \ldots, \mu^{k-1}x_k, \mu^{n-1}x_n)^t$, et soit λ est valeur propre de J associée au vecteur propre x. Calculons $(\mu E + \frac{1}{\mu}F)x_\mu$:
$((\mu E + \frac{1}{\mu}F)x_\mu)_i = \mu a_{i,i-1}\mu^{i-2}x_{i-1} + \frac{1}{\mu}a_{i,i+1}\mu^i x_{i+1} = \mu^{i-1}(a_{i,i-1}x_{i-1} + a_{i,i+1}x_{i+1} = \mu^{i-1}((E+F)x)_i = \lambda(Dx_\mu)_i$.
On a donc $(\mu E + \frac{1}{\mu}F)x_\mu = \lambda D x_\mu$. En prenant $\mu = \lambda$ dans légalité précédente, on obtient : $\frac{1}{\lambda}Fx_\lambda = \lambda(D-E)x_\lambda$, et donc $(D-E)^{-1}Fx_\lambda = \lambda^2 x_\lambda$. déduire que si $\lambda \neq 0$ est valeur propre de $B_{\rm J}$ alors λ^2 est valeur propre de G (associée au vecteur propre x_λ).

1.b. Réciproquement, supposons maintenant que λ^2 est valeur propre non nulle de G, alors il existe $y \in \mathbb{R}^n, y \neq 0$ tel que $(D-E)^{-1}Fy = \lambda^2 y$. Soit $x \in \mathbb{R}^n$ tel que $y = x_\lambda$, c'est-à-dire $x_i = \lambda^{1-i} y_i$, pour $i = 1, \ldots, n$. On en déduit que $\frac{1}{\lambda} F x_\lambda = \lambda^2 (D-E) x_\lambda$, et donc $(\lambda E + \frac{1}{\lambda} F) x_\lambda = \lambda D x_\lambda$.
Il est alors facile de vérifier (calculs similaires à ceux de la question 1.a) que $(E+F)x = \lambda D x$, d'où on déduit que λ est valeur propre de $B_{\rm J}$.

2. De par la question 1, on a finalement que $\lambda \in \mathcal{C}, \lambda \neq 0$ est valeur propre de $B_{\rm J}$ si et seulement si $\lambda^2 \in \mathcal{C}, \lambda \neq 0$ est valeur propre de $B_{\rm GS}$.
On en déduit que $\rho(B_{\rm GS}) = \rho(B_{\rm J})^2$.
On a donc en particuler que $\rho(B_{\rm GS}) < 1$ si et seulement si $\rho(B_{\rm GS}) < 1$, ce qui prouve que les méthodes de Jacobi et Gauss-Seidel convergent ou divergent simultanément.
De plus, on a vu à l'exercice 59 page 113 que si B désigne la matrice d'itération d'une méthode iérative $x^{(k+1)} = Bx^{(k)} + c$ pour la résolution de $Ax = b$, alors

$$\frac{\|x^{(k+1)} - x\|}{\|x^{(k)} - x\|} \to \rho(B) \text{ lorsque } n \to +\infty.$$

On en déduit que lorsqu'elle converge, la méthode de Gauss-Seidel converge plus rapidement que la méthode de Jacobi.

3.1 Soit B_ω la matrice d'itération de la méthode SOR associée à A, et soit ν_ω une valeur propre de $B_\omega = (\frac{1}{\omega} D - E)^{-1}(\frac{1-\omega}{\omega} D + F)$. Il existe donc $y \in \mathbb{C}^n, y \neq 0$, tel que

$$(1 - \omega D + \omega F) y = \nu_\omega (D - \omega E) y.$$

Ceci s'écrit encore : $(\omega F + \nu_\omega \omega E) y = (\nu_\omega - 11 + \omega) D y$, et aussi, en notant λ une valeur propre non nulle de $B_{\rm J}$,

$$(\frac{\lambda\omega}{\nu_\omega - 1 + \omega} F + \frac{\lambda \nu_\omega \omega}{\nu_\omega - 1 + \omega} E) y = \lambda D y,$$

soit encore

$$(\mu E + \frac{1}{\mu} F) y = \lambda D y, \tag{1.124}$$

avec

$$\mu = \frac{\lambda \nu_\omega \omega}{\nu_\omega - 1 + \omega} \text{ et } \frac{1}{\mu} = \frac{\lambda\omega}{\nu_\omega - 1 + \omega}.$$

Ceci est possible si $\nu_\omega - 1 + \omega \neq 0$, $\lambda\omega \neq 0$, et

$$\frac{\nu_\omega - 1 + \omega}{\lambda \nu_\omega \omega} = \frac{\lambda\omega}{\nu_\omega - 1 + \omega}. \tag{1.125}$$

Remarquons tout d'abord qu'on a forcément $\nu_\omega \neq 1 - \omega$. En effet, sinon, le vecteur propre y associé à ν_ω vérifie $\omega F y = -\omega E y$, ce qui est impossible pour $\omega \in]0, 2[$ et $y \neq 0$.
On a également $\lambda\omega \neq 0$ car $\lambda \neq 0$ et $\omega \neq 0$.
Voyons maintenant pour quelles valeurs de ν_ω la relation (1.125) est vérifiée. La relation (1.125) est équivalente à $(\nu_\omega - 1 + \omega)^2 = (\lambda \nu_\omega \omega)(\lambda\omega)$ ce qui revient à dire que $\nu_\omega = \mu_\omega^2$, où μ_ω est solution de l'équation

$$\mu_\omega^2 - \omega\lambda\mu_\omega + \omega - 1 = 0. \tag{1.126}$$

La relation (1.125) est donc vérifiée pour $\nu_\omega = \mu_\omega^2$, où μ_ω est racine de l'équation $\mu_\omega^2 - \omega\lambda\mu_\omega + \omega - 1 = 0$. Soit donc μ_ω^+ et μ_ω^- les racines (ou éventuellement la racine double) de cette équation (qui en admet toujours car on la résout dans $\mathcal{C}$).
Donc si $\lambda \neq 0$ est valeur propre de $B_{\rm J}$ associée au vecteur propre x, en vertu de (1.124) et de la question 1.a, les valeurs ν_ω telles que $\nu_\omega = (\mu_\omega)^2$ où μ_ω est solution de (1.126) sont valeurs propres de la matrice B_ω associés au vecteurs propres $x_(\mu_\omega)$.

Réciproquement, si $\nu_\omega = \mu_\omega^2$, où μ_ω est solution de l'équation (1.126), est valeur propre de B_ω, alors il existe un vecteur $y \neq 0$ tel que $B_\omega y = \nu_\omega y$. Soit $x \in \mathbb{R}^n$ tel que $x_{\mu_\omega} = y$ (i.e. $x_i = \mu_\omega^{1-i} y_i$ pour $i = 1, \ldots, n$). On a alors : $((1-\omega)D + \omega F)x_{\mu_\omega} = \mu_\omega^2 (D - \omega E)x_{\mu_\omega}$, soit encore $\omega(E+F)x_{\mu_\omega} = (\mu_\omega^2 - (1-\omega))Dx_{\mu_\omega}$. Or $\mu_\omega^2 - (1-\omega) = \omega\lambda\mu_\omega$ grâce a (1.126), et donc $(E+F)x_{\mu_\omega} = \lambda\mu_\omega D x_{\mu_\omega}$. On vérifie facilement que ceci entraîne $(E+F)x = \lambda D x$. on a ainsi montré que λ est valeur propre de B_{J}.
On a montré que $\lambda \neq 0$ est valeur propre de B_{J} si et seulement si $\nu_\omega = \mu_\omega^2$, où μ_ω est solution de l'équation (1.126). On en déduit que

$$\rho(B_\omega) = \max_{\lambda \text{ valeur propre de } B_{\mathrm{J}}} \{|\mu_\omega|; \mu_\omega^2 - \lambda\omega\mu_\omega + \omega - 1 = 0\}$$

est valeur propre de B_ω ν_ω telles que $= (\mu_\omega^+)^2$ et $\nu_\omega^- = (\mu_\omega^-)^2$ sont valeurs propres de la matrice B_ω associés au vecteurs propres $x_(\mu_\omega^+)$ et $x_{\mu_\omega^-}$. En déduire que

$$\rho(matsor) = \max_{\lambda \text{ valeur propre de } B_{\mathrm{J}}} \{|\mu_\omega|; \mu_\omega^2 - \lambda\omega\mu_\omega + \omega - 1 = 0\}.$$

Exercice 55 page 92 (Méthode de Jacobi pour des matrices particulières)

1. Soit $x \in \mathbb{R}^n$, supposons que
$$\|x\|_A = \sum_{i=1}^n a_{i,i}|x_i| = 0.$$
Comme $a_{i,i} > 0$, $\forall i = 1, \ldots, n$, on en déduit que $x_i = 0$, $\forall i = 1, \ldots, n$. D'autre part, il est immédiat de voir que $\|x + y\|_A \leq \|x\| + \|y\|_A$ pour tout $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$ et que $\|\lambda x\|_A = |\lambda|\|x\|_A$ pour tout $(x, \lambda) \in \mathbb{R}^n \times \mathbb{R}$. On en déduit que $\|\cdot\|_A$ est une norme sur $\mathbb{R}^n$.
2. Posons $\tilde{A} = \lambda Id + A$ et notons $\tilde{a}_{i,j}$ ses coefficients. Comme $\lambda \in \mathbb{R}_+^*$,et grâce aux hypothèses (1.115)–(1.117), ceux-ci vérifient :
$$\tilde{a}_{i,j} \leq 0, \forall i, j = 1, \ldots, n, \ i \neq j, \tag{1.127}$$
$$\tilde{a}_{i,i} > \sum_{\substack{i=1 \\ j \neq i}}^n a_{i,j}, \forall i = 1, \ldots, n. \tag{1.128}$$
La matrice $\tilde{A}$ est donc à diagonale dominante stricte, et par l'exercice 49 page 89, elle est donc inversible.
3. La méthode de Jacobi pour la résolution du système (1.118) s'écrit :
$$\tilde{D}u^{(k+1)} = (E + F)u^{(k)} + b, \tag{1.129}$$
avec $\tilde{D} = \lambda Id + D$, et $A = D - E - F$ est la décomposition habituelle de A en partie diagonale, triangulaire inférieure et triangulaire supérieure. Comme $a_{i,i} \geq 0$ et $\lambda \in \mathbb{R}_+^*$, on en déduit que $\tilde{D}$ est inversible, et que donc la suite $(u^{(k)})_{k \in \mathbb{N}}$ est bien définie dans $\mathbb{R}$.
4. Par définition de la méthode de Jacobi, on a :
$$u_i^{(k+1)} = \frac{1}{a_{i,i} + \lambda}\Big(\sum_{\substack{j=1,n \\ j \neq i}} a_{i,j} u_j^{(k)} + b_i\Big).$$
On en déduit que
$$u_i^{(k+1)} - u_i^{(k)} = \frac{1}{a_{i,i} + \lambda} \sum_{\substack{j=1,n \\ j \neq i}} a_{i,j}(u_j^{(k)} - u_j^{(k-1)}).$$

et donc

$$\|u^{(k+1)} - u^{(k)}\|_A \leq \sum_{i=1}^{n} \frac{a_{i,i}}{a_{i,i}+\lambda} \sum_{\substack{j=1,n \\ j\neq i}} a_{i,j}(u_j^{(k)} - u_j^{(k-1)}).$$

Or $\frac{a_{i,i}}{a_{i,i}+\lambda} \leq \frac{1}{1+\frac{\lambda}{a_{i,i}}} \leq \frac{1}{1+\alpha}$. On a donc

$$\|u^{(k+1)} - u^{(k)}\|_A \leq \frac{1}{1+\alpha} \sum_{j=1}^{n} (u_j^{(k)} - u_j^{(k-1)}) \sum_{\substack{j=1,n \\ j\neq i}} a_{i,j}.$$

Et par hypothèse, $\sum_{\substack{j=1,n \\ j\neq i}} a_{i,j} = a_{j,j}$. On en déduit que

$$\|u^{(k+1)} - u^{(k)}\|_A \leq \frac{1}{1+\alpha} \|u^{(k)} - u^{(k-1)}\|_A.$$

On en déduit le résultat par une récurrence immédiate.

5. Soient p et $q = p + m \in \mathbb{N}$, avec $m \geq 0$. Par le résultat de la question précédente, on a :

$$\begin{aligned} \|u^{(q)} - u^{(p)}\|_A &\leq \sum_{i=1}^{m} \|u^{(p+i)} - u^{(p-i-1)}\|_A \\ &\leq \|u^{(1)} - u^{(0)}\|_A (\frac{1}{1+\alpha})^p \sum_{i=0}^{m} (\frac{1}{1+\alpha})^i \end{aligned}$$

Or $\alpha > 0$ donc la série de terme général $(\frac{1}{1+\alpha})^i$, et on a :

$$\begin{aligned} \|u^{(q)} - u^{(p)}\|_A &\leq \|u^{(1)} - u^{(0)}\|_A (\frac{1}{1+\alpha})^p \sum_{i=0}^{+\infty} (\frac{1}{1+\alpha})^i \\ &\leq (1+\frac{1}{\alpha}) \|u^{(1)} - u^{(0)}\|_A (\frac{1}{1+\alpha})^p \\ &\to 0 \text{ lorsque } p \to +\infty. \end{aligned}$$

On en déduit que pour tout $\epsilon > 0$, il existe n tel que si $p, q > n$ alors $\|u^{(q)} - u^{(p)}\|_A \leq \epsilon$, ce qui montre que la suite est de Cauchy, et donc qu'elle converge. Soit $\overline{u}$ sa limite. En passant à la limite dans (1.129), on obtient que $\overline{u}$ est solution de (1.118).

Exercice 57 page 93 (Une méthode itérative particulière)

1. Det $(A) = -1$ et donc A est inversible.

2. Det $(\frac{1}{\omega} Id - E) = \frac{1}{\omega} \left(\frac{1}{\omega^2} - 2 \right)$. Or $\omega \in]0, 2[$. Donc la matrice $\frac{1}{\omega} Id - E$ est inversible si $\omega \neq \frac{\sqrt{2}}{2}$.

3. Les valeurs propres de B_ω sont les complexes λ tels qu'il existe $x \in C^3, x \neq 0$, t.q : $B_\omega x = \lambda x$, c'est–à–dire :

$$\left(F + \frac{1-\omega}{\omega} Id \right) x = \lambda \left(\frac{1}{\omega} Id - E \right) x,$$

soit encore $M_{\lambda,\omega} x = 0$, avec $M_{\lambda,\omega} = \omega F + \lambda\omega E + (1 - \omega - \lambda) Id$.

Or

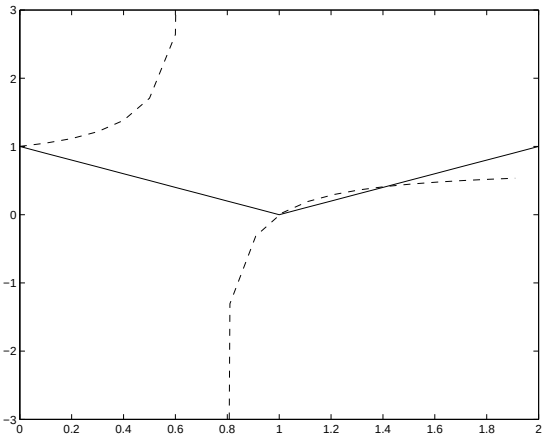

FIGURE 1.8: Graphe des valeurs propres λ_1 et λ_3

$$\begin{aligned} Det\,(M_{\lambda,\omega}) &= (1-\omega-\lambda)((1-\omega-\lambda)^2-2\lambda^2\omega^2) \\ &= (1-\omega-\lambda)(1-\omega-(1+\sqrt{2}\omega)\lambda)(1-\omega-(1-\sqrt{2}\omega)\lambda) \end{aligned}$$

Les valeurs propres de B_ω sont donc réelles, et égales à

$$\lambda_1 = 1-\omega, \lambda_2 = \frac{1-\omega}{1+\sqrt{2}\omega} \text{ et } \lambda_3 = \frac{1-\omega}{1-\sqrt{2}\omega}.$$

Par définition, le rayon spectral $\rho(B_\omega)$ de la matrice B_ω est égal à $\max(|\lambda_1|, |\lambda_2|, |\lambda_3|)$. Remarquons tout d'abord que $|1+\sqrt{2}\omega| > 1, \forall \omega \in]0,2[$, et donc $|\lambda_1| > |\lambda_2|, \forall \omega \in]0,2[$. Il ne reste donc plus qu'à comparer $|\lambda_1|$ et $|\lambda_3|$. Une rapide étude des fonctions $|\lambda_1|$ et $|\lambda_3|$ permet d'établir le graphe représentatif ci-contre.
On a donc :

$$\rho(B_\omega) = |\lambda_3(\omega)| = \left|\frac{1-\omega}{1-\sqrt{2}\omega}\right| \text{ si } \omega \in]0, \sqrt{2}]$$

$$\rho(B_\omega) = \lambda_1(\omega) = |1-\omega| \text{ si } \omega \in [\sqrt{2}, 2[.$$

4. La méthode est convergente si $\rho(B_\omega) < 1$; Si $\omega \in [\sqrt{2}, 2[$, $\rho(B_\omega) = \omega - 1 < 1$; si $\omega \in]0, \sqrt{2}[$,

$$\rho(B_\omega) = \left|\frac{1-\omega}{1-\sqrt{2}\omega}\right| < 1$$

dès que $\dfrac{1-\omega}{\sqrt{2}\omega - 1} < 1$, c'est–à–dire $\omega > \dfrac{2}{1+\sqrt{2}}$.
Le minimum de $\rho(B_\omega)$ est atteint pour $\omega_0 = 1$, on a alors $\rho(B_\omega) = 0$.

Exercice 58 page 94 (Méthode des directions alternées)

1. On a vu en cours qu'une méthode itérative définie par

$$\begin{aligned} &u^{(0)} \in \mathbb{R}^n, \\ &u^{(k+1)} = Bu^{(k)} + c \end{aligned} \tag{1.130}$$

converge si et seulement si $\rho(B) < 1$. Mettons donc l'algorithme (1.120) sous la forme (1.130). On a :

$$(Y+\alpha Id)u^{(k+1)} = -X[(X+\alpha Id)^{-1}(-Yu^{(k)}+b)]$$

soit encore

$$u^{(k+1)} = (Y+\alpha Id)^{-1}X(X+\alpha Id)^{-1}Yu^{(k)} - (Y+\alpha Id)^{-1}X(X+\alpha Id)^{-1}b + (Y+\alpha Id)^{-1}b.$$

On peut donc bien écrire la méthode (1.120) sous la forme (1.130) avec

$$B = (Y+\alpha Id)^{-1}X(X+\alpha Id)^{-1}Y,$$

et la méthode définie par (1.120) converge si et seulement si $\rho(B) < 1$. Il reste à montrer qu'elle converge vers u solution de $Au = b$. Soit $u = \lim_{u\to+\infty} u^{(k)}$. On veut montrer que $Au = b$. Comme $u^{(k)}$ converge et que $u^{(k+1/2)}$ est défini par (1.120), on a aussi que $u^{(k+1/2)}$ converge. Soit $v = \lim_{h\to+\infty} u^{(k+1/2)}$. En passant à la limite dans (1.120), on obtient :

$$\begin{aligned}(X+\alpha Id)v &= -Yu + b,\\ (Y+\alpha Id)u &= -Xv + b.\end{aligned}$$

En additionnant et retranchant ces deux équations, on obtient :

$$Xv + Yu + \alpha Id(u+v) = -Yu - Xv + 2b, \tag{1.131a}$$
$$Xv - Yu + \alpha Id(v-u) = -Yu + Xv. \tag{1.131b}$$

L'équation (1.131b) entraîne $\alpha Id(v-u) = 0$, c'est–à–dire $v = u$ car $\alpha \neq 0$, et en reportant dans (1.131a), on obtient :

$$(X+Y)u + 2\alpha u = -(X+Y)u + b,$$

soit encore

$$(X+Y+\alpha Id)u = b, \text{ c'est–à–dire } Au = b.$$

2. On veut montrer que si $X + \dfrac{\alpha}{2}Id$ et $Y + \dfrac{\alpha}{2}Id$ sont définies positives, alors

$$\rho((X+\alpha Id)^{-1}Y(Y+\alpha Id)^{-1}X) < 1.$$

On utilise la méthode proposée par l'énoncé.

a) Grâce à l'exercice 29 sur les valeurs propres d'un produit de matrices, on sait que les valeurs propres de $(Y+\alpha Id)^{-1}X(X+\alpha Id)^{-1}Y$ sont égales aux valeurs propres de $Y(Y+\alpha Id)^{-1}X(X+\alpha Id)^{-1}$. On a donc $\rho((Y+\alpha Id)^{-1}X(X+\alpha Id)^{-1}Y) = \rho(X(X+\alpha Id)^{-1}Y(Y+\alpha Id)^{-1})$.

b) Comme les matrices $X(X+\alpha Id)^{-1}$ et $Y(Y+\alpha Id)^{-1}$ sont symétriques, en posant

$$Z = Y(Y+\alpha Id)^{-1}X(X+\alpha Id)^{-1},$$

on a :

$$\begin{aligned}\rho(Z) &= \|Y(Y+\alpha Id)^{-1}X(X+\alpha Id)^{-1}\|_2\\ &\leq \|Y(Y+\alpha Id)^{-1}\|_2\,\|X(X+\alpha Id)^{-1}\|_2\end{aligned}$$

et donc

$$\rho(Z) \leq \rho(X(X+\alpha Id)^{-1})\rho(Y(Y+\alpha Id)^{-1}).$$

c) Soit λ valeur propre de X, associée au vecteur propre w. On a $Xw = \lambda w$ et $(X+\alpha Id)w = (\lambda+\alpha)w$, soit encore $w = (\lambda+\alpha)(X+\alpha Id^{-1}w$ Donc

$$Xw = \frac{\lambda}{\lambda+\alpha}(X+\alpha Id)w, \text{ soit encore } (X+\alpha Id)^{-1}Xw = \frac{\lambda}{\lambda+\alpha}w.$$

On en déduit que $\mu = \dfrac{\lambda}{\lambda+\alpha}$ est valeur propre de $X(X+\alpha Id)^{-1}$ associé au vecteur propre w. Pour que $\rho(X(X+\alpha Id)^{-1}) < 1$, il faut et il suffit donc que $|\dfrac{\lambda}{\lambda+\alpha}| < 1$ pour toute valeur propre de λ. Comme $\alpha > 0$, si $\lambda \geq 0$, $|\dfrac{\lambda}{\lambda+\alpha}| = \dfrac{\lambda}{\lambda+\alpha} < 1$. Si $\lambda < 0$, il faut distinguer le cas $\lambda \leq -\alpha$, auquel cas $|\dfrac{\lambda}{\lambda+\alpha}| = \dfrac{\lambda}{\lambda+\alpha} < 1$ du cas $\lambda \in]-\alpha, 0[$. Remarquons qu'on ne peut pas avoir $\lambda = -\alpha$ car la matrice $X+\alpha Id$ est supposée définie positive. Donc on a dans ce dernier cas :

$$|\frac{\lambda}{\lambda+\alpha}| = \frac{-\lambda}{\lambda+\alpha}$$

et la condition $\rho(X(X+\alpha Id)^{-1})$ entraîne

$$-\lambda < \lambda + \alpha$$

c'est–à–dire

$$\lambda > -\frac{\alpha}{2}$$

ce qui est équivalent à dire que la matrice $X + \dfrac{\alpha}{2}Id$ est définie positive.

d) On peut donc conclure que si les matrices $(X + \dfrac{\alpha}{2}Id)$ et $(Y + \dfrac{\alpha}{2}Id)$ sont définies positives, alors $\rho(\beta) < 1$ (grâce à b) et c)) et donc la méthode (1.120) converge.

3. Soit $f \in C([0,1]\times[0,1])$ et soit A la matrice carrée d'ordre $n = M \times M$ obtenue par discrétisation de l'équation $-\Delta u = f$ sur le carré $[0,1]\times[0,1]$ avec conditions aux limites de Dirichlet homogènes $u = 0$ sur $\partial\Omega$, par différences finies avec un pas uniforme $h = \frac{1}{M}$, et $\boldsymbol{b}$ le second membre associé.

(a) On rappelle que l'opérateur Laplacien est défini pour $u \in C^2(\Omega)$, où Ω est un ouvert de $\mathbb{R}^2$, par

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}.$$

La discrétisation de $-\Delta u = f$ est donnée en section 1.2.2 13. Définissons une discrétisation uniforme du carré par les points (x_i, y_j), pour $i = 1, \ldots, M$ et $j = 1, \ldots, M$ avec $x_i = ih$, $y_j = jh$ et $h = 1/(M+1)$.

En reprenant la technique exposée page 11 dans le cas 1D pour l'approximation des dérivées secondes, et en utilisant l'ordre "*lexicographique*" pour numéroter les inconnues, on obtient un système linéaire $A\boldsymbol{u} = \boldsymbol{b}$. Pour fixer les idées, nous prenons ici $M = 3$, et donc $h = \frac{1}{4}$, comme décrit sur la figure ci-contre. Dans ce cas, on a 9 inconnues, et le système s'écrit $A\boldsymbol{u} = \boldsymbol{b}$ où A est une matrice 9×9 et $\boldsymbol{b} = (b_1, \ldots, b_9) \in \mathbb{R}^9$, avec

$$A = \frac{1}{h^2}\begin{bmatrix} 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 4 & 0 & 0 & -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 4 & -1 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & -1 & 4 & -1 & 0 & -1 & 0 \\ 0 & 0 & -1 & 0 & -1 & 4 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & 0 & 0 & 4 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 \end{bmatrix} \text{ et } \boldsymbol{b} = \begin{bmatrix} f(h,h) \\ f(2h,h) \\ f(3h,h) \\ f(h,2h) \\ f(2h,2h) \\ f(3h,2h) \\ f(h,3h) \\ f(2h,3h) \\ f(3h,3h) \end{bmatrix}$$

Dans le cas général, on aurait une matrice avec la même structure, mais avec des blocs diagonaux de taille $(n-1)\times(n-1)$.

(b) On fait une itération dans la direction x et une autre dans la direction y, en choisissant les matrices X et Y avec $X + Y = A$ et

$$X = \frac{1}{h^2}\begin{bmatrix} 2 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 2 \end{bmatrix}$$

$$Y = \frac{1}{h^2}\begin{bmatrix} 2 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 2 & 0 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & 0 & 2 & 0 & 0 & -1 & 0 \\ 0 & 0 & -1 & 0 & 0 & 2 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & -1 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 2 \end{bmatrix}$$

Calculons les valeurs propres de la matrice X : c'est une matrice diagonale par blocs identiques, et chaque bloc est la matrice de discrétisation du laplacien unidimensionnel sur un maillage uniforme de $M + 1$ mailles de pas $h = \frac{1}{M+1}$ de l'intervalle $[0, 1]$. Les valeurs propres de chaque bloc ont ʻté calculées à l'exercice 41) :

$$\lambda_k = \frac{2}{h^2}(1 - \cos k\pi h) = \frac{2}{h^2}(1 - \cos\frac{k\pi}{n+1}), k = 1, \ldots, n,$$

Il est facile de voir que les valeurs propres de X sont égales à ces valeurs propres et que la matrice X est donc symétrique définie positive.

La matrice Y est obtenue à partir de la matrice X par des permutations de ligne et colonne. En effet, la matrice Y est aussi la matrice de discrétisation du laplacien en une dimension d'espace (c'est-à-dire de $-u'' = f$), mais dans la direction y, et donc avec une numérotation qui n'est pas la numérotation naturelle en une dimension d'espace. On obtient donc Y à partir de X en écrivant :

$$Y = \prod_{i,j=1,M} C_{ij} X \prod_{i,j=1,M} C_{ij} X,$$

où C_{ij} la matrice de permutation dont les coefficients $(C_{ij})_{k,\ell}$ sont donnés par :

$$(C_{ij})_{k,\ell} = \begin{cases} 1 \text{ si } k = \ell \text{ et } k \notin \{(M-1)j + i, (M-1)i + j\}, \\ 1 \text{ si } k = (M-1)j + i, \text{ et } \ell = (M-1)i + j \\ 0 \text{ dans tous les autres cas.} \end{cases}$$

On en déduit que Y a les mêmes valeurs propres que X, et qu'elle est donc symétrique définie positive. On en conclut que la méthode des directions alternées avec les choix de X et Y données ci dessus et $\alpha > 0$ (en l'occurence ici la direction x puis la direction y converge.

1.6 Valeurs propres et vecteurs propres

Les techniques de recherche des éléments propres, c.à.d. des valeurs et vecteurs propres (voir Définition 1.2 page 7) d'une matrice sont essentielles dans de nombreux domaines d'application, par exemple en dynamique des struc-

tures : la recherche des modes propres d'une structure peut s'avérer importante pour le dimensionnement de structures sous contraintes dynamiques ; elle est essentielle dans la compréhension des phénomènes acoustiques. On peut se demander pourquoi on parle dans ce chapitre, intitulé "systèmes linéaires" du problème de recherche des valeurs propres : il s'agit en effet d'un problème non linéaire, les valeurs propres étant les solutions du polynôme caractéristique, qui est un polynôme de degré n, où n est la dimension de la matrice. Il n'est malheureusement pas possible de calculer numériquement les valeurs propres comme les racines du polynôme caractéristique, car cet algorithme est instable : une petite perturbation sur les coefficients du polynôme peut entraîner une erreur très grande sur les racines, et en général (voir par exemple le chapitre 5 du polycopié d' E. Hairer, en ligne sur le web (voir l'introduction de ce cours). De nombreux algorithmes ont été développés pour le calcul des valeurs propres et vecteurs propres. Ces méthodes sont en fait assez semblables aux méthodes de résolution de systèmes linéaires. Dans le cadre de ce cours, nous nous restreignons à deux méthodes très connues : la méthode de la puissance (et son adaptation de la puissance inverse), et la méthode dite QR.

1.6.1 Méthode de la puissance et de la puissance inverse

L'idé de la méthode de la puissance est très simple. Soit A une matrice à coefficients complexes de taille $n \times n$. On note $\lambda_1, \cdots, \lambda_n$ ses valeurs propres et on suppose qu'elles satisfont

$$|\lambda_n| > |\lambda_i|,\ i = 2, \cdots, n.$$

On appellera alors λ_n la valeur propre dominante de A.
Soit $\boldsymbol{x} \in \mathbb{C}^n$ tel que $\boldsymbol{x} \neq 0$ au hasard, et considérons la suite de vecteurs unitaires

$$\boldsymbol{x}^{(0)} = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|},\ \ \boldsymbol{x}^{(1)} = \frac{A\boldsymbol{x}^{(0)}}{\|A\boldsymbol{x}^{(0)}\|}, \cdots, \boldsymbol{x}^{(k+1)} = \frac{A\boldsymbol{x}^{(k)}}{\|A\boldsymbol{x}^{(k)}\|}$$

Prenons par exemple la matrice $A = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$ dont les valeurs propres sont 1 et 3, et les vecteurs propres associés $\boldsymbol{f}^{(1)} = \frac{\sqrt{2}}{2}\begin{bmatrix} -1 \\ -1 \end{bmatrix}$ et $\boldsymbol{f}^{(2)} = \frac{\sqrt{2}}{2}\begin{bmatrix} -1 \\ 1 \end{bmatrix}$. Partons de $\boldsymbol{x} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ et faisons tourner scilab :

```
-->x =  A * x
 x  =

    2.
  - 1.

-->x = x/norm(x)
 x  =

    0.8944272
  - 0.4472136

-->x =  A * x
 x  =

    2.236068
  - 1.7888544

-->x = x/norm(x)
 x  =

    0.7808688
  - 0.6246950

-->x =  A * x
 x  =

    2.1864327
  - 2.0302589

-->x = x/norm(x)
 x  =

    0.7327935
  - 0.6804511

-->x =  A * x
 x  =

    2.1460381
  - 2.0936957

-->x = x/norm(x)
 x  =

    0.7157819
  - 0.6983239

-->x =  A * x
 x  =

    2.1298877
  - 2.1124297

-->x = x/norm(x)
 x  =

    0.7100107
  - 0.7041909

-->x =  A * x
 x  =

    2.1242123
  - 2.1183925

-->x = x/norm(x)
 x  =

    0.7080761
  - 0.7061361

-->x =  A * x
 x  =

    2.1222883
  - 2.1203484

-->x = x/norm(x)
 x  =

    0.7074300
  - 0.7067834

-->x =  A * x
 x  =

    2.1216434
  - 2.1209968

-->x = x/norm(x)
 x  =

    0.7072145
  - 0.706999

-->x =  A * x
 x  =

    2.1214281
  - 2.1212125

-->x = x/norm(x)
 x  =

    0.7071427
  - 0.7070709
```

On voit clairement sur cet exemple que la suite $\boldsymbol{x}^{(k)}$ converge vers $\lambda_2 \boldsymbol{f}_2$ lorsque $k \to +\infty$.

Théorème 1.60 (Convergence de la méthode de la puissance)**.** *Soit A une matrice de $\mathcal{M}_n(\mathbb{C})$. On note $\lambda_1, \cdots, \lambda_n$ les valeurs propres de A, $(\boldsymbol{f}_1, \cdots, \boldsymbol{f}_n)$ une base orthonormée de trigonalisation de A telle que $A\boldsymbol{f}_n = \lambda_n \boldsymbol{f}_n$. On suppose que la valeur propre λ_n est dominante, c.à.d. que*

$$|\lambda_n| > |\lambda_{n-1}| \geq \cdots \geq |\lambda_1|,$$

et on suppose de plus que $\lambda_n \in \mathbb{R}$. Alors si $\boldsymbol{x} \notin Vect(\boldsymbol{f}_1, \cdots, \boldsymbol{f}_{n-1})$, la suite de vecteurs $\boldsymbol{x}_{2k}$ converge vers un vecteur unitaire qui est vecteur propre de A pour la valeur propre dominante λ_n.
De plus, la suite $(Ax_k, x_k)_{n\in\mathbb{N}}$ converge vers λ_n lorsque $k \to +\infty$.

Démonstration. La démonstration de ce résultat fait l'objet de l'exercice 59 dans le cas plus simple où A est une matrice symétrique, et donc diagonalisable dans $\mathbb{R}$. □

Remarque 1.61 (Sur la vitesse de convergence de l'algorithme de la puissance)**.** *Supposons pour simplifier que $\lambda_n > 0$, on voit que*

$$\begin{aligned} \|x_k - \tilde{\boldsymbol{f}}_1\|^2 &= 2 - \frac{2}{\|\tilde{\boldsymbol{f}}_1 + B_k\|}(\tilde{\boldsymbol{f}}_1 + B_k, \tilde{\boldsymbol{f}}_1) \\ &\leq \underbrace{2(1 - \|\tilde{\boldsymbol{f}}_1 + B_k\|)}_{\leq \|B_k\|} + \frac{2}{\|\tilde{\boldsymbol{f}}_1 + B_k\|}\|B_k\|^2 \leq C\|B_k\| \end{aligned}$$

avec $\|B_k\| \sim C\left(\frac{|\lambda_2|}{|\lambda_1|}\right)^k$. La méthode de la puissance converge d'autant plus vite que l'écart entre les deux plus grandes valeurs propres est grand.

La méthode de la puissance souffre de plusieurs inconvénients :

1. Elle ne permet de calculer que la plus grande valeur propre. Or très souvent, on veut pouvoir calculer la plus petite valeur propre.
2. De plus, elle ne peut converger que si cette valeur propre est simple.
3. Enfin, même dans le cas où elle est simple, si le rapport des des deux plus grandes valeurs propres est proche de 1, la méthode va converger trop lentement.

Mais de manière assez miraculeuse, il existe un remède à chacun de ces maux :

1. Pour calculer plusieurs valeurs propres simulatanément, on procède par blocs : on part de p vecteurs orthogonaux $x_1^0, \ldots x_p^0$ (au lieu d'un seul). Une itération de la méthode consiste alors à multiplier les p vecteurs par A et à les orthogonaliser par Gram-Schmidt. En répétant cette itération, on approche, si tout se passe bien, p valeurs propres et vecteurs propres de A, et la vitesse de convergence de la méthode est maintenant $\frac{\lambda_{n-p}}{\lambda_n}$.
2. Si l'on veut calculer la plus petite valeur propre, on applique la méthode de la puissance à A^{-1}. On a alors convergence (toujours si tout se passe bien) de $\frac{\|x_{k+1}\|}{\|x_k\|}$ vers $1/|\lambda_1|$. Bien sûr, la mise en oeuvre effective ne s'effectue pas avec l'inverse de A, mais en effectuant une décomposition LU de A qui permet ensuite la résolution du système linéaire $Ax_{k+1} = Ax_k$.
3. Enfin, pour accélérer la convergence de la méthode, on utilise une translation sur A, qui permet de se rapprocher de la valeur propre que l'on veut effectivement calculer. Voir à ce propos l'exercice 60.

1.6.2 Méthode QR

Toute matrice A peut se décomposer sous la forme $A = QR$, où Q est une matrice orthogonale et R une matrice triangulaire supérieure. Dans le cas où A est inversible, cette décomposition est unique. On a donc le théorème suivant :

Théorème 1.62 (Décomposition QR d'une matrice)**.** *Soit $A \in \mathcal{M}_n(\mathbb{R})$. Alors il existe Q matrice orthogonale et R matrice triangulaire supérieure à coefficients diagonaux positifs ou nuls tels que $A = QR$. Si la matrice A est inversible, alors cette décomposition est unique.*

La démonstration est effectuée dans le cas inversible la question 1 de l'exercice 63. La décomposition QR d'une matrice A inversible s'obtient de manière très simple par la méthode de Gram-Schmidt, qui permet de construire une base orthonormée $q_1, \ldots, q_n$ (les colonnes de la matrice Q), à partir de n vecteurs vecteurs indépendants $a_1, \ldots, a_n$ (les colonnes de la matrice A). On se reportera à l'exercice 61 pour un éventuel raffraichissement de mémoire sur Gram-Schmidt... Dans le cas où A n'est pas inversible (et même non carrée), la décomposition existe mais n'est pas unique. la démonstration dans le cadre général se trouve dans le livre de Ph. Ciarlet conseillé en début de ce cours.
L'algorithme QR pour la recherche des valeurs propres d'une matrice est extêmement simple : Si A est une matrice inversible, on pose $A_0 = A$, on effectue la décomposition QR de A : $A = A_0 = Q_0R_0$ et on calcule $A_1 = R_0Q_0$. Comme le produit de matrices n'est pas commutatif, les matrices A_0 et A_1 ne sont pas égales, mais en revanche elles sont semblables ; en effet, graâce à l'associativité du produit matriciel, on a :

$$A_1 = R_0Q_0 = (Q_0^{-1}Q_0)R_0Q_0 = Q_0^{-1}(Q_0R_0)Q_0 = Q_0^{-1}AQ_0.$$

Les matrices A_0 et A_1 ont donc même valeurs propres.
On recommence alors l'opération : à l'itération n, on.effectue la décomposition QR de A_n : $A_n = Q_nR_n$ et on calcule $A_{n+1} = R_nQ_n$.
Par miracle, pour la plupart des matrices, les coefficients diagonaux de la matrice R_n tendent vers les valeurs propres de la matrice A, et les colonnes de la matrice Q_n vers les vecteurs propres associés. Notons que la convergence de l'algorithme QR est un problème ouvert. On retiendra que l'algorithme converge pour une large classe de matrices, et on pourra trouver dans les livres de Serre ou Hubbard-Hubert la démonstration sous une hypothèse assez technique et difficile à vérifier en pratique ; l'exercice 63 donne la démonstration (avec la même hypothèse technique) pour le cas plus simple d'une matrice symétrique définie positive.
Pour améliorer la convergence de l'algorithme QR, on utilise souvent la technique dite de "shift" (translation en français). A l'itération n, au lieu d'effectuer la décomposition QR de la matrice A_n, on travaille sur la matrice $A_n - bI$, où b est choisi proche de la plus grande valeur propre. En général on choisit le coefficient $b = a_{nn}^{(k)}$.
L'exercice 62 donne un exemple de l'application de la méthode QR avec shift.

1.6.3 Exercices

Exercice 59 (Méthode de la puissance)**.** *Suggestions en page 117, corrigé en page 117*

1. Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique. Soit $\lambda_n \in \mathbb{R}$ valeur propre de A t.q. $|\lambda_n| = \rho(A)$ et soit $x^{(0)} \in \mathbb{R}^n$. On suppose que $-\lambda_n$ n'est pas une valeur propre de A et que $x^{(0)}$ n'est pas orthogonal à $\mathrm{Ker}(A - \lambda_n Id)$, ce qui revient à dire que . lorsqu'on écrit le vecteur propre $x^{(0)}$ dans la base des vecteurs propres, la composante sur le vecteur propre associé à λ_n est non nulle. On définit la suite $(x^{(k)})_{n\in\mathbb{N}}$ par $x^{(k+1)} = Ax^{(k)}$ pour $n \in \mathbb{N}$. Montrer que

   (a) $\frac{x^{(k)}}{(\lambda_n)^n} \to x$, quand $k \to \infty$, avec $x \neq 0$ et $Ax = \lambda_n x$.

   (b) $\frac{\|x^{(k+1)}\|}{\|x^{(k)}\|} \to \rho(A)$ quand $n \to \infty$.

   Cette méthode de calcul de la plus grande valeur propre s'appelle "méthode de la puissance".

2. Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible et $b \in \mathbb{R}^n$. Pour calculer x t.q. $Ax = b$, on considère un méthode itérative : on se donne un choix initial $x^{(0)}$, et on construit la suite $x^{(k)}$ telle que $x^{(k+1)} = Bx^{(k)} + c$ avec $c = (Id - B)A^{-1}b$, et on suppose B symétrique. On rappelle que si $\rho(B) < 1$, la suite tend vers x. Montrer que, sauf cas particuliers à préciser,

   (a) $\frac{\|x^{(k+1)}-x\|}{\|x^{(k)}-x\|} \to \rho(B)$ quand $k \to \infty$ (ceci donne une estimation de la vitesse de convergence de la méthode itérative).

   (b) $\frac{\|x^{(k+1)}-x^{(k)}\|}{\|x^{(k}-x^{(k-1)}\|} \to \rho(B)$ quand $k \to \infty$ (ceci permet d'estimer $\rho(B)$ au cours des itérations).

Exercice 60 (Méthode de la puissance inverse avec shift)**.** *Suggestions en page 117.*

Soient $A \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique et $\lambda_1, \ldots, \lambda_p$ ($p \leq n$) les valeurs propres de A. Soit $i \in \{1, \ldots, p\}$, on cherche à calculer λ_i. Soit $x^{(0)} \in \mathbb{R}^n$. On suppose que $x^{(0)}$ n'est pas orthogonal à $\mathrm{Ker}(A - \lambda_i Id)$. On suppose également connaître $\mu \in \mathbb{R}$ t.q. $0 < |\mu - \lambda_i| < |\mu - \lambda_j|$ pour tout $j \neq i$. On définit la suite $(x^{(k)})_{n\in\mathbb{N}}$ par $(A - \mu Id)x^{(k+1)} = x^{(k)}$ pour $n \in \mathbb{N}$.

1. Vérifier que la construction de la suite revient à appliquer la méthode de la puissance à la matrice $A - \mu Id)^{-1}$.
2. Montrer que $x^{(k)}(\lambda_i - \mu)^k \to x$, quand $k \to \infty$, où x est un vecteur propre associé à la valeur propre λ_i, c.à.d. $x \neq 0$ et $Ax = \lambda_i x$.
3. Montrer que $\frac{\|x^{(k+1)}\|}{\|x^{(k)}\|} \to \frac{1}{|\mu-\lambda_i|}$ quand $k \to \infty$.

Exercice 61 (Orthogonalisation de Gram-Schmidt)**.** *Corrigé en page 118*

Soient u et v deux vecteurs de $\mathbb{R}^n$. On rappelle que la projection orthogonale $\mathrm{proj}_u(v)$ du vecteur v sur la droite vectorielle engendrée par u peut s'écrire de la manière suivante :

$$\mathrm{proj}_u(v) = \frac{v \cdot u}{u \cdot u} u,$$

où $u \cdot v$ désigne le produit scalaire des vecteurs u et v. On note $\|\cdot\|$ la norme euclidienne sur $\mathbb{R}^n$.

1. Soient $(a_1, \ldots, a_n)$ une base de $\mathbb{R}^n$. On rappelle qu'à partir de cette base, on peut obtenir une base orthogonale $(v_1, \ldots, v_n)$ et une base orthonormale $(q_1, \ldots, q_n)$ par le procédé de Gram-Schmidt qui s'écrit :

$$\begin{aligned}
v_1 &= a_1, & q_1 &= \frac{a_1}{\|a_1\|} \\
v_2 &= a_2 - \mathrm{proj}_{v_1}(a_2), & q_2 &= \frac{v_2}{\|v_2\|} \\
v_3 &= a_3 - \mathrm{proj}_{v_1}(a_3) - \mathrm{proj}_{v_2}(a_3), & q_3 &= \frac{v_3}{\|v_3\|} \\
v_4 &= a_4 - \mathrm{proj}_{v_1}(a_4) - \mathrm{proj}_{v_2}(a_4) - \mathrm{proj}_{v_3}(a_4), & q_4 &= \frac{v_4}{\|v_4\|} \\
&\vdots & &\vdots \\
v_k &= a_k - \sum_{j=1}^{k-1} \mathrm{proj}_{v_j}(a_k), & q_k &= \frac{v_k}{\|v_k\|}
\end{aligned}$$

On a donc

$$v_k = a_k - \sum_{j=1}^{k-1} \frac{a_k \cdot v_j}{v_j \cdot v_j} v_j, \qquad q_k = \frac{v_k}{\|v_k\|}. \tag{1.132}$$

1. Montrer par récurrence que la famille $(v_1, \ldots, v_n)$ est une base orthogonale de $\mathbb{R}^n$.

2. Soient A la matrice carrée d'ordre n dont les colonnes sont les vecteurs a_j et Q la matrice carrée d'ordre N dont les colonnes sont les vecteurs q_j définis par le procédé de Gram-Schmidt (1.132), ce qu'on note :

$$A = \begin{bmatrix} a_1 & a_2 & \ldots & a_n \end{bmatrix}, \qquad Q = \begin{bmatrix} q_1 & q_2 & \ldots & q_n \end{bmatrix}.$$

Montrer que

$$a_k = \|v_k\| q_k + \sum_{j=1}^{k-1} \frac{a_k \cdot v_j}{\|v_j\|} q_j.$$

En déduire que $A = QR$, où R est une matrice triangulaire supérieure dont les coefficients diagonaux sont positifs.

3. Montrer que pour toute matrice $A \in \mathcal{M}_n(\mathbb{R})$ inversible, on peut construire une matrice orthogonale Q (c.à. d. telle que $QQ^t = Id$) et une matrice triangulaire supérieure R à coefficients diagonaux positifs telles que $A = QR$.

4. Donner la décomposition QR de $A = \begin{bmatrix} 1 & 4 \\ 1 & 0 \end{bmatrix}$.

5. On considère maintenant l'algorithme suivant (où l'on stocke la matrice Q orthogonale cherchée dans la matrice A de départ (qui est donc écrasée)

Algorithme 1.63 (Gram-Schmidt modifié)**.**

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

Calcul de la norme de a_k

$r_{kk} := (\sum_{i=1}^n a_{ik}^2)^{\frac{1}{2}}$

Normalisation

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

$a_{\ell k} := a_{\ell k}/r_{kk}$

Fin pour ℓ

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

Produit scalaire correspondant à $q_k \cdot a_j$

$r_{kj} := \sum_{i=1}^n a_{ik} a_{ij}$

On soustrait la projection de a_k *sur* q_j *sur tous les vecteurs de* A *après* k.

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

$a_{ij} := a_{ij} - a_{ik} r_{kj}$

Fin pour i

Fin pour j

Montrer que la matrice A résultant de cet algorithme est identique à la matrice Q donnée par la méthode de Gram-Schmidt, et que la matrice R est celle de Gram-Schmidt. (Cet algorithme est celui qui est effectivement implanté, car il est plus stable que le calcul par le procédé de Gram-Schmidt original.)

Exercice 62 (Méthode QR avec shift)**.** *Soit* $A = \begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & 0 \end{bmatrix}$

1. *Calculer les valeurs propres de la matrice* A.
2. *Effectuer la décomposition* QR *de la matrice* A.
3. *Calculer* $A_1 = RQ$ *et* $\tilde{A}_1 = RQ - bId$ *où* b *est le terme* a_{22}^1 *de la matrice* A_1
4. *Effectuer la décomposition* QR *de* A_1 *et* $\tilde{A}_1$*, et calculer les matrices* $A_2 = R_1 Q_1$ *et* $\tilde{A}_2 = \tilde{R}_1 \tilde{Q}_1$.

Exercice 63 (Méthode QR pour la recherche de valeurs propres)**.** *Corrigé en page 119*

Soit A une matrice inversible. Pour trouver les valeurs propres de A, on propose la méthode suivante, dite "méthode QR" : On pose $A_1 = A$ et on construit une matrice orthogonale Q_1 et une matrice triangulaire supérieure R_1 telles que $A_1 = Q_1 R_1$ (par exemple par l'algorithme de Gram-Schmidt). On pose alors $A_2 = R_1 Q_1$, qui est aussi une matrice inversible. On construit ensuite une matrice orthogonale Q_2 et une matrice triangulaire supérieure R_2 telles que $A_2 = Q_2 R_2$ et on pose $A_3 = R_3 Q_3$. On continue et on construit une suite de matrices A_k telles que :

$$A_1 = A = Q_1 R_1,\ R_1 Q_1 = A_2 = Q_2 R_2,\ \ldots,\ R_k Q_k = A_k = Q_{k+1} R_{k+1}. \tag{1.133}$$

Dans de nombreux cas, cette construction permet d'obenir les valeurs propres de la matrice A sur la diagonale des matrices A_k. Nous allons démontrer que ceci est vrai pour le cas particulier des matrices symétriques définies positives dont les valeurs propres sont simples (on peut le montrer pour une classe plus large de matrices).

On suppose à partir de maintenant que A est une matrice symétrique définie positive qui admet N valeurs propres (strictement positives) vérifiant $\lambda_1 < \lambda_2 < \ldots < \lambda_n$. On a donc :

$$A = P\Lambda P^t, \text{ avec } \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda n), \text{ et } P \text{ est une matrice orthogonale.} \tag{1.134}$$

(La notation $\mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ désigne la matrice diagonale dont les termes diagonaux sont λ_1,... ,λ_n).

On suppose de plus que

$$P^t \text{ admet une décomposition } LU \text{ et que les coefficients diagonaux de } U \text{ sont strictement positifs.} \tag{1.135}$$

On va montrer que A_k tend vers $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$.

2. Soient Q_i et R_i les matrices orthogonales et triangulaires supérieures définies par (1.133).
2.1 Montrer que $A^2 = \tilde{Q}_2\tilde{R}_2$ avec $\tilde{Q}_k = Q_1Q_2$ et $\tilde{R}_k = R_2R_1$.

2.2 Montrer, par récurrence sur k, que

$$A^k = \tilde{Q}_k\tilde{R}_k, \tag{1.136}$$

avec

$$\tilde{Q}_k = Q_1Q_2 \ldots Q_{k-1}Q_k \text{ et } \tilde{R}_k = R_kR_{k-1} \ldots R_2R_1. \tag{1.137}$$

2.3 Justifier brièvement le fait que $\tilde{Q}_k$ est une matrice orthogonale et $\tilde{R}_k$ est une matrice triangulaire à coefficients diagonaux positifs.

3. Soit $M_k = \Lambda^k L \Lambda^{-k}$.
3.1 Montrer que $PM_k = \tilde{Q}_kT_k$ où $T_k = \tilde{R}_kU^{-1}\Lambda^{-k}$ est une matrice triangulaire supérieure dont les coefficients diagonaux sont positifs.

3.2 Calculer les coefficients de M_k en fonction de ceux de L et des valeurs propres de A.

3.3 En déduire que M_k tend vers la matrice identité et que $\tilde{Q}_kT_k$ tend vers P lorsque $k \to +\infty$.

4. Soient $(B_k)_{k\in\mathbb{N}}$ et $(C_k)_{k\in\mathbb{N}}$ deux suites de matrices telles que les matrices B_k sont orthogonales et les matrices C_k triangulaires supérieures et de coefficients diagonaux positifs. On va montrer que si B_kC_k tend vers la matrice orthogonale B lorsque k tend vers l'infini alors B_k tend vers B et C_k tend vers l'identité lorsque k tend vers l'infini.
On suppose donc que B_kC_k tend vers la matrice orthogonale B. On note $b_1, b_2, \ldots, b_n$ les colonnes de la matrice B et $b_1^{(k)}, b_2^{(k)}, \ldots, b_n^{(k)}$ les colonnes de la matrice B_k, ou encore :

$$B = \begin{bmatrix} b_1 & b_2 & \ldots & b_n \end{bmatrix}, \qquad B_k = \begin{bmatrix} b_1^{(k)} & b_2^{(k)} & \ldots & b_n^{(k)} \end{bmatrix}.$$

et on note $c_{i,j}^{(k)}$ les coefficients de C_k.

4.1 Montrer que la première colonne de B_kC_k est égale à $c_{1,1}^{(k)}b_1^{(k)}$. En déduire que $c_{1,1}^{(k)} \to 1$ et que $b_1^{(k)} \to b_1$.

4.2 Montrer que la seconde colonne de B_kC_k est égale à $c_{1,2}^{(k)}b_1^{(k)} + c_{2,2}^{(k)}b_2^{(k)}$. En déduire que $c_{1,2}^{(k)} \to 0$, puis que $c_{2,2}^{(k)} \to 1$ et que $b_2^{(k)} \to b_2$.

4.3 Montrer que lorsque $k \to +\infty$, on a $c_{i,j}^{(k)} \to 0$ si $i \neq j$, puis que $c_{i,i}^{(k)} \to 1$ et $b_i^{(k)} \to b_i$.

4.4 En déduire que B_k tend B et C_k tend vers l'identité lorsque k tend vers l'infini.

5. Déduire des questions 3 et 4 que $\tilde{Q}_k$ tend vers P et T_k tend vers Id lorsque $k \to +\infty$.

6. Montrer que $\tilde{R}_k(\tilde{R}_{k-1})^{-1} = T_k\Lambda T_{k-1}$. En déduire que R_k et A_k tendent vers Λ.

1.6.4 Suggestions

Exercice 59 page 113 (Méthode de la puissance pour calculer le rayon spectral de A.)

1. Décomposer x_0 sur une base de vecteurs propres orthonormée de A, et utiliser le fait que $-\lambda_n$ n'est pas valeur propre.
2. a/ Raisonner avec $y^{(k)} = x^{(k)} - x$ où x est la solution de $Ax = b$ et appliquer la question 1.
b/ Raisonner avec $y^{(k)} = x^{(k+1)} - x^{(k)}$.

Exercice 60 page 113 (Méthode de la puissance inverse)

Appliquer l'exercice précédent à la matrice $B = (A - \mu Id)^{-1}$.

1.6.5 Corrigés

Exercice 59 page 113 (Méthode de la puissance pour calculer le rayon spectral de A)

1. Comme A est une matrice symétrique, A est diagonalisable dans $\mathbb{R}$. Soit $(f_1, \ldots, f_n) \in (\mathbb{R}^n)^n$ une base orthonormée de vecteurs propres de A associée aux valeurs propres $(\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^n$. On décompose $x^{(0)}$ sur $(f_i)_{i=1,\ldots,n}$: $x^{(0)} = \sum_{i=1}^n \alpha_i f_i$. On a donc $Ax^{(0)} = \sum_{i=1}^n \lambda_i \alpha_i f_i$ et $A^n x^{(0)} = \sum_{i=1}^n \lambda_i^n \alpha_i f_i$.
On en déduit :
$$\frac{x^{(n)}}{\lambda_n^n} = \sum_{i=1}^n \left(\frac{\lambda_i}{\lambda_n}\right)^n \alpha_i f_i.$$
Comme $-\lambda_n$ n'est pas valeur propre,
$$\lim_{n\to+\infty} (\frac{\lambda_i}{\lambda_n})^n = 0 \text{ si } \lambda_i \neq \lambda_n. \tag{1.138}$$
Soient $\lambda_1, \ldots, \lambda_p$ les valeurs propres différentes de λ_n, et $\lambda_{p+1}, \ldots, \lambda_n = \lambda_n$. On a donc
$\lim_{n\to+\infty} \frac{x^{(n)}}{\lambda_n^n} = \sum_{i=p+1}^n \alpha_i f_i = x$, avec $Ax = \lambda_n x$.
De plus, $x \neq 0$: en effet, $x^{(0)} \notin (Ker(A - \lambda_n Id))^\perp = Vect\{f_1, \ldots, f_p\}$, et donc il existe $i \in \{p+1, \ldots, n\}$ tel que $\alpha_i \neq 0$.
Pour montrer (b), remarquons que :
$$\|x^{(n+1)}\| = \sum_{i=1}^n \lambda_i^{n+1} \alpha_i \text{ et } \|x^{(n)}\| = \sum_{i=1}^n \lambda_i^n \alpha_i$$
car $(f_1, \ldots, f_n)$ est une base orthonormée. On a donc
$$\frac{\|x^{(n+1)}\|}{\|x^{(n)}\|} = \lambda_n^n \frac{\|\frac{x^{(n+1)}}{\lambda_n^{n+1}}\|}{\|\frac{x^{(n)}}{\lambda_n}\|} \to \lambda_n \frac{\|x\|}{\|x\|} = \lambda_n \text{ lorsque } n \to +\infty.$$

2. a) La méthode I s'écrit à partir de $x^{(0)}$ connu : $x^{(n+1)} = Bx^{(n)} + c$ pour $n \geq 1$, avec $c = (I - B)A^{-1}b$.
On a donc
$$\begin{aligned} x^{(n+1)} - x &= Bx^{(n)} + (Id - B)x - x \\ &= B(x^{(n)} - x). \end{aligned} \tag{1.139}$$
Si $y^{(n)} = x^{(n)} - x$, on a donc $y^{(n+1)} = By^{(n)}$, et d'après la question 1a) si $y^{(0)} \not\perp \mathrm{Ker}(B - \mu_n Id)$ où μ_n est la plus grande valeur propre de B, (avec $|\mu_n| = \rho(B) et - \mu_n$ non valeur propre), alors
$$\frac{\|y^{(n+1)}\|}{\|y^{(n)}\|} \longrightarrow \rho(B) \text{ lorsque } n \to +\infty,$$

c'est–à–dire

$$\frac{\|x^{(n+1)} - x\|}{\|x^{(n)} - x\|} \longrightarrow \rho(B) \text{ lorsque } n \to +\infty.$$

b) On applique maintenant 1a) à $y^{(n)} = x^{(n+1)} - x^{(n)}$ avec

$$y^{(0)} = x^{(1)} - x^{(0)} \text{ où } x^{(1)} = Ax^{(0)}.$$

On demande que $x^{(1)} - x^{(0)} \notin \mathrm{Ker}(B - \mu_n Id)^\perp$ comme en a), et on a bien $y^{(n+1)} = By^{(n)}$, donc $\frac{\|y^{(n+1)}\|}{\|y^{(n)}\|} \longrightarrow \rho(B)$ lorsque $n \to +\infty$.

Exercice 61 page 114 (Orthogonalisation par Gram-Schmidt)

1. Par définition de la projection orthogonale, on a $v_1 \cdot v_2 = a_1 \cdot (a_2 - \mathrm{proj}_{a_1}(a_2)) = 0$.
Supposons la récurrence vraie au rang $N-1$ et montrons que v_n est orthogonal à tous les v_i pour $i = 1, \ldots, N-1$.
Par définition, $v_n = a_n - \sum_{j=1}^{n-1} \frac{a_k \cdot v_j}{v_j \cdot v_j} v_j$, et donc

$$v_n \cdot v_i = a_n \cdot v_i - \sum_{j=1}^{n-1} \frac{a_n \cdot v_j}{v_j \cdot v_j} v_j \cdot v_i = a_n \cdot v_i - \frac{a_n \cdot v_i}{v_i \cdot v_i}$$

par hypothèse de récurrence. On en déduit que $v_n \cdot v_i = 0$ et donc que la famille $(v_1, \ldots v_n)$ est une base orthogonale.

2. De la relation (1.132), on déduit que :

$$a_k = v_k + \sum_{j=1}^{k-1} \frac{w_k \cdot v_j}{v_j \cdot v_j} v_j, \qquad q_k = \frac{v_k}{\|v_k\|},$$

et comme $v_j = \|v_j\| a_j$, on a bien :

$$a_k = \|v_k\| q_k + \sum_{j=1}^{k-1} \frac{a_k \cdot v_j}{\|v_j\|} q_j.$$

La k-ième colonne de A est donc une combinaison linéaire de la k-ème colonne de Q affectée du poids $\|v_k\|$ et des $k-1$ premières affectées des poids $\frac{a_k \cdot v_j}{\|l v_j\|}$. Ceci s'écrit sous forme matricielle $A = QR$ où R est une matrice carrée dont les coefficients sont $R_{k,k} = \|v_k\|$, $R_{j,k} = \frac{a_k \cdot v_j}{\|v_j\|}$ si $j < k$, et $R_{j,k} = 0$ si $j > k$. La matrice R est donc bien triangulaire supérieure et à coefficients diagonaux positifs.

3. Si A est inversible, par le procédé de Gram-Schmidt (1.132) on construit la matrice $Q = \begin{bmatrix} q_1 & q_2 & \ldots & q_n \end{bmatrix}$, et par la question 1.b, on sait construire une matrice R triangulaire supérieure à coefficients diagonaux positifs $A = QR$.

4. On a $a_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ et donc $q_1 = \frac{1}{2} \begin{bmatrix} \sqrt{2} \\ \sqrt{2} \end{bmatrix}$

Puis $a_2 = \begin{bmatrix} 4 \\ 0 \end{bmatrix}$ et donc $v_2 = a_2 - \frac{a_2 \cdot v_1}{v_1 \cdot v_1} v_1 = \begin{bmatrix} 4 \\ 0 \end{bmatrix} - \frac{4}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ -2 \end{bmatrix}$. Donc $\boldsymbol{q}_2 = \frac{1}{2} \begin{bmatrix} \sqrt{2} \\ -\sqrt{2} \end{bmatrix}$, et $Q = \frac{1}{2} \begin{bmatrix} \sqrt{2} & \sqrt{2} \\ \sqrt{2} & -\sqrt{2} \end{bmatrix}$.

Enfin, $R = \begin{bmatrix} \|v_1\| & \frac{a_2 \cdot v_1}{\|v_1\|} \\ 0 & \|v_1\| \end{bmatrix} = \begin{bmatrix} \sqrt{2} & 2\sqrt{2} \\ 0 & 2\sqrt{2} \end{bmatrix}$, et $Q = \frac{1}{2} \begin{bmatrix} \sqrt{2} & \sqrt{2} \\ \sqrt{2} & -\sqrt{2} \end{bmatrix}$.

Exercice 63 page 115 (Méthode QR pour la recherche de valeurs propres)

1.1 Par définition et associativité du produit des matrices,

$$A^2 = (Q_1R_1)(Q_1R_1) = Q_1(R_1Q_1)R_1 = Q_1(R_1Q_1)R_1 = Q_1(Q_2R_2)R_1 = (Q_1Q_2)(R_2R_1) = \tilde{Q}_2\tilde{R}_2$$

avec $\tilde{Q}_2 = Q_1Q_2$ et $\tilde{R}_2 = R_1R_2$.

1.2 La propriété est vraie pour $k = 2$. Supposons la vraie jusqu'au rang $k-1$ et montrons là au rang k. Par définition, $A^k = A^{k-1}A$ et donc par hypothèse de récurrence, $A^k = \tilde{Q}_{k-1}\tilde{R}_{k-1}A$. On en déduit que :

$$\begin{aligned}
A^k &= \tilde{Q}_{k-1}\tilde{R}_{k-1}Q_1R_1 \\
&= Q_1 \dots Q_{k-1}R_{k-1} \dots R_2(R_1Q_1)R_1 \\
&= Q_1 \dots Q_{k-1}R_{k-1} \dots R_2(Q_2R_2)R_1 \\
&= Q_1 \dots Q_{k-1}R_{k-1} \dots (R_2Q_2)R_2R_1 \\
&= Q_1 \dots Q_{k-1}R_{k-1} \dots R_3(Q_3R_3)R_2R_1 \\
&\vdots \qquad\qquad\qquad \vdots \\
&= Q_1 \dots Q_{k-1}R_{k-1} \dots R_j(Q_jR_j)R_{j-1} \dots R_2R_1 \\
&= Q_1 \dots Q_{k-1}R_{k-1} \dots R_{j+1}(R_jQ_j)R_{j-1} \dots R_2R_1 \\
&= Q_1 \dots Q_{k-1}R_{k-1} \dots R_{j+1}(Q_{j+1}R_j)R_{j-1} \dots R_2R_1 \\
&= Q_1 \dots Q_{k-1}R_{k-1}(Q_{k-1}R_{k-1})R_{k-2} \dots R_2R_1 \\
&= Q_1 \dots Q_{k-1}(R_{k-1}Q_{k-1})R_{k-1}R_{k-2} \dots R_2R_1 \\
&= Q_1 \dots Q_{k-1}(Q_kR_k)R_{k-1}R_{k-2} \dots R_2R_1 \\
&= \tilde{Q}_k\tilde{R}_k
\end{aligned}$$

1.3 La matrice $\tilde{Q}_k$ est un produit de matrices orthogonales et elle est donc orthogonale. (On rappelle que si P et Q sont des matrices orthogonales, c.à.d. $P^{-1} = P^t$ et $Q^{-1} = Q^t$, alors $(PQ)^{-1} = Q^{-1}P^{-1} = Q^tP^t = (PQ)^t$ et donc PQ est orthogonale.)

De même, le produit de deux matrices triangulaires supérieures à coefficients diagonaux positifs est encore une matrice triangulaire supérieure à coefficients diagonaux positifs.

2.1 Par définition, $PM_k = P\Lambda^kL\Lambda^{-k} = P\Lambda^kP^tP^{-t}L\Lambda^{-k} = A^kP^{-t}L\Lambda^{-k}$.
Mais $A^k = \tilde{Q}_k\tilde{R}_k$ et $P^t = LU$, et donc :
$PM_k = \tilde{Q}_k\tilde{R}_kU^{-1}\Lambda^{-k} = \tilde{Q}_kT_k$ où $T_k = \tilde{R}_kU^{-1}\Lambda^{-k}$. La matrice T_k est bien triangulaire supérieure à coefficients diagonaux positifs, car c'est un produit de matrices triangulaires supérieures à coefficients diagonaux positifs.
2.2

$$(M_k)_{i,j} = (\Lambda^kL\Lambda^{-k})_{i,j} = \begin{cases} L_{i,i} & \text{si } i = j, \\ \frac{\lambda_j^k}{\lambda_i^k}Li,j & \text{si } i > j, \\ 0 & \text{sinon.} \end{cases}$$

2.3 On déduit facilement de la question précédente que, lorsque $k \to +\infty$, $(M_k)_{i,j} \to 0$ si $i \neq j$ et $(M_k)_{i,j} \to 1$ et donc que M_k tend vers la matrice identité et que $\tilde{Q}_kT_k$ tend vers P lorsque $k \to +\infty$.

3.1 Par définition, $(B_kC_k)_{i,1} = \sum_{\ell=1,n}(B_k)_{i,\ell}(C_k)_{\ell,1} = (B_k)_{i,1}(C_k)_{1,1}$ car C_k est triangulaire supérieure. Donc la première colonne de B_kC_k est bien égale à $c_{1,1}^{(k)}\boldsymbol{b}_1^{(k)}$.

Comme B_kC_k tend vers B, la première colonne b_1 de B_kC_k tend vers la première colonne de B, c'est -à-dire

$$c_{1,1}^{(k)}b_1^{(k)} \to lb_1 \text{ lorsque } k \to \infty.$$

Comme les matrices B et B_k sont des matrices orthogonales, leurs vecteurs colonnes sont de norme 1, et donc

$$|c_{1,1}^{(k)}| = \|c_{1,1}^{(k)} l b_1^{(k)}\| \to \|b_1\| = 1 \text{ lorsque } k \to \infty.$$

On en déduit que $|c_{1,1}^{(k)}| \to 1$ lorsque $k \to +\infty$, et donc $\lim_{k\to+\infty} c_{1,1}^{(k)} = \pm 1$. Or, par hypothèse, la matrice $C^{(k)}$ a tous ses coefficients diagonaux positifs, on a donc bien $c_{1,1}^{(k)} \to 1$ lorsque $k \to +\infty$. Par conséquent, on a $\boldsymbol{b}_1^{(k)} \to \boldsymbol{b}_1$ lorsque $k \to \infty$.

3.2 Comme C_k est triangulaire supérieure, on a :

$$(B_k C_k)_{i,2} = \sum_{\ell=1,n} (B_k)_{i,\ell}(C_k)_{\ell,2} = (B_k)_{i,1}(C_k)_{1,1} + (B_k)_{i,2}(C_k)_{2,1},$$

et donc la seconde colonne de $B_k C_k$ est bien égale à $c_{1,2}^{(k)} l b_1^{(k)} + c_{2,2}^{(k)} b_2^{(k)}$.
On a donc

$$c_{1,2}^{(k)} b_1^{(k)} + c_{2,2}^{(k)} l b_2^{(k)} \to b_2 \text{ lorsque } k \to +\infty. \tag{1.140}$$

La matrice B_k est orthogonale, et donc $b_1^{(k)} \cdot b_1^{(k)} = 1$ et $b_1^{(k)} \cdot b_2^{(k)} = 0$. De plus, par la question précédente, $b_1^{(k)} \to b_1$ lorsque $k \to +\infty$, On a donc, en prenant le produit scalaire du membre de gauche de (1.140) avec $l b_1^{(k)}$,

$$c_{1,2}^{(k)} = \left(c_{1,2}^{(k)} b_1^{(k)} + c_{2,2}^{(k)} b_2^{(k)} \right) \cdot b_1^{(k)} \to b_2 \cdot b_1 = 0 \text{ lorsque } k \to +\infty.$$

Comme $c_{1,2}^{(k)} \to 0$ et $b_1^{(k)} \to b_1$ on obtient par (1.140) que

$$c_{2,2}^{(k)} b_2^{(k)} \to b_2 \text{ lorsque } k \to +\infty.$$

Le même raisonnement que celui de la question précédente nous donne alors que $c_{2,2}^{(k)} \to 1$ et $b_2^{(k)} \to b_2$ lorsque $k \to +\infty$.

3.3 On sait déjà par les deux questions précédentes que ces assertions sont vraies pour $i = 1$ et 2. Supposons qu'elles sont vérifiées jusqu'au rang $i-1$, et montrons que $c_{i,j}^{(k)} \to 0$ si $i \neq j$, puis que $c_{i,i}^{(k)} \to 1$ et $\boldsymbol{b}_i^{(k)} \to \boldsymbol{b}_i$.
Comme C_k est triangulaire supérieure, on a :

$$(B_k C_k)_{i,j} = \sum_{\ell=1,n} (B_k)_{i,\ell}(C_k)_{\ell,j} = \sum_{\ell=1}^{j-1} (B_k)_{i,\ell}(C_k)_{\ell,j} + (B_k)_{i,j}(C_k)_{j,j},$$

et donc la j-ème colonne de $B_k C_k$ est égale à $\sum_{\ell=1}^{j-1} c_{\ell,1}^{(k)} \boldsymbol{b}_\ell^{(k)} + c_{j,j}^{(k)} \boldsymbol{b}_j^{(k)}$. On a donc

$$\sum_{\ell=1}^{j-1} c_{\ell,1}^{(k)} \boldsymbol{b}_\ell^{(k)} + c_{j,j}^{(k)} \boldsymbol{b}_j^{(k)} \to \boldsymbol{b}_j \text{ lorsque } k \to +\infty. \tag{1.141}$$

La matrice B_k est orthogonale, et donc $\boldsymbol{b}_i^{(k)} \cdot \boldsymbol{b}_i^{(k)} = \delta_{i,j}$. De plus, par hypothèse de récurrence, on sait que $\boldsymbol{b}_\ell^{(k)} \to \boldsymbol{b}_\ell$ pour tout $\ell \leq j-1$. En prenant le produit scalaire du membre de gauche de (1.141) avec $\boldsymbol{b}_m^{(k)}$, pour $m < i$, on obtient

$$c_{m,j}^{(k)} = \left(\sum_{\ell=1}^{j-1} c_{\ell,1}^{(k)} \boldsymbol{b}_\ell^{(k)} + c_{j,j}^{(k)} \boldsymbol{b}_j^{(k)} \right) \cdot \boldsymbol{b}_m^{(k)} \to \boldsymbol{b}_m \cdot \boldsymbol{b}_j = 0 \text{ lorsque } k \to +\infty.$$

On déduit alors de (1.141) que $c_{j,j}^{(k)} \boldsymbol{b}_j^{(k)} \to \boldsymbol{b}_j$ lorsque $k \to +\infty$, et le même raisonnement que celui de la question 4.1 nous donne alors que $c_{j,j}^{(k)} \to 1$ et $\boldsymbol{b}_j^{(k)} \to \boldsymbol{b}_j$ lorsque $k \to +\infty$.

ce qui conclut le raisonnement par récurrence.

3.4 En déduire que B_k tend B et C_k tend vers l'identité lorsque k tend vers l'infini.
On a montré aux trois questions précédentes que la j-ième colonne de B_k tend vers la j-ième colonne de B, et que $c_{i,j}^{(k)} \to \delta_{i,j}$ lorque k tend vers $+\infty$. On a donc bien le résultat demandé.

4. D'après la question 3, $\tilde{Q}_k T_k$ tend vers P, et d'après la question 4, comme $\tilde{Q}_k$ est orthogonale et T_k triangulaire supérieure à coefficients positifs, on a bien $\tilde{Q}_k$ qui tend vers P et T_k qui tend vers Id lorsque $k \to +\infty$.

5. On a $\tilde{R}_k = T_k \Lambda^k U$ et donc $\tilde{R}_k (\tilde{R}_{k-1})^{-1} = T_k \Lambda^k U U^{-1} \Lambda^{-k+1} T_{k-1} = T_k \Lambda T_{k-1}$.
Comme T_k tend vers Id, on a $R_k = \tilde{R}_k (\tilde{R}_{k-1})^{-1}$ qui tend vers Λ. De plus, $A_k = Q_k R_k$, où $Q_k = \tilde{Q}_k (\tilde{Q}_{k-1})^{-1}$ tend vers Id et R_k tend vers Λ. Donc A_k tend vers Λ.

Chapitre 2

Systèmes non linéaires

Dans le premier chapitre, on a étudié quelques méthodes de résolution de systèmes linéaires en dimension finie. L'objectif est maintenant de développer des méthodes de résolution de systèmes non linéaires, toujours en dimension finie. On se donne $g \in C(\mathbb{R}^n, \mathbb{R}^n)$ et on cherche x dans $\mathbb{R}^n$ solution de :

$$\left\{ \begin{array}{l} x \in \mathbb{R}^n \\ g(x) = 0. \end{array} \right. \tag{2.1}$$

Au Chapitre I on a étudié des méthodes de résolution du système (2.1) dans le cas particulier $g(x) = Ax - b$, $A \in \mathcal{M}_n(\mathbb{R})$, $b \in \mathbb{R}^n$. On va maintenant étendre le champ d'étude au cas où g n'est pas forcément affine. On étudiera deux familles de méthodes pour la résolution approchée du système (2.1) :

— les méthodes de point fixe : point fixe de contraction et point fixe de monotonie
— les méthodes de type Newton [1].

2.1 Les méthodes de point fixe

2.1.1 Point fixe de contraction

Soit $g \in C(\mathbb{R}^n, \mathbb{R}^n)$, on définit la fonction $f \in C(\mathbb{R}^n, \mathbb{R}^n)$ par $f(x) = x - g(x)$. On peut alors remarquer que $g(x) = 0$ si et seulement si $f(x) = x$. Résoudre le système non linéaire (2.1) revient donc à trouver un point fixe de f. Encore faut-il qu'un tel point fixe existe... On rappelle le théorème de point fixe bien connu :

Théorème 2.1 (Point fixe). *Soit E un espace métrique complet, d la distance sur E, et $f : E \to E$ une fonction strictement contractante, c'est–à–dire telle qu'il existe $k \in]0, 1[$ tel que $d(f(x), f(y)) \leq kd(x, y)$ pour tout $x, y \in E$. Alors il existe un unique point fixe $\bar{x} \in E$ qui vérifie $f(\bar{x}) = \bar{x}$. De plus si $x^{(0)} \in E$, et $x^{(k+1)} = f(x^{(k)})$, $\forall k \geq 0$, alors $x^{(k)} \to \bar{x}$ quand $n + \infty$.*

DÉMONSTRATION – *Etape 1 : Existence de $\bar{x}$ et convergence de la suite*

Soit $x^{(0)} \in E$ et $(x^{(k)})_{n \in \mathbb{N}}$ la suite définie par $x^{(k+1)} = f(x^{(k)})$ pour $n \geq 0$. On va montrer que :

1. $(x^{(k)})_n$ est de Cauchy (donc convergente car E est complet),
2. $\lim_{n \to +\infty} x^{(k)} = \bar{x}$ est point fixe de f.

1. Isaac Newton (1643 - 1727, né d'une famille de fermiers, est un philosophe, mathématicien, physicien, alchimiste et astronome anglais. Figure emblématique des sciences, il est surtout reconnu pour sa théorie de la gravitation universelle et la création, en concurrence avec Leibniz, du calcul infinitésimal.

Par hypothèse, on sait que pour tout $n \geq 1$,

$$d(x^{(k+1)}, x^{(k)}) = d(f(x^{(k)}), f(x^{(k-1)})) \leq k d(x^{(k)}, x^{(k-1)}).$$

Par récurrence sur n, on obtient que

$$d(x^{(k+1)}, x^{(k)}) \leq k^n d(x^{(1)}, x^{(0)}),\ \forall n \geq 0.$$

Soit $n \geq 0$ et $p \geq 1$, on a donc :

$$\begin{aligned} d(x^{(n+p)}, x^{(k)}) &\leq d(x^{(n+p)}, x^{(n+p-1)}) + \cdots + d(x^{(k+1)}, x^{(k)}) \\ &\leq \sum_{q=1}^{p} d(x^{(n+q)}, x^{(n+q-1)}) \\ &\leq \sum_{q=1}^{p} k^{n+q-1} d(x^{(1)}, x^{(0)}) \\ &\leq d(x^{(1)}, x^{(0)}) k^n (1 + k + \ldots + k^{p-1}) \\ &\leq d(x^{(1)}, x^{(0)}) \frac{k^n}{1-k} \longrightarrow 0 \text{ quand } n \to +\infty \text{ car } k < 1. \end{aligned}$$

La suite $(x^{(k)})_{n \in \mathbb{N}}$ est donc de Cauchy, *i.e.* :

$$\forall \varepsilon > 0,\ \ \exists n_\varepsilon \in \mathbb{N}\,;\ \ \forall n \geq n_\varepsilon,\ \ \forall\ \ p \geq 1\ \ d(x^{(n+p)}, x^{(k)}) \leq \varepsilon.$$

Comme E est complet, on a donc $x^{(k)} \longrightarrow \bar{x}$ dans E quand $n \to +\infty$. Comme la fonction f est strictement contractante, elle est continue, donc on a aussi $f(x^{(k)}) \longrightarrow f(\bar{x})$ dans E quand $n + \infty$. En passant à la limite dans l'égalité $x^{(k+1)} = f(x^{(k)})$, on en déduit que $\bar{x} = f(\bar{x})$.

Etape 2 : Unicité

Soit $\bar{x}$ et $\bar{y}$ des points fixes de f, qui satisfont donc $\bar{x} = f(\bar{x})$ et $\bar{y} = f(\bar{y})$. Alors $d(f(\bar{x}), f(\bar{y})) = d(\bar{x}, \bar{y}) \leq k d(\bar{x}, \bar{y})$; comme $k < 1$, ceci est impossible sauf si $\bar{x} = \bar{y}$.

■

La méthode du point fixe s'appelle aussi méthode des itérations successives. Dans le cadre de ce cours, nous prendrons $E = \mathbb{R}^n$, et la distance associée à la norme euclidienne, que nous noterons $|\cdot|$.

$$\forall (\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^n \times \mathbb{R}^n \text{ avec } \boldsymbol{x} = (x_1, \ldots, x_n), \boldsymbol{y} = (y_1, \ldots, y_n), d(\boldsymbol{x}, \boldsymbol{y}) = |\boldsymbol{x} - \boldsymbol{y}| = \left(\sum_{i=1}^{n} (x_i - y_i)^2 \right)^{\frac{1}{2}}.$$

A titre d'illustration, essayons de la mettre en oeuvre pour trouver les points fixes des fonctions $x \mapsto x^2$ et $x \mapsto \sqrt{x}$.

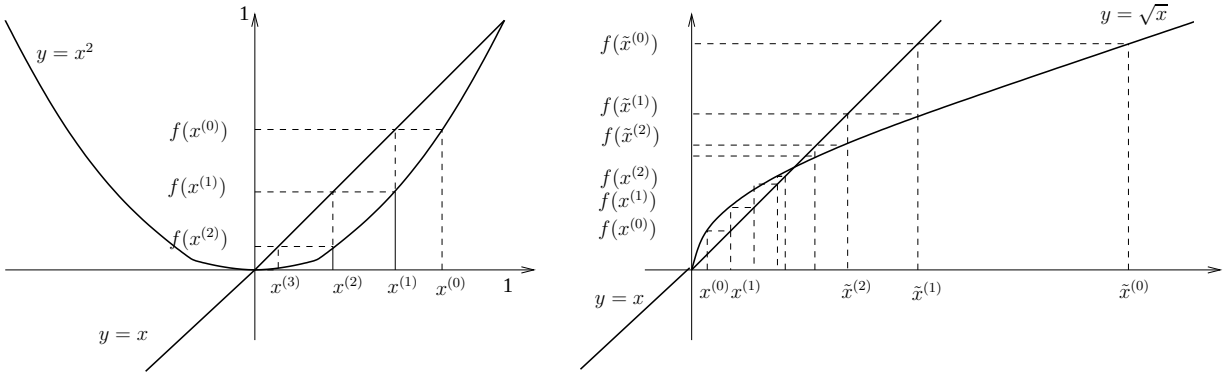

FIGURE 2.1: Comportement des itérés successifs du point fixe – A gauche : $x \mapsto x^2$, à droite : $x \mapsto \sqrt{x}$.

Pour la fonction $x \mapsto x^2$, on voit sur la figure de gauche que si l'on part de $x = x^{(0)} < 1$, la méthode converge rapidement vers 0 ; de fait, cette fonction n'est strictement contractante que sur l'intervalle $]-\frac{1}{2}, \frac{1}{2}[$. Donc si

$x = x^{(0)} \in]-\frac{1}{2}, \frac{1}{2}[$, on est dans les conditions d'application du théorème du point fixe. Mais en fait, la suite $(x^{(k)})_{n\in\mathbb{N}}$ définie par le point fixe converge pour tout $x^{(0)} \in]-1, 1[$; ceci est très facile à voir car $x^{(k)} = x^2 k$ et on a donc convergence vers 0 si $|x| < 1$.
Par contre si l'on partait de $x^{(0)} > 1$ (non représenté sur la figure), on divergerait rapidement : mais rien de surprenant à cela, puisque la fonction $x \mapsto x^2$ n'est pas contractante sur $[1, +\infty[$
Dans le cas de la fonction $x \mapsto \sqrt{x}$, on voit sur le graphique que les itérés convergent vers 1 que l'on parte à droite ou à gauche de $x = 1$; on peut même démontrer (exercice) que si $x^{(0)} > 0$, la suite $(x)_{n\in\mathbb{N}}$ converge vers 1 lorsque $k \to +\infty$. Pourtant la fonction $x \mapsto \sqrt{x}$ n'est contractante que pour $x > \frac{1}{4}$; mais on n'atteint jamais le point fixe 0, ce qui est moral, puisque la fonction n'est pas contractante en 0. On se rend compte encore sur cet exemple que le théorème du point fixe donne une condition suffisante de convergence, mais que cette condition n'est pas nécessaire.

Comportement des itérés successifs du point fixe pour $x \mapsto \frac{1}{x}$

Remarquons que l'hypothèse que f envoie E dans E est cruciale. Par exemple la fonction $f : x \mapsto \frac{1}{x}$ est lipschitzienne de rapport $k < 1$ sur $[1 + \varepsilon, +\infty[$ pour tout $\varepsilon > 0$ mais elle n'envoie pas $[1+\varepsilon, +\infty[$ dans $[1+\varepsilon, +\infty[$. La méthode du point fixe à partir du choix initial $x \neq 1$ donne la suite $x, \frac{1}{x}, x, \frac{1}{x}, \ldots, x, \frac{1}{x}$ qui ne converge pas, voir figure ci contre.

Remarque 2.2.

1. *Sous les hypothèses du théorème 2.1, $d(x^{(k+1)}, \bar{x}) = d(f(x^{(k)}), f(\bar{x})) \leq kd(x^{(k)}, \bar{x})$; donc si $x^{(k)} \neq \bar{x}$ alors $\frac{d(x^{(k+1)}, \bar{x})}{d(x^{(k)}, \bar{x})} \leq k \ (< 1)$, voir à ce sujet la définition 2.11. La convergence est donc au moins linéaire (même si de fait, cette méthode converge en général assez lentement).*
2. *Le théorème 2.1 se généralise en remplaçant l'hypothèse " f strictement contractante" par " il existe $n > 0$ tel que $f^{(k)} = \underbrace{f \circ f \circ \ldots \circ f}_{n \text{ fois}}$ est strictement contractante " (reprendre la démonstration du théorème pour le vérifier).*

La question qui vient alors naturellement est : que faire pour résoudre $g(x) = 0$ si la méthode du point fixe appliquée à la fonction $x \mapsto x - g(x)$ ne converge pas ? Dans ce cas, f n'est pas strictement contractante ; une idée possible est de pondérer la fonction g par un paramètre $\omega \neq 0$ et d'appliquer les itérations de point fixe à la fonction $f_\omega(x) = x - \omega g(x)$; on remarque là encore que x est encore solution du système (2.1) si et seulement si x est point fixe de $f_\omega(x)$. On aimerait dans ce cas trouver ω pour que f_ω soit strictement contractante, c.à.d. pour que $|f_\omega(x) - f_\omega(y)| = |x - y - \omega(g(x) - g(y))| \leq k|x - y|$ pour $(x, y) \in \mathbb{R}^n$, avec $k < 1$. Or

$$\begin{aligned} |x - y - \omega(g(x) - g(y))|^2 &= \big(x - y - \omega(g(x) - g(y))\big) \cdot \big(x - y - \omega(g(x) - g(y))\big) \\ &= |x - y|^2 - 2(x - y) \cdot (\omega(g(x) - g(y))) + \omega^2 |g(x) - g(y)|^2. \end{aligned}$$

Supposons que g soit lipschitzienne, et soit $M > 0$ sa constante de Lipschitz :

$$|g(x) - g(y)| \leq M|x - y|, \ \forall x, y \in \mathbb{R}^n. \tag{2.2}$$

On a donc

$$|x - y - \omega(g(x) - g(y))|^2 \leq (1 + \omega^2 M^2)|x - y|^2 - 2(x - y) \cdot (\omega(g(x) - g(y)))$$

Or on veut $|x-y-\omega(g(x)-g(y))|^2 \leq k|x-y|^2$, avec $k<1$. On a donc intérêt à ce que le terme $-2(x-y)\cdot(\omega(g(x)-g(y)))$ soit de la forme $-a|x-y|^2$ avec a strictement positif. Pour obtenir ceci, on va supposer de plus que :

$$\exists \alpha>0 \text{ tel que } (g(x)-g(y))\cdot(x-y)\geq \alpha|x-y|^2, \forall x,y\in \mathbb{R}^n, \tag{2.3}$$

On obtient alors :

$$|x-y-\omega(g(x)-g(y))|^2 \leq (1+\omega^2M^2-2\omega\alpha)|x-y|^2.$$

Et donc si $\omega \in]0, \frac{2\alpha}{M^2}[$, le polynôme $\omega^2M^2-2\omega\alpha$ est strictement négatif : soit $-\mu$ (noter que $\mu\in]0,1[$) et on obtient que

$$|x-y-\omega(g(x)-g(y))|^2 \leq (1-\mu)|x-y|^2.$$

On peut donc énoncer le théorème suivant :

Théorème 2.3 (Point fixe de contraction avec relaxation)**.** *On désigne par $|\cdot|$ la norme euclidienne sur $\mathbb{R}^n$. Soit $g\in C(\mathbb{R}^n,\mathbb{R}^n)$ lipschitzienne de constante de Lipschitz $M>0$, et telle que* (2.3) *est vérifiée : alors la fonction $f_\omega : x\mapsto x-\omega g(x)$ est strictement contractante si $0<\omega<\dfrac{2\alpha}{M^2}$. Il existe donc un et un seul $\bar{x}\in\mathbb{R}^n$ tel que $g(\bar{x})=0$ et $x^{(k)}\to\bar{x}$ quand $n\to+\infty$ avec $x^{(k+1)}=f_\omega(x^{(k)})=x^{(k)}-\omega g(x^{(k)})$.*

Remarque 2.4. *Le théorème 2.3 permet de montrer que sous les hypothèses* (2.3) *et* (2.2)*, et pour $\omega\in]0,\frac{2\alpha}{M^2}[$, on peut obtenir la solution de* (2.1) *en construisant la suite :*

$$\begin{cases} x^{(k+1)}=x^{(k)}-\omega g(x^{(k)}) \ n\geq 0, \\ x^{(0)}\in\mathbb{R}^n. \end{cases} \tag{2.4}$$

Or on peut aussi écrire cette suite de la manière suivante :

$$\begin{cases} \tilde{x}^{(k+1)} &= f(x^{(k)}), \ \forall n\geq 0 \\ x^{(k+1)} &= \omega\tilde{x}^{(k+1)}+(1-\omega)x^{(k)}, \quad x^{(0)}\in\mathbb{R}^n. \end{cases} \tag{2.5}$$

En effet si $x^{(k+1)}$ est donné par la suite (2.5)*, alors*

$$x^{(k+1)}=\omega\tilde{x}^{(k+1)}+(1-\omega)x^{(k)}=\omega f(x^{(k)})+(1-\omega)x^{(k)}=-\omega g(x^{(k)})+x^{(k)}.$$

Le procédé de construction de la suite (2.5) *est l'algorithme de relaxation sur f.*

Remarque 2.5 (Quelques rappels de calcul différentiel)**.**
Soit $h\in C^2(\mathbb{R}^n,\mathbb{R})$. La fonction h est donc en particulier différentiable, c.à.d. que pour tout $x\in\mathbb{R}^n$, il existe $Dh(x)\in\mathcal{L}(\mathbb{R}^n,\mathbb{R})$ telle que

$$h(x+\eta)=h(x)+Dh(x)(\eta)+|\eta|\varepsilon(\eta), \text{ avec } \varepsilon(\eta)\underset{\eta\to 0}{\to}0\ .$$

On a dans ce cas, par définition du gradient, $Dh(x)(\eta)=\nabla h(x)\cdot\eta$ où $\nabla h(x)=(\partial_1 h(x),\cdots,\partial_n h(x))^t\in\mathbb{R}^n$ est le gradient de h au point x (on désigne par $\partial_i h$ la dérivée partielle de f par rapport à sa i-ème variable). Comme on suppose $h\in C^2(\mathbb{R}^n,\mathbb{R})$, on a donc $g=\nabla h\in C^1(\mathbb{R}^n,\mathbb{R}^n)$, et g est continûment différentiable, c'est–à–dire

$$Dg(x)\in\mathcal{L}(\mathbb{R}^n,\mathbb{R}^n), \text{ et } g(x+\eta)=g(x)+Dg(x)(\eta)+|\eta|\varepsilon(\eta),$$

avec $\varepsilon(\eta)\underset{\eta\to 0}{\to}0$.
Comme $Dg(x)\in\mathcal{L}(\mathbb{R}^n,\mathbb{R}^n)$, on peut représenter $Dg(x)$ par une matrice de $\mathcal{M}_n(\mathbb{R})$, on confond alors l'application linéaire et la matrice qui la représente dans la base canonique, et on écrit par abus de notation $Dg(x)\in\mathcal{M}_n(\mathbb{R})$. On peut alors écrire, grâce à cet abus de notation,

$$Dg(x)(\eta)=Dg(x)\eta \text{ avec } (Dg(x)\eta)_i=\sum_{i,j=1,n}\partial^2_{i,j}h_j(x) \text{ où } \partial^2_{i,j}h=\partial_i(\partial_j h)(x).$$

Comme h est de classe C^2, la matrice $Dg(x)$ est symétrique, et donc diagonalisable dans $\mathbb{R}$. Pour $x \in \mathbb{R}^n$, on note $(\lambda_i(x))_{1\leq i\leq n}$ les valeurs propres de $Dg(x)$, qui sont donc réelles.

La proposition suivante donne une condition suffisante pour qu'une fonction vérifie les hypothèses (2.3) et (2.2).

Proposition 2.6. *Soit $h \in C^2(\mathbb{R}^n, \mathbb{R})$, et $(\lambda_i)_{i=1,n}$ les valeurs propres de la matrice hessienne de h. On suppose qu'il existe des réels strictement positifs β et γ tels que $-\underline{\lambda} \leq \lambda_i(x) \leq -\overline{\lambda},\ \forall i \in \{1 \dots n\},\ \forall x \in \mathbb{R}^n$. (Notons que cette hypothèse est plausible puisque les valeurs propres de la matrice hessienne sont réelles). Alors la fonction $g = \nabla h$ (gradient de h) vérifie les hypothèses* (2.3) *et* (2.2) *du théorème 2.3 avec $\alpha = \gamma$ et $M = \beta$.*

DÉMONSTRATION – Montrons d'abord que l'hypothèse (2.3) est vérifiée. Soit $(x, y) \in (\mathbb{R}^n)^2$, on veut montrer que $(g(x) - g(y)) \cdot (x - y) \leq -\gamma|x - y|^2$. On introduit pour cela la fonction $\varphi \in C^1(\mathbb{R}, \mathbb{R}^n)$ définie par :

$$\varphi(t) = g(x + t(y - x)).$$

On a donc $\varphi(1) - \varphi(0) = g(y) - g(x) = \int_0^1 \varphi'(t)dt$. Or $\varphi'(t) = Dg(x + t(y - x))(y - x)$. Donc $g(y) - g(x) = \int_0^1 Dg(x + t(y - x))(y - x)dt$. On en déduit que :

$$(g(y) - g(x)) \cdot (y - x) = \int_0^1 (Dg(x + t(y - x))(y - x) \cdot (y - x))\, dt.$$

Comme $\lambda_i(x) \in [-\underline{\lambda}, -\overline{\lambda}]\ \forall i \in \{1, \dots, n\}$, on a

$$-\underline{\lambda}|y|^2 \leq Dg(z)y \cdot y \leq -\overline{\lambda}|y|^2 \text{ pour tout } z \in \mathbb{R}^n$$

On a donc :

$$(g(y) - g(x)) \cdot (y - x) \leq \int_0^1 -\overline{\lambda}|y - x|^2 dt = -\overline{\lambda}|y - x|^2$$

ce qui montre que l'hypothèse (2.3) est bien vérifiée.

Montrons maintenant que l'hypothèse (2.2) est vérifiée. On veut montrer que $|g(y) - g(x)| \leq \underline{\lambda}|y - x|$. Comme

$$g(y) - g(x) = \int_0^1 Dg(x + t(y - x))(y - x)dt,$$

on a

$$\begin{aligned} |g(y) - g(x)| &\leq \int_0^1 |Dg(x + t(y - x))(y - x)|dt \\ &\leq \int_0^1 |Dg(x + t(y - x))||y - x|dt, \end{aligned}$$

où $|.|$ est la norme sur $\mathcal{M}_n(\mathbb{R})$ induite par la norme euclidienne sur $\mathbb{R}^n$.

Or, comme $\lambda_i(x) \in [-\underline{\lambda}, -\overline{\lambda}]$ pour tout $i = 1, \dots, N$, la matrice $-Dg(x + t(y - x))$ est symétrique définie positive et donc, d'après la proposition 1.29 page 51, son rayon spectral est égal à sa norme, pour la norme induite par la norme euclidienne. On a donc :

$$|Dg(x + t(y - x)| = \rho(Dg(x + t(y - x)) \leq \beta.$$

On a donc ainsi montré que : $|g(y) - g(x)| \leq \underline{\lambda}|y - x|$, ce qui termine la démonstration. ∎

Remarque 2.7 (Un cas particulier)**.** *Dans de nombreux cas issus de la discrétisation d'équations aux dérivées partielles, le problème de résolution d'un problème non linéaire apparaît sous la forme $Ax = R(x)$ où A est une matrice carrée d'ordre n inversible, et $R \in C(\mathbb{R}^n, \mathbb{R}^n)$. On peut le réécrire sous la forme $x = A^{-1}R(x)$ et appliquer l'algorithme de point fixe sur la fonction $f : x \mapsto A^{-1}Rx$, ce qui donne comme itération : $x^{(k+1)} = A^{-1}R(x^{(k)})$. Si on pratique un point fixe avec relaxation, dont le paramètre de relaxation $\omega > 0$, alors l'itération s'écrit :*

$$\tilde{x}^{(k+1)} = A^{-1}R(x^{(k)}), \qquad x^{(k+1)} = \omega\tilde{x}^{(k+1)} + (1 - \omega)x^{(k)}.$$

2.1.2 Point fixe de monotonie

Théorème 2.8 (Point fixe de monotonie)**.**
Soient $A \in \mathcal{M}_n(\mathbb{R})$ et $R \in C(\mathbb{R}^n, \mathbb{R}^n)$. On suppose que :

1. *la matrice A est une matrice d'inverse positive, ou IP–matrice voir exercice 6), c.à.d. que A est inversible et tous les coefficients de A^{-1} sont positifs ou nuls, ce qui est équivalent à dire que :*

$$Ax \geq 0 \Rightarrow x \geq 0,$$

au sens composante par composante, c'est-à-dire $\big((Ax)_i \geq 0,\ \forall i = 1, \ldots, n\big) \Rightarrow \big(x_i \geq 0,\ \forall i = 1, \ldots, n\big)$.

2. *R est monotone, c.à.d. que si $x \geq y$ (composante par composante) alors $R(x) \geq R(y)$ (composante par composante).*
3. *0 est une sous-solution du problème, c'est-à-dire que $0 \leq R(0)$ et il existe $\tilde{x} \in \mathbb{R}^n$; $\tilde{x} \geq 0$ tel que $\tilde{x}$ est une sur-solution du problème, c'est-à-dire que $A\tilde{x} \geq R(\tilde{x})$.*

On pose $x^{(0)} = 0$ et $Ax^{(k+1)} = R(x^{(k)})$. On a alors :

1. $0 \leq x^{(k)} \leq \tilde{x},\ \ \forall n \in \mathbb{N}$,
2. $x^{(k+1)} \geq x^{(k)},\ \ \forall n \in \mathbb{N}$,
3. $x^{(k)} \to \bar{x}$ *quand* $n \to +\infty$ *et* $A\bar{x} = R(\bar{x})$.

DÉMONSTRATION – Comme A est inversible la suite $(x^{(k)})_{n\in\mathbb{N}}$ vérifiant

$$\left\{ \begin{array}{l} x^{(0)} = 0, \\ Ax^{(k+1)} = R(x^{(k)}),\ \ n \geq 0 \end{array} \right.$$

est bien définie. On va montrer par récurrence sur n que $0 \leq x^{(k)} \leq \tilde{x}$ pour tout $n \geq 0$ et que $x^{(k)} \leq x^{(k+1)}$ pour tout $n \geq 0$.

1. Pour $n = 0$, on a $x^{(0)} = 0$ et donc $0 \leq x^{(0)} \leq \tilde{x}$ et $Ax^{(1)} = R(0) \geq 0$. On en déduit que $x^{(1)} \geq 0$ grâce aux hypothèses 1 et 3 et donc $x^{(1)} \geq x^{(0)} = 0$.
2. On suppose maintenant (hypothèse de récurrence) que $0 \leq x^{(p)} \leq \tilde{x}$ et $x^{(p)} \leq x^{(p+1)}$ pour tout $p \in \{0, \ldots, n-1\}$. On veut montrer que $0 \leq x^{(k)} \leq \tilde{x}$ et que $x^{(k)} \leq x^{(k+1)}$. Par hypothèse de récurrence pour $p = n-1$, on sait que $x^{(k)} \geq x^{(k-1)}$ et que $x^{(k-1)} \geq 0$. On a donc $x^{(k)} \geq 0$. Par hypothèse de récurrence, on a également que $x^{(k-1)} \leq \tilde{x}$ et grâce à l'hypothèse 2, on a donc $R(x^{(k-1)}) \leq R(\tilde{x})$. Par définition de la suite $(x^{(k)})_{n\in\mathbb{N}}$, on a $Ax^{(k)} = R(x^{(k-1)})$ et grâce à l'hypothèse 3, on sait que $A\tilde{x} \geq R(\tilde{x})$. On a donc : $A(\tilde{x} - x^{(k)}) \geq R(\tilde{x}) - R(x^{(k-1)}) \geq 0$. On en déduit alors (grâce à l'hypothèse 1) que $x^{(k)} \leq \tilde{x}$.
   De plus, comme $Ax^{(k)} = R(x^{(k-1)})$ et $Ax^{(k+1)} = R(x^{(k)})$, on a $A(x^{(k+1)} - x^{(k)}) = R(x^{(k)}) - R(x^{(k-1)}) \geq 0$ par l'hypothèse 2, et donc grâce à l'hypothèse 1, $x^{(k+1)} \geq x^{(k)}$.

On a donc ainsi montré (par récurrence) que

$$\begin{array}{l} 0 \leq x^{(k)} \leq \tilde{x}, \quad \forall n \geq 0 \\ x^{(k)} \leq x^{(k+1)}, \quad \forall n \geq 0. \end{array}$$

Ces inégalités s'entendent composante par composante, c.à.d. que si $x^{(k)} = (x_1^{(k)} \ldots x_n^{(k)})^t \in \mathbb{R}^n$ et $\tilde{x} = (\tilde{x}_1 \ldots \tilde{x}_n)^t \in \mathbb{R}^n$, alors $0 \leq x_i^{(k)} \leq \tilde{x}_i$ et $x_i^{(k)} \leq x_i^{(k+1)}$, $\forall i \in \{1, \ldots, n\}$, et $\forall n \geq 0$.

Soit $i \in \{1, \ldots, n\}$; la suite $(x_i^{(k)})_{n\in\mathbb{N}} \subset \mathbb{R}$ est croissante et majorée par $\tilde{x}_i$ donc il existe $\bar{x}_i \in \mathbb{R}$ tel que $\bar{x}_i = \lim_{n\to+\infty} x_i^{(k)}$. Si on pose $\bar{x} = (\bar{x}_1 \ldots \bar{x}_n)^t \in \mathbb{R}^n$, on a donc $x^{(k)} \longrightarrow \bar{x}$ quand $n \to +\infty$.

Enfin, comme $Ax^{(k+1)} = R(x^{(k)})$ et comme R est continue, on obtient par passage à la limite lorsque $n \to +\infty$ que $A\bar{x} = R(\bar{x})$ et que $0 \leq \bar{x} \leq \tilde{x}$. ∎

L'hypothèse 1 du théorème 2.8 est vérifiée par exemple par les matrices A qu'on a obtenues par discrétisation par différences finies des opérateurs $-u''$ sur l'intervalle $]0,1[$ (voir page 11 et l'exercice 43) et Δu sur $]0,1[\times]0,1[$ (voir page 14).

Théorème 2.9 (Généralisation du précédent)**.**
Soit $A \in \mathcal{M}_n(\mathbb{R})$, $R \in C^1(\mathbb{R}^n, \mathbb{R}^n)$, $R = (R_1, \ldots, R_n)^t$ tels que

1. *Pour tout $\beta \geq 0$ et pour tout $x \in \mathbb{R}^n$, $Ax + \beta x \geq 0 \Rightarrow x \geq 0$*
2. *$\dfrac{\partial R_i}{\partial x_j} \geq 0$, $\forall i, j$ t.q. $i \neq j$ (R_i est monotone croissante par rapport à la variable x_j si $j \neq i$) et $\exists \gamma > 0$, $-\gamma \leq \dfrac{\partial R_i}{\partial x_i} \leq 0$, $\forall x \in \mathbb{R}^n$, $\forall i \in \{1, \ldots, n\}$ (R_i est monotone décroissante par rapport à la variable x_i).*
3. *$0 \leq R(0)$ (0 est sous-solution) et $\exists \tilde{x} \geq 0$ tel que $A(\tilde{x}) \geq R(\tilde{x})$ ($\tilde{x}$ est sur-solution).*

Soient $x^{(0)} = 0$, $\beta \geq \gamma$, et $(x^{(k)})_{n \in \mathbb{N}}$ la suite définie par $Ax^{(k+1)} + \beta x^{(k+1)} = R(x^{(k)}) + \beta x^{(k)}$. Cette suite converge vers $\bar{x} \in \mathbb{R}^n$ et $A\bar{x} = R(\bar{x})$. De plus, $0 \leq x^{(k)} \leq \tilde{x}$ $\forall n \in \mathbb{N}$ et $x^{(k)} \leq x^{(k+1)}$, $\forall n \in \mathbb{N}$.

DÉMONSTRATION – On se ramène au théorème précédent avec $A + \beta Id$ au lieu de A et $R + \beta$ au lieu de R. ∎

Remarque 2.10 (Point fixe de Brouwer)**.** *On s'est intéressé ici uniquement à des théorèmes de point fixe "constructifs", i.e. qui donnent un algorithme pour le déterminer. Il existe aussi un théorème de point fixe dans $\mathbb{R}^n$ avec des hypothèses beaucoup plus générales (mais le théorème est non constructif), c'est le théorème de Brouwer[2] : si f est une fonction continue de la boule unité de $\mathbb{R}^n$ dans la boule unité, alors elle admet un point fixe dans la boule unité.*

2.1.3 Vitesse de convergence

Définition 2.11 (Vitesse de convergence)**.** *Soit $(x^{(k)})_{n \in \mathbb{N}} \in \mathbb{R}^n$ et $\bar{x} \in \mathbb{R}^n$. On suppose que $x^{(k)} \to \bar{x}$ lorsque $n \to +\infty$, que la suite est non stationnaire, c.à.d. que $x^{(k)} \neq \bar{x}$ pour tout $n \in \mathbb{N}$, et que*

$$\lim_{n \to +\infty} \frac{\|x^{(k+1)} - \bar{x}\|}{\|x^{(k)} - \bar{x}\|} = \beta \in [0, 1]. \tag{2.6}$$

On s'intéresse à la "vitesse de convergence" de la suite $(x^{(k)})_{n \in \mathbb{N}}$. On dit que :

1. *La convergence est* **sous–linéaire** *si $\beta = 1$.*
2. *La convergence est* **au moins linéaire** *si $\beta \in [0, 1[$.*
3. *La convergence* **est linéaire** *si $\beta \in]0, 1[$.*
4. *La convergence est* **super linéaire** *si $\beta = 0$. Dans ce cas, on dit également que :*
   (a) *La convergence est* **au moins quadratique** *s'il existe $\gamma \in \mathbb{R}_+$ et il existe $n_0 \in \mathbb{N}$ tels que si $n \geq n_0$ alors $\|x^{(k+1)} - \bar{x}\| \leq \gamma \|x^{(k)} - \bar{x}\|^2$.*
   (b) *La convergence est* **quadratique** *si*

$$\lim_{n \to +\infty} \frac{\|x^{(k+1)} - \bar{x}\|}{\|x^{(k)} - \bar{x}\|^2} = \gamma > 0.$$

Plus généralement, on dit que :
   (a) *La convergence est* **au moins d'ordre** *k s'il existe $\gamma \in \mathbb{R}_+$ et il existe $n_0 \in \mathbb{N}$ tels que si $n \geq n_0$ alors $\|x^{(k+1)} - \bar{x}\| \leq \gamma \|x^{(k)} - \bar{x}\|^k$.*
   (b) *La convergence est* **d'ordre** *k si*

$$\lim_{n \to +\infty} \frac{\|x^{(k+1)} - \bar{x}\|}{\|x^{(k)} - \bar{x}\|^k} = \gamma > 0.$$

2. Luitzen Egbertus Jan Brouwer (1881-1966), mathématicien néerlandais.

Remarque 2.12 (Sur la vitesse de convergence des suites)**.**

- *Remarquons d'abord que si une suite $(x^{(k)})_{n\in\mathbb{N}} \in \mathbb{R}^n$ converge vers $\bar{x} \in \mathbb{R}^n$ lorsque n tend vers l'infini, alors on a forcément $\beta \leq 1$ dans* (2.6)*. En effet, si la suite vérifie* (2.6) *avec $\beta > 1$, alors il existe $n_0 \in \mathbb{N}$ tel que si $n \geq n_0$, $|x_n - \bar{x}| \geq |x_{n_0} - \bar{x}|$ pour tout $n \geq n_0$, ce qui contredit la convergence.*
- *Quelques exemples de suites sui convergent sous-linéairement : $x_n = \frac{1}{\sqrt{n}}$, $x_n = \frac{1}{n}$, mais aussi, de manière moins intuitive : $x_n = \frac{1}{n^2}\ldots$. Toutes ces suites vérifient l'égalité* (2.6) *avec $\beta = 1$.*
- *Attention donc, contrairement à ce que pourrait suggérer son nom, la convergence linéaire (au sens donné ci-dessus), est déjà une convergence très rapide. Les suites géométriques définies par $x_n = \beta^n$ avec $\beta \in]0,1[$ sont des suites qui convergent linéairement (vers 0), car elles verifient évidemment bien* (2.6) *avec $\beta \in]0,1[$.*
- *la convergence quadratique est encore plus rapide ! Par exemple la suite définie par $x_{n+1} = x_n^2$ converge de manière quadratique pour un choix initial $x_0 \in]-1,1[$. Mais si par malheur le choix initial est en dehors de cet intervalle, la suite diverge alors très vite... de manière exponentielle, en fait, puisque $x_n = exp(n \ln x_0)$. C'est le cas de la méthode de Newton, que nous allons introduire maintenant. Lorsqu'elle converge, elle converge très vite (nous démontrerons que la vitesse de convergence est quadratique). Mais lorsqu'elle diverge, elle diverge aussi très vite...*

Pour construire des méthodes itératives qui convergent "super vite", nous allons donc essayer d'obtenir des vitesses de convergence super linéaires. C'est dans cet esprit que nous étudions dans la proposition suivante des conditions suffisantes de convergence de vitesse quadratique pour une méthode de type point fixe, dans le cas d'une fonction f de $\mathbb{R}$ dans $\mathbb{R}$.

Proposition 2.13 (Vitesse de convergence d'une méthode de point fixe)**.** *Soit $f \in C^1(\mathbb{R},\mathbb{R})$; on suppose qu'il existe $\bar{x} \in \mathbb{R}$ tel que $f(\bar{x}) = \bar{x}$. On construit la suite*

$$x^{(0)} \in \mathbb{R}$$
$$x^{(k+1)} = f(x^{(k)}).$$

1. *Si on suppose que $f'(\bar{x}) \neq 0$ et $|f'(\bar{x})| < 1$, alors il existe $\alpha > 0$ tel que si $x^{(0)} \in I_\alpha = [\bar{x} - \alpha, \bar{x} + \alpha]$ on a $x^{(k)} \to \bar{x}$ lorsque $n \to +\infty$. De plus si $x^{(k)} \neq \bar{x}$ pour tout $n \in \mathbb{N}$, alors*

$$\frac{|x^{(k+1)} - \bar{x}|}{|x^{(k)} - \bar{x}|} \to |f'(\bar{x})| = \beta \text{ avec } \beta \in]0,1[.$$

   La convergence est donc linéaire.

2. *Si on suppose maintenant que $f'(\bar{x}) = 0$ et $f \in C^2(\mathbb{R},\mathbb{R})$, alors il existe $\alpha > 0$ tel que si $x^{(0)} \in I_\alpha = [\bar{x} - \alpha, \bar{x} + \alpha]$, alors $x^{(k)} \to \bar{x}$ quand $n + \infty$, et si $x^{(k)} \neq \bar{x}$, $\forall n \in \mathbb{N}$ alors*

$$\frac{|x^{(k+1)} - \bar{x}|}{|x^{(k)} - \bar{x}|^2} \to \beta = \frac{1}{2}|f''(\bar{x})|.$$

   Dans ce cas, la convergence est donc au moins quadratique.

DÉMONSTRATION –

1. Supposons que $|f'(\bar{x})| < 1$, et montrons qu'il existe $\alpha > 0$ tel que si $x^{(0)} \in I_\alpha$ alors $x^{(k)} \to \bar{x}$. Comme $f \in C^1(\mathbb{R},\mathbb{R})$ il existe $\alpha > 0$ tel que $\gamma = \max_{x\in I_\alpha} |f'(x)| < 1$ (par continuité de f').
   On va maintenant montrer que $f : I_\alpha \to I_\alpha$ est strictement contractante, on pourra alors appliquer le théorème du point fixe à $f_{|I_\alpha}$, (I_α étant fermé), pour obtenir que $x^{(k)} \to \bar{x}$ où $\bar{x}$ est l'unique point fixe de $f_{|I\alpha}$.
   Soit $x \in I_\alpha$; montrons d'abord que $f(x) \in I_\alpha$: comme $f \in C^1(\mathbb{R},\mathbb{R})$, il existe $\xi \in]x,\bar{x}[$ tel que $|f(x) - \bar{x}| = |f(x) - f(\bar{x})| = |f'(\xi)||x - \bar{x}| \leq \gamma|x - \bar{x}| < \alpha$, ce qui prouve que $f(x) \in I_\alpha$.
   On vérifie alors que $f_{|I_\alpha}$ est strictement contractante en remarquant que pour tous $x, y \in I_\alpha$, $x < y$, il existe $\xi \in]x, y[(\subset I_\alpha)$ tel que $|f(x) - f(y)| = |f'(\xi)||x - y| \leq \gamma|x - y|$ avec $\gamma < 1$.

On a ainsi montré que $x^{(k)} \to \bar{x}$ si $x^{(0)} \in I_\alpha$.

Cherchons maintenant la vitesse de convergence de la suite. Supposons que $f'(\bar{x}) \neq 0$ et $x^{(k)} \neq \bar{x}$ pour tout $n \in \mathbb{N}$. Comme$x^{(k+1)} = f(x^{(k)})$ et $\bar{x} = f(\bar{x})$, on a $|x^{(k+1)} - \bar{x}| = |f(x^{(k)}) - f(\bar{x})|$. Comme $f \in C^1(\mathbb{R}, \mathbb{R})$, il existe $\xi_n \in]x^{(k)}, \bar{x}[$ ou $]\bar{x}, x^{(k)}[$, tel que $f(x^{(k)}) - f(\bar{x}) = f'(\xi_n)(x^{(k)} - \bar{x})$. On a donc

$$\frac{|x^{(k+1)} - \bar{x}|}{|x^{(k)} - \bar{x}|} = |f'(\xi_n)| \longrightarrow |f'(\bar{x})| \text{ car } x^{(k)} \to \bar{x} \text{ et } f' \text{ est continue.}$$

On a donc une convergence linéaire.

2. Supposons maintenant que $f \in C^2(\mathbb{R}, \mathbb{R})$ et $f'(\bar{x}) = 0$. On sait déjà par ce qui précède qu'il existe $\alpha > 0$ tel que si $x^{(0)} \in I_\alpha$ alors $x^{(k)} \to \bar{x}$ lorsque $n \to +\infty$. On veut estimer la vitesse de convergence ; on suppose pour cela que $x^{(k)} \neq \bar{x}$ pour tout $n \in \mathbb{N}$. Comme $f \in C^2(\mathbb{R}, \mathbb{R})$, il existe $\xi_n \in]x^{(k)}, \bar{x}[$ tel que

$$f(x^{(k)}) - f(\bar{x}) = f'(\bar{x})(x^{(k)} - \bar{x}) + \frac{1}{2} f''(\xi_n)(x^{(k)} - \bar{x})^2.$$

On a donc : $x^{(k+1)} - \bar{x} = \frac{1}{2} f''(\xi_n)(x^{(k)} - \bar{x})^2$ ce qui entraîne, par continuité de f'', que

$$\frac{|x^{(k+1)} - \bar{x}|}{|x^{(k)} - \bar{x}|^2} = \frac{1}{2}|f''(\xi_n)| \longrightarrow \frac{1}{2}|f''(\bar{x})| \text{ quand } n \to +\infty.$$

La convergence est donc quadratique.

■

2.1.4 Méthode de Newton dans $\mathbb{R}$

On va étudier dans le paragraphe suivant la méthode de Newton pour la résolution d'un système non linéaire. (En fait, il semble que l'idée de cette méthode revienne plutôt à Simpson [3] Donnons l'idée de la méthode de Newton dans le cas $n = 1$ à partir des résultats de la proposition précédente. Soit $g \in C^3(\mathbb{R}, \mathbb{R})$ et $\bar{x} \in \mathbb{R}$ tel que $g(\bar{x}) = 0$. On cherche une méthode de construction d'une suite $(x^{(k)})_n \in \mathbb{R}^n$ qui converge vers $\bar{x}$ de manière quadratique. On pose

$$f(x) = x + h(x)g(x) \text{ avec } h \in C^2(\mathbb{R}, \mathbb{R}) \text{ tel que } h(x) \neq 0 \ \forall x \in \mathbb{R}, .$$

et on a donc

$$f(x) = x \Leftrightarrow g(x) = 0.$$

Si par miracle $f'(\bar{x}) = 0$, la méthode de point fixe sur f va donner (pour $x^{(0)} \in I_\alpha$ donné par la proposition 2.13) $(x^{(k)})_{n \in \mathbb{N}}$ tel que $x^{(k)} \to \bar{x}$ de manière au moins quadratique. Or on a $f'(x) = 1 + h'(x)g(x) + g'(x)h(x)$ et donc $f'(\bar{x}) = 1 + g'(\bar{x})h(\bar{x})$. Il suffit donc de prendre h tel que $h(\bar{x}) = -\dfrac{1}{g'(\bar{x})}$. Ceci est possible si $g'(\bar{x}) \neq 0$.

En résumé, si $g \in C^3(\mathbb{R}, \mathbb{R})$ est telle que $g'(\overline{x}) \neq 0 \ \forall x \in \mathbb{R}$ et $g(\bar{x}) = 0$, on peut construire, pour x assez proche de $\bar{x}$, la fonction $f \in C^2(\mathbb{R}, \mathbb{R})$ définie par

$$f(x) = x - \frac{g(x)}{g'(x)}.$$

Grâce à la proposition 2.13, il existe $\alpha > 0$ tel que si $x^{(0)} \in I_\alpha$ alors la suite définie par $x^{(k+1)} = f(x^{(k)}) = x^{(k)} - \frac{g(x^{(k)})}{g'(x^{(k)})}$ converge vers $\bar{x}$ de manière au moins quadratique.

3. Voir Nick Kollerstrom (1992). *Thomas Simpson and "Newton's method of approximation" : an enduring myth*, The British Journal for the History of Science, 25, pp 347-354 doi :10.1017/S0007087400029150 – Thomas Simpson est un mathématicien anglais du 18-ème siècle à qui on attribue généralement la méthode du même nom pour le calcul approché des intégrales, probablement à tort car celle-ci apparaît déjà dans les travaux de Kepler deux siècles plus tôt !

Remarquons que dans le cas $n = 1$, la suite de Newton peut s'obtenir naturellement en remplaçant l'équation $g(\overline{x}) = 0$ par $g(x^{(k+1)}) = 0$, et $g(x^{(k+1)})$ par le développement limité en x^k :

$$g(x^{(k+1)}) = g(x^{(k)})g'(x^{(k)})(x^{(k+1)} - x^{(k)}) + |x^{(k+1)} - x^{(k)}|\epsilon(x^{(k+1)} - x^{(k)}).$$

C'est le plus sûr moyen mnémotechnique pour retrouver l'itération de Newton :

$$g(x^{(k)}) + g'(x^{(k)})(x^{(k+1)} - x^{(k)}) = 0 \text{ ou encore } g'(x^{(k)})(x^{(k+1)} - x^{(k)}) = -g(x^{(k)}). \tag{2.7}$$

Comparons sur un exemple les méthodes de point fixe et de Newton. On cherche le zéro de la fonction $g : x \mapsto x^2 - 3$ sur $\mathbb{R}_+$. Notons en passant que la construction de la suite $x^{(k)}$ par point fixe ou Newton permet l'approximation effective de $\sqrt{3}$. Si on applique le point fixe standard, la suite $x^{(k)}$ s'écrit

$$\begin{aligned} x^{(0)} &\text{ donné ,} \\ x^{(k+1)} &= x^{(k)} - (x^{(k)})^2 + 3. \end{aligned}$$

Si on applique le point fixe avec paramètre de relaxation ω, la suite $x^{(k)}$ s'écrit

$$\begin{aligned} x^{(0)} &\text{ donné ,} \\ x^{(k+1)} &= -x^{(k)} + \omega(-x^{(k)})^2 + 3) \end{aligned}$$

Si maintenant on applique la méthode de Newton, la suite $x^{(k)}$ s'écrit

$$\begin{aligned} x^{(0)} &\text{ donné ,} \\ x^{(k+1)} &= -\frac{(x^{(k)})^2 - 3}{2x^{(k)}} b \end{aligned}$$

Comparons les suites produites par scilab à partir de $x^{(0)} = 1$ par le point fixe standard, le point fixe avec relaxation ($\omega = .1$) et la méthode de Newton.

— **point fixe standard :**
1 3 - 3 - 9 - 87 - 7653 - 58576059 - 3.431D+15 - 1.177D+31

— **point fixe avec relaxation :**
1. 1.2 1.356 1.4721264 1.5554108 1.6134805 1.6531486 1.6798586 1.6976661 1.7094591
1.717234 1.7223448 1.7256976 1.7278944 1.7293325 1.7302734 1.7308888 1.7312912
1.7315543 1.7317263 1.7318387 1.7319122 1.7319602 1.7319916 1.7320121 1.73204
1.7320437 1.7320462 1.7320478 1.7320488 1.7320495 1.73205 1.7320503 1.7320504
1.7320506 1.7320507 1.7320507 1.7320507 1.7320508

— **Newton :**
1. 2. 1.75 1.7321429 1.7320508 1.7320508

Remarque 2.14 (Attention à l'utilisation du théorème des accroissements finis…)**.** *On a fait grand usage du théorème des accroissements finis dans ce qui précède. Rappelons que sous la forme qu'on a utilisée, ce théorème n'est valide que pour les fonctions de* $\mathbb{R}$ *dans* $\mathbb{R}$*. On pourra s'en convaincre en considérant la fonction de* $\mathbb{R}$ *dans* $\mathbb{R}^2$ *définie par :*

$$\varphi(x) = \begin{bmatrix} \sin x \\ \cos x \end{bmatrix}.$$

On peut vérifier facilement qu'il n'existe pas de $\xi \in \mathbb{R}$ *tel que* $\varphi(2\pi) - \varphi(0) = 2\pi\varphi'(\xi)$.

2.1.5 Exercices

Enoncés

Exercice 64 (Calcul différentiel)**.** *Suggestions en page 133, corrigé détaillé en page 134*

Soit $f \in C^2(\mathbb{R}^n, \mathbb{R})$.
1. Montrer que pour tout $x \in \mathbb{R}^n$, il existe un unique vecteur $a(x) \in \mathbb{R}^n$ tel que $Df(x)(h) = a(x) \cdot h$ pour tout $h \in \mathbb{R}^n$.
Montrer que $(a(x))_i = \partial_i f(x)$.

2. On pose $\nabla f(x) = (\partial_1 f(x), \ldots, \partial_1 f(x))^t$. Soit φ l'application définie de $\mathbb{R}^n$ dans $\mathbb{R}^n$ par $\varphi(x) = \nabla f(x)$. Montrer que $\varphi \in C^1(\mathbb{R}^n, \mathbb{R}^n)$ et que $D\varphi(x)(y) = A(x)y$, où $A(x))_{i,j} = \partial^2_{i,j} f(x)$.

Exercice 65 (Calcul différentiel, suite)**.**

1. Soit $f \in C^2(\mathbb{R}^2, \mathbb{R})$ la fonction définie par $f(x_1, x_2) = ax_1 + bx_2 + cx_1x_2$, où a, b, et c sont trois réels fixés. Donner la définition et l'expression de $Df(x)$, $\nabla f(x)$,Df, $D^2f(x)$, $H_f(x)$.
2. Même question pour la fonction $f \in C^2(\mathbb{R}^3, \mathbb{R})$ définie par $f(x_1, x_2, x_3) = x_1^2 + x_1^2x_2 + x_2 \sin(x_3)$.

Exercice 66 (Point fixe)**.**

Soit $I = [0, 1]$, et $f : x \mapsto x^4$. Montrer que la suite des itérés de point fixe converge pour tout $x \in [0, 1]$ et donner la limite de la suite en fonction du choix initial $x^{(0)}$.

Exercice 67 (Point fixe)**.** *corrigé détaillé en page* **??**.

1. On veut résoudre l'équation $2xe^x = 1$.
   (a) Vérifier que cette équation peut s'écrire sous forme de point fixe : $x = \frac{1}{2}e^{-x}$.
   (b) Ecrire l'algorithme de point fixe, et calculer les itérés x_0, x_1, x_2 et x_3 en partant depuis $x_0 = 1$.
   (c) Justifier la convergence de la méthode.
2. On veut résoudre l'équation $x^2 - 2 = 0$.
   (a) Vérifier que cette équation peut s'écrire sous forme de point fixe : $x = \frac{2}{x}$.
   (b) Ecrire l'algorithme de point fixe, et tracer sur un graphique les itérés x_0, x_1, x_2 et x_3 en partant de $x_0 = 1$ et $x_0 = 2$.
   (c) Essayer ensuite le point fixe sur $x = \frac{x^2+2}{2x}$. Pas très facile à deviner, n'est ce pas ?
   (d) Pour suivre les traces de Newton (ou plutôt Simpson, semble-t-il) : à x_n connu, écrire le développement limité de $g(x) = x^2 - 2$ entre $x^{(n)}$ et $x^{(n+1)}$, remplacer l'équation $g(\overline{x}) = 0$ par $g(x^{(n+1)}) = 0$, et $g(x^{(n+1)})$ par le développement limité en x^{n+1} , et en déduire l'approximation $x^{(n+1)} = x^{(n)} - \frac{g(x^{(n)})}{g'(x^{(n)})}$. Retrouver ainsi l'itération de la question précédente (pour $g(x) = x^2 - 2$).

Exercice 68 (Méthode de monotonie)**.** *Suggestions en page 133, corrigé détaillé en page* **??**.

On suppose que $f \in C^1(\mathbb{R}, \mathbb{R})$, $f(0) = 0$ et que f est croissante. On s'intéresse, pour $\lambda > 0$, au système non linéaire suivant de n équations à n inconnues (notées $u_1, \ldots, u_n$) :

$$\begin{array}{l} (Au)_i = \alpha_i f(u_i) + \lambda b_i \ \forall i \in \{1, \ldots, n\}, \\ u = (u_1, \ldots, u_n)^t \in \mathbb{R}^n, \end{array} \tag{2.8}$$

où $\alpha_i > 0$ pour tout $i \in \{1, \ldots, n\}$, $b_i \geq 0$ pour tout $i \in \{1, \ldots, n\}$ et $A \in \mathcal{M}_n(\mathbb{R})$ est une matrice vérifiant

$$u \in \mathbb{R}^n, \ Au \geq 0 \Rightarrow u \geq 0. \tag{2.9}$$

On suppose qu'il existe $\mu > 0$ t.q. (2.8) ait une solution, notée $u^{(\mu)}$, pour $\lambda = \mu$. On suppose aussi que $u^{(\mu)} \geq 0$. Soit $0 < \lambda < \mu$. On définit la suite $(v^{(k)})_{n\in\mathbb{N}} \subset \mathbb{R}^n$ par $v^{(0)} = 0$ et, pour $n \geq 0$,

$$(Av^{(k+1)})_i = \alpha_i f(v_i^{(k)}) + \lambda b_i \ \forall i \in \{1, \ldots, n\}. \tag{2.10}$$

Montrer que la suite $(v^{(k)})_{n\in\mathbb{N}}$ est bien définie, convergente (dans $\mathbb{R}^n$) et que sa limite, notée $u^{(\lambda)}$, est solution de (2.8) (et vérifie $0 \leq u^{(\lambda)} \leq u^{(\mu)}$).

Exercice 69 (Point fixe amélioré)**.** *Suggestions en page 133, Corrigé en page* **??**

Soit $g \in C^3(\mathbb{R},\mathbb{R})$ et $\overline{x} \in \mathbb{R}$ tels que $g(\overline{x}) = 0$ et $g'(\overline{x}) \neq 0$.
On se donne $\varphi \in C^1(\mathbb{R},\mathbb{R})$ telle que $\varphi(\overline{x}) = \overline{x}$.
On considère l'algorithme suivant :

$$\begin{cases} x_0 \in \mathbb{R}, \\ \\ x_{n+1} = h(x_n), n \geq 0. \end{cases} \tag{2.11}$$

avec $h(x) = x - \dfrac{g(x)}{g'(\varphi(x))}$

1) Montrer qu'il existe $\alpha > 0$ tel que si $x_0 \in [\overline{x}-\alpha, \overline{x}+\alpha] = I_\alpha$, alors la suite donnée par l'algorithme (2.11) est bien définie ; montrer que $x_n \to \overline{x}$ lorsque $n \to +\infty$.

On prend maintenant $x_0 \in I_\alpha$ où α est donné par la question 1.

2) Montrer que la convergence de la suite $(x_n)_{n\in\mathbb{N}}$ définie par l'algorithme (2.11) est au moins quadratique.

3) On suppose que φ' est lipschitzienne et que $\varphi'(\overline{x}) = \dfrac{1}{2}$. Montrer que la convergence de la suite $(x_k)_{k\in\mathbb{N}}$ définie par (2.11) est au moins cubique, c'est-à-dire qu'il existe $c \in \mathbb{R}_+$ tel que

$$|x_{k+1} - \overline{x}| \leq c|x_k - \overline{x}|^3, \qquad \forall k \geq 1.$$

4) Soit $\beta \in \mathbb{R}_+^*$ tel que $g'(x) \neq 0 \quad \forall x \in I_\beta =]\overline{x}-\beta, \overline{x}+\beta[$; montrer que si on prend φ telle que :

$$\varphi(x) = x - \frac{g(x)}{2g'(x)} \qquad \text{si } x \in I_\beta,$$

alors la suite définie par l'algorithme (2.11) converge de manière cubique.

Suggestions

Exercice 64 page 131 (Calcul différentiel) 1. Utiliser le fait que $Df(x)$ est une application linéaire et le théorème de Riesz. Appliquer ensuite la différentielle à un vecteur h bien choisi.
2. Mêmes idées...

Exercice 68 page 132 (Méthode de monotonie) Pour montrer que la suite $(v^{(k)})_{n\in\mathbb{N}}$ est bien définie, remarquer que la matrice A est inversible. Pour montrer qu'elle est convergente, montrer que les hypothèses du théorème du point fixe de monotonie vu en cours sont vérifiées.

Exercice 69 page 133 (Point fixe amélioré)
1) Montrer qu'on peut choisir α de manière à ce que $|h'(x)| < 1$ si $x \in I_\alpha$, et en déduire que $g'(\varphi(x_n) \neq 0$ si x_0 est bien choisi.
2) Remarquer que

$$|x_{k+1} - \overline{x}| = (x_k - \overline{x})(1 - \frac{g(x_k) - g(\overline{x})}{(x_k - \overline{x})g'(\varphi(x_k))}. \tag{2.12}$$

En déduire que

$$|x_{n+1} - \overline{x}| \leq \frac{1}{\varepsilon}|x_n - \overline{x}|^2 \sup_{x\in I_\alpha} |\varphi'(x)| \sup_{x\in I_\alpha} |g''(x)|.$$

3) Reprendre le même raisonnement avec des développements d'ordre supérieur.

4) Montrer que φ vérifie les hypothèses de la question 3).

Corrigés

Exercice 64 page 131 1. Par définition, $T = Df(x)$ est une application linéaire de $\mathbb{R}^n$ dans $\mathbb{R}^n$, qui s'écrit donc sous la forme : $T(h) = \sum_{i=1}^n a_i h_i = a \cdot h$. Or l'application T dépend de x, donc le vecteur a aussi.

Montrons maintenant que $(a(x))_i = \partial_i f(x)$, pour $1 \leq i \leq n$ Soit $h^{(i)} \in \mathbb{R}^n$ défini par $h^{(i)}_j = h\delta_{i,j}$, où $h > 0$ et $\delta_{i,j}$ désigne le symbole de Kronecker, i.e. $\delta_{i,j} = 1$ si $i = j$ et $\delta_{i,j} = 0$ sinon. En appliquant la définition de la différentielle avec $h^{(i)}$, on obtient :

$$f(x + h^{(i)}) - f(x) = Df(x)(h(i)) + \|h^{(i)}\|\varepsilon(h^{(i)}),$$

c'est–à–dire :

$$f(x_1, \ldots, x_{i-1}, x_i + h, x_{i-1}, \ldots, x_n) - f(x_1, \ldots, x_n) = (a(x))_i h + h\varepsilon(h^{(i)}).$$

En divisant par h et en faisant tendre h vers 0, on obtient alors que $(a(x))_i = \partial_i f(x)$.

2. Comme $f \in C^2(\mathbb{R}^n, \mathbb{R})$, on a $(\partial_i f(x) \in C^1(\mathbb{R}^n, \mathbb{R})$, et donc $\varphi \in C^1(\mathbb{R}^n, \mathbb{R}^n)$. Comme $D\varphi(x)$ est une application linéaire de $\mathbb{R}^n$ dans $\mathbb{R}^n$, il existe une matrice $A(x)$ carrée d'ordre n telle que $D\varphi(x)(y) = A(x)y$ pour tout $y \in \mathbb{R}^n$. Il reste à montrer que $(A(x))_{i,j} = \partial^2_{i,j} f(x)$. Soit $h^{(i)} \in \mathbb{R}^n$ défini à la question préceédente, pour $i,\ j = 1, \ldots, n$, on a

$$(D\varphi(x)(h^{(j)}))_i = (A(x)h^{(j)})_i = \sum_{k=1}^n a_{i,k}(x)h^{(j)})_k = ha_{i,j}(x).$$

Or par définition de la différentielle,

$$\varphi_i(x + h^{(j)}) - \varphi_i(x) = (D\varphi(x)(h^{(j)}))_i + \|h^{(j)}\|\varepsilon_i(h^{(j)}),$$

ce qui entraîne, en divisant par h et en faisant tendre h vers 0 : $\partial_j \varphi_i(x) = a_{i,j}(x)$. Or $\varphi_i(x) = \partial_i f(x)$, et donc $(A(x))_{i,j} = a_{i,j}(x) = \partial^2_{i,j} f(x)$.

Corrigé de l'exercice 64 (Point fixe et Newton)

1. Résolution de l'équation $2xe^x = 1$.
   (a) Comme e^x ne s'annule pas, l'équation $2xe^x = 1$ est équivalente à l'équation $x = \frac{1}{2}e^{-x}$, qui est sous forme point fixe $x = f(x)$ avec $f(x) = \frac{1}{2}e^{-x}$.
   (b) L'algorithme de point fixe s'écrit

$$x^{(0)} \text{ donné} \tag{2.13a}$$
$$x^{(k+1} = f(x^{(k}). \tag{2.13b}$$

   Scilab donne :

```
x  =  1.
x  =  0.1839397
x  =  0.4159930
x  =  0.3298425
```

   Notons que la suite n'est pas monotone.
   (c) On a $f'(x) = -\frac{1}{2}e^{-x}$ et donc $f'(x) \leq \frac{1}{2}$. L'application $x \mapsto f(x) = \frac{1}{2}e^{-x}$ est contractante de $[0, 1]$ dans $[0, 1]$, et elle admet donc un point fixe, qui est limite de la suite construite par l'algorithme précédent. —
2. Résolution de l'équation $x^2 - 2 = 0$.
   (a) On se place sur l'intervalle $]0, 4[$, de manière à ce que x ne s'annule pas, auquel cas l'équation $x^2 - 2 = 0$ est manifestement équivalente à l'équation $x = \frac{2}{x}$, qui est sous forme point fixe $x = f(x)$ avec $f(x) = \frac{2}{x}$.

(b) L'algorithme de point fixe s'écrit toujours (2.13), mais si on part de $x_0 = 1$ ou $x_0 = 2$, on obtient une suite cyclique (1, 2, 1, 2,1, 2...) ou (2, 1, 2, 1, 2,1, 2...) qui ne converge pas.

(c) Scilab donne

```
% x  =    1.
%  x  =     1.5
%  x  =     1.4166667
%  x  =     1.4142157
```

(d) Le développement limité de $g(x) = x^2 - 2$ entre $x^{(n)}$ et $x^{(n+1)}$ s'écrit :

$$g(x^{(n+1)}) = g(x^{(n)}) + (x^{(n+1)} - x^{(n)})g'(x^{(n)}) + (x^{(n+1)} - x^{(n)})\varepsilon(x^{(n+1)} - x^{(n)}),$$

avec $\varepsilon(x) \to 0$ lorsque $x \to 0$. En écrivant qu'on cherche $x^{(n+1)}$ tel que $g(x^{(n+1)}) = 0$ et en négligeant le terme de reste du développement limité, on obtient :

$$0 = g(x^{(n)}) + (x^{(n+1)} - x^{(n)})g'(x^{(n)}),$$

Pour $g(x) = x^2 - 2$, on a $g'(x) = 2x$ et donc l'équation précédente donne bien l'itération de la question précédente.