Cours d'analyse numérique

Raphaèle Herbin, Thierry Gallouët

Université Aix Marseille

Licence de mathématiques

Cours d'Analyse numérique

Thierry Gallouët et Raphaèle Herbin

4 août 2023

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, `https://www.i2m.univ-amu.fr/perso/`
— Introduction to linear algebra, Gilbert Strang, Wellesley Cambridge Press, 2008

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 $b \in \mathbb{R}^n$, l'objectif est de résoudre le système linéaire $A\boldsymbol{x} = b$, c'est-à-dire de trouver $\boldsymbol{x}$ solution de :

$$\begin{cases} \boldsymbol{x} \in \mathbb{R}^n \\ A\boldsymbol{x} = b \end{cases} \tag{1.1}$$

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 paragraphes 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

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

On peut écrire les matrices A et B sous forme de lignes (notées $\boldsymbol{\ell}_i$) et colonnes (notées $\boldsymbol{c}_j$) :

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

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 :

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

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 une 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.
Il est intéressant pour la suite de ce cours de voir ce que donne la multiplication d'une matrice par une matrice de permutation.
Commençons par une exemple. Soit P et A des matrices carrées d'ordre 2 définies par

$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},\ A = \begin{bmatrix} a & b \\ c & d \end{bmatrix},\ PA = \begin{bmatrix} c & d \\ a & b \end{bmatrix},\ AP = \begin{bmatrix} b & a \\ d & c \end{bmatrix}.$$

La multiplication de A par la matrice P échange les lignes de A lorqu'on multiplie A par P à gauche, et elle échange les colonnes de A lorqu'on multiplie A par P à droite. Noter que ceci montre d'ailleurs bien que le produit matriciel n'est pas commutatif... La matrice P s'appelle matrice de permutation. Les matrices de permutation auront un fort rôle à jouer dans l'élaboration d'algorithmes de résolution des systèmes linéaires (voir l'algorithme de Gauss avec pivot partiel).
De manière plus générale, on peut définir une matrice de permutation de la façon suivante :

Définition 1.1 (Matrice de permutation)**.** *Soit $n \in \mathbb{N}$ et soient $i, j \in \{1, \ldots, n\}$. On notera $P^{(i \leftrightarrow j)} \in \mathcal{M}_n(\mathbb{R})$ la matrice telle que :*

1. *Si $i = j$, $P^{(i \leftrightarrow j)} = \mathrm{Id}_n$,*
2. *Si $i \neq j$, $p_{i,i}^{(i \leftrightarrow j)} = p_{j,j}^{(i \leftrightarrow j)} = 0$, $p_{i,j}^{(i \leftrightarrow j)} = p_{j,i}^{(i \leftrightarrow j)} = 1$, et pour tout $k, l \in \{1, \ldots, n\}$ tel que $(k,l) \notin \{(i,i),(i,j),(j,i),(j,j)\}$, si $k = l$, $p_{k,l}^{(i \leftrightarrow j)} = 1$ sinon $p_{k,l}^{(i \leftrightarrow j)} = 0$.*

La matrice $P^{(i \leftrightarrow j)}$ est alors appelée matrice de permutation élémentaire. Une matrice de permutation est définie comme le produit d'un nombre fini de permutations élémentaires.

Remarquons qu'une matrice de permutation possède alors n termes égaux à 1, et tous les autres égaux à 0, tels que chaque ligne et chaque colonne comprenne exactement l'un des termes égaux à 1 (pour les amateurs de jeu d'échecs, ces termes sont disposés comme n tours sur un échiquier de taille $n \times n$ telles qu'aucune tour ne peut en prendre une autre).
Pour toute matrice $A \in \mathcal{M}_n(\mathbb{R})$ et toute matrice de permutation P, la matrice PA est obtenue à partir de A par permutation des lignes de A, et la matrice AP est obtenue à partir de A par permutation des colonnes de A. Dans un système linéaire $Ax = b$, on remarque qu'on ne change pas la solution x si on permute des lignes, c'est à dire si l'on résout $PAx = Pb$. Notons que le produit de matrices de permutation est évidemment une matrice de permutation, et que toute matrice de permutation P est inversible et $P^{-1} = P^t$ (voir exercice 2).

Le tableau 1.1 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.
On rappelle pour une bonne lecture de ce tableau les quelques définitions suivantes (pour le cas ou il y aurait des notions que vous avez oubliées ou que vous ne maîtrisez pas bien).

Définition 1.2 (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).*

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

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 pivots (non nuls)	A a $r < n$ pivots
A est de rang maximal : $\text{rang}(A) = n$.	$\text{rang}(A) = r < n$
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	Zéro est valeur propre de A
A^tA est symétrique définie positive	A^tA n'est que semi-définie

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

Définition 1.3 (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 1.4 (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 $\det\tilde{A} = -\det A$.

(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 :

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

Définition 1.5 (Matrices symétriques et symétriques définies positives)**.** *Soit $A = (a_{i,j})_{1\leq i,j\leq n} \in \mathcal{M}_n(\mathbb{R})$ une matrice carrée d'ordre n.*
On dit que la matrice A est symétrique si $a_{i,j} = a_{j,i}$ pour tout (i,j) $1 \leq i,j \leq n$.
On dit que la matrice A est symétrique définie positive (s.d.p.) si elle est symétrique et si elle vérifie de plus $A\boldsymbol{x}\cdot\boldsymbol{x} \geq 0$ pour tout $\boldsymbol{x} \in \mathbb{R}^n$ et $[A\boldsymbol{x}\cdot\boldsymbol{x} = 0 \Longrightarrow \boldsymbol{x} = \mathbf{0}$ pour tout $\boldsymbol{x} \in \mathbb{R}^n]$.
On dit que la matrice A est symétrique semi-définie positive (s.d.p.) si elle est symétrique et si elle vérifie $A\boldsymbol{x}\cdot\boldsymbol{x} \geq 0$ pour tout $\boldsymbol{x} \in \mathbb{R}^n$.

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 1.6 (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 des 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 distinctes. 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 $E_i = \ker(A - \lambda_i\mathrm{Id})$ (appelée multiplicité géométrique) est égale a la multiplicité algébrique de λ_i, c'est-à-dire son ordre de multiplicité en tant que racine du polynôme caractéristique.
Par exemple la matrice $A = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$ 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 1.7. *Soit A une matrice réelle carrée d'ordre n, diagonalisable dans $\mathbb{R}$. Alors*

$$A = P\mathrm{diag}(\lambda_1,\ldots,\lambda_n)P^{-1},$$

où P est la matrice dont les vecteurs colonnes sont égaux à des vecteurs propres $\boldsymbol{u}_1,\ldots,\boldsymbol{u}_n$ associées aux 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 $i = 1, \dots n$, et donc, en notant P la matrice dont les colonnes sont les vecteurs propres $\boldsymbol{u}_i$,

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

et donc

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

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 est aussi la matrice définie (de manière unique) par $P\boldsymbol{e}_i = \boldsymbol{u}_i$, où $(e_i)_{i=1,\dots,n}$ 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 $(\boldsymbol{e}_i)_{i=1,\dots,n}$ à la base $(\boldsymbol{u}_i)_{i=1,\dots,n}$; (il est bien clair que la i-ème colonne de P est constituée des composantes de $\boldsymbol{u}_i$ dans la base canonique $(\boldsymbol{e}_1, \dots, \boldsymbol{e}_n)$.

La matrice P est inversible car les vecteurs propres forment une base, et on peut donc aussi écrire :

$$P^{-1}AP = \text{diag}(\lambda_1, \dots, \lambda_n) \text{ ou } A = P\text{diag}(\lambda_1, \dots, \lambda_n)P^{-1}.$$

■

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 1.8 (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*

$$E \times E \to \mathbb{R},$$
$$(\boldsymbol{x}, \boldsymbol{y}) \to (\boldsymbol{x} \mid \boldsymbol{y})_E,$$

qui vérifie :

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

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

Conséquence immédiate : Dans le cas où $E = \mathbb{R}^n$, le produit scalaire canonique de $\boldsymbol{x} = (x_1, \dots, x_n)^t$ et $\boldsymbol{y} = (y_1, \dots, y_n)^t$ est défini par $(\boldsymbol{x} \mid \boldsymbol{y})_E = \boldsymbol{x} \cdot \boldsymbol{y} = \sum_{i=1}^n x_i y_i$. Si $A \in \mathcal{M}_n(\mathbb{R})$ est une matrice symétrique, alors l'application T définie de E dans E par : $T(\boldsymbol{x}) = A\boldsymbol{x}$ est linéaire, et :

$$(T(\boldsymbol{x})|\boldsymbol{y}) = A\boldsymbol{x} \cdot \boldsymbol{y} = \boldsymbol{x} \cdot A^t\boldsymbol{y} = \boldsymbol{x} \cdot A\boldsymbol{y} = (\boldsymbol{x} \mid T(\boldsymbol{y})).$$

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 $\boldsymbol{f}_i \cdot \boldsymbol{f}_j = \delta_{i,j}, \forall \, (i,j) \in \{1, \dots, n\}^2$.

Interprétation algébrique : Il existe une matrice de passage P de $(\boldsymbol{e}_1, \dots, \boldsymbol{e}_n)$ base canonique de $\mathbb{R}^n$ dans la base $(\boldsymbol{f}_1, \dots, \boldsymbol{f}_n)$ dont la i-ème colonne de P est constituée des coordonnées de $\boldsymbol{f}_i$ dans la base $(\boldsymbol{e}_1 \dots \boldsymbol{e}_n)$. On a :

$P\boldsymbol{e}_i = \boldsymbol{f}_i$. On a alors $P^{-1}AP\boldsymbol{e}_i = P^{-1}A\boldsymbol{f}_i = P^{-1}(\lambda_i \boldsymbol{f}_i) = \lambda_i \boldsymbol{e}_i = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)\boldsymbol{e}_i$, où $\mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ désigne la matrice diagonale de coefficients diagonaux $\lambda_1, \ldots, \lambda_n$. 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,

$$P^tP\boldsymbol{e}_i \cdot \boldsymbol{e}_j = P\boldsymbol{e}_i \cdot P\boldsymbol{e}_j = (\boldsymbol{f}_i|\boldsymbol{f}_j) = \delta_{i,j},\ \forall i, j \in \{1 \ldots n\},$$

et donc $(P^tP\boldsymbol{e}_i - \boldsymbol{e}_i)\cdot\boldsymbol{e}_j = 0, \forall j \in \{1 \ldots n\}, \forall i \in \{1, \ldots n\}$. On en déduit que $P^tP\boldsymbol{e}_i = \boldsymbol{e}_i$ pour tout $i = 1, \ldots n$, *i.e.* $P^tP = PP^t = \mathrm{Id}$.

DÉMONSTRATION *du lemme 1.8* Cette démonstration se fait par récurrence sur la dimension de E. On note $(\cdot|\cdot)$ le produit scalaire dans E et $\|\cdot\|$ la norme associée.

1ère étape. On suppose $dimE = 1$. Soit $\boldsymbol{e} \in E$, $e \neq 0$, alors $E = \mathbb{R}\boldsymbol{e} = \mathbb{R}\boldsymbol{f}_1$ avec $\boldsymbol{f}_1 = \dfrac{1}{\|e\|}\boldsymbol{e}$. Soit $T : E \to E$ linéaire. 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 dim $E < n$. 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 :

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

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

$$\frac{1}{\|\boldsymbol{e}+t\boldsymbol{y}\|}(\boldsymbol{e}+t\boldsymbol{y}) \in S_1 \text{ et donc } \varphi(\boldsymbol{e}) = \lambda \geq \left(T\left(\frac{1}{\|\boldsymbol{e}+t\boldsymbol{y}\|}(\boldsymbol{e}+t\boldsymbol{y})\right) | \frac{1}{\|\boldsymbol{e}+t\boldsymbol{y}\|}(\boldsymbol{e}+t\boldsymbol{y}))\right)_E$$

donc $\lambda(\boldsymbol{e}+t\boldsymbol{y} \mid \boldsymbol{e}+t\boldsymbol{y})_E \geq (T(\boldsymbol{e}+t\boldsymbol{y}) \mid \boldsymbol{e}+t\boldsymbol{y})$. En développant on obtient :

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

Comme $t > 0$, ceci donne :

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

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

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

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

$$-u''(x) = f(x) \tag{1.5a}$$

$$u(0) = u(1) = 0. \tag{1.5b}$$

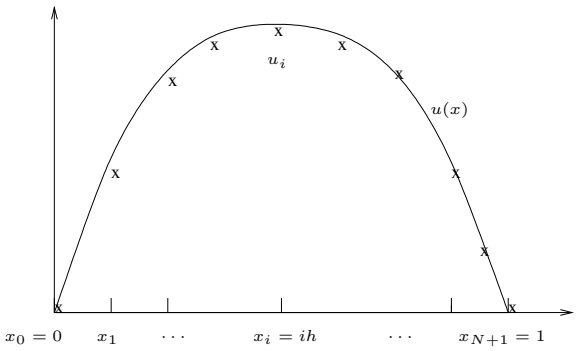


FIGURE 1.1: Solution exacte et approchée de $-u'' = f$

Remarque 1.9 (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 dérivées 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 :

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

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 :

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

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 : on 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, avec $b_i = f(x_i)$,

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

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 15 page 21), 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 :

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

Si on appelle $\boldsymbol{e}$ le vecteur de composantes e_i et R le vecteur de composantes R_i on déduit de la définition (1.7) 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 K_n dépend de n (c'est-à-dire 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 66, 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 $i = 1, \ldots, M$ et $j = 1, \ldots, M$ 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 12), 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 $b = (b_1, \ldots, b_n)^t$. Les composantes $b_1, \ldots, b_n$ sont définies par :pour $i, j = 1, \ldots, M$, on pose $k = j + (i-1)M$ et $b_k = f(x_i, y_j)$.

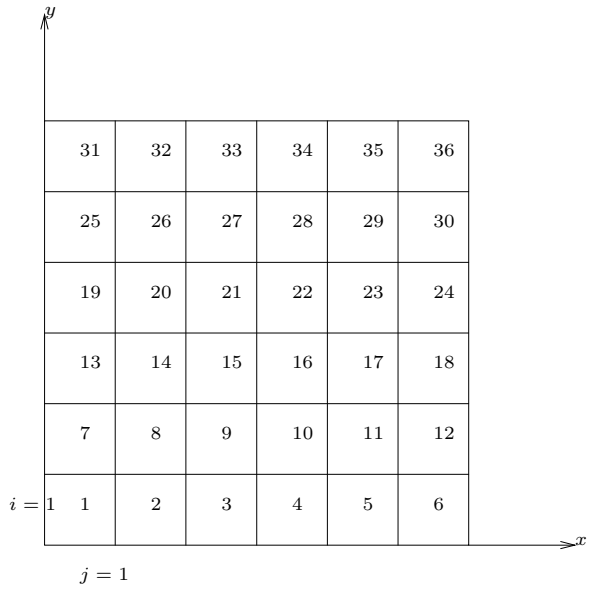


FIGURE 1.2: Ordre lexicographique des inconnues, exemple dans le cas $M = 6$

Les coefficients de $A = (a_{k,\ell})_{k,l=1,n}$ peuvent être calculés de la manière suivante :

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

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

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

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

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

où Id désigne la matrice identité d'ordre M, et 0 la matrice nulle d'ordre M.

Matrices monotones, matrices inversibles dont l'inverse est à coefficients positifs 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 évocateur... Dans ce cours, on lui préfèrera le terme "à inverse à coefficients positifs", ou ICP-matrice ; en effet, on montre à la proposition 1.11 qu'une matrice A est monotone si et seulement si elle est inversible et que son inverse a tous ses coefficients positifs.

Définition 1.10 (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 :

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

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

Proposition 1.11 (Caractérisation des matrices monotones, ICP-matrice)**.** *Une matrice A est monotone si et seulement si elle inversible et à inverse à coefficients positifs (ou ICP) (c.à.d. dont l'inverse a tous ses coefficients sont positifs).*

La démonstration de ce résultat est l'objet de l'exercice 14. Retenez que toute matrice monotone est une ICP-matrice. Cette propriété de monotonie peut être 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. Elle est aussi importante pour montrer que des bornes physiques du modèle sont respectées par le schéma numérique.

1.2.3 Exercices (matrices, exemples)

Exercice 1 (A faire sans calcul !)**.** Effectuer le produit matriciel

$$\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

Exercice 2 (Permutations et matrices)**.** Pour $n \geq 1$, on note Σ_n l'ensemble des bijections de $\{1, \ldots, n\}$ dans lui-même (ces bijections s'appellent des permutations), et pour $i = 1, \ldots,$ on note $E_i \in \mathcal{M}_{n,1}(\mathbb{R})$ la matrice colonne dont tous les éléments sont nuls sauf le i-ème, qui est égal à 1. A tout élément $\sigma \in \Sigma_n$, on associe la matrice $P_\sigma \in \mathcal{M}_n(\mathbb{R})$ dont les colonnes sont $E_{\sigma(1)}, \ldots, E_{\sigma(n)}$.

1. Dans cette question seulement, on suppose $n = 2$. Ecrire toutes les matrices de la forme P_σ.
2. Même question avec $n = 3$.
3. Montrer que pour tout $\sigma \in \Sigma_n$, P_σ est une matrice de permutation.
4. Montrer que si P est une matrice de permutation, alors il existe $\sigma \in \Sigma_n$ tel que $P = P_\sigma$.
5. Montrer que
$$P_\sigma \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x_{\sigma^{-1}(1)} \\ \vdots \\ x_{\sigma^{-1}(n)} \end{bmatrix}.$$
6. Montrer que si $\sigma, \tilde{\sigma} \in \Sigma_n$, alors $P_\sigma P_{\tilde{\sigma}} = P_{\sigma \circ \tilde{\sigma}}$. En déduire que le produit de 2 matrices de permutation est une matrice de permutation.
7. Montrer que pour tout $\sigma \in \Sigma_n$, $P_{\sigma^{-1}} = (P_\sigma)^t$. En déduire que toute matrice de permutation est inversible, d'inverse sa transposée.

Exercice 3 (Théorème du rang)**.** *Corrigé en page 23.*
Soit $A \in \mathcal{M}_{n,p}(\mathbb{R})$ $(n, p \geq 1)$. On rappelle que $\ker(A) = \{\boldsymbol{x} \in \mathbb{R}^p;\ A\boldsymbol{x} = 0\}$, $\mathrm{Im}(A) = \{A\boldsymbol{x},\ \boldsymbol{x} \in \mathbb{R}^p\}$ et $\mathrm{rang}(A) = \dim(\mathrm{Im}(A))$. Noter que $\ker(A) \subset \mathbb{R}^p$ et $\mathrm{Im}(A) \subset \mathbb{R}^n$.
Soit $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_r$ une base de $\mathrm{Im}(A)$ (donc $r \leq n$) et, pour $i \in \{1, \ldots, r\}$, $\boldsymbol{a}_i$ tel que $A\boldsymbol{a}_i = f_i$.

1. Montrer que la famille $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_r$ est une famille libre de $\mathbb{R}^p$ (et donc $r \leq p$).
2. On note G le sous espace vectoriel de $\mathbb{R}^p$ engendré par $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_r$. Montrer que $\mathbb{R}^p = G \oplus \ker(A)$. En déduire que (théorème du rang)
$$p = \dim(\ker(A)) + \dim(\mathrm{Im}(A)).$$

3. On suppose ici que $n = p$. Montrer que l'application $\boldsymbol{x} \mapsto A\boldsymbol{x}$ (de $\mathbb{R}^n$ dans $\mathbb{R}^n$) est injective si et seulement si elle est surjective.

Exercice 4 (rang(A)=rang(A^t))**.** *Corrigé en page 23.*
Soit $A \in \mathcal{M}_{n,p}(\mathbb{R})$ $(n, p \geq 1)$.

1. Soient P une matrice inversible de $\mathcal{M}_n(\mathbb{R})$ et Q une matrice inversible de $\mathcal{M}_p(\mathbb{R})$. Montrer que $\dim(\mathrm{Im}(PA)) = \dim(\mathrm{Im}(AQ)) = \dim(\mathrm{Im}(A))$. Montrer aussi que les matrices P^t et Q^t sont inversibles.

Soit $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_r$ une base de $\mathrm{Im}(A)$ (donc $r \leq p$) et, pour $i \in \{1, \ldots, r\}$, $\boldsymbol{a}_i$ tel que $A\boldsymbol{a}_i = \boldsymbol{f}_i$. Soit $\boldsymbol{a}_{r+1}, \ldots, \boldsymbol{a}_p$ une base de $\ker(A)$ (si $\ker(A) \neq \{0\}$). La famille $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ est une base de $\mathbb{R}^p$ (voir question 1. de l'exercice 3). De même, on complète (si $r < n$) $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_r$ par $\boldsymbol{f}_{r+1}, \ldots, \boldsymbol{f}_n$ de manière à avoir une base $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n$ de $\mathbb{R}^n$.

2. Montrer qu'il existe deux matrices $P \in \mathcal{M}_p(\mathbb{R})$ et $Q \in \mathcal{M}_n(\mathbb{R})$ telles que $P\boldsymbol{e}_i = \boldsymbol{a}_i$ (pour tout $i = 1, \ldots, p$) et $Q\boldsymbol{f}_j = \bar{\boldsymbol{e}}_j$ (pour tout $j = 1, \ldots, n$) ou $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p$ est la base canonique de $\mathbb{R}^p$ et $\bar{\boldsymbol{e}}_1, \ldots, \bar{\boldsymbol{e}}_n$ est la base canonique de $\mathbb{R}^n$. Montrer que P et Q sont inversibles.

On pose $J = QAP$.

3. calculer les colonnes de J et de J^t et en déduire que les matrices J et J^t sont de même rang.
4. Montrer que A et A^t sont de même rang.
5. On suppose maintenant que $n = p$. Montrer que les vecteurs colonnes de A sont liés si et seulement si les vecteurs lignes de A sont liés.

Exercice 5 (Décomposition de $\mathbb{R}^n$ à partir d'une matrice)**.** Soit $n \geq 1$ et $A \in \mathcal{M}_n(\mathbb{R})$.

1. On suppose que la matrice A est diagonalisable. Montrer que $\mathbb{R}^n = \ker(A) \oplus \mathrm{Im}(A)$.
2. Donner un exemple pour lequel $\mathbb{R}^n \neq \ker(A) \oplus \mathrm{Im}(A)$ (on pourra se limiter au cas $n = 2$).

Exercice 6 (Vrai ou faux ? Motiver les réponses…)**.** *Suggestions en page 22, corrigé en page 24*

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 (comptées avec leur multiplicité et éventuellement complexes).
7. Soit A une matrice carrée telle que $A\boldsymbol{x} = \boldsymbol{0} \Longrightarrow \boldsymbol{x} = \boldsymbol{0}$, alors A est inversible.
8. Soit A une matrice carrée telle que $A\boldsymbol{x} \geq \boldsymbol{0} \Longrightarrow \boldsymbol{x} \geq \boldsymbol{0}$, alors A est inversible.
9. Une matrice symétrique est inversible.
10. Une matrice symétrique définie positive est inversible.
11. Le système linéaire
$$\sum_{j=1}^{n+1} a_{i,j} x_j = 0 \text{ pour tout } i = 1, \ldots, n$$
admet toujours une solution non nulle.
12. La fonction $A \mapsto A^{-1}$ est continue de $GL_n(\mathbb{R})(\mathbb{R})$ dans $GL_n(\mathbb{R})(\mathbb{R})$ ($GL_n(\mathbb{R})$ désigne l'ensemble des matrices carrées inversibles d'ordre n).

Exercice 7 (Sur quelques notions connues)**.** *Corrigé en page 24*

1. Soit A une matrice carrée d'ordre n et $\boldsymbol{b} \in \mathbb{R}^n$. 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 $AB = \mathrm{Id}$ et $BC = \mathrm{Id}$, alors $A = C$.
3. Combien y a -t-il 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 $B^{1024} = \mathrm{Id}$.

Exercice 8 (A propos de $BB^t = I$)**.** Pour $n \geq 1$, on note I_n la matrice identité d'ordre n.

1. Existe-t-il $B \in \mathcal{M}_{2,1}(\mathbb{R})$ telle que $BB^t = I_2$ (justifier la réponse) ?
2. Soit $n > 2$, Existe-t-il $B \in \mathcal{M}_{n,1}(\mathbb{R})$ telle que $BB^t = I_n$ (justifier la réponse) ?

Exercice 9 (Matrices symétriques)**.** Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique. Pour tout $k = 1, \ldots, n$, on note A_k la matrice principale d'ordre k de la matrice A.

1. Dans cette question seulement, on suppose que A est symétrique définie positive. Prouver que, pour tout $k = 1, \ldots, n$, la matrice A_k est symétrique définie positive.
2. On note $\underline{\lambda}_k$ la plus petite valeur propre de A_k et $\overline{\lambda}_k$ sa plus grande valeur propre. Prouver que la suite $(\underline{\lambda}_k)_{k=1,\ldots,n}$ est décroissante et que la suite $(\overline{\lambda}_k)_{k=1,\ldots,n}$ est croissante . [*Indication : on pourra décomposer les vecteurs propres de A_k, après prolongement par 0, dans une base de vecteurs propres de A_{k+1}*].

Exercice 10 (La matrice K_3)**.** *Suggestions en page 22. Corrigé en page 25*
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 problèmes 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. On suppose que u est de classe C^4 (et donc f est de calsse C^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. Lorsque f est la fonction identiquement égale à 1, 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 solutions, soit pas de solution.

(b) Ecrire la discrétisation du problème (1.18), toujours avec $h = \frac{1}{4}$, sous la forme $\widetilde{K}\boldsymbol{u} = \widetilde{\boldsymbol{b}}$ en explicitant $\widetilde{K}$ et $\widetilde{\boldsymbol{b}}$.

(c) Montrer que la matrice $\widetilde{K}$ n'est pas inversible : on part d'un problème continu mal posé, et on obtient par discrétisation un problème discret mal posé...

Exercice 11 (Matrices symétriques définies positives)**.** *Suggestions en page 22, corrigé en page 26.*
On rappelle que toute matrice $A \in \mathcal{M}_n(\mathbb{R})$ symétrique est diagonalisable dans $\mathbb{R}$ (cf. lemme 1.8 page 10). 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 $\{\boldsymbol{f}_1, \dots, \boldsymbol{f}_n\}$, et il existe $\lambda_1, \dots, \lambda_n \in \mathbb{R}$ t.q. $A\boldsymbol{f}_i = \lambda_i \boldsymbol{f}_i$, pour tout $i \in \{1, \dots, n\}$, et $\boldsymbol{f}_i \cdot \boldsymbol{f}_j = \delta_{i,j}$ pour tout $i, j \in \{1, \dots, n\}$ ($x \cdot y$ 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 12 (Résolution d'un système sous forme particulière)**.** *Suggestions en page 22.*
Soit $n \geq 1$, $p \geq 1$, $A \in \mathcal{M}_n(\mathbb{R})$ et $B \in \mathcal{M}_{n,p}(\mathbb{R})$. On suppose que A est une matrice symétrique définie positive et que $\text{rang}(B) = p$ (justifier que ceci implique que $p \leq n$).
Pour $i \in \{1, \dots, p\}$, on pose $\boldsymbol{z}_i = A^{-1}B\boldsymbol{e}_i$ où $\boldsymbol{e}_1, \dots, \boldsymbol{e}_p$ désigne la base canonique de $\mathbb{R}^p$ ($B\boldsymbol{e}_i$ est donc la i-ieme colonne de B).

1. Montrer que $\{B\boldsymbol{e}_i, i \in \{1, \dots, p\}\}$ est une base de $\text{Im}(B)$.
2. Montrer que A^{-1} est une matrice symétrique définie positive et que $\ker(B^tA^{-1}B) = \ker(B) = \{0\}$. En déduire que $\{B^t\boldsymbol{z}_i, i \in \{1, \dots, p\}\}$ est une base de $\mathbb{R}^p$.

Soient $\boldsymbol{b} \in \mathbb{R}^n$ et $\boldsymbol{c} \in \mathbb{R}^p$. On cherche le couple $(\boldsymbol{x}, \boldsymbol{y})$, avec $\boldsymbol{x} \in \mathbb{R}^n$ et $\boldsymbol{y} \in \mathbb{R}^p$, solution du système suivant (écrit sous forme de blocs) :

$$\begin{bmatrix} A & B \\ B^t & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{c} \end{bmatrix}. \tag{1.19}$$

On pose $\boldsymbol{u} = A^{-1}\boldsymbol{b}$ et on note $y_1, \dots, y_p$ les composantes de $\boldsymbol{y}$.

3. Montrer que $(\boldsymbol{x}, \boldsymbol{y})$ est solution de (1.19) si et seulement si

$$\sum_{i=1}^{p} y_i B^t \boldsymbol{z}_i = B^t\boldsymbol{u} - \boldsymbol{c}, \tag{1.20}$$

$$\boldsymbol{x} = \boldsymbol{u} - \sum_{i=1}^{p} y_i \boldsymbol{z}_i. \tag{1.21}$$

En déduire que le système (1.19) a une unique solution.

4. Montrer que la matrice (symétrique) $\begin{bmatrix} A & B \\ B^t & 0 \end{bmatrix}$ est inversible mais n'est pas symétrique définie positive.

Exercice 13 (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 1.8 page 10, qu'il existe une base de E, notée $\{\boldsymbol{f}_1, \dots, \boldsymbol{f}_n\}$ et une famille $\{\lambda_1, \dots, \lambda_n\} \subset \mathbb{R}$ telles que $T^{-1}S\boldsymbol{f}_i = \lambda_i \boldsymbol{f}_i$ pour tout $i \in \{1, \dots, n\}$ et t.q. $(T\boldsymbol{f}_i, \boldsymbol{f}_j) = \delta_{i,j}$ pour tout $i, j \in \{1, \dots, n\}$.

Exercice 14 (ICP-matrice)**.** *Corrigé en page 27*

L'objet de cet exercice est de démontrer la proposition 1.11. 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 rappelle qu'une A est une matrice est monotone (voir définition 1.10 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 $A = (a_{i,j})_{i,j=1,\dots,n} \in \mathcal{M}_n(\mathbb{R})$. Montrer que A est une matrice monotone si et seulement si A est une ICP-matrice, *i.e.* 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 ICP-matrice si et seulement si :

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

   En déduire que les matrices $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ et $\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$ sont des ICP-matrices.
3. Montrer que si $A \in \mathcal{M}_n(\mathbb{R})$ est une ICP-matrice alors A^t (la transposée de A) est une ICP-matrice.
4. Montrer que si A est telle que

$$a_{i,j} \leq 0, \text{ pour tout } i, j = 1, \dots, n,\ i \neq j, \text{ et } a_{i,i} > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{i,j}|, \text{ pour tout } i = 1, \dots, n, \tag{1.23}$$

   alors A est une ICP-matrice ; en déduire que si A^t satisfait (1.23), alors A est une ICP-matrice.
5. Soit A une matrice **inversible** telle que

$$a_{i,j} \leq 0, \text{ pour tout } i, j = 1, \dots, n,\ i \neq j, \text{ et } a_{i,i} \geq \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{i,j}|, \text{ pour tout } i = 1, \dots, n. \tag{1.24}$$

   Pour tout $\varepsilon \geq 0$, on définit la matrice $A_\varepsilon = A + \varepsilon \mathrm{Id}$, où Id désigne la matrice identité.
   (a) Prouver que, pour $\varepsilon > 0$, la matrice A_ε est une ICP-matrice.
   (b) Prouver que la matrice A_ε est inversible pour tout $\varepsilon \geq 0$, et que les coefficients de A_ε^{-1} sont des fonctions continues de ε.
   (c) En déduire que A est une ICP-matrice.
6. Montrer que si $A \in \mathcal{M}_n(\mathbb{R})$ est une ICP-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 $\{x \in \mathbb{R}^n \text{ t.q. } Ax > 0\} \subset \{x \in \mathbb{R}^n \text{ t.q. } x > 0\}$ et ne pas être une ICP-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 ICP-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 15 (Matrice du Laplacien discret 1D)**.** *Corrigé détaillé en page 28.*
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 13, 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 16 (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 17 (Réaction diffusion 1d.)**.** *Corrigé détaillé en page 29.*
On s'intéresse à la discrétisation par Différences Finies du problème aux limites suivant :

$$\begin{array}{l} -u''(x) + u(x) = f(x),\ x \in]0,1[, \\ u(0) = u(1) = 0. \end{array} \tag{1.25}$$

Soit $n \in \mathbb{N}^\star$. On note $U = (u_j)_{j=1,\ldots,n}$ une "valeur approchée" de la solution u du problème (1.25) 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 18 (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 $[0,1]$?

Exercice 19 (Déterminant d'une matrice sous forme de blocs)**.**
Soient $A \in \mathcal{M}_n(\mathbb{R})$ $(n > 1)$, $b, c \in \mathbb{R}^n$ et $\lambda \in \mathbb{R}$. On s'intéresse à la matrice $\bar{A} \in \mathcal{M}_{n+1}(\mathbb{R})$ définie sous forme de blocs de la manière suivante :

$$\bar{A} = \begin{bmatrix} A & b \\ c^t & \lambda \end{bmatrix} \tag{1.26}$$

On montre dans cet exercice que les deux assertions suivantes sont, sauf cas particuliers, fausses :

A1 $\det(\bar{A}) = \lambda \det(A) - \det(bc^t)$,

A2 $\det(\bar{A}) = \lambda \det(A) - c^t\, b$,

1. Dans cette question, on prend $n \geq 2$, $A = 0$, $b = c$ et on suppose que $b \neq 0$.
   (a) Montrer que $\text{rang}(\bar{A}) \leq 2$ et en déduire que $\bar{A}$ n'est pas inversible.
   (b) En déduire que l'assertion A2 est fausse pour cet exemple.
2. Dans cette question, on suppose que A est symétrique définie positive, $\lambda = 0$, $b = c$ et que $b \neq 0$.
   (a) Montrer que $\bar{A}$ est inversible et que $\text{rang}(bb^t) = 1$.
   (b) En déduire que l'assertion A1 est fausse pour cet exemple.

Exercice 20 (Résolution d'un système linéaire particulier)**.** Soient $A \in \mathcal{M}_n(\mathbb{R})$ une matrice de rang $(n-1)$ et $\boldsymbol{b} \in \mathrm{Im}(A)$. Cet exercice donne une méthode pour calculer une solution $\boldsymbol{u} \in \mathbb{R}^n$ du système linéaire $A\boldsymbol{u} = \boldsymbol{b}$.
On se donne $\boldsymbol{a} \notin \ker(A)^\perp$, $\boldsymbol{c} \notin \mathrm{Im}(A)$ et on définit la matrice $\bar{A}$ de $\mathcal{M}_n(\mathbb{R})$ par $\bar{A} = A + \boldsymbol{c}\,\boldsymbol{a}^t$.

1. Montrer que $\dim(\ker(A)) = 1$.
2. Montrer qu'il existe un et un seul $\boldsymbol{u}$ solution de
$$A\boldsymbol{u} = \boldsymbol{b}, \tag{1.27}$$
$$\boldsymbol{u} \cdot \boldsymbol{a} = 0. \tag{1.28}$$
3. Montrer que l'unique solution de (1.27)-(1.28) est aussi l'unique solution de $\bar{A}\boldsymbol{u} = \boldsymbol{b}$. En déduire que $\bar{A}$ est inversible.
4. On suppose dans cette question que A est symétrique.
   (a) Montrer que pour toute matrice carrée M d'ordre n symétrique, on a $\mathrm{Im}(M) \subset \ker(M)^\perp$.
   (b) Montrer $\mathrm{Im}(A) = \ker(A)^\perp$ et en déduire qu'un choix possible est $\boldsymbol{a} = \boldsymbol{c} \notin \mathrm{Im}(A)$. [*Suggestion* : montrer que $\dim(\ker(A)^\perp) = n-1$.]
5. On suppose dans cette question que $\boldsymbol{a} \in \ker(A)^\perp$ et $\boldsymbol{c} \in \mathbb{R}^n$. Montrer que la matrice $\bar{A} = A + \boldsymbol{c}\,\boldsymbol{a}^t$ n'est pas inversible.
6. On suppose dans cette question que $\boldsymbol{c} \in \mathrm{Im}(A)$ et $\boldsymbol{a} \in \mathbb{R}^n$. Montrer que la matrice $\bar{A} = A + \boldsymbol{c}\,\boldsymbol{a}^t$ n'est pas inversible.

1.2.4 Suggestions pour les exercices

Exercice 6 page 17 (Vrai ou faux ?)

1. Considérer la matrice ZZ^t.
12. Ecrire que $A^{-1} = \frac{1}{\det}(A)\,\mathrm{com}(A)^t$ où $\det(A)$ est le déterminant (non nul) de A et $\mathrm{com}(A)$ la comatrice de A.

Exercice 10 page 18 (La matrice K_3)

2. Ecrire le développement de Taylor de $u(x_i + h)$ et $u(x_i - h)$.
3. Pour l'erreur de discrétisation, se souvenir qu'elle dépend de l'erreur de consistance, et regarder sa majoration.
4. Pour tenir compte de la condition limite en 1, écrire un développement limité de $u(1-h)$.
5.1 Distinguer les cas $c = 0$ et $c \neq 0$.

Exercice 11 page 19 (Matrices symétriques définies positives)

3. Utiliser la diagonalisation sur les opérateurs linéaires associés.

Exercice 12 page 19 (Résolution d'un système sous forme particulière)

1. Utiliser le fait que $Im(B)$ est l'ensemble des combinaisons linéaires des colonnes de B.
2. Utiliser le caractère s.d.p. de A puis le théorème du rang.

1.2.5 Corrigés des exercices

Exercice 3 page 16 (Théorème du rang)

1. Soit $\boldsymbol{a}_1, \dots, \boldsymbol{a}_r$ dans $\mathbb{R}$ tel que $\sum_{i=1}^r \alpha_i \boldsymbol{a}_i = 0$. On a donc

$$0 = A(\sum_{i=1}^r \alpha_i \boldsymbol{a}_i) = \sum_{i=1}^r \alpha_i A\boldsymbol{a}_i = \sum_{i=1}^r \alpha_i \boldsymbol{f}_i.$$

   Comme la famille $\boldsymbol{f}_1, \dots, \boldsymbol{f}_r$ est une famille libre, on en déduit que $\alpha_i = 0$ pour tout $i \in \{1, \dots, r\}$ et donc que la famille $\boldsymbol{a}_1, \dots, \boldsymbol{a}_r$ est libre.

2. Soit $\boldsymbol{x} \in \mathbb{R}^p$. Comme $\boldsymbol{f}_1, \dots, \boldsymbol{f}_r$ est une base de $\mathrm{Im}(A)$, il existe $\alpha_1, \dots, \alpha_r$ tel que $A\boldsymbol{x} = \sum_{i=1}^r \alpha_i \boldsymbol{f}_i$. On pose $\boldsymbol{y} = \sum_{i=1}^r \alpha_i \boldsymbol{a}_i$. On a $A\boldsymbol{y} = A\boldsymbol{x}$ et $\boldsymbol{x} = (\boldsymbol{x} - \boldsymbol{y}) + \boldsymbol{y}$. Comme $\boldsymbol{y} \in G$ et $A(\boldsymbol{x} - \boldsymbol{y}) = 0$, on en déduit que $\mathbb{R}^p = G + \ker A$.

   Soit maintenant $\boldsymbol{x} \in \ker A \cap G$. Comme $\boldsymbol{x} \in G$, il existe $\alpha_1, \dots, \alpha_r$ tel que $\boldsymbol{x} = \sum_{i=1}^r \alpha_i \boldsymbol{a}_i$. On a donc $A\boldsymbol{x} = \sum_{i=1}^r \alpha_i \boldsymbol{f}_i$. Comme $\boldsymbol{f}_1, \dots, \boldsymbol{f}_r$ est une famille libre et que $A\boldsymbol{x} = 0$, on en déduit que $\alpha_i = 0$ pour tout $i \in \{1, \dots, r\}$ et donc $\boldsymbol{x} = 0$. Ceci montre que $\mathbb{R}^p = G \oplus \ker(A)$. Enfin, comme $\dim G = r = \dim(\mathrm{Im}A)$, on en déduit bien que $p = \dim(\ker(A)) + \dim(\mathrm{Im}(A))$.

3. On suppose ici $p = n$. Comme $n = \dim(\ker(A)) + \dim(\mathrm{Im}(A))$, on a $\dim(\ker(A)) = 0$ si et seulement si $\dim(\mathrm{Im}(A)) = n$. Ceci montre que l'application $\boldsymbol{x} \mapsto A\boldsymbol{x}$ (de $\mathbb{R}^n$ dans $\mathbb{R}^n$) est injective si et seulement si elle est surjective.

Exercice 4 page 17 (rang(A)=rang(A^t))

1. On remarque tout d'abord que le noyau de PA est égal au noyau de A. En effet, soit $\boldsymbol{x} \in \mathbb{R}^p$. Il est clair que $A\boldsymbol{x} = 0$ implique $PA\boldsymbol{x} = 0$. D'autre part, comme P est inversible, $PA\boldsymbol{x} = 0$ implique $A\boldsymbol{x} = 0$. On a donc bien $\ker(PA) = \ker(A)$. On en déduit que $\dim(\ker(PA)) = \dim(\ker(A))$ et donc, avec le théorème du rang (exercice 3), que $\dim(\mathrm{Im}(PA)) = \dim(\mathrm{Im}(A))$.

   Pour montrer que $\dim(\mathrm{Im}(AQ)) = \dim(\mathrm{Im}(A))$, on remarque directement que $\mathrm{Im}(AQ) = \mathrm{Im}(A)$. En effet, on a, bien sûr, $\mathrm{Im}(AQ) \subset \mathrm{Im}(A)$ (l'inversibilité de Q est inutile pour cette inclusion). D'autre part, si $\boldsymbol{z} \in \mathrm{Im}(A)$, il existe $\boldsymbol{x} \in \mathbb{R}^p$ tel que $A\boldsymbol{x} = \boldsymbol{z}$. Comme Q est inversible, il existe $\boldsymbol{y} \in \mathbb{R}^p$ tel que $\boldsymbol{x} = Q\boldsymbol{y}$. On a donc $\boldsymbol{z} = AQ\boldsymbol{y}$, ce qui prouve que $\mathrm{Im}(A) \subset \mathrm{Im}(AQ)$. Finalement, on a bien $\mathrm{Im}(AQ) = \mathrm{Im}(A)$ et donc $\dim(\mathrm{Im}(AQ)) = \dim(\mathrm{Im}(A))$.

   Pour montrer que P^t est inversible, il suffit de remarquer que $(P^{-1})^t P^t = (PP^{-1})^t = I_n$ (où I_n désigne la matrice Identité de $\mathbb{R}^n$). Ceci montre que P^t est inversible (et que $(P^t)^{-1} = (P^{-1})^t$). Bien sûr, un raisonnement analogue donne l'inversibilité de Q^t.

2. Par définition du produit matrice vecteur, $P\boldsymbol{e}_i = \boldsymbol{c}_i(P)$, i-ème colonne de P ; l suffit de prendre pour P la matrice dont les colonnes sont les vecteurs $\boldsymbol{a}_1, \dots, \boldsymbol{a}_p$; l'image de P est égale à $\mathbb{R}^p$ car la famille $\boldsymbol{a}_1, \dots, \boldsymbol{a}_p$ est une base de $\mathbb{R}^p$, ce qui prouve que P est inversible (on a $\mathrm{Im}(P) = \mathbb{R}^p$ et $\ker P = \{0\}$ par le théorème du rang).

   Soit maintenant $R \in \mathcal{M}_n(\mathbb{R})$ dont les colonnes sont les vecteurs $\boldsymbol{f}_j$; la matrice R est bien inversible car la famille $\boldsymbol{f}_1, \dots, \boldsymbol{f}_n$ est une base $\mathbb{R}^n$. On a donc, toujours par définition du produit matrice vecteur, $R\bar{\boldsymbol{e}}_j = \boldsymbol{c}_j(R) = \boldsymbol{f}_j$ pour $j = 1, n$. Posons $Q = R^{-1}$; on a alors $QR\bar{\boldsymbol{e}}_j = \bar{\boldsymbol{e}}_j = Q\boldsymbol{f}_j$, et la matrice Q est évidemment inversible.

3. Pour $i \in \{1, \dots, p\}$, la i-ème colonne de J est donnée par $\boldsymbol{c}_i(J) = QAPe_i = QA\boldsymbol{a}_i$. Si $i \in \{1, \dots, r\}$, on a donc $\boldsymbol{c}_i(J) = Q\boldsymbol{f}_i = \bar{\boldsymbol{e}}_i$. Si $i \in \{r+1, \dots, p\}$, on a $\boldsymbol{c}_i(J) = 0$ (car $\boldsymbol{a}_i \in \ker A$). Ceci montre que $\mathrm{Im}(J)$ est l'espace vectoriel engendré par $\bar{\boldsymbol{e}}_1, \dots, \bar{\boldsymbol{e}}_r$ et donc que le rang de J est r.

   La matrice J appartient à $\mathcal{M}_{n,p}(\mathbb{R})$, sa transposée appartient donc à $\mathcal{M}_{p,n}(\mathbb{R})$. En transposant la matrice J, on a, pour tout $i \in \{1, \dots, r\}$, $\boldsymbol{c}_i(J^t) = e_i$ et, pour tout $i \in \{r+1, \dots, n\}$, $\boldsymbol{c}_i(J^t) = 0$. Ceci montre que $\mathrm{Im}(J^t)$ est l'espace vectoriel engendré par $\boldsymbol{e}_1, \dots, \boldsymbol{e}_r$ et donc que le rang de J^t est aussi r.

4. Il suffit maintenant d'appliquer la première question, elle donne que le rang que A est le même que le rang de J et, comme $J^t = P^t A^t Q^t$, que le rang que A^t est le même que le rang de J^t. Finalement le rang de A et de A^t est r.
5. Les vecteurs colonnes de A sont liés si et seulement si le rang de A est strictement inférieur à n. Les vecteurs colonnes de A^t sont liés si et seulement si le rang de A^t est strictement inférieur à n. Comme les vecteurs colonnes de A^t sont les vecteurs lignes de A, on obtient le résultat désiré grâce au fait que A et A^t ont même rang.

Exercice 6 page 17 (Vrai ou faux ?)

1. Faux : La matrice ZZ^t est de rang 1 et donc non inversible.
2. Faux : La matrice inverse d'une matrice triangulaire inférieure est triangulaire inférieure.
3. Vrai : le polynôme caractéristique d'une matrice A est le déterminant de $A - \lambda \mathrm{Id}$.
4. Faux : la matrice $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ est inversible et non diagonalisable dans $\mathbb{R}$.
5. Faux : la matrice $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ est inversible et non diagonalisable dans $\mathbb{C}$.
6. Vrai : c'est le terme de degré 0 du polynôme caractéristique.
7. Vrai : si $\mathrm{Ker}(A) = \{0\}$, alors A est inversible.
8. Vrai : on va montrer que $\mathrm{Ker}(A) = \{0\}$, Supposons que $A\boldsymbol{x} = \mathbf{0}$, alors $A\boldsymbol{x} \geq \mathbf{0}$ et $A\boldsymbol{x} \leq \mathbf{0}$, ou encore $A(-\boldsymbol{x}) \geq \mathbf{0}$ Donc par hypothèse, $\boldsymbol{x} \geq \mathbf{0}$ et $-\boldsymbol{x} \geq \mathbf{0}$, et donc $\boldsymbol{x} = 0$, ce qui montre que $\mathrm{Ker}(A) = \{0\}$.
9. Faux : la matrice nulle est symétrique.
10. Vrai : Si A est s.d.p.alors $A\boldsymbol{x} = \mathbf{0}$ entraîne $A\boldsymbol{x} \cdot \boldsymbol{x} = 0$ et donc $\boldsymbol{x} = \mathbf{0}$, ce qui montre que $\mathrm{Ker}(A) = \{0\}$ et donc que A est inversible.
11. Vrai : l'ensemble des solutions est le noyau de la matrice $A \in \mathcal{M}_{n,n+1}(\mathbb{R})$ qui est de dimension au moins un par le théorème du rang.
12. Vrai : on peut écrire que $A^{-1} = \frac{1}{\det}(A)\mathrm{com}(A)^t$ où $\det(A)$ est le déterminant (non nul) de A et $\mathrm{com}(A)$ la comatrice de A, c.à.d. la matrice des cofacteurs des coefficients de A ; on rappelle que le cofacteur $c_{i,j}$ de l'élément $a_{i,j}$ est défini par $c_{i,j} = (-1)^{i+j}\Delta_{i,j}$ où $\Delta_{i,j}$ est le mineur relatif à (i, j), *i.e.* le déterminant de la sous matrice carrée d'ordre $n-1$ obtenue à partir de A en lui retirant sa i-ème ligne et sa j-ème colonne). On peut vérifier facilement que les applications $A \mapsto \det(A)$ et $A \mapsto c_{i,j}$ sont continues de $GL_n(\mathbb{R})(\mathbb{R})$ dans $\mathbb{R}^*$ et $\mathbb{R}$ respectivement (comme polynôme en les éléments de la matrice A), et que donc $A \mapsto A^{-1}$ est continue.

Exercice 7 page 17 (Sur quelques notions connues)

1. Supposons qu'il existe deux solutions distinctes $\boldsymbol{x}_1$ et $\boldsymbol{x}_2$ au système $A\boldsymbol{x} = \boldsymbol{b}$. Soit $\boldsymbol{z} = \boldsymbol{x}_1 - \boldsymbol{x}_2$. On a donc $A\boldsymbol{z} = \mathbf{0}$ et $z \neq 0$.
   — Si A est inversible, on a donc $\boldsymbol{z} = \mathbf{0}$ en contradiction avec $x_1 \neq x_2$.
   — Si A est non inversible, alors $A(t\boldsymbol{z}) = \mathbf{0}$ pour tout $t \in \mathbb{R}$, et donc il y a une infinité de solutions au système $A\boldsymbol{x} = \boldsymbol{b}$.
2. $C = (AB)C = A(BC) = A$.
3. Les matrices carrées d'ordre 2 ont quatre coefficients, et donc il y a $2^4 = 16$ matrices ne comportant que des 1 ou des 0 comme coefficients. Une matrice $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ est inversible si $ad - bc \neq 0$. Dans le cas de matrices ne comportant que des 1 ou des 0 comme coefficients, les valeurs non nulles possibles de $ad - bc$ sont 1 et -1, obtenues respectivement pour ($ad = 1$, $bc = 0$) et ($ad = 0$, $bc = 1$), c.à.d pour les matrices

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$$

et

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$$

4. Les valeurs propres de B sont i et $-i$ (car la trace de B est nulle et son déterminant est égal à 1). Donc $B^{1024} = \mathrm{Id}$

Exercice 10 page 18 (La matrice K_3)

1. La solution est $-\frac{1}{2}x(x-1)$, qui est effectivement positive.
2. Avec les développements limités vus en cours, on obtient :

$$K_3 = \frac{1}{h^2}\begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix},\ \boldsymbol{b} = \begin{bmatrix} f(h) \\ f(2h) \\ f(3h) \end{bmatrix},\ \text{où } h = \frac{1}{4}$$

3. L'échelonnement du système $K_3\boldsymbol{x} = \boldsymbol{b}$ sur la matrice augmentée (ou la méthode de Gauss) donne :

$$\frac{1}{h^2}\left[\begin{array}{ccc|c} 2 & -1 & 0 & b_1 \\ -0 & \frac{3}{2} & -1 & b_2 + \frac{1}{2}b_1 \\ 0 & 0 & \frac{4}{3} & b_3 + \frac{2}{3}b_2 + \frac{1}{3}b_1 \end{array}\right]$$

Donc pour $h = \frac{1}{4}$ et $b_1 = b_2 = b_3 = 1$ on obtient

$$u_1 = \frac{3}{32}, u_2 = \frac{1}{8} \text{ et } u_3 = \frac{3}{32}.$$

On a $u_i = u(x_i)$, ce qui veut dire que l'erreur de discrétisation est nulle. On a vu en cours (formule (1.8)) que l'erreur de consistance R peut être majorée par $\frac{h^2}{12}\|u^{(4)}\|_\infty$.. Ici u est un polynôme de degré 2, et donc $R = 0$. Or par l'inégalité (1.12), l'erreur de discrétisation $\boldsymbol{e} = (u(x_1) - u_1, u(x_2) - u_2, u(x_3) - u_3)^t$ satisfait $\boldsymbol{e} = K_3^{-1}R$. On en déduit que cette erreur de discrétisation est nulle.
Notons qu'il s'agit là d'un cas tout à fait particulier dû au fait que la solution exacte est un polynôme de degré inférieur ou égal à 3.

4. Avec la condition limite (1.17), la solution exacte du problème pour $f \equiv 1$ est maintenant $u(x) = -\frac{1}{2}x(x-2)$.
Pour prendre en compte la condition limite (1.17), on effectue un développement limité de u à l'ordre 2 en $x = 1$

$$u(1-h) = u(1) - hu'(1) + \frac{1}{2}h^2u''(\zeta) \text{ avec } \zeta \in [1-h, 1].$$

Les inconnues discrètes sont maintenant les valeurs approchées recherchées aux points x_i, $i \in \{1,2,3,4\}$, notées u_i, $i \in \{1,2,3,4\}$. Comme $u'(1) = 0$, l'égalité précédente suggère de prendre comme équation discrète $u_3 = u_4 - (1/2)f(1)$ (on rappelle que $x_4 = 1$).
Le système discret à résoudre est donc :

$$\begin{aligned} 2u_1 - u_2 &= h^2 f(x_1), \\ -u_1 + 2u_2 - u_3 &= h^2 f(x_2) \\ -u_2 + 2u_3 - u_4 &= h^2 f(x_3) \\ -u_3 + u_4 &= \frac{1}{2}h^2 f(x_4) \end{aligned}$$

Le système linéaire à résoudre est donc $K\boldsymbol{u} = \boldsymbol{b}$, avec

$$K = \frac{1}{h^2}\begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix},\ \boldsymbol{b} = \begin{bmatrix} f(h) \\ f(2h) \\ f(3h) \\ \frac{1}{2}f(4h) \end{bmatrix}.$$

En notant $b_i = f(x_i)$, l'échelonnement du système $h^2 K\boldsymbol{x} = h^2\boldsymbol{b}$ sur la matrice augmentée donne :

$$\left[\begin{array}{cccc|c} 2 & -1 & 0 & 0 & h^2 b_1 \\ 0 & \frac{3}{2} & -1 & 0 & h^2(b_2 + \frac{1}{2}b_1) \\ 0 & 0 & \frac{4}{3} & -1 & h^2(b_3 + \frac{2}{3}b_2 + \frac{1}{3}b_1) \\ 0 & 0 & 0 & \frac{1}{4} & h^2(\frac{1}{2}b_4 + \frac{1}{2}b_2 + \frac{1}{4}b_1 + \frac{3}{4}b_3) \end{array}\right]$$

Donc pour $h = \frac{1}{4}$ et $b_1 = b_2 = b_3 = b_4 = 1$ on obtient

$$u_1 = \frac{7}{32}, u_2 = \frac{3}{8}, u_3 = \frac{15}{32} \text{ et } u_4 = \frac{1}{2}.$$

La solution exacte aux points de discrétisation est :

$$u(x_1) = \frac{1}{2}\frac{1}{4}(2 - \frac{1}{4}) = \frac{7}{32}, u(x_2) = \frac{1}{2}\frac{1}{2}(2 - \frac{1}{2}) = \frac{3}{8}, u(x_3) = \frac{1}{2}\frac{3}{4}(2 - \frac{3}{4}) = \frac{15}{32}, u(x_4) = \frac{1}{2}.$$

On a donc $u(x_i) = u_i$ pour tout $i \in \{1, 2, 3, 4\}$, ce qu'on aurait pu deviner sans calculs car ici aussi l'erreur de discrétisation est nulle car l'erreur de consistance est nulle en raison du traitement que nous avons fait de la condition aux limites de Neumann ($u'(1) = 0$) et du fait que la solution exacte est un polynôme de degré au plus égal à 2.

5.

(a) Il est facile de voir que si $c \neq 0$, aucune fonction ne peut satisfaire le problème (1.18), alors que si $c = 0$, toutes les fonctions constantes conviennent.

(b) On a maintenant une condition de Neumann en 0 et en 1.
Un raisonnement similaire aux questions précédentes nous conduit à introduire 5 inconnues discrètes u_i, $i \in \{1, \ldots, 5\}$. Le système à résoudre est maintenant :

$$\widetilde{K} = \frac{1}{h^2}\begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 1 \end{bmatrix},\ \widetilde{\boldsymbol{b}} = \begin{bmatrix} \frac{1}{2}f(0) \\ f(h) \\ f(2h) \\ f(3h) \\ \frac{1}{2}f(4h) \end{bmatrix}.$$

(c) La matrice $\widetilde{K}$ n'est pas inversible car la somme de ses colonnes est égale au vecteur nul : on part d'un problème continu mal posé, et on obtient effectivement par discrétisation un problème discret mal posé.

‘

Exercice 11 page 19 (Matrices symétriques définies positives)

1. On note $e_1, \ldots, e_n$ la base canonique de $\mathbb{R}^n$. Pour tout $i \in \{1, \ldots, n\}$, on a $a_{i,i} = Ae_i \cdot e_i$ et donc, comme A est définie positive, on en déduit $a_{i,i} > 0$.

2. On utilise le rappel donné dans l'énoncé. Les λ_i sont les valeurs propres de A. Soit $x \in \mathbb{R}^n$, décomposons x sur la base orthonormée $(\boldsymbol{f}_i)_{i=1,n}$: $x = \sum_{i=1}^n \alpha_i \boldsymbol{f}_i$. On a donc :

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

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. On note T l'application (linéaire) de $\mathbb{R}^n$ dans $\mathbb{R}^n$ définie par $T(x) = Ax$. On prouve tout d'abord l'existence de B. 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 $B^2 = A$.
Pour montrer l'unicité de B, on peut remarquer que, si B^2=A, on a, pour tout $i \in \{1, \ldots, n\}$,

$$(B + \sqrt{\lambda_i}I)(B - \sqrt{\lambda_i}I)f_i = (B^2 - \lambda_i I)f_i = (A - \lambda_i I)f_i = 0,$$

où I désigne la matrice identité. On a donc $(B - \sqrt{\lambda_i}I)f_i \in \ker(B + \sqrt{\lambda_i}I)$. Mais, comme B est s.d.p., les valeurs propres de B sont des réels strictement positifs, on a donc $\ker(B + \sqrt{\lambda_i}I) = \{0\}$ et donc $Bf_i = \sqrt{\lambda_i}f_i$. Ce qui détermine complètement B.

Exercice 14 page 20 (ICP-matrice)

1. Supposons d'abord que A est une ICP-matrice, c.à.d. 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 matrice monotone, 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 $A^{-1}e_i \geq 0$, 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

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

   Dans le premier cas, on a forcément $bc \neq 0$: en effet sinon on aurait $ad < 0$, or $a \leq 0$ et $d \leq 0$ donc $ad \geq 0$.
   Dans le second cas, on a forcément $ad \neq 0$: en effet sinon on aurait $bc < 0$, or $b \leq 0$ et $c \leq 0$ donc $bc \geq 0$.
   Les conditions précédentes sont donc équivalentes aux conditions (1.22).
3. La matrice A^t est une ICP-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 (1.23), 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. (a) Puisque la matrice A vérifie l'hypothèse (1.24) et puisque $\varepsilon > 0$, la matrice A_ε vérifie l'hypothèse (1.23), et c'est donc une ICP-matrice par la question précédente.

   (b) Pour $\varepsilon > 0$, la matrice A_ε est une ICP-matrice donc inversible, et pour $\varepsilon = 0$, $A_\varepsilon = A$ et A est inversible par hypothèse. La fonction $\varepsilon \mapsto A + \varepsilon \mathrm{Id}$ est continue de $\mathbb{R}$ dans $\mathcal{M}_n(\mathbb{R})$, et la fonction $M \mapsto M^{-1}$ est continue de $\mathcal{M}_n(\mathbb{R})$ dans $\mathcal{M}_n(\mathbb{R})$. Par composition, les coefficients de A_ε^{-1} sont donc des fonctions continues de ε.

   (c) Comme la matrice A_ε est une ICP-matrice, les coefficients de A_ε^{-1} sont tous positifs ou nuls. Par continuité, les coefficients de A^{-1} sont donc aussi tous positifs ou nuls, et donc A est une ICP-matrice.

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 ICP-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 ICP-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 $Tg > 0$.

Exercice 15 page 21 (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 64 ; 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 11).

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

$$\begin{aligned} h^2 Ax \cdot x &= 2x_1^2 - x_1x_2 - \sum_{i=2}^{n-1} \left(x_i x_{i-1} + 2x_i^2\right) - \sum_{i=3}^{n} x_i x_{i-1} + 2x_n^2 - x_{n-1}x_n \\ &= \sum_{i=1}^{n} x_i^2 + \sum_{i=2}^{n} x_{1-i}^2 + x_n^2 - 2\sum_{i=1}^{n} x_i x_{i-1} \\ &= \sum_{i=2}^{n} (x_i - x_{i-1})^2 + x_1^2 + x_n^2 \geq 0. \end{aligned}$$

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 17 page 21 (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 $A = (N+1)^2 K_n + \mathrm{Id}$, Id désigne la matrice identité et

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

1.3 Les méthodes directes

1.3.1 Définition

Définition 1.12 (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 : addition, soustraction, mutiplication, division, et extraction de racine carrée pour la methode de choleski.*

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 $\boldsymbol{b} \in \mathbb{R}^n$. On cherche à calculvoir à ce sujeter $\boldsymbol{x} \in \mathbb{R}^n$ tel que $A\boldsymbol{x} = \boldsymbol{b}$.
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 $A\boldsymbol{x} = \boldsymbol{b}$, avec

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

On écrit la **matrice augmentée**, constituée de la matrice A et du second membre $\boldsymbol{b}$.

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

Gauss et opérations matricielles On pose $A^{(1)} = A$, $\boldsymbol{b}^{(1)} = \boldsymbol{b}$ et $\widetilde{A}^{(1)} = A$, $\boldsymbol{b}^{(1)} = \boldsymbol{b}$.
La première ligne a un 1 en première position (en gras dans la matrice), ce coefficient est non nul, et c'est ce qu'on appelle 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 (ce qui revient à multiplier la matrice A par $E_2^{(1)} = \mathrm{Id}$). On veut ensuite annuler le premier coefficient de la troisième ligne. On retranche donc (-1) fois la première ligne à la troisième [2] :

$$\begin{bmatrix} \mathbf{1} & 0 & 1 & 2 \\ 0 & 2 & -1 & 1 \\ -1 & 1 & -2 & -2 \end{bmatrix} \overset{\boldsymbol{\ell}_3 \leftarrow \boldsymbol{\ell}_3 + \boldsymbol{\ell}_1}{\longrightarrow} \begin{bmatrix} \mathbf{1} & 0 & 1 & 2 \\ 0 & \mathbf{2} & -1 & 1 \\ 0 & 1 & -1 & 0 \end{bmatrix}$$

Ceci revient à multiplier la matrice $\tilde{A}$ à gauche par la matrice

$$E_3^{(1)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}.$$

On appelle *matrices d'élimination* les matrices $E_2^{(1)}$ et $E_3^{(1)}$.
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 :

$$\begin{bmatrix} \mathbf{1} & 0 & 1 & 2 \\ 0 & \mathbf{2} & -1 & 1 \\ 0 & 1 & -1 & 0 \end{bmatrix} \overset{\boldsymbol{\ell}_3 \leftarrow \boldsymbol{\ell}_3 - 1/2\boldsymbol{\ell}_2}{\longrightarrow} \begin{bmatrix} \mathbf{1} & 0 & 1 & 2 \\ 0 & \mathbf{2} & -1 & 1 \\ 0 & 0 & -\mathbf{\frac{1}{2}} & -\frac{1}{2} \end{bmatrix}.$$

Ceci revient à multiplier la matrice précédente à gauche par la matrice d'élimination

$$E_3^{(2)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -\frac{1}{2} & 1 \end{bmatrix}.$$

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.
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

$$A\boldsymbol{x} = \boldsymbol{b} \text{ en } U\boldsymbol{x} = E_2 E_1 b, \text{ où } U = E_2 E_1 A$$

est une matrice triangulaire supérieure.

2. 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.

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[3]. 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).

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 donné au th{eorème 1.22 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 1.22.

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 $\boldsymbol{b}^{(1)} = \boldsymbol{b}$. 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,

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

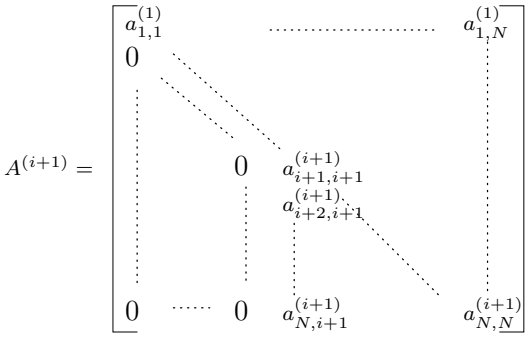


FIGURE 1.3: Allure de la matrice de Gauss à l'étape $i+1$

il faut également modifier le second membre $\boldsymbol{b}$ en conséquence, donc on peut effectuer les manipulations sur la matrice augmentée $\tilde{A}^{(i)} = \left[A^{(i)} \quad \boldsymbol{b}^{(i)}\right]$. L'étape de factorisation et descente s'écrit donc de la manière suivante : pour $k > i$, **si** $a_{i,i}^{(i)} \neq 0$, on pose :

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

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

ce qui revient à multiplier à gauche la matrice augmentée $\tilde{A}^{(i)}$ par la matrice $E^{(i)}$ dont l'expression, et celle de son inverse, sont données par

$$E^{(i)} = \begin{bmatrix} 1 & 0 & \ldots & & & & & 0 \\ 0 & 1 & 0 & \ldots & & & & 0 \\ 0 & 0 & 1 & 0 & \ldots & & & 0 \\ & & \ldots & & & & & \\ & & \ldots & & & & & \\ 0 & 0 & \ldots & 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 0 & \ldots & -\ell_{i+1,i} & 1 & \ldots & 0 \\ & & & & & & & \\ & & \ldots & & & & & \\ 0 & 0 & 0 & \ldots & -\ell_{n,i} & 0 & \ldots & 1 \end{bmatrix} \quad (E^{(i)})^{-1} = \begin{bmatrix} 1 & 0 & \ldots & & & & & 0 \\ 0 & 1 & 0 & \ldots & & & & 0 \\ 0 & 0 & 1 & 0 & \ldots & & & 0 \\ & & \ldots & & & & & \\ & & \ldots & & & & & \\ 0 & 0 & \ldots & 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 0 & \ldots & \ell_{i+1,i} & 1 & \ldots & 0 \\ & & & & & & & \\ & & \ldots & & & & & \\ 0 & 0 & 0 & \ldots & \ell_{n,i} & 0 & \ldots & 1 \end{bmatrix}$$

On obtient donc, en posant $L^{(i+1)} = L^{(i)}(E^{(i)})^{-1}$ et $A^{(i+1)} = E^{(i)}A^{(i)}$,

$$A = L^{(i)}(E^{(i)})^{-1}E^{(i)}A^{(i)} = L^{(i+1)}A^{(i+1)} \text{ et } A^{(i+1)}x = b^{(i+1)},$$

avec les matrices $L^{(i+1)}$ et $A^{(i+1)}$ données par

$$L^{(i+1)} = \begin{bmatrix} 1 & 0 & \ldots & & & & \\ \ell_{2,1} & 1 & 0 & \ldots & & & \\ \ell_{3,1} & \ell_{3,2} & 1 & 0 & \ldots & & \\ & & \ldots & & & & \\ \ell_{i,1} & \ell_{i,2} & \ldots & 1 & 0 & \ldots & \\ \ell_{i+1,1} & \ell_{i+1,2} & \ldots & \ell_{i+1,i} & 1 & 0 & \ldots \\ & & & & & & \\ & & \ldots & & & & \\ \ell_{n,1} & \ell_{n,2} & \ldots & \ell_{n,i} & 0 & \ldots & 1 \end{bmatrix}$$

$$A^{(i+1)} = \begin{bmatrix} a_{1,1}^{(1)} & a_{1,2}^{(1)} & \dots & & & a_{1,n}^{(1)} \\ 0 & a_{2,2}^{(2)} \dots & & & & a_{2,n}^{(2)} \\ & & \dots & & & \\ \dots & 0 & a_{i,i}^{(i)} & a_{i,i+1}^{(i)} & \dots & a_{i,n}^{(i)} \\ 0 & \dots & 0 & a_{i+1,i+1}^{(i+1)} & \dots & a_{i+1,n}^{(i+1)} \\ & & \dots & & & \\ 0 & \dots & 0 & a_{n,i+1}^{(i+1)} & \dots & a_{n,n}^{(i+1)} \end{bmatrix} \qquad b^{(i+1)} = \begin{bmatrix} b_1^{(1)} \\ b_2^{(2)} \\ \dots \\ b_i^{(i)} \\ b_{i+1}^{(i+1)} \\ \dots \\ \\ b_n^{(i+1)} \end{bmatrix}$$

La matrice $A^{(i+1)}$ est de la forme annoncée. Elle vérifie les propriétés suivantes :

1. $a_{k,j}^{(i+1)} = 0$ pour tout $j = 1, \dots, i$ et $k > j$,
2. le système $A^{(i+1)}\boldsymbol{x} = \boldsymbol{b}^{(i+1)}$ est bien équivalent au système $A^{(i)}\boldsymbol{x} = \boldsymbol{b}^{(i)}$,
3. la matrice $A^{(n)}$ est triangulaire supérieure.

La matrice $L^{(i+1)}$ est également de la forme annoncée. Elle vérifie les propriétés suivantes :

1. la matrice $L^{(i+1)}$ est triangulaire inférieure, avec tous les coefficients de la diagonale égaux à 1,
2. les colonnes 1 à $i-1$ de la matrice $L^{(i+1)}$ sont celles de la matrice $L^{(i)}$, car elles n'ont pas modifiées par la multiplication à droite par $(E^{(i)})^{-1}$,
3. la colonne i de la matrice $L^{(i+1)}$ est celle de la matrice $E^{(i)}$,
4. et on a bien $A = L^{(i+1)}A^{(i+1)}$.

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 $A^{(n)}\boldsymbol{x} = \boldsymbol{b}^{(n)}$ équivalent au système $A\boldsymbol{x} = \boldsymbol{b}$, 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 $A^{(n)}\boldsymbol{x} = \boldsymbol{b}^{(n)}$; ceci est une étape facile. Comme $A^{(n)}$ est une matrice inversible, on a $a_{i,i}^{(i)} \neq 0$ pour tout $i = 1, \dots, n$, et comme $A^{(n)}$ est une matrice triangulaire supérieure, on peut donc calculer les composantes de $\boldsymbol{x}$ en "remontant", c'est–à–dire de la composante x_n à la composante x_1 :

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

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 1.13 (Gauss sans permutation)**.**

1. *(Factorisation et descente) Pour commencer, on pose $u_{i,j} = a_{i,j}$ et $y_i = b_i$ pour pour $i, j \in \{1, \dots, n\}$. Puis, pour i allant de 1 à $n-1$, on effectue les calculs suivants :*
   (a) *On ne change pas la i-ème ligne (qui est la ligne du pivot)*
   (b) *On modifie les lignes $i+1$ à n et le second membre $\boldsymbol{y}$ en utilisant la ligne i.*
   Pour k allant de $i+1$ à n :
   $\ell_{k,i} = \frac{u_{k,i}}{u_{i,i}}$ *(si $u_{i,i} = 0$, prendre la méthode avec pivot partiel).*
   Pour j allant de $i+1$ à n,
   $u_{k,j} = u_{k,j} - \ell_{k,i} u_{i,j}$

Fin pour

$y_k = y_k - \ell_{k,i} y_i$

Fin pour

2. *(Remontée) On calcule x :*

$x_n = \dfrac{y_n}{u_{n,n}}$

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

$x_i = \dfrac{1}{u_{i,i}} x_i$

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'il 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 $(2/3)n^3$.
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 $\boldsymbol{b}^{(n)}$ et $\boldsymbol{x}$) 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 $A\boldsymbol{x} = \boldsymbol{b}$. En effet, comme $A^{(n)}\boldsymbol{x} = \boldsymbol{b}^{(n)}$ et $\boldsymbol{b}^{(n)} = L^{-1}\boldsymbol{b}$, on en déduit que $LU\boldsymbol{x} = \boldsymbol{b}$, et comme A et LU sont inversibles, on en déduit que $A^{-1}\boldsymbol{b} = (LU)^{-1}\boldsymbol{b}$ pour tout $\boldsymbol{b} \in \mathbb{R}^n$. 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 1.14 (LU simple (sans permutation))**.**

1. *(Factorisation)*

On pose $u_{i,j} = a_{i,j}$ pour pour $i, j \in \{1, \ldots, n\}$.

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

(a) *On ne change pas la i-ème ligne*

(b) *On modifie les lignes $i+1$ à n ((mais pas le second membre) en utilisant la ligne i.*

Pour k allant de $i+1$ à n :

$\ell_{k,i} = \frac{u_{k,i}}{u_{i,i}}$ *(si $u_{i,i} = 0$, prendre la méthode avec pivot partiel).*

Pour j allant de $i+1$ à n,

$u_{k,j} = u_{k,j} - \ell_{k,i} u_{i,j}$

Fin pour

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 $\ell_{i,i} = 1$)*
   Fin pour
3. *(Remontée) On calcule x (avec $Ux = y$)*
   Pour i allant de n à 1,
   $x_i = \frac{1}{u_{i,i}}(y_i - \sum_{j=i+1}^{n} u_{i,j} x_j)$
   Fin pour

Remarque 1.15 (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. A la fin de la factorisation, U est stockée dans la partie supérieure de A (y compris la diagonale) et L dans la partie strictement inférieure de A (c'est-à-dire sans la diagonale, la diagonale de L est connue car toujours formée de 1). Dans l'algorithme précédent, on remplace donc tous les "u" et "ℓ" par "a". De même, on remplace tous les "x" et "y" par "b". A la fin des étapes de descente et de remontée, la solution du problème est alors stockée dans $\boldsymbol{b}$.*
L'introduction de L, U, $\boldsymbol{x}$ et $\boldsymbol{y}$ peut toutefois aider à comprendre la méthode.

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. Cette CNS fait intervenir les matrices principales d'ordre k et leurs déterminants.
Commençons par une définition, puis un lemme de décomposition par blocs qui va nous permettre de prouver cette CNS.

Définition 1.16 (Matrice principale d'ordre k et mineur principal)**.** *Soit $n \in \mathbb{N}$, $A \in \mathcal{M}_n(\mathbb{R})$ et $k \in \{1, \dots, n\}$. 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 $i = 1, \dots, k$ et $j = 1, \dots, k$. Le* mineur principal d'ordre k *de A est le déterminant de la matrice principale d'ordre k.*

Lemme 1.17 (Décomposition LU de la matrice principale d'ordre k)**.** *Soit $n \in \mathbb{N}$, $A \in \mathcal{M}_n(\mathbb{R})$, $k \in \{1, \dots, n\}$ et A_k la matrice principale d'ordre k de A. 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 $A_k = L_k U_k$. 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 matrice 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

$$A_{k+1} = \begin{bmatrix} L_k & 0_{1\times k} \\ \boldsymbol{c}_1(C_k) & 1 \end{bmatrix} \begin{bmatrix} U_k & \boldsymbol{\ell}_1(B_k) \\ 0_{k\times 1} & (D_k)_{1,1} \end{bmatrix} \tag{1.33}$$

où $\boldsymbol{\ell}_1(B_k) \in \mathcal{M}_{k,1}(\mathbb{R})$ est la première colonne de la matrice B_k, $\boldsymbol{c}_1(C_k) \in \mathcal{M}_{1,k}$ 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 :

$$A = \begin{bmatrix} A_k & P_k \\ R_k & S_k \end{bmatrix},$$

avec $A_k \in \mathcal{M}_k(\mathbb{R})$, $P_k \in \mathcal{M}_{k,n-k}(\mathbb{R})$, $R_k \in \mathcal{M}_{n-k,k}(\mathbb{R})$ et $S_k \in \mathcal{M}_{n-k,n-k}(\mathbb{R})$. Par hypothèse, on a $A_k = L_k U_k$. De plus L_k et U_k sont inversibles, et il existe donc une unique matrice $B_k \in \mathcal{M}_{k,n-k}(\mathbb{R})$ (resp. $C_k \in \mathcal{M}_{n-k,k}(\mathbb{R})$) telle que $L_k B_k = P_k$ (resp $C_k U_k = R_k$). On pose alors $D_k = S_k - C_k B_k$, on obtient (1.32). L'égalité (1.33) en découle immédiatement. ∎

Proposition 1.18 (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, telles que $A = LU$.

(P2) Les mineurs principaux 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, matrice principale d'ordre k de A, vérifie $A_k = L_k U_k$ 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

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

et donc (P1) $\Rightarrow$ (P2).

Montrons maintenant la réciproque. On suppose que les mineurs prinicipaux de A sont non nuls, et on va montrer que $A = LU$. On va en fait montrer par récurrence que pour tout $k = 1, \dots n$, on a $A_k = L_k U_k$ où L_k triangulaire inférieure de coefficients égaux à 1 et U_k inversible triangulaire supérieure. Le premier mineur est non nul, donc $a_{1,1} = 1 \times a_{1,1}$, et la récurrence est bien initialisée. On la suppose vraie à l'étape k. Par le lemme 1.17, on a donc A_{k+1} qui est de la forme (1.33), c.à.d. $A_{k+1} = L_{k+1}U_{k+1}$. 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) $\Rightarrow$ (P1) (l'unicité de L et U est laissée en exercice). ∎

Que faire en cas de pivot nul : la technique de permutation ou de "pivot partiel" 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 on doit 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 1.18 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 :

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

On a donc en particulier

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

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 1.19 (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 1.20 (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 $k=i+1,\dots,$ (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,
   $a_{t(k),j}=a_{t(k),j}-\ell_{t(k),i}u_{t(i),j}$
   Fin pour
   Fin pour
3. *(Descente) On calcule y*
   Pour i allant de 1 à n,
   $y_i=b_{t(i)}-\sum_{j=1}^{i-1}\ell_{t(i),j}y_j$
   Fin pour
4. *(Remontée) On calcule x*
   Pour i allant de n à 1,
   $x_i=\frac{1}{u_{t(i),i}}(y_i-\sum_{j=i+1}^{n}u_{t(i),j}x_j)$
   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). On a donc aussi changé l'ordre dans lequel interviennent les composantes du second membre : le système $A\boldsymbol{x}=\boldsymbol{b}$ est devenu $PA\boldsymbol{x}=P\boldsymbol{b}$. Par contre, on n'a pas touché à l'ordre dans lequel interviennent les composantes de $\boldsymbol{x}$ et $\boldsymbol{y}$.

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 $\{1,\dots,n\}$ dans $\{1,\dots,n\}$ (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 1.21 (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 une matrice inversible.

Théorème 1.22 (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" 1.20). Posons $A^{(0)} = A$.
À chaque étape i de l'algorithme 1.20 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 commutent en général 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 montre ainsi par récurrence l'existence de la décomposition LU (voir aussi l'exercice 29 page 49).

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 $L_2^{-1}L_1 = U_2U_1^{-1} = \mathrm{Id}$, et donc que $L_1 = L_2$ et $U_1 = U_2$. ∎

Remarque 1.23 (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 40.*

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 :

$$\begin{aligned}
& \ell_{n,n}\, x_n = y_n \text{ donc } x_n = \frac{y_n}{\ell_{n,n}} \\
& \ell_{n-1,n-1}x_{n-1} + \ell_{n,n-1}x_n = y_{n-1} \text{ donc } x_{n-1} = \tfrac{y_{n-1} - \ell_{n,n-1}x_n}{\ell_{n-1,n-1}} \\
& \vdots \\
& \sum_{j=1,n} \ell_{j,1}x_j = y_1 \text{ donc } x_1 = \frac{y_1 - \sum_{j=2,n} \ell_{j,1}x_j}{\ell_{1,1}}.
\end{aligned}$$

   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 1.22 page 38, 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 :

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

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 1.24 (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 $i \in \{1, \ldots, n\}$,*
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 :

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

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 :

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

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 :

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

Pour que (1.37) soit vérifiée, il faut que

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

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 :

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

On a donc $Az \cdot z = \alpha - a^tB^{-1}a > 0$ ce qui montre que (1.38) est vérifiée. On peut ainsi choisir $\lambda = \sqrt{\alpha - a^tB^{-1}a}$ (> 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 $i = q+2, \ldots, n$; 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 1.25 (Choleski et LU)**.** *Considérons une matrice A symétrique définie positive. Alors une matrice P de permutation dans le théorème 1.24 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 $t = \mathrm{Id}$ dans l'algorithme 1.20, celle-ci n'est jamais modifiée et donc on a $P = \mathrm{Id}$. Les théorèmes d'existence et d'unicité 1.22 et 1.24 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 41 page 52.*

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

Proposition 1.26 (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 1.24 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 1.22 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 $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, voir aussi exercice 43.

Remarque 1.27 (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 :

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

(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 1.28 (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 :

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

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.
Le calcul du déterminant d'une matrice carrée d'ordre n en utilisant les formules "usuelles" (c'est-à-dire en développant par rapport à une ligne ou une colonne) 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 Choleski environ 350 et la méthode de Cramer (avec les formules usuelles de calcul du déterminant) 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 [4] 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 $j_i = \min\{j \in \{1, \dots, n\} \text{ tels que } a_{i,j} \neq 0\}$. 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 $k = i - j_i, \dots, i$, ce qui peut permettre un large gain de place mémoire.
Une propriété intéressante de la méthode de Choleski est de conserver le profil. On peut montrer (en reprenant les calculs effectués dans la deuxième partie de la démonstration du théorème 1.24) 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 Choleski. Ceci est possible en remarquant que : $Ax = b \Leftrightarrow A^tAx = A^tb$ car $\det(A) = \det(A^t) \neq 0$. 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 $Bx \cdot y = x \cdot B^ty$ 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 33).

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 :

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

4. 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 scalaires.

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 (méthodes directes)

Exercice 21 (Vrai ou faux ?)**.** *Corrigé en page 54*

Les propositions suivantes sont-elles vraies ou fausses ?

1. La matrice $\begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$ admet une décomposition de Choleski.
2. La matrice $B = \begin{bmatrix} 1 & -2 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & 3 \end{bmatrix}$ est symétrique définie positive.
3. La matrice B ci-dessus admet une décomposition LU.
4. La matrice $\begin{bmatrix} 1 & -1 \\ 1 & 3 \end{bmatrix}$ s'écrit C^tC.
5. La matrice $A = \begin{bmatrix} 1 & 1 \\ 1 & 5 \end{bmatrix}$ admet une décomposition de Choleski $A = C^tC$ avec $C = \begin{bmatrix} -1 & -1 \\ 0 & -2 \end{bmatrix}$.
6. Soit $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}$
   (a) La matrice AA^t admet une décomposition de Choleski.
   (b) La matrice A^tA admet une décomposition de Choleski.

Exercice 22 (Elimination de Gauss)**.** On cherche la solution du système linéaire $A\boldsymbol{x} = \boldsymbol{b}$ avec

$$A = \begin{bmatrix} 1 & 0 & 6 & 2 \\ 8 & 0 & -2 & -2 \\ 2 & 9 & 1 & 3 \\ 2 & 1 & -3 & 10 \end{bmatrix} \text{ et } \boldsymbol{b} = \begin{bmatrix} 6 \\ -2 \\ -8 \\ -4 \end{bmatrix}.$$

1. Pourquoi la méthode de Gauss sans permutation ne fonctionne-t-elle pas pour résoudre ce système linéaire ?
2. Donner une permutation de lignes de A permettant d'utiliser ensuite la méthode de Gauss.
3. Donner la solution de ce système linéaire. (NB : La solution prend ses valeurs dans $\mathbb{Z}\ldots$)

Exercice 23 (Factorisation LU sur un exemple)**.**

1. Calculer la factorisation LU, où L est une matrice triangulaire inférieure dont les éléments diagonaux sont tous égaux à 1, et U est une matrice triangulaire supérieure inversible,
   de la matrice $A = \begin{bmatrix} 2 & -1 & 0 & \ldots & \\ -1 & 2 & -1 & 0 & \\ 0 & -1 & 2 & & 0 \\ \ldots & & & & -1 \\ 0 & & & -1 & 2 \end{bmatrix} \in \mathcal{M}_n(\mathbb{R})$.
2. En notant $(m_i)_{i=1,\ldots,n}$ les mineurs principaux de A, donner l'expression de m_i en fonction de i.

Exercice 24 (Factorisation LU sur un autre exemple)**.**

1. Trouver la factorisation $A = LU$, L triangulaire inférieure de diagonale égale à la matrice identité, U triangulaire supérieure inversible, pour

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

2. En déduire les valeurs des mineurs principaux de A.

Exercice 25 (LU)**.** *Corrigé en page 54*

1. Donner la décomposition LU de la matrice $A = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 2 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 2 & 1 & 0 \end{bmatrix}$.
2. Montrer que la matrice $A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$ 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 $\begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix}$

Exercice 26 (Décomposition LU et mineurs principaux)**.** *Suggestions en page 53.*
L'objet de cet exercice est de montrer sur un exemple comment prouver qu'une décomposition LU sans permutation existe *sans l'effectuer*.
Soit $n \geq 1$ et soit $A \in \mathcal{M}_n(\mathbb{R})$ la matrice dont les coefficients sont :

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

1. Montrer que $\det(A) = 2^{n-1}$.
2. Montrer que A admet une décomposition LU sans permutation et calculer les coefficients diagonaux de la matrice U.

Exercice 27 (Décomposition LU d'une matrice particulière)**.**
Soient $\alpha, \beta \in \mathbb{R}$ tels que $\alpha\beta \neq 1$.

1. Soit $A_3 = \begin{bmatrix} 1 & \beta & \beta^2 \\ \alpha & 1 & \beta \\ \alpha^2 & \alpha & 1 \end{bmatrix}$.
   (a) Montrer que A_3 admet une unique décomposition LU, c'est-à-dire, $A_3 = LU$, avec L triangulaire inférieure avec des 1 sur la diagonale et U triangulaire supérieure, on demande ici une réponse sans calcul explicite de la décomposition.
   (b) Donner l'expression des matrices d'élimination de la procédure de décomposition LU pour A_3. Calculer L^{-1} et L à l'aide de ces matrices. Comparer L avec A. Donner l'expression de U.
2. Soit $n \geq 2$ un entier. On définit de manière plus générale $A_n = (a_{ij})_{i,j=1}^n$, avec $a_{ii} = 1$, $a_{ij} = \alpha^{i-j},\ i > j$ et $a_{ij} = \beta^{j-i},\ j > i$.
   (a) Montrer que $\det(A_n) = (1 - \alpha\beta)^{n-1}$ puis que A_n admet une unique décomposition LU (ici encore sans calculer les matrices L et U).

   On décompose $A_n = \mathrm{Id}_n - E_n - F_n$, où Id_n est la matrice identité de taille $n \times n$ et $-E_n$ (resp. $-F_n$), la partie triangulaire inférieure (resp. supérieure) stricte de A_n
   (b) Montrer que

$$(\mathrm{Id}_n - E_n)^{-1} = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ -\alpha & 1 & 0 & \dots & 0 \\ 0 & -\alpha & 1 & 0 & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & -\alpha & 1 \end{bmatrix}.$$

(c) Calculer $(\mathrm{Id}_n - E_n)^{-1}A_n$ et en déduire l'expression de L_n dans la décomposition LU de A_n.

Exercice 28 (Matrices symétriques définies positives, mineurs principaux et décomposition LU)**.** *On rappelle que les mineurs principaux d'une matrice $A \in \mathcal{M}_n(\mathbb{R})$, sont les déterminants Δ_k des matrices principales $A_k = A(1:k, 1:k)$ extraites de la matrice A.*

1. *Soit A une matrice $n \times n$. On suppose que les pivots de l'élimination de Gauss sont tous non nuls pour cette matrice. Expliquer pourquoi, à chaque étape k de l'élimination de Gauss pour trouver la décomposition LU, le mineur principal Δ_k reste inchangé.*
2. *Montrer qu'une matrice symétrique est définie positive si et seulement tous ses mineurs pricipaux strictement positifs. Cette CNS s'appelle critère de Sylvester.*
3. *En déduire que toute matrice symétrique définie positive admet une décomposition LU sans permutation.*
4. *On considère la matrice $A = \begin{bmatrix} a & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}$*
   (a) Pour quelles valeurs de a cette matrice est-elle définie positive ?
   (b) Ecrire l'algorithme de Gauss pour obtenir la décomposition LU de A pour $a = 2$.

Exercice 29 (Existence de la décomposition LU à une permutation près)**.** *Suggestions en page 53, corrigé en page 55*

L'objet de cet exercice est de démontrer par récurrence le résultat suivant (voir aussi théorème 1.22) :

Lemme 1.29 (Décomposition LU d'une matrice inversible par technique du pivot partiel)**.** *Soit $n \in \mathbb{N}$, $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible ; il existe une matrice de permutation $P \in \mathcal{M}_n(\mathbb{R})$ au sens de la définition 1.1, une matrice $L \in \mathcal{M}_n(\mathbb{R})$ triangulaire inférieure inversible et une matrice triangulaire supérieure $U \in \mathcal{M}_n(\mathbb{R})$ de coefficients diagonaux tous égaux à 1, telles que l'on ait la relation $PA = LU$ (décomposition LU de la matrice PA).*

Pour cela, nous allons démontrer par récurrence la propriété suivante : pour tout $k \in \{1, \dots, n\}$, il existe une matrice de permutation $P^{(k)} \in \mathcal{M}_n(\mathbb{R})$, une matrice $L_k \in \mathcal{M}_k(\mathbb{R})$ triangulaire inférieure inversible et une matrice triangulaire supérieure $U_k \in \mathcal{M}_k(\mathbb{R})$ de coefficients diagonaux tous égaux à 1, telles que la matrice $A^{(k)} = P^{(k)}A$ vérifie $A_k^{(k)} = L_k U_k$, en notant $A_k^{(k)} \in \mathcal{M}_k(\mathbb{R})$ la matrice définie par $(A_k^{(k)})_{i,j} = a_{i,j}^{(k)}$ pour $i = 1, \dots, k$ et $j = 1, \dots, k$.

1. Montrer que l'hypothèse de récurrence est vrai au rang $k = 1$.

On suppose maintenant que la propriété de récurrence est vérifiée au rang $k \in \{1, \dots, n-1\}$, et on va prouver qu'elle est encore vraie au rang $k+1$.

2. Montrer que la matrice $A^{(k)} = P^{(k)}A$ peut s'écrire sous la forme par blocs suivante :

$$A^{(k)} = \begin{bmatrix} L_k & 0_{k\times(n-k)} \\ C & D \end{bmatrix} \begin{bmatrix} U_k & V \\ 0_{(n-k)\times k} & \mathrm{Id}_{n-k} \end{bmatrix}, \tag{1.45}$$

   où $0_{p,q}$ désigne la matrice nulle de dimension $p \times q$, $V \in \mathcal{M}_{k,n-k}(\mathbb{R})$ et $C \in \mathcal{M}_{n-k,k}(\mathbb{R})$ et $D \in \mathcal{M}_{n-k,n-k}(\mathbb{R})$.

On appelle $\boldsymbol{c}_1(D)$, $\boldsymbol{c}_1(V)$, $\boldsymbol{c}_1(E)$ et $\boldsymbol{c}_1(G)$ les premières colonnes respectives des matrices D, V, E et G.

3. Montrer que $\boldsymbol{c}_1(D) \neq 0_{(n-k)\times 1}$.

Soit $i^* \in \{k+1, \dots, n\}$ t.q. $|d_{i^*,1}| = \max\{|d_{i,1}|, 1 \in \{k+1, \dots, n\}\}$. On pose $P^{(k+1)} = P^{(i^* \leftrightarrow k+1)}P^{(k)}$, $A^{(k+1)} = P^{(i^* \leftrightarrow k+1)}A^{(k)} = P^{(k+1)}A$, et

$$L_{k+1} = \begin{bmatrix} L_k & 0_{k\times 1} \\ \boldsymbol{\ell}_{i^*}(C) & d_{i^*,1} \end{bmatrix}, \; U_{k+1} = \begin{bmatrix} U_k & \boldsymbol{c}_1(V) \\ 0_{1\times k} & 1 \end{bmatrix}, \; A_{k+1}^{(k+1)} = \begin{bmatrix} A_k^{(k)} & \boldsymbol{c}_1(E) \\ \boldsymbol{\ell}_{i^*}(F) & g_{i^*,1} \end{bmatrix}, \tag{1.46}$$

où $\boldsymbol{\ell}_{i^*}(C)$ (resp. $\boldsymbol{\ell}_{i^*}(F)$) désigne la i^*-ème ligne de la matrice C (resp. F).

4. Montrer que les matrices $P^{(k+1)}$, L_{k+1} et U_{k+1} vérifient l'hypothèse de récurrence par construction, et conclure la démonstration du lemme 1.29.

Exercice 30 (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 $a_{i,j} = 0$)

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

Pour chacune de ces matrices, quels sont les coefficients nuls (notés 0 dans les matrices) qui resteront nécessairement nuls dans les matrices L et U de la factorisation LU sans permutation (si elle existe) ?

Exercice 31 (Un système linéaire par blocs)**.** Soit $n \geq 1$, $A \in \mathcal{M}_n(\mathbb{R})$, $\boldsymbol{b}$ et $\boldsymbol{c} \in \mathbb{R}^n$ et $d \in \mathbb{R}$. On note M la matrice appartenant à $\mathcal{M}_{n+1}(\mathbb{R})$ définie (par blocs) par :

$$M = \begin{bmatrix} A & \boldsymbol{b} \\ \boldsymbol{c}^t & d \end{bmatrix}$$

(noter qu'on a identifié $\mathbb{R}^n$ à $\mathcal{M}_{n,1}(\mathbb{R})$). On suppose que la matrice A est inversible. On note $\boldsymbol{x_b}$ le vecteur de $\mathbb{R}^n$ tel que $A\boldsymbol{x_b} = \boldsymbol{b}$.

1. Montrer que M est inversible si et seulement si $d - \boldsymbol{c}^t\boldsymbol{x_b} \neq 0$.
2. On suppose maintenant que M est inversible. Soit $\boldsymbol{\alpha} \in \mathbb{R}^n$ et $\beta \in \mathbb{R}$. On note $\boldsymbol{x_\alpha}$ le vecteur de $\mathbb{R}^n$ tel que $A\boldsymbol{x_\alpha} = \boldsymbol{\alpha}$. Soit $\boldsymbol{x} = \begin{bmatrix} \boldsymbol{y} \\ z \end{bmatrix} \in \mathbb{R}^{n+1}$ tel que $M\boldsymbol{x} = \begin{bmatrix} \boldsymbol{\alpha} \\ \beta \end{bmatrix}$. Donner l'expression de $\boldsymbol{y}$ et z en fonction de $\boldsymbol{x_\alpha}$ et $\boldsymbol{x_b}$.

Exercice 32 (Matrices symétriques 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 33 (Sur la méthode LL^t)**.** *Corrigé détaillé en page 57.*
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 34 (Décomposition LU d'une matrice à paramètres)**.** *Corrigé en page 57.*
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 (si elle existe).
Donner les conditions sur a, b, c et d pour que la matrice A soit inversible.

Exercice 35 (Echelonnement et factorisation LU et LDU)**.** *Corrigé en page 58.*
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 36 (Méthode de Choleski sur un exemple)**.** Soit la matrice

$$A = \begin{bmatrix} 4 & 2 & 0 \\ 2 & 4 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

1. Montrer que A est symétrique définie positive, en effectuant sa décomposition de Choleski.
2. Que valent les mineurs principaux de A ?
3. A l'aide de la décomposition de Choleski de A, résoudre $AX = B$ où $B = {}^t(8, 13, 5)$.

Exercice 37 (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 38 (LU et Choleski sur un exemple)**.**
Soit $M = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 8 & 10 \\ 1 & 10 & 18 \end{bmatrix}$.

1. Calculer les mineurs principaux de M. En déduire que M admet des décompositions LU et de Choleski.
2. Donner la décomposition LU de M.
3. Donner la décomposition de Choleski de M.

Exercice 39 (Factorisation de Choleski)**.**
1. *Pouver, au moyen de la factorisation de Choleski, que la matrice A définie par*

$$A = \begin{pmatrix} 1 & -1 & -1 \\ -1 & 2 & 1 \\ -1 & 1 & 5 \end{pmatrix}$$

est symétrique définie positive.

2. *Quelles sont les valeurs des mineurs principaux de A ?*

Exercice 40 (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 une matrice triangulaire inférieure dont les coefficients diagonaux sont égaux à 1, et $\widetilde{U}$ une matrice 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 la matrice A de rang égal à $n - 1$.)

Exercice 41 (Décomposition LL^t "pratique")**.** *Corrigé en page 59.*

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 1.25.)

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 42 (Factorisation de Choleski sur un exemple)**.** Calculer la factorisation de Choleski de la matrice suivante :

$$A = \begin{bmatrix} 4 & 4 & 2 & 0 \\ 4 & 5 & 0 & 0 \\ 2 & 0 & 6 & 1 \\ 0 & 0 & 1 & 2 \end{bmatrix}$$

Exercice 43 (Décomposition LDL^t et LL^t)**.** *Corrigé en page 61*

1. Soit $A = \begin{bmatrix} 2 & 1 \\ 1 & 0 \end{bmatrix}$. 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 $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ admet-elle une décomposition LDL^t ?

Exercice 44 (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

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

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 :

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

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

Exercice 45 (Choleski pour matrice bande)**.** Suggestions en page 54, corrigé en page 63
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 14.

1.3.6 Suggestions

Exercice 26 page 48 (Décomposition LU et mineurs principaux)

1. On pourra par exemple raisonner par récurrence et remarquer que $\det A = \det B$ où B est obtenue en ajoutant, pour tout $i \in \{2, \dots, n\}$, la première ligne de A à la i-ème ligne de A, ce qui correspond à la première étape de l'algorithme de décomposition LU.

2. Utiliser la caractérisation par les mineurs (proposition 1.18).

Exercice 29 page 49 (Existence de la décomposition LU à une permutation près)

2. Ecrire $A^{(k)} = P^{(k)}A$ sous une forme par blocs.

3. Procéder par contradiction.

Exercice 45 page 53

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 21 page 47 (Vrai ou faux ?)

1. La matrice A est symétrique, sa trace est égale à 3 et son déterminant à 1, donc elle est s.d.p. et donc elle admet une décomposition de Choleski.
   Autre argument, ses deux mineurs principaux sont strictement positifs.
   Autre argument, A admet une décomposition LU avec 2 pivots strictement positifs
2. La matrice B n'est pas symétrique.
3. 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.
4. Non car elle n'est pas symétrique.
5. 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
$$A = \begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix}.$$
6. (a) FAUX. La matrice est d'ordre 3, mais de rang au plus 2, donc elle n'est pas inversible.
   (b) VRAI. $A^tA = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$ qui est symétrique définie positive (trace et déterminants strictement positifs, par exemple).

Exercice 25 page 48 (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. 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 28 page 49 (Matrices symétriques définies positives, mineurs principaux et décomposition LU)

1. Lorsqu'on fait l'élimination de Gauss à l'étape k, on remplace la ligne k par une combinaison linéaire de cette ligne avec la ligne $k-1$ de A_k, et donc on ne change pas le déterminant de A_k.
2. Montrons d'abord que la condition est nécessaire. Comme A est symétrique définie positive, $\boldsymbol{x}A\boldsymbol{x}^t > 0$ pour tout $\boldsymbol{x} \in \mathbb{R}^n$. En prenant $\boldsymbol{x} = \begin{bmatrix} \boldsymbol{y} \\ \mathbf{0} \end{bmatrix}$, avec $\boldsymbol{y} \in \mathbb{R}^k$ et $\mathbf{0} \in \mathbb{R}^{n-k}$, on obtient donc que $\boldsymbol{y}A_k\boldsymbol{y}^t > 0$ pour tout $\boldsymbol{y} \in \mathbb{R}^k$, pour tout $k = 1, \ldots, n$ ce qui montre bien que les matrices A_k sont toutes symétriques définies positives.

   Réciproquement, pour montrer que si tous les mineurs pricipaux d'une matrice sont strictement positifs, alors la matrice est définie positive, on va raisonner par récurrence sur n. Soit A une matrice $n \times n$, symétrique, dont tous les mineurs principaux sont strictement positifs. Pour $n = 1$, il est clair que A est s.d.p. Supposons le résultat vrai pour toute matrice $p \times p$ avec $p \leq n-1$, et appliquons l'algorithme de Gauss à la matrice A. Comme les déterminants mineurs sont tous strictement positifs au départ et qu'ils restent constants pendant l'algorithme de Gauss, les déterminants mineurs de la matrice U obtenue à la fin de l'algorithme de Gauss sont strictement positifs. Or les matrices principales de U sont triangulaires supérieures, donc leur déterminant est le produit des valeurs propres. En conséquence, toutes les valeurs propres de A sont strictement positives, ce qui montre bien que A est symétrique est définie positive.
3. Ceci découle directement de la proposition 1.18 du cours (CNS pour LU sans permutation)
4. (a) Appliquons le critère de Sylvester. En notant m_i le i-ème mineur principal, on a $m_1 = a$, $m_2 = \begin{vmatrix} a & 1 \\ 1 & 2 \end{vmatrix} = 2a - 1$ et $m_2 = \begin{vmatrix} a & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{vmatrix} = 3a - 2$. Il en résulte que A est définie positive si et seulement si $a > 2/3$.

   (b)
$$\begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix} \xrightarrow{\boldsymbol{\ell}_2 \leftarrow \boldsymbol{\ell}_2 - \frac{1}{2}\boldsymbol{\ell}_1, \boldsymbol{\ell}_3 \leftarrow \boldsymbol{\ell}_3 - \frac{1}{2}\boldsymbol{\ell}_1} \begin{bmatrix} 2 & 1 & 1 \\ 0 & 3/2 & 1/2 \\ 0 & 1/2 & 3/2 \end{bmatrix}$$
$$\xrightarrow{\boldsymbol{\ell}_3 \leftarrow \boldsymbol{\ell}_3 - \frac{1}{3}\boldsymbol{\ell}_2} \begin{bmatrix} 2 & 1 & 1 \\ 0 & 3/2 & 1/2 \\ 0 & 0 & 4/3 \end{bmatrix}$$

   Donc $A = LU$ avec $L = \begin{bmatrix} 1 & 0 & 0 \\ 1/2 & 1 & 0 \\ 1/2 & 1/3 & 1 \end{bmatrix}$ et $U = \begin{bmatrix} 2 & 1 & 1 \\ 0 & 3/2 & 1/2 \\ 0 & 0 & 4/3 \end{bmatrix}$

Exercice 29 page 49 (Existence de la décomposition LU à une permutation près)

1. Vérifions la propriété de récurrence au rang $k = 1$. Soit $i^* \in \{1, \ldots, n\}$ t.q. $|a_{i^*,1}| = \max\{|a_{i,1}|, 1 \in \{1, \ldots, n\}\}$ (noter que ce max est forcément non nul car la matrice est inversible). Soit $P^{(1)} = P^{(1 \leftrightarrow i^*)}$ (voir Définition 1.1). On a alors $A_1^{(1)} = \begin{bmatrix} a_{i^*,1} \end{bmatrix}$, $L_1 = A_1^{(1)}$ et $U_1 = \begin{bmatrix} 1 \end{bmatrix}$.

2. Il suffit d'écrire la décomposition par blocs de $A^{(k)}$:
$$A^{(k)} = \begin{bmatrix} A_k^{(k)} & E \\ F & G \end{bmatrix},$$
avec $A_k^{(k)} \in \mathcal{M}_k(\mathbb{R})$, $E \in \mathcal{M}_{k,n-k}(\mathbb{R})$, $F \in \mathcal{M}_{n-k,k}(\mathbb{R})$ et $G \in \mathcal{M}_{n-k,n-k}(\mathbb{R})$. Par hypothèse de récurrence, on a $A_k^{(k)} = L_k U_k$. De plus L_k et U_k sont inversibles, et il existe donc une unique matrice $V \in \mathcal{M}_{k,n-k}(\mathbb{R})$ (resp. $C \in \mathcal{M}_{n-k,k}(\mathbb{R})$) telle que $L_k V = E$ (resp $CU_k = F$). En posant $D = G - CV$, on obtient l'égalité (1.45).

3. En effet, si $\boldsymbol{c}_1(D) = 0_{(n-k)\times 1}$, alors $\boldsymbol{c}_1(G) = C\boldsymbol{c}_1(V) = FU^{-1}\boldsymbol{c}_1(V)$ et en même temps $\boldsymbol{c}_1(E) = L\boldsymbol{c}_1(V) = A_k^{(k)}U^{-1}\boldsymbol{c}_1(V)$. On obtient alors que la colonne $k+1$ de la matrice $A^{(k)}$, composée des deux vecteurs $\boldsymbol{c}_1(E)$ et $\boldsymbol{c}_1(G)$, est obtenue par la combinaison linéaire avec les coefficients $U^{-1}\boldsymbol{c}_1(V)$ des k premières colonnes de la matrice $A^{(k)}$, constituées des matrices $A_k^{(k)}$ et F. C'est impossible, puisque la matrice $A^{(k)}$ est le produit des deux matrices inversibles $P^{(k)}$ et A.

4. On a bien

1. $L_k v_{\cdot,1} = \boldsymbol{c}_1(E)$,
2. $\boldsymbol{\ell}_{i^*}(C)U_k = \boldsymbol{\ell}_{i^*}(F)$,
3. $\boldsymbol{\ell}_{i^*}(C)\boldsymbol{c}_1(V) + d_{i^*,1} = g_{i^*,1}$.

La conclusion du lemme est alors obtenue pour $k = n$.

Exercice 33 page 50 (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 34 page 50 (Décomposition LU d'une matrice à paramètres)

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 :

$$E_1A=\begin{bmatrix}a&a&a&a\\0&b-a&b-a&b-a\\0&b-a&c-a&c-a\\0&b-a&c-a&d-a\end{bmatrix}.$$

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 :

$$E_2E_1A=\begin{bmatrix}a&a&a&a\\0&b-a&b-a&b-a\\0&0&c-b&c-b\\0&0&c-b&d-b\end{bmatrix}.$$

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 :

$$E_3E_2E_1A = \begin{bmatrix} a & a & a & a \\ 0 & b-a & b-a & b-a \\ 0 & 0 & c-b & c-b \\ 0 & 0 & 0 & d-c \end{bmatrix}.$$

On $A = LU$ avec $L = (E_3E_2E_1)^{-1} = (E_1)^{-1}(E_2)^{-1}(E_3)^{-1}$; 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 (sans calculs !) :

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

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

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 35 page 51 (Echelonnement et factorisation LU et LDU)

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 $k = 1$

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

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

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

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 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 l'élimination de Gauss dans la matrice L.

Pour la seconde matrice, l'élimination donne :

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

Exercice 41 page 52 (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 $k = 1$

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

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 $k = 2$

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

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 1.22 et la remarque 1.25 qu'il existe une unique décomposition LU : $A = LU$. Le théorème 1.24 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 1.24 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 43 page 52 (Décompositions LL^t et LDL^t)

1. On pose $L = \begin{bmatrix} 1 & 0 \\ \gamma & 1 \end{bmatrix}$ et $D = \begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix}$. 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 $L = \begin{bmatrix} a & 0 \\ b & c \end{bmatrix}$, 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 $\boldsymbol{x} = (x_1, x_2)^t \in \mathbb{R}^2$, alors $A\boldsymbol{x} \cdot \boldsymbol{x} = 2x_1(x_1 + x_2)$, et en prenant $\boldsymbol{x} = (1, -2)^t$, on a $A\boldsymbol{x} \cdot \boldsymbol{x} < 0$.

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 :

$$A = \left[\begin{array}{c|c} B & a \\ \hline a^t & \alpha \end{array}\right] \tag{1.47}$$

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 $\tilde{D} = \mathrm{diag}(d_{1,1}, d_{2,2}, \ldots, d_{n,n})$ 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 :

$$L = \left[\begin{array}{c|c} M & 0 \\ \hline b^t & 1 \end{array}\right], \; D = \left[\begin{array}{c|c} \tilde{D} & 0 \\ \hline 0 & \lambda \end{array}\right], \tag{1.48}$$

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 (1.48) 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 :

$$a_{i,j} = \sum_{k=1}^{n} \ell_{i,k} d_{k,k} \ell_{j,k}, \quad \forall\, (i, j) \in \{1, \ldots, n\}^2. \tag{1.49}$$

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

$$d_{n+1,n+1} = a_{n+1,n+1} - \sum_{k=1}^{n} \ell_{n+1,k}^2 d_{k,k}. \tag{1.50}$$

   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 :

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

3. Procédons par identification, en posant comme à la première question $L = \begin{bmatrix} 1 & 0 \\ \gamma & 1 \end{bmatrix}$ et $D = \begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix}$. Pour que $A = LDL^t$, il faut $\alpha = 0$, $\beta = 0$ et $\alpha\gamma = 1$ ce qui est impossible. Cet exemple montre qu'une matrice symétrique (non définie positive) n'admet pas forcément une décomposition LDL^t, voir à ce propos la proposition 1.26.

Exercice 45 page 53 (Décomposition LL^t d'une matrice bande)

On utilise le résultat de conservation du profil de la matrice énoncé dans le cours, voir aussi exercice 30. 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. En reprenant l'algorithme de construction de la matrice L (1.39)-(1.41), on remarque que pour le calcul de la colonne $j+1$, avec $0 \leq j < n-1$, on a le nombre d'opérations suivant :

— Calcul de $\ell_{j+1,j+1} = (a_{j+1,j+1} - \sum_{k=1}^{n} \ell_{j+1,k} \ell_{j+1,k})^{1/2} > 0$: une multiplication, une soustraction (en effet comme la matrice est tridiagonale, la conservation du profil entraîne que $\ell_{j,k} = 0$ si $k < j$), une extraction de racine, soit trois opérations élémentaires.

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

On en déduit que le nombre d'opérations élémentaires pour le calcul de la colonne $j+1$, avec $1 \leq j < n-1$, 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 rappelle les formules donnant L :
Pour $i = 1, \ldots, n$,

$$\ell_{i,i} = (a_{i,i} - \sum_{k=\max\{1,i-q\}}^{i-1} \ell_{i,k}\ell_{i,k})^{1/2}, \tag{1.52}$$

et pour $j = \{\max\{1, i-q\}, \ldots, i-1\}$,

$$\ell_{i,j} = \left(a_{j,i} - \sum_{k=\max\{1,i-q\}}^{j-1} \ell_{i,k}\ell_{j,k}\right) \frac{1}{\ell_{j,j}}. \tag{1.53}$$

Pour $i \in \{1, \ldots, n\}$, on note $N(i)$ le nombre d'opérations pour calculer $\ell_{i,i}$ et, pour $j \in \{\max\{1, i-q\}, \ldots, i-1\}$, $N(i,j)$ le nombre d'opérations pour calculer $\ell_{i,j}$. On note aussi

$$M(i) = N(i) + \sum_{\max\{1,i-q\}}^{i-1} N(i,j).$$

Pour calculer $M(i)$, on distingue les cas $i \leq q$ et $i > q$.
Cas $i \leq q$
Dans ce cas $\{\max\{1, i-q\} = 1$ et

$$N(i) = 2(i-1)+1 = 2i-1,\ N(i,j) = 2(j-1)+1 = 2j-1,\ \sum_{j=1}^{i-1} N(i,j) = (i-1)^2.$$

Ce qui donne $M(i) = i^2$ et donc $\sum_1^q M(i) = q(q+1)(2q+1)/6$.
Cas $i > q$
Dans ce cas $\{\max\{1, i-q\} = i-q$ et

$$\begin{aligned}
&N(i) = 2(i-1-(i-q)+1)+1 = 2q+1,\\
&N(i,j) = 2(j-1-(i-q)+1)+1 = 2(j-1+q)+1,\\
&\sum_{j=i-q}^{i-1} N(i,j) = \sum_{j=i-q}^{i-1} 2(j-i+q)+1 = 2\sum_{k=1}^{q-1} k + (i-1-(i-q)+1) = q(q-1)+q = q^2,\\
&M(i) = 2q+1+q^2 = (q+1)^2.
\end{aligned}$$

On en déduit $Z_p(n)$:

$$Z_p(n) = \sum_{i=1}^{n} M(i) = (n-q)(q+1)^2 + \frac{q(q+1)(2q+1)}{6} = n(q+1)^2 - \frac{q(q+1)(4q+5)}{6}.$$

Remarquons qu'on retrouve bien le nombre obtenu pour $q=1$, $Z_2(n) = 4n-3$.

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 14), 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).

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 1.30 (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.*
$$\|AB\| \leq \|A\|\|B\|, \ \forall A, B \in \mathcal{M}_n(\mathbb{R}) \tag{1.54}$$
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 :*
$$\|A\| = \sup\{\|A\boldsymbol{x}\|; \ \boldsymbol{x} \in \mathbb{R}^n, \|\boldsymbol{x}\| = 1\}, \ \forall A \in \mathcal{M}_n(\mathbb{R}) \tag{1.55}$$

Proposition 1.31 (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 $S_1 = \{\boldsymbol{x} \in \mathbb{R}^n \mid \|\boldsymbol{x}\| = 1\}$ qui est un compact de $\mathbb{R}^n$. Donc φ est bornée et atteint ses bornes : il existe $\boldsymbol{x}_0 \in S_1$ 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 1.32 (Rayon spectral)**.** *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice. 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 1.33 (Caractérisation de normes induites)**.** *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 $\|\cdot\|_\infty$. Alors*
$$\|A\|_\infty = \max_{i\in\{1,\dots,n\}} \sum_{j=1}^{n} |a_{i,j}|. \tag{1.56}$$

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*
$$\|A\|_1 = \max_{j\in\{1,\dots,n\}} \sum_{i=1}^{n} |a_{i,j}| \tag{1.57}$$

3. *On munit $\mathbb{R}^n$ de la norme $\|\cdot\|_2$ et $\mathcal{M}_n(\mathbb{R})$ de la norme induite correspondante, notée aussi $\|\cdot\|_2$.*
$$\|A\|_2 = (\rho(A^tA))^{\frac{1}{2}}. \tag{1.58}$$
En particulier, si A est symétrique, $\|A\|_2 = \rho(A)$.

DÉMONSTRATION – La démonstration des points 1 et 2 fait l'objet de l'exercice 47 page 75. 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. On rappelle d'abord qu'une matrice unitaire $Q \in \mathcal{M}_n(\mathbb{C})$ est une matrice inversible telle que $Q^* = Q^{-1}$; ceci est équivalent à dire que les colonnes de Q forment une base orthonormale de $\mathbb{C}^n$. Une matrice carrée orthogonale est une matrice unitaire à coefficients réels ; on a dans ce cas $Q^* = Q^t$, et les colonnes de Q forment une base orthonormale de $\mathbb{R}^n$.

Théorème 1.34 (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 unitaire (c.à.d. une matrice telle que $Q^* = Q^{-1}$ 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 de dimension finie n sur $\mathbb{C}$, muni d'un produit scalaire. Alors il existe une base orthonormée $(\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n)$ de $\mathbb{C}^n$ et une famille de complexes $(t_{i,j})_{i=1,\ldots,n, j=1,\ldots,n, j\geq i}$ 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 $i \in \{1, \ldots, n\}$.
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 $(t_{i,j})_{i,j=1,\ldots,n, j\geq i}$ et Q la matrice unitaire 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 $n+1$. Soit donc E un espace vectoriel sur $\mathbb{C}$ de dimension $n+1$, muni d'un produit scalaire. Soit ψ une application linéaire de E dans E. 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 un sous espace vectoriel de E supplémentaire 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, F est de dimension n et le produit scalaire sur E induit un produit scalaire sur F). Il existe donc une base orthonormée $\boldsymbol{f}_2, \ldots, \boldsymbol{f}_{n+1}$ de F et $(t_{i,j})_{j\geq i\geq 2}$ tels que

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

On en déduit que

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

Le fait que l'ensemble des $t_{i,i}$ est l'ensemble des valeurs propores de A, comptées avec leur multiplicité. vient de l'égalité $\det(A - \lambda I) = \det(T - \lambda I)$ pour tout $\lambda \in \mathbb{C}$.

■

L'objet du théorème suivant est de montrer 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 1.35 (Approximation du rayon spectral par une norme induite)**.**
1. Soit $\|\cdot\|$ une norme induite. Alors

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

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

DÉMONSTRATION – 1 Soit $\lambda \in \mathbb{C}$ valeur propre de A telle que $|\lambda| = \rho(A)$.
On suppose tout d'abord que $\lambda \in \mathbb{R}$. Il existe alors un vecteur non nul de $\mathbb{R}^n$, noté $\boldsymbol{x}$, tel que $A\boldsymbol{x} = \lambda\boldsymbol{x}$. Comme $\|\cdot\|$ est une norme induite, on a

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

On en déduit que $|\lambda| \leq \|A\|$ et donc $\rho(A) \leq \|A\|$.

Si $\lambda \in \mathbb{C} \setminus \mathbb{R}$, la démonstration est un peu plus compliquée car la norme considérée est une norme dans $\mathbb{R}^n$ (et non dans $\mathbb{C}^n$). On montre tout d'abord que $\rho(A) < 1$ si $\|A\| < 1$.

En effet, Il existe $x \in \mathbb{C}^n$, $x \neq 0$, tel que $Ax = \lambda x$. En posant $x = y + iz$, avec $y, z \in \mathbb{R}^n$, on a donc pour tout $k \in \mathbb{N}$, $\lambda^k x = A^k x = A^k y + iA^k z$. Comme $\|A^k y\| \leq \|A\|^k \|y\|$ et $\|A^k z\| \leq \|A\|^k \|z\|$, on a, si $\|A\| < 1$, $A^k y \to 0$ et $A^k z \to 0$ (dans $\mathbb{R}^n$) quand $k \to +\infty$. On en déduit que $\lambda^k x \to 0$ dans $\mathbb{C}^n$. En choisissant une norme sur $\mathbb{C}^n$, notée $\|\cdot\|_a$, on a donc $|\lambda|^k \|x\|_a \to 0$ quand $k \to +\infty$, ce qui montre que $|\lambda| < 1$ et donc $\rho(A) < 1$.

Pour traiter le cas général (A quelconque dans $\mathcal{M}_n(\mathbb{R})$), il suffit de remarquer que la démonstration précédente donne, pour tout $\eta > 0$, $\rho(A/(\|A\| + \eta)) < 1$ (car $\|A/(\|A\| + \eta)\| < 1$). On a donc $\rho(A) < \|A\| + \eta$ pour tout $\eta > 0$, ce qui donne bien $\rho(A) \leq \|A\|$.

2. Soit $A \in \mathcal{M}_n(\mathbb{R})$, alors par le théorème de triangularisation de Schur (théorème 1.34 ppécédent), il existe une base $(\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n)$ de $\mathbb{C}^n$ et une famille de complexes $(t_{i,j})_{i,j=1,\ldots,n, j\geq i}$ telles que $A\boldsymbol{f}_i = \sum_{j\leq i} t_{j,i}\boldsymbol{f}_j$. Soit $\eta \in]0,1[$, qu'on choisira plus précisément plus tard. Pour $i = 1, \ldots, n$, on définit $\boldsymbol{e}_i = \eta^{i-1}\boldsymbol{f}_i$. La famille $(\boldsymbol{e}_i)_{i=1,\ldots,n}$ 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 $(\boldsymbol{e}_i)_{i=1,\ldots,n}$. Notons que cette norme dépend de A et de η. Soit $\varepsilon > 0$; montrons que pour η bien choisi, on a $\|A\| \leq \rho(A) + \varepsilon$. Remarquons d'abord que

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

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

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

On en déduit que

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

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

$$\sum_{j=1}^n \sum_{\substack{k,\ell\geq j\\(k,\ell)\neq(j,j)}} \eta^{k+\ell-2j}t_{j,k}\overline{t_{j,\ell}} \leq \eta C_T n^3,$$

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

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

on a donc, pour tout x dans $\mathbb{C}^n$, $x \neq 0$,

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

On en déduit que $\|A\|^2 \leq \rho(A)^2 + \eta C_T n^3$ et donc

$$\|A\| \leq \rho(A)\big(1 + \frac{\eta C_T n^3}{\rho(A)^2}\big)^{\frac{1}{2}} \leq \rho(A)(1 + \frac{\eta C_T n^3}{\rho(A)^2}).$$

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

■

Corollaire 1.36 (Convergence et rayon spectral)**.** *Soit $A \in \mathcal{M}_n(\mathbb{R})$. Alors :*

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

DÉMONSTRATION – Si $\rho(A) < 1$, il existe $\varepsilon > 0$ tel que $\rho(A) < 1 - 2\varepsilon$; grâce au résultat d'approximation du rayon spectral du théorème 1.35, il existe donc une norme induite $\|\cdot\|_{A,\varepsilon}$ telle que $\|A\|_{A,\varepsilon} = \mu \leq \rho(A) + \varepsilon = 1 - \varepsilon < 1$. Comme $\|\cdot\|_{A,\varepsilon}$ 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 1.37 (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 1.38 (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*

$$\rho(A) = \lim_{k\to\infty} \|A^k\|^{\frac{1}{k}}. \tag{1.59}$$

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

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

En effet, si i $\rho(A) < 1$, d'après le corollaire 1.36 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

$$\liminf_{k\to\infty} \|A^k\|^{\frac{1}{k}} < 1 \Rightarrow \rho(A) < 1. \tag{1.61}$$

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 $n \to +\infty$. 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 $n \to +\infty$. Soit $\eta \in]l, 1[$ il existe donc n_0 tel que pour $n \geq n_0$, $\|A^{k_n}\|^{1/k_n} \leq \eta$. On en déduit que pour $n \geq n_0$, $\|A^{k_n}\| \leq \eta^{k_n}$, 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 (1.60),

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

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

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

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 (1.61), $\rho(\frac{1}{\beta}A) < 1$, 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

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

De (1.62) et (1.63), on déduit que

$$\limsup_{k\to+\infty} \|A^k\|^{\frac{1}{k}} = \liminf_{k\to+\infty} \|A^k\|^{\frac{1}{k}} = \lim_{k\to+\infty} \|A^k\|^{\frac{1}{k}} = \rho(A). \tag{1.64}$$

■

Un corollaire important de la proposition 1.38 est le suivant.

Corollaire 1.39 (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 $\|\cdot\|$ est une norme matricielle, alors $\|A^k\| \leq \|A\|^k$ et donc par la caractérisation (1.59) 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 1.40 (Matrices de la forme $Id + A$)**.**

1. *Soit $\|\cdot\|$ une norme matricielle,* 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 53 page 77.
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 1.39, 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 12), 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 :

$$\begin{cases} \boldsymbol{x} + \boldsymbol{\delta_x} \in \mathbb{R}^n \\ (A + \delta_A)(\boldsymbol{x} + \boldsymbol{\delta_x}) = \boldsymbol{b} + \boldsymbol{\delta_b}. \end{cases} \tag{1.65}$$

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 1.41 (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 1.42 (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 1.43 (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 $\lambda_1, \dots, \lambda_n$, alors A^{-1} a comme valeurs propres $\lambda_1^{-1}, \dots, \lambda_n^{-1}$ et A^t a comme valeurs propres $\lambda_1, \dots, \lambda_n$.
1. Par définition, on a $\mathrm{cond}_2(A) = \|A\|_2\|A^{-1}\|_2$. Or par le point 3. de la proposition 1.33 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 1.44 (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 1.44 fait l'objet de l'exercice 56 page 77.
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 1.45 (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*

$$A(\boldsymbol{x} + \boldsymbol{\delta_x}) = \boldsymbol{b} + \boldsymbol{\delta_b}, \tag{1.66}$$

alors

$$\frac{\|\boldsymbol{\delta_x}\|}{\|\boldsymbol{x}\|} \leq \mathrm{cond}(A)\frac{\|\boldsymbol{\delta_b}\|}{\|\boldsymbol{b}\|} \tag{1.67}$$

DÉMONSTRATION – En retranchant (1.1) à (1.66), on obtient :

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

et donc

$$\|\boldsymbol{\delta_x}\| \leq \|A^{-1}\|\|\boldsymbol{\delta_b}\|. \tag{1.68}$$

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 (1.68), on obtient le résultat souhaité. ∎

Remarquons que l'estimation (1.67) est optimale. En effet, on va démontrer qu'on peut avoir égalité dans (1.67). 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 (1.65), 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 1.31 page 65). 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 1.31, 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 :

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

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 (1.67) 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 1.46 (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 $\delta_A \in \mathcal{M}_n(\mathbb{R})$; 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*

$$(A + \delta_A)(\boldsymbol{x} + \boldsymbol{\delta_x}) = \boldsymbol{b} \tag{1.69}$$

alors

$$\frac{\|\boldsymbol{\delta_x}\|}{\|\boldsymbol{x} + \boldsymbol{\delta_x}\|} \leq \mathrm{cond}(A)\frac{\|\delta_A\|}{\|A\|} \tag{1.70}$$

DÉMONSTRATION – En retranchant (1.1) à (1.69), on obtient :

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

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 1.47 (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 $\boldsymbol{b} \in \mathbb{R}^n$, $\boldsymbol{b} \neq \boldsymbol{0}$. 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* (1.65)*, alors*

$$\frac{\|\boldsymbol{\delta_x}\|}{\|\boldsymbol{x}\|} \leq \frac{\mathrm{cond}(A)}{1 - \|A^{-1}\|\ \|\delta_A\|}\left(\frac{\boldsymbol{\delta_b}\|}{\|\boldsymbol{b}\|} + \frac{\|\delta_A\|}{\|A\|}\right). \tag{1.71}$$

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 1.40 page 70 et l'exercice 53 page 77) $(\mathrm{Id} + B)$ 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 64 page 80) 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 (1.67) 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 (1.67) 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[5]. 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 15 page 21), 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 64 page 80) 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 (1.67) 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 (1.67) 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 66 page 81. 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

$$\frac{\|\boldsymbol{\delta_u}\|}{\|\boldsymbol{u}\|} \leq C \frac{\|\boldsymbol{\delta_b}\|}{\|\boldsymbol{b}\|} \text{ avec } \boldsymbol{b} = \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix}. \tag{1.72}$$

L'estimation (1.72) est évidemment bien meilleure que l'estimation (1.67) 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 (1.67) est peut-être optimale dans le cas d'une matrice et d'un second membre quelconques, (on a montré ci-dessus qu'il peut y avoir égalité dans (1.67)) 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 (normes et conditionnement)

Exercice 46 (Normes de l'Identité)**.** Soit Id la matrice "Identité" de $\mathcal{M}_n(\mathbb{R})$. Montrer que pour toute norme induite on a $\|\mathrm{Id}\| = 1$ et que pour toute norme matricielle on a $\|\mathrm{Id}\| \geq 1$.

Exercice 47 (Normes induites particulières)**.** *Suggestions en page 82, corrigé détaillé en page 83.*

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

5. 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 $\mathbb{R}$

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 48 (Normes subordonnées)**.** *Soit $n \in \mathbb{N}^\star$, et soit $(\omega_i)_{i=1,\dots,n}$ une famille de réels strictement positifs et soit $\|\cdot\|$ la norme vectorielle sur $\mathbb{R}^n$ définie par*
$$\forall x \in \mathbb{R}^n,\ \|x\| = \sum_{i=1}^{n} \omega_i |x_i|.$$
Nous notons $\|\cdot\|$ la norme matricielle subordonnée à cette norme vectorielle.

1. *Prouver que, pour toute matrice $A \in \mathcal{M}_n(\mathbb{R})$ avec $A = (a_{i,j})_{i,j=1,\dots,n}$, on a*
$$\|A\| \leq \max_{j=1,\dots,n} \sum_{i=1}^{n} |a_{i,j}| \frac{\omega_i}{\omega_j}.$$
2. *Choisir $x \in \mathbb{R}^n$ avec $\|x\| = 1$ pour que*
$$\|Ax\| = \max_{j=1,\dots,n} \sum_{i=1}^{n} |a_{i,j}| \frac{\omega_i}{\omega_j}.$$
3. *Donner l'expression de $\|A\|$.*

Exercice 49 (Exemple de 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 $\|A\|_s^2 = tr(A^t A)$. 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 50 (Valeurs propres d'un produit de matrices)**.** Soient p et n des entiers naturels non nuls, 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. On suppose que $n \leq p$. Montrer que si 0 est valeur propre de AB alors 0 est valeur propre de BA. *(Il est conseillé de distinguer les cas $Bx \neq 0$ et $Bx = 0$, où x est un vecteur propre associé à la valeur propre nulle de AB. Pour le deuxième cas, on pourra distinguer selon que $ImA = \mathbb{R}^n$ ou non.)*
3. Montrer en donnant un exemple que 0 peut être une valeur propre 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 51 (Matrice diagonalisable et rayon spectral)**.** *Corrigé en page 83.*

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 52 (Le rayon spectral est-il une norme ou une semi-norme ?)**.** On définit les matrices carrées d'ordre 2 suivantes :

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

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 53 (Série de Neumann)**.** *Suggestions en page 82, corrigé détaillé en page 84.*

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 $\|\cdot\|$ est 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 54 (Norme induite et rayon spectral)**.** Soit $\|\cdot\|$ 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 55 (Calcul de conditionnement)**.** *Corrigé détaillé en page 85.*

Calculer le conditionnement pour la norme 2 de la matrice $\begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix}$.

Exercice 56 (Propriétés générales du conditionnement)**.** *Corrigé détaillé en page 85.*
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 57 (Conditionnement de la matrice transposée)**.** On suppose que $A \in \mathcal{M}_n(\mathbb{R})$ est inversible.

1. Montrer que si $B \in \mathcal{M}_n(\mathbb{R})$, on a pour tout $\lambda \in \mathbb{C}$, $\det(AB - \lambda \mathrm{Id}) = \det(BA - \lambda \mathrm{Id})$.
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 $\mathrm{cond}_2(A) = \mathrm{cond}_2(A^t)$.
5. A-t-on $\|A^t\|_1 = \|A\|_1$?
6. Montrer que dans le cas $n = 2$, on a toujours $\mathrm{cond}_1(A) = \mathrm{cond}_1(A^t), \forall A \in M_2(\mathbb{R})$.

7. Calculer $\text{cond}_1(A)$ pour $A = \begin{bmatrix} 2 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$ et conclure.

Exercice 58 (Conditionnement et normes $\|\cdot\|_1$ et $\|\cdot\|_\infty$)**.**

1. On considère la matrice $B = (B_{ij})$ de $\mathcal{M}_n(\mathbb{R})$ définie par $B_{ii} = 1$, $B_{ij} = -1$ $i < j$, $B_{ij} = 0$ sinon.
   (a) Calculer B^{-1}.
   (b) En déduire $\text{cond}_1(B)$ et $\text{cond}_\infty(B)$.
2. Soit A une matrice carrée de taille $n \times n$. L'objectif de cette question est de montrer que
$$\frac{1}{n^2}\text{cond}_\infty(A) \leq \text{cond}_1(A) \leq n^2\text{cond}_\infty(A).$$
   (a) Montrer que pour tout $x \in \mathbb{R}^n$,
$$\|x\|_\infty \leq \|x\|_1 \leq n\|x\|_\infty.$$
   (b) En déduire que pour toute matrice carrée de taille $n \times n$
$$\frac{1}{n}\|A\|_\infty \leq \|A\|_1 \leq n\|A\|_\infty.$$
   (c) Conclure.

Exercice 59 (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 $x \in \mathbb{R}^n$, $\lambda \in \mathbb{R}$) :
$$\begin{aligned} &Ax + \lambda b = f, \\ &c \cdot x + \alpha\lambda = \gamma. \end{aligned} \tag{1.73}$$

1. Ecrire le système (1.73) 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 (1.73) soit inversible. Dans toute la suite, on supposera que cette relation est vérifiée.
3. On propose la méthode suivante pour la résolution du système (1.73) :
   (a) On calcule z solution de $Az = b$, et h solution de $Ah = f$.
   (b) On pose $x = h - \dfrac{\gamma - c \cdot h}{\alpha - c \cdot z}z, \qquad \lambda = \dfrac{\gamma - c \cdot h}{\alpha - c \cdot z}.$

   Montrer que $x \in \mathbb{R}^n$ et $\lambda \in \mathbb{R}$ ainsi calculés sont bien solutions du système (1.73).
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.

Exercice 60 (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 $\text{cond}\,(A) \geq \sup_{\substack{B\in\mathcal{M}_n(\mathbb{R})\\ \det B=0}} \frac{\|A\|}{\|A-B\|}$

Exercice 61 (Minoration du conditionnement)**.** *Corrigé détaillé en page 86.*

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

1. Montrer que si $A + \delta_A$ est singulière, alors
$$\text{cond}(A) \geq \frac{\|A\|}{\|\delta_A\|}. \tag{1.74}$$
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 (1.74) 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 (1.74).
[*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 (1.74) pour trouver un minorant, qui tend vers $+\infty$ lorsque α tend vers 0, de $\text{cond}(A)$ pour la matrice
$$A = \begin{pmatrix} 1 & -1 & 1 \\ -1 & \alpha & -\alpha \\ 1 & \alpha & \alpha \end{pmatrix}.$$

Exercice 62 (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 $\text{cond}\,(A^2)$ et $(\text{cond}\,A)^2$?
2. On suppose que A symétrique. Montrer que $\text{cond}_2\,(A^2) = (\text{cond}_2\,A)^2$.
3. On suppose que $\text{cond}_2\,(A^2) = (\text{cond}_2\,A)^2$. Peut-on conclure que A est symétrique ? (justifier la réponse.)

Exercice 63 (Calcul de l'inverse d'une matrice et conditionnement)**.** *Corrigé détaillé en page 86.*

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 :
$$\left\{\begin{array}{lllll} e_1 &=& \dfrac{\|B-A^{-1}\|}{\|A^{-1}\|}, & e_2 &= \dfrac{\|B^{-1}-A\|}{\|A\|} \\ \\ e_3 &=& \|AB-\text{Id}\|, & e_4 &= \|BA-\text{Id}\| \end{array}\right. \tag{1.75}$$
   (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
$$\varepsilon\text{cond}(A) < 1.$$
   Montrer que dans ce cas,
$$e_1 \leq \varepsilon,\ e_2 \leq \frac{\varepsilon\text{cond}(A)}{1-\varepsilon\text{cond}(A)},\ e_3 \leq \varepsilon\text{cond}(A) \text{ et } e_4 \leq \varepsilon\text{cond}(A).$$

(c) On suppose maintenant que $AB - \mathrm{Id} = E'$ 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 64 (Conditionnement du Laplacien discret 1D)**.** *Suggestions en page 82, corrigé détaillé en page 88.*
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 13, 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 65 (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 (1.25) étudié à l'exercice 17, qu'on rappelle :

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

Soit $n \in \mathbb{N}^\star$. On note $U = (u_j)_{j=1,\ldots,n}$ une "valeur approchée" de la solution u du problème (1.25) 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 $A = (N+1)^2B + \mathrm{Id}$, 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

$$\begin{array}{l} -u''(x) = \lambda u(x),\ x \in]0,1[, \\ u(0) = u(1) = 0. \end{array} \tag{1.77}$$

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 66 (Conditionnement "efficace"). *Suggestions en page 82.*

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 13, 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 :

$$\begin{cases} \frac{1}{h^2}(u_i - u_{i-1}) + \frac{1}{h^2}(u_i - u_{i+1}) = b_i,\ \forall i \in \{1, \ldots, n\}, \\ u_0 = u_{n+1} = 0. \end{cases} \tag{1.78}$$

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

   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 $\boldsymbol{\phi} = (\phi_1, \ldots \phi_n) \in \mathbb{R}^n$ par $\phi_i = \varphi(ih)$ pour tout $i \in \{1, \ldots, n\}$. Montrer que $(A\boldsymbol{\phi})_i = 1$ pour tout $i \in \{1, \ldots, n\}$.

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 $\boldsymbol{\phi}$ 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 $i \in \{1, \ldots, n\}$. 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 47 page 75 (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 53 page 77 (Série de Neumann)

1. Montrer que si $\rho(A) < 1$, alors 0 n'est pas valeur propre de $Id + A$ et $\text{Id} - A$.

2. Utiliser le corollaire 1.36.

Exercice 56 page 77 (Propriétés générales du conditionnement)

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.). 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 64 page 80(Conditionnement du Laplacien discret 1D)

2. Chercher les vecteurs propres $\Phi \in \mathbb{R}^n$ de A sous la forme $\Phi_j = \varphi(x_j)$, $j = 1, \dots, n$ 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 66 page 81 (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 $\|A^{-1}\| = \frac{1}{8}$, 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

$$\varphi_h(x) = \begin{cases} \varphi(ih) = \phi_i \text{ si } x \in]x_i - \frac{h}{2},\ x_i + \frac{h}{2}[,\ i = 1, \dots, n, \\ 0 \text{ si } x \in [0, \frac{h}{2}] \text{ ou } x \in]1 - \frac{h}{2}, 1]. \end{cases} \quad f_h(x) = \begin{cases} f(ih) = b_i \text{ si } x \in]x_i - \frac{h}{2},\ x_i + \frac{h}{2}[, \\ 0 \text{ si } x \in [0, \frac{h}{2}] \text{ ou } x \in]1 - \frac{h}{2}, 1]. \end{cases}$$

2. Utiliser le fait que $A\phi = (1 \dots 1)^t$.

1.4.7 Corrigés

Exercice 47 page 75 (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

$$\|A\boldsymbol{x}\|_\infty = \max_{i=1,\dots,n} |\sum_{j=1,\dots,n} a_{i,j}x_j| \leq \max_{i=1,\dots,n} \sum_{j=1,\dots,n} |a_{i,j}||x_j|.$$

Or $\|\boldsymbol{x}\|_\infty = 1$ donc $|x_j| \leq 1$ et

$$\|A\boldsymbol{x}\|_\infty \leq \max_{i=1,\dots,n} \sum_{j=1,\dots,n} |a_{i,j}|.$$

Montrons maintenant que la valeur $\alpha = \max_{i=1,\dots,n} \sum_{j=1,\dots,n} |a_{i,j}|$ 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 $\text{sign}(s)$ le signe de s, c'est–à–dire

$$\text{sign}(s) = \begin{cases} s/|s| \text{ si } s \neq 0, \\ 0 \text{ si } s = 0. \end{cases}$$

Choisissons $\boldsymbol{x} \in \mathbb{R}^n$ défini par $x_j = \text{sign}(a_{i_0,j})$ où i_0 est tel que $\sum_{j=1,\dots,n} |a_{i_0,j}| \geq \sum_{j=1,\dots,n} |a_{i,j}|$, $\forall i = 1,\dots,n$. On a bien $\|\boldsymbol{x}\|_\infty = 1$, et

$$\|A\boldsymbol{x}\|_\infty = \max_{i=1,\dots,n} |\sum_{j=1}^{n} a_{i,j}\text{sign}(a_{i_0,j})|.$$

Or, par choix de $\boldsymbol{x}$, on a

$$\sum_{j=1,\dots,n} |a_{i_0,j}| = \max_{i=1,\dots,n} \sum_{j=1,\dots,n} |a_{i,j}|.$$

On en déduit que pour ce choix de $\boldsymbol{x}$, on a bien $\|A\boldsymbol{x}\| = \max_{i=1,\dots,n} \sum_{j=1,\dots,n} |a_{i,j}|$.

2. Par définition, $\|A\|_1 = \sup_{\substack{\boldsymbol{x}\in\mathbb{R}^n \\ \|\boldsymbol{x}\|_1=1}} \|A\boldsymbol{x}\|_1$, et

$$\|A\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |\sum_{j=1}^{n} a_{i,j}x_j| \leq \sum_{j=1}^{n} |x_j| \left(\sum_{i=1}^{n} |a_{i,j}| \right) \leq \max_{j=1,\dots,n} \sum_{i=1}^{n} |a_{i,j}| \sum_{j=1,\dots,n} |x_j|.$$

Et comme $\sum_{j=1}^{n} |x_j| = 1$, on a bien que $\|A\|_1 \leq \max_{j=1,\dots,n} \sum_{i=1,\dots,n} |a_{i,j}|$.
Montrons maintenant qu'il existe $\boldsymbol{x} \in \mathbb{R}^n$, $\|\boldsymbol{x}\|_1 = 1$, tel que $\|A\boldsymbol{x}\|_1 = \sum_{i=1,\dots,n} |a_{i,j}|$. 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 $\sum_{i=1,\dots,n} |a_{i,j_0}| = \max_{j=1,\dots,n} \sum_{i=1,\dots,n} |a_{i,j}|$. On vérifie alors facilement qu'on a bien $\|A\boldsymbol{x}\|_1 = \max_{j=1,\dots,n} \sum_{i=1,\dots,n} |a_{i,j}|$.

Exercice 51 page 76 (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 53 page 77 (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 $\rho(A) < 1$. Remarquons que

$$(\sum_{k=0}^{n} A^k)(Id - A) = \mathrm{Id} - A^{n+1}. \tag{1.79}$$

Comme $\rho(A) < 1$, d'après le corollaire 1.36, on a $A^k \to 0$ lorsque $k \to 0$. De plus, $Id - A$ est inversible. En passant à la limite dans (1.79) et on a donc

$$(Id - A)^{-1} = \sum_{k=0}^{+\infty} A^k. \tag{1.80}$$

Réciproquement, si $\rho(A) \geq 1$, la série ne peut pas converger en raison du corollaire 1.36.
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 (1.80). On en déduit que si $\|A\| < 1$,

$$\|(Id - A)^{-1}\| \leq \sum_{k=0}^{+\infty} \|A^k\| \leq \sum_{k=0}^{+\infty} \|A\|^k = \frac{1}{1 - \|A\|}.$$

On a de même

$$(Id + A)^{-1} = \sum_{k=0}^{+\infty} (-1)^k A^k,$$

d'où on déduit de manière similaire que

$$\|(Id + A)^{-1}\| \leq \sum_{k=0}^{+\infty} \|A^k\| \leq \sum_{k=0}^{+\infty} \|A\|^k = \frac{1}{1 - \|A\|}.$$

Exercice 55 page 77 (Calcul de conditionnement)

On a $A^tA = \begin{pmatrix} 4 & 2 \\ 2 & 2 \end{pmatrix}$. Les valeurs propres de cette matrice sont $3 \pm \sqrt{5}$ et donc $\mathrm{cond}_2(A) = \sqrt{\frac{3+\sqrt{5}}{3-\sqrt{5}}} \neq 2$.

Exercice 56 page 77 (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 $A^tA = \sigma \mathrm{Id}\ A^t = \alpha^2 A^{-1}$ avec $\alpha = \sqrt{\sigma}$. En posant $Q = \dfrac{1}{\alpha}A$, on a donc $Q^t = \dfrac{1}{\alpha}A^t = \alpha A^{-1} = Q^{-1}$.
Réciproquement, si $A = \alpha Q$, alors $A^tA = \alpha^2 \mathrm{Id}$, $\frac{\sigma_n}{\sigma_1} = 1$, 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 $cond_2(A) = \sqrt{\dfrac{\sigma_n}{\sigma_1}}$ 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 $cond_2(A+B) = \dfrac{\nu_n}{\nu_1}$, où $0 < \nu_1 \leq \ldots \leq \nu_n$ sont les valeurs propres de $A + B$.

a) On va d'abord montrer que

$$cond_2(A+B) \leq \frac{\lambda_n + \mu_n}{\lambda_1 + \mu_1}.$$

On sait que si A est s.d.p., alors

$$cond_2(A) = \frac{\lambda_n}{\lambda_1}.$$

Or, si A est s.d.p., alors $\sup\limits_{\|x\|_2=1} Ax \cdot x = \lambda_n$; il suffit pour s'en rendre compte de décomposer x sur la base $(\boldsymbol{f}_i)_{i=1\dots n}$. Soit $x = \sum\limits_{i=1}^{n} \alpha_i \boldsymbol{f}_i$, alors :

$$Ax \cdot x = \sum_{i=1}^{n} \alpha_i^2 \lambda_i \leq \lambda_n \sum_{i=1}^{n} \alpha_i^2 = \lambda_n.$$

Et $A\boldsymbol{f}_n \cdot \boldsymbol{f}_n = \lambda_n$.
De même, $Ax \cdot x \geq \lambda_1 \sum_{i=1}^{n} \alpha_i^2 = \lambda_1$ et $Ax \cdot x = \lambda_1$ si $x = f_1$. Donc $\inf\limits_{\|x\|=1} Ax \cdot x = \lambda_1$.
On en déduit que si A est s.d.p.,

$$\mathrm{cond}_2(A) = \frac{\sup_{\|x\|=1} Ax \cdot x}{\inf_{\|x\|=1} Ax \cdot x}.$$

Donc $cond_2(A+B) = \dfrac{\sup_{\|x\|=1}(A+B)x \cdot x}{\inf_{\|x\|=1}(A+B)x \cdot x}$. Or

$$\sup_{\|x\|=1}(Ax \cdot x + Bx \cdot x) \leq \sup_{\|x\|=1} Ax \cdot x + \sup_{\|x\|=1} Bx \cdot x = \lambda_n + \mu_n,$$
$$\inf_{\|x\|=1}(Ax \cdot x + Bx \cdot x) \geq \inf_{\|x\|=1} Ax \cdot x + \inf_{\|x\|=1} Bx \cdot x = \lambda_1 + \mu_1,$$

et 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 $\dfrac{a+b}{c+d} \geq \dfrac{a}{c}$ 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 61 page 79 (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 $\rho(B) < 1$. 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

$$(A+\delta_A)x = Ax - \frac{-y\ x^t}{x^t\ x}x = y - \frac{-y\ x^t x}{x^t\ x} = 0.$$

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 (1.74).

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

$$\mathrm{cond}(A) \geq \frac{\|A\|}{\|\delta_A\|}. \tag{1.81}$$

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 63 page 79 (Calcul de l'inverse d'une matrice et conditionnement)

1. (a) L'inverse de la matrice A vérifie les quatre équations suivantes :

$$\begin{cases} X - A^{-1} = 0, & X^{-1} - A = 0, \\ AX - \mathrm{Id} = 0, & XA - \mathrm{Id} = 0. \end{cases}$$

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, $\|AE\| \leq \|A\|\|E\| \leq \text{cond}(A)\varepsilon < 1$; on en déduit, en utilisant le théorème 1.11, que :

$$\|(Id + AE))^{-1}\| \leq \frac{1}{1 - \|AE\|}, \text{ et donc } e_2 \leq \frac{\varepsilon \text{cond}(A)}{1 - \varepsilon \text{cond}(A)}.$$

(γ) Par définition, $e_3 = \|AB - \text{Id}\| = \|A(A^{-1} + E) - \text{Id}\| = \|AE\| \leq \|A\|\|E\| \leq \|A\|\varepsilon\|A^{-1}\| = \varepsilon \text{cond}(A)$.

(δ) Enfin, $e_4 = \|BA - \text{Id}\| = \|(A^{-1} + E)A - \text{Id}\| \leq \|EA\| \leq \|E\|\|A\| \leq \varepsilon \text{cond}(A)$.

(c) (α) Comme $B = A^{-1}(Id + E')$, on a

$$e_1 = \frac{\|A^{-1}(Id + E') - A^{-1}\|}{\|A^{-1}\|} \leq \|\text{Id} + E' - \text{Id}\| \leq \varepsilon.$$

(β) Par définition,

$$\begin{aligned}e_2 &= \frac{\|(Id+E')^{-1}A - A\|}{\|A\|} \\ &= \frac{\|(Id+E')^{-1}(A - (Id+E')A)\|}{\|A\|} \\ &\leq \|(Id + E')^{-1}\|\|\text{Id} - (Id + E')\| \leq \frac{\varepsilon}{1-\varepsilon}\end{aligned}$$

car $\varepsilon < 1$ (théorème 1.1).

(γ) Par définition, $e_3 = \|AB - \text{Id}\| = \|AA^{-1}(Id + E') - \text{Id}\| = \|E'\| \leq \varepsilon$.

(δ) Enfin, $e_4 = \|BA - \text{Id}\| = \|A^{-1}(Id + E')A - \text{Id}\| = \|A^{-1}(A + E'A - A)\| \leq \|A^{-1}\|\|AE'\| \leq \varepsilon \text{cond}(A)$.

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 $\|(A + \delta_A)^{-1} - A^{-1}\| = \|(A + \delta_A)^{-1}(\text{Id} - (A + \delta_A)A^{-1}\| \leq \|(A + \delta_A)^{-1}\|\|\text{Id} - \text{Id} - \delta_A A^{-1}\| \leq \|(A + \delta_A)^{-1}\|\delta_A\|\|A^{-1}\|$. On en déduit le résultat.

Exercice 64 page 80 (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

$$\begin{cases} -\varphi''(x) = \lambda\varphi(x) & x \in]0,1[\\ \varphi(0) = \varphi(1) = 0 \end{cases} \tag{1.82}$$

(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 (1.82) 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 $A\Phi^{(k)}$:

$$(A\Phi^{(k)})_i = -\sin k\pi(i-1)h + 2\sin k\pi(ih) - \sin k\pi(i+1)h.$$

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

$$\lambda_k = \frac{2}{h^2}(1 - \cos k\pi h) = \frac{2}{h^2}(1 - \cos\frac{k\pi}{n+1}) \tag{1.83}$$

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

$$\lambda_k^{(h)} = \frac{2}{h^2}\left(1 - 1 + \frac{k^2\pi^2h^2}{2} + O(h^4)\right) = k^2\pi^2 + O(h^2)$$

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 66 page 81 (Conditionnement "efficace")

Partie I
1. Soit $u = (u_1, \ldots, u_n)^t$. On a

$$Au = b \Leftrightarrow \begin{cases} \dfrac{1}{h^2}(u_i - u_{i-1}) + \dfrac{1}{h^2}(u_i - u_{i+1}) = b_i, \ \forall i = 1, \ldots, n, \\ u_0 = u_{n+1} = 0. \end{cases}$$

Supposons $b_i \geq 0, \forall i = 1, \ldots, n$, et soit

$$p = \min\{k \in \{0, \ldots, n+1\}; u_k = \min\{u_i, \ i = 0, \ldots, n+1\}\}.$$

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 $p \in \{1, \ldots, n\}$, 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),

$$u_i + \|b\|\phi_i \geq 0 \quad \forall i = 1 \ldots n.$$

Si maintenant on prend $\bar{b}_i = b_i - \|b\| \leq 0$, alors

$$u_i - \|b\|\phi_i \leq 0 \quad \forall i = 1, \ldots, n.$$

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 $\|A^{-1}\| = \frac{1}{8}$ 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

$$\begin{array}{ll} \varphi_h(x) & = \left\{ \begin{array}{l} \varphi(ih) = \phi_i \text{ si } x \in]x_i - \frac{h}{2},\ x_i + \frac{h}{2}[,\ i = 1, \ldots, n, \\ 0 \text{ si } x \in [0, \frac{h}{2}] \text{ ou } x \in]1 - \frac{h}{2}, 1]. \end{array} \right. \quad \text{et} \\ f_h(x) & = \left\{ \begin{array}{l} f(ih) = b_i \text{ si } x \in]x_i - \frac{h}{2},\ x_i + \frac{h}{2}[, \\ f(ih) = 0 \text{ si } x \in [0, \frac{h}{2}] \text{ ou } x \in]1 - \frac{h}{2}, 1]. \end{array} \right. \end{array}$$

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, $A\varphi = (1 \ldots 1)^t$ donc $u \cdot A\varphi = \displaystyle\sum_{i=1}^{n} u_i$; or $u \cdot A\varphi = A^t u \cdot \varphi = Au \cdot \varphi$ car A est symétrique. Donc $u \cdot A\varphi = \displaystyle\sum_{i=1}^{n} b_i\varphi_i \geq \frac{\alpha}{h}$ d'après la question 1. Comme $\delta_u = A^{-1}\delta_b$, on a donc $\|\delta_u\| \leq \|A^{-1}\|\ \|\delta_b\|$; et comme $n\|u\| \geq \dfrac{\alpha}{h}$, on obtient : $\dfrac{\|\delta_u\|}{\|u\|} \leq \dfrac{1}{8}\dfrac{hn}{\alpha}\|\delta_b\|\dfrac{\|f\|}{\|b\|}$. 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 46). Si on utilise une méthode directe genre Choleski, on aura donc besoin de la place mémoire pour stocker la structure bande.
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. Par exemple, si on a seulement 5 diagonales non nulles dans la matrice du système à résoudre, système de n équations et n inconnues, la place mémoire nécessaire pour un produit matrice vecteur est $6n$. 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 1.48 (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, découverte dans les années 50[6]*. Elle est miraculeuse en ce sens qu'elle donne la solution exacte du système $A\boldsymbol{x} = \boldsymbol{b}$ 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 : $f(\boldsymbol{x}) = \frac{1}{2}A\boldsymbol{x} \cdot \boldsymbol{x} - \boldsymbol{b} \cdot \boldsymbol{x}$. Or on peut montrer que lorsque A est symétrique définie positive, la recherche de $\boldsymbol{x}$ minimisant f dans $\mathbb{R}^n$ est équivalent à la résolution du système $A\boldsymbol{x} = \boldsymbol{b}$ (Voir paragraphe 3.2.2 page 197). Malheureusement, 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 général 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. Ces problèmes ont été résolus grâce On a alors recours aux techniques dites 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 68 page 103 et du théorème 3.19 page 205.

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.

6. Hestenes, Magnus R. ; Stiefel, Eduard (December 1952). "Methods of Conjugate Gradients for Solving Linear Systems". Journal of Research of the National Bureau of Standards. 49 (6).

Définition 1.49 (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 1.50 (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 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

$$P\boldsymbol{x} = (P-A)\boldsymbol{x}+\boldsymbol{b} = N\boldsymbol{x}+\boldsymbol{b}. \tag{1.84}$$

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 :

$$\begin{aligned} P\boldsymbol{x}^{(k+1)} &= (P-A)\boldsymbol{x}^{(k)}+\boldsymbol{b} \\ &= N\boldsymbol{x}^{(k)}+\boldsymbol{b}, \end{aligned} \tag{1.85}$$

ce qui peut également s'écrire :.

$$\boldsymbol{x}^{(k+1)} = B\boldsymbol{x}^{(k)}+\boldsymbol{c}, \text{ avec } B=P^{-1}(P-A)=\mathrm{Id}-P^{-1}A=P^{-1}N \text{ et } \boldsymbol{c}=P^{-1}\boldsymbol{b}. \tag{1.86}$$

Remarque 1.51 (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

$$\boldsymbol{e}^{(k)} = \boldsymbol{x}^{(k)}-\boldsymbol{x},\ k\in\mathbb{N} \tag{1.87}$$

où $\boldsymbol{x}^{(k)}$ est construit par (1.86) 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 1.52. *La suite* $(\boldsymbol{e}^{(k)})_{k\in\mathbb{N}}$ *définie par* (1.87) *est également définie par*

$$\begin{aligned} \boldsymbol{e}^{(0)} &= \boldsymbol{x}^{(0)}-\boldsymbol{x} \\ \boldsymbol{e}^{(k)} &= B^k\boldsymbol{e}^{(0)} \end{aligned} \tag{1.88}$$

DÉMONSTRATION – Comme $\boldsymbol{c}=P^{-1}\boldsymbol{b}=P^{-1}A\boldsymbol{x}$, on a

$$\boldsymbol{e}^{(k+1)} = \boldsymbol{x}^{(k+1)}-\boldsymbol{x} = B\boldsymbol{x}^{(k)}-\boldsymbol{x}+P^{-1}A\boldsymbol{x} \tag{1.89}$$

$$= B(\boldsymbol{x}^{(k)}-\boldsymbol{x}). \tag{1.90}$$

Par récurrence sur k,

$$\boldsymbol{e}^{(k)} = B^k(\boldsymbol{x}^{(0)}-\boldsymbol{x}),\quad \forall k\in\mathbb{N}. \tag{1.91}$$

■

Théorème 1.53 (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* (1.86)*.*

1. *La suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ converge, quel que soit $\boldsymbol{x}^{(0)}$, 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, quel que soit $\boldsymbol{x}^{(0)}$, 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 (1.86) converge vers $\boldsymbol{x} = A^{-1}\boldsymbol{b}$ si et seulement si la suite $\boldsymbol{e}^{(k)}$ définie par (1.88) tend vers $\mathbf{0}$. On en déduit par le lemme 1.36 que la suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ converge (vers $\boldsymbol{x}$), pour tout $x^{(0)}$, 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 1.36, $\rho(B) < 1$ et donc la méthode converge pour tout $x^{(0)}$.
   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 1.35 : 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 (1.86). 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 90 page 127 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 choisie 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 90 page 127) 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

$$A = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}. \tag{1.92}$$

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 $(\boldsymbol{e}^{(k)})_{k\in\mathbb{N}}$ 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 B 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 (1.92), 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

$$\begin{aligned} &\boldsymbol{x}^{(0)} \in \mathbb{R}^n \text{ donné}, \\ &\boldsymbol{x}^{(k+1)} = B\boldsymbol{x}^{(k)} + \boldsymbol{c}, \text{ avec } \boldsymbol{c} = \alpha\boldsymbol{b}. \end{aligned} \tag{1.93}$$

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 $1 - \alpha = 3\alpha - 1$, ce qui donne $\alpha = \frac{1}{2}$. Cette méthode est connue sous le nom de *méthode de Richardson*[7]. 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 (1.93) qu'on vient d'étudier.

La méthode de Jacobi

Dans le cas de l'exemple de la matrice A donné par (1.92), 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 matrice diagonale dont les coefficients sont les coefficients situés sur la diagonale de A. La *méthode de Jacobi*[8] consiste justement à prendre $P = D$, et ce même si la diagonale de A n'est pas constante.

7. 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 des guerres et de leur prévention.

8. 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....

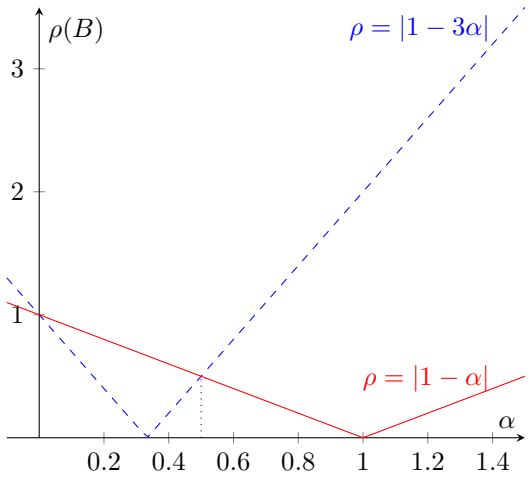


FIGURE 1.4: Rayon spectral de la matrice B de Richardson en fonction du coefficient α.

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 (1.92), et donc dans ce cas la méthode de Jacobi s'écrit

$$\begin{aligned} &\boldsymbol{x}^{(0)} = \begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \end{bmatrix} \in \mathbb{R}^2 \text{ donné }, \\ &\boldsymbol{x}^{(k+1)} = \begin{bmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \end{bmatrix} = B_J \boldsymbol{x}^{(k)} + \boldsymbol{c}, \text{ avec } B_J = \begin{bmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0, \end{bmatrix} \text{ et } \boldsymbol{c} = \frac{1}{2}\boldsymbol{b}. \end{aligned} \tag{1.94}$$

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 :

$$D = \begin{bmatrix} a_{1,1} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & & 0 & a_{n,n} \end{bmatrix}, \; -E = \begin{bmatrix} 0 & 0 & \dots & 0 \\ a_{2,1} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & 0 \\ a_{n,1} & \dots & a_{n-1,n} & 0 \end{bmatrix} \text{ et } -F = \begin{bmatrix} 0 & a_{1,2} & \dots & a_{1,n} \\ \vdots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & a_{n,n-1} \\ 0 & \dots & 0 & -0 \end{bmatrix}. \tag{1.95}$$

La méthode de Jacobi s'écrit donc :

$$\begin{cases} \boldsymbol{x}^{(0)} \in \mathbb{R}^n \\ D\boldsymbol{x}^{(k+1)} = (E+F)\boldsymbol{x}^{(k)} + \boldsymbol{b}. \end{cases} \tag{1.96}$$

Lorsqu'on écrit la méthode de Jacobi comme sous la forme (1.86) on a $B = D^{-1}(E+F)$; on notera B_J cette matrice :

$$B_J = \begin{bmatrix} 0 & -\frac{a_{1,2}}{a_{1,1}} & \dots & -\frac{a_{1,n}}{a_{1,1}} \\ -\frac{a_{2,1}}{a_{2,2}} & \ddots & & -\frac{a_{2,n}}{a_{2,2}} \\ \vdots & \ddots & \ddots & \vdots \\ -\frac{a_{n,1}}{a_{n,n}} & \dots & -\frac{a_{n,n-1}}{a_{n,n}} & 0 \end{bmatrix}.$$

La méthode de Jacobi s'écrit aussi :

$$\begin{cases} \boldsymbol{x}^{(0)} \in \mathbb{R}^n \\ a_{i,i} x_i^{(k+1)} = -\sum_{j<i} a_{i,j} x_j^{(k)} - \sum_{j>i} a_{i,j} x_j^{(k)} + b_i \;\; i = 1, \dots, n. \end{cases} \tag{1.97}$$

La méthode de Gauss-Seidel

Dans l'écriture (1.97) 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[9] 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 "nouvelle" valeur $x_1^{(k+1)}$ qu'on vient de calculer plutôt que la valeur $x_1^{(k)}$ comme dans (1.97) ; 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 (1.97) $x_j^{(k)}$ par $x_j^{(k+1)}$ si $j < i$. On obtient donc l'algorithme suivant :

$$\left\{ \begin{array}{l} \boldsymbol{x}^{(0)} \in \mathbb{R}^n \\ a_{i,i} x_i^{(k+1)} = -\sum_{j<i} a_{i,j} x_j^{(k+1)} - \sum_{i<j} a_{i,j} x_j^{(k)} + b_i, \quad i = 1, \ldots, n. \end{array} \right. \tag{1.98}$$

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 $P - A = F$:

$$\left\{ \begin{array}{l} \boldsymbol{x}_0 \in \mathbb{R}^n \\ (D - E)\boldsymbol{x}^{(k+1)} = F\boldsymbol{x}^{(k)} + \boldsymbol{b}. \end{array} \right. \tag{1.99}$$

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 (1.92) : 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 :

$$\begin{array}{l} \boldsymbol{x}^{(0)} = \begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \end{bmatrix} \in \mathbb{R}^2 \text{ donné} , \\ \boldsymbol{x}^{(k+1)} = \begin{bmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \end{bmatrix} = B_{GS}\boldsymbol{x}^{(k)} + \boldsymbol{c}, \text{ avec } B_{GS} = \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{4} \end{bmatrix} \text{ et } \boldsymbol{c} = \begin{bmatrix} \frac{1}{2} & 0 \\ \frac{1}{4} & \frac{1}{2} \end{bmatrix} \boldsymbol{b}. \end{array} \tag{1.100}$$

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 à chaque itération 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 1.54 (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, voir exercice 70

9. 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.

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 :

$$\left\{ \begin{array}{l} x_0 \in \mathbb{R}^n \\ a_{i,i}\tilde{x}_i^{(k+1)} = -\sum_{j<i} a_{i,j}x_j^{(k+1)} - \sum_{i<j} a_{i,j}x_j^{(k)} + b_i \\ x_i^{(k+1)} = \omega\tilde{x}_i^{(k+1)} + (1-\omega)x_i^{(k)}, \quad i = 1, \ldots, n. \end{array} \right. \tag{1.101}$$

(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 (1.101)) :

$$\left\{ \begin{array}{l} x^{(0)} \in \mathbb{R}^n \\ a_{i,i}x_i^{(k+1)} = \omega\left[-\sum_{j<i} a_{i,j}x_j^{(k+1)} - \sum_{j>i} a_{i,j}x_j^{(k)} + b_i\right] \\ \qquad\qquad +(1-\omega)a_{i,i}x_i^{(k)}. \end{array} \right. \tag{1.102}$$

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 1.55 (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 (1.95)*; 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 1.62 page 101). 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 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 1.56 (Condition suffisante de convergence d'une méthode itérative)**.** *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* (1.86) *converge.*

DÉMONSTRATION – On rappelle (voir le corollaire (1.39) page 70) 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 :

$$\|P^{-1}N\boldsymbol{x}\|_* < \|x\|_*,\ \forall \boldsymbol{x} \in \mathbb{R}^n,\ \boldsymbol{x} \neq 0. \tag{1.103}$$

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 (1.103) est vérifiée par cette norme. Soit $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{x} \neq 0$, on a : $\|P^{-1}N\boldsymbol{x}\|_*^2 = AP^{-1}N\boldsymbol{x}\cdot P^{-1}N\boldsymbol{x}$. Or $N = P - A$, et donc : $\|P^{-1}N\boldsymbol{x}\|_*^2 = A(\mathrm{Id} - P^{-1}A)\boldsymbol{x}\cdot(Id - P^{-1}A)\boldsymbol{x}$. 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 $\rho(P^{-1}N) < 1$), 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 : $P\boldsymbol{y}\cdot\boldsymbol{y} = \boldsymbol{y}\cdot P^t\boldsymbol{y} = P^t\boldsymbol{y}\cdot\boldsymbol{y}$, 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 1.57 (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 (1.95)*; 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 1.55 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 1.56 page 99, 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 1.58 (Comparaison Gauss–Seidel/Jacobi)**.** *On a vu (théorème 1.57) 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 69 page 104.*

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 75 page 106, et un résultat de comparaison des méthodes pour les matrices tridiagonales par blocs, voir le théorème 1.59 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 1.57) tel que

$$\rho(B_{\omega_0}) \leq \rho(B_\omega), \forall \omega \in]0,2[.$$

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 1.59 (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 1.104 page 101 ; 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. On suppose de plus que toutes les valeurs propres de la matrice d'itération J de la méthode de Jacobi sont réelles et que $\rho(B_{\mathrm{J}}) < 1$. 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 1.60 (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 71 page 104). 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 alors de cette structure lors de la résolution par une méthode itérative.

Définition 1.61. *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*

$$A = \begin{bmatrix} A_{1,1} & A_{1,2} & \dots & & \dots & A_{1,S} \\ A_{2,1} & \ddots & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & A_{S-1,S} \\ A_{S,1} & \dots & & \dots & A_{S,S-1} & A_{S,S} \end{bmatrix} \tag{1.104}$$

Remarque 1.62.

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. Les méthodes que nous allons décrire maintenant sont alors celles que nous avons vu dans le cas de matrices sans structure particulière.*
2. *Si A est symétrique définie positive, la condition $A_{i,i}$ inversible dans la définition 1.61 est inutile car $A_{i,i}$ est nécessairement symétrique définie positive donc inversible. Pour s'en convaincre, 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* (1.104) *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 cherche une matrice P tel que 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). On avait pris pour P une matrice diagonale dans la méthode de Jacobi. La méthode de Jacobi par blocs 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 :

$$E = \begin{bmatrix} 0 & 0 & \dots & & \dots & 0 \\ -A_{2,1} & \ddots & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & 0 \\ -A_{S,1} & \dots & & \dots & -A_{S,S-1} & 0 \end{bmatrix}, \; F = \begin{bmatrix} 0 & -A_{1,2} & \dots & & \dots & -A_{1,S} \\ 0 & \ddots & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & -A_{S-1,S} \\ 0 & \dots & & \dots & 0 & 0 \end{bmatrix}.$$

On a bien $A = P - N$ et avec D, E et F définies comme ci-dessus, la méthode de Jacobi s'écrit :

$$\begin{cases} x^{(0)} \in \mathbb{R}^n \\ Dx^{(k+1)} = (E+F)x^{(k)} + b. \end{cases} \tag{1.105}$$

Lorsqu'on écrit la méthode de Jacobi comme sous la forme (1.86) 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 :

$$\begin{cases} x_0 \in \mathbb{R}^n \\ A_{i,i}x_i^{(k+1)} = -\displaystyle\sum_{j<i} A_{i,j}x_j^{(k)} - \sum_{j>i} A_{i,j}x_j^{(k)} + b_i \;\; i = 1, \dots, S. \end{cases} \tag{1.106}$$

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 :

$$\begin{cases} x_0 \in \mathbb{R}^n \\ a_{i,i}x_i^{(k+1)} = -\displaystyle\sum_{j<i} a_{i,j}x_j^{(k)} - \sum_{j>i} a_{i,j}x_j^{(k)} + b_i \;\; i = 1, \dots, n. \end{cases} \tag{1.107}$$

Méthode de Gauss-Seidel

La même procédure que dans le cas $S = n$ et $n_i = 1$ donne :

$$\begin{cases} x^{(0)} \in \mathbb{R}^n \\ A_{i,i}x_i^{(k+1)} = -\sum_{j<i} A_{i,j}x_j^{(k+1)} - \sum_{i<j} A_{i,j}x_j^{(k)} + b_i, \quad i = 1, \dots, S. \end{cases} \tag{1.108}$$

La méthode de Gauss–Seidel s'écrit donc sous forme la forme $Px^{(k+1)} = (P-A)x^{(k)} + b$, $P = D - E$ et $P - A = F$:

$$\begin{cases} x_0 \in \mathbb{R}^n \\ (D-E)x^{(k+1)} = Fx^{(k)} + b. \end{cases} \tag{1.109}$$

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 :

$$\begin{cases} x_0 \in \mathbb{R}^n \\ A_{i,i}\tilde{x}_i^{(k+1)} = -\displaystyle\sum_{j<i} A_{i,j}x_j^{(k+1)} - \sum_{i<j} A_{i,j}x_j^{(k)} + b_i \\ x_i^{(k+1)} = \omega\tilde{x}_i^{(k+1)} + (1-\omega)x_i^{(k)}, \quad i = 1, \dots, S. \end{cases} \tag{1.110}$$

(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 (1.101)) :

$$\left\{ \begin{array}{l} x^{(0)} \in \mathbb{R}^n \\ A_{i,i}x_i^{(k+1)} = \omega \left[-\sum_{j<i} A_{i,j}x_j^{(k+1)} - \sum_{j>i} A_{i,j}x_j^{(k)} + b_i \right] \\ \qquad\qquad +(1-\omega)A_{i,i}x_i^{(k)}. \end{array} \right. \tag{1.111}$$

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

$$B = \left(\frac{D}{\omega} - E\right)^{-1} (F + \left(\frac{1-\omega}{\omega}\right) D).$$

Remarque 1.63 (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 71 page 104). 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 (méthodes itératives)

Exercice 67 (Convergence de suites)**.** *Corrigé en page 114*

Etudier la convergence de la suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}} \subset \mathbb{R}^n$ 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 68 (Méthode de Richardson)**.** *Suggestions en page 113, corrigé en page 114*

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 Optimisation. 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 $\nabla f(\boldsymbol{x}) = A\boldsymbol{x} - \boldsymbol{b}$, et annuler le gradient consiste à résoudre le système linéaire $A\boldsymbol{x} = \boldsymbol{b}$.

Exercice 69 (Non convergence de la méthode de Jacobi)**.** *Suggestions en page 113. Corrigé en page 115.*
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 70 (Jacobi et Gauss–Seidel : cas des matrices tridiagonales)**.** L'objet de cet exercice est de démontrer le théorème 1.54 sur la comparaison des méthodes de Jacobi et Gauss-Seidel pour les matrices tridiagonales.
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 particulière $A = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$ et calculer leurs rayons spectraux. Montrer que les méthodes convergent.
2. Montrer que λ est valeur propre de B_J si et seulement s'il existe un vecteur complexe $\boldsymbol{x} = (x_1, \dots x_n) \in \mathbb{C}^n$, $\boldsymbol{x} \neq \mathbf{0}$, 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
$$-\lambda^2 a_{p,p-1}y_{p-1} - a_{p,p+1}y_{p+1} = \lambda^2 a_{p,p}y_p, \quad p = 1, \dots, n.$$
4. Montrer que μ est valeur propre de B_{GS} associée à un vecteur propre $\boldsymbol{z} \neq \mathbf{0}$ si et seulement si
$$(F - \mu(D - E))\,\boldsymbol{z} = \mathbf{0}.$$
5. Montrer que 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 $\rho(B_{GS}) \neq \rho(B_J)^2$. Quelle est l'hypothèse mise en défaut ici ?

Exercice 71 (Méthode de Jacobi et relaxation)**.** *Suggestions en page 114, corrigé en page 116*
Soit $n \geq 1$. Soit $A = (a_{i,j})_{i,j=1,\dots,n} \in \mathcal{M}_n(\mathbb{R})$ 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 :

$$\begin{array}{l} D = (d_{i,j})_{i,j=1,\dots,n},\ d_{i,j} = 0 \text{ si } i \neq j,\ d_{i,i} = a_{i,i}, \\ E = (e_{i,j})_{i,j=1,\dots,n},\ e_{i,j} = 0 \text{ si } i \leq j,\ e_{i,j} = -a_{i,j} \text{ si } i > j, \\ F = (f_{i,j})_{i,j=1,\dots,n},\ f_{i,j} = 0 \text{ si } i \geq j,\ f_{i,j} = -a_{i,j} \text{ si } i < j. \end{array}$$

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 72 (Une méthode itérative pour un système linéaire)**.**
Soient $n \in \mathbb{N}$ tel que $n \geq 3$ et $\boldsymbol{b} \in \mathbb{R}^n$, de composantes $(b_1, \ldots, b_n)$. On cherche $\boldsymbol{x} \in \mathbb{R}^n$, de composantes $(x_1, \ldots, x_n)$, solution de

$$\begin{cases} 4x_1 + x_2 = b_1, \\ x_{i-1} + 4x_i + x_{i+1} = b_i, \ i = 2, \ldots, n-1, \\ x_{n-1} + 4x_n = b_n. \end{cases} \tag{1.112}$$

1. On suppose, dans cette question uniquement, que $\boldsymbol{b} = \boldsymbol{0}$.
   (a) Montrer que $4|x_i| \leq 2\|\boldsymbol{x}\|_\infty$ pour $i \in \{1, \ldots, n\}$.
   (b) En déduire que $\boldsymbol{x} = \boldsymbol{0}$.
2. Montrer que dans le cas d'un second membre quelconque $\boldsymbol{b}$, il existe une unique $\boldsymbol{x} \in \mathbb{R}^n$ solution du système linéaire (1.112).
3. Afin de résoudre le système, on considère la méthode itérative suivante : $x^{(0)} = 0 \in \mathbb{R}^n$ et

$$\begin{cases} x_1^{(k+1)} = \alpha x_1^{(k)} + \frac{\alpha-1}{4}(x_2^{(k)} - b_1), \\ x_i^{(k+1)} = \alpha x_i^{(k)} + \frac{\alpha-1}{4}(x_{i-1}^{(k)} + x_{i+1}^{(k)} - b_i), \ i = 2, \ldots, n-1, \\ x_n^{(k+1)} = \alpha x_n^{(k)} + \frac{\alpha-1}{4}(x_{n-1}^{(k)} - b_n). \end{cases} \tag{1.113}$$

   qui dépend donc du paramètre $\alpha \in \mathbb{R}$. On cherche maintenant le paramètre α qui assure une convergence optimale.

(a) Montrer que pour tout $\alpha \in \mathbb{R}$, on a

$$\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}\|_\infty \leq \left(|\alpha| + \left|\frac{\alpha-1}{2}\right|\right) \|\boldsymbol{x}^{(k)} - \boldsymbol{x}\|_\infty.$$

(b) Trouver $\alpha_{\min}, \alpha_{\max} \in \mathbb{R}$, tels que $\alpha \in]\alpha_{\min}, \alpha_{\max}[$ si et seulement si $|\alpha| + \left|\frac{\alpha-1}{2}\right| < 1$.

(c) Montrer que la méthode itérative converge vers x pour $\alpha \in]\alpha_{\min}, \alpha_{\max}[$.

(d) Lorsque $\alpha = 0$, de quelle méthode itérative (vue en cours) s'agit-il ?

Exercice 73 (Jacobi pour une matrice 3×3 particulière)**.**

Soit $A = \begin{bmatrix} a & 0 & \alpha \\ 0 & b & 0 \\ \alpha & 0 & c \end{bmatrix}$. On suppose que A est symétrique définie positive. Montrer que la méthode de Jacobi converge pour n'importe quel second membre et n'importe quel choix initial.

Exercice 74 (Une matrice cyclique)**.** *Suggestions en page 114*

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 75 (Jacobi pour les matrices à diagonale dominante stricte)**.** *Suggestions en page 114, corrigé en page 118*

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 76 (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 :

$$\begin{cases} -u''(x) + \alpha u(x) = f(x), \ x \in [0,1], \\ u(0) = 0, u(1) = 1, \end{cases} \tag{1.114}$$

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 77 (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 :

$$\text{pour tous ensembles d'indices } I \subset \{1,\dots,n\}, I \neq \emptyset, \text{ et } J = \{1\dots,n\} \setminus I, J \neq \emptyset, \ \exists i \in I, \exists j \in J; a_{i,j} \neq 0. \tag{1.115}$$

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 :

$$\forall i = 1,\dots,n, a_{i,i} \geq \sum_{j \neq i} |a_{i,j}| \tag{1.116}$$

(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 (1.116). 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 $\rho(B_{\text{J}})$ son rayon spectral. On suppose que A vérifie la propriété supplémentaire suivante :

$$\exists i_0; a_{i_0,i_0} > \sum_{j \neq i_0} |a_{i,j}|. \tag{1.117}$$

   (a) Montrer que $\rho(B_{\text{J}}) \leq 1$.
   (b) Montrer que si $Jx = \lambda x$ avec $|\lambda| = 1$, alors $|x_i| = \|x\|_\infty$, $\forall i = 1,\dots,n$, où $\|x\|_\infty = \max_{k=1,\dots,N} |x_k|$. En déduire que $x = 0$ et que la méthode de Jacobi converge.
   (c) Retrouver ainsi le résultat de la question 2 de l'exercice 76.

5. En déduire que si A est une matrice qui vérifie les propriétés (1.115), (1.116) et (1.117), alors A est inversible.

6. Montrer que la matrice A suivante est symétrique définie positive et vérifie les propriétés (1.116) et (1.117).

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

La méthode de Jacobi converge-t-elle pour la résolution d'un système linéaire dont la matrice est A ?

Exercice 78 (Méthodes de Jacobi et Gauss Seidel pour une matrice 3×3)**.** *Corrigé détaillé en page 119*

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 $\boldsymbol{x}^{(0)}$:

$$(i)\ \ \boldsymbol{x}^{(0)} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \qquad (ii)\ \ \boldsymbol{x}^{(0)} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}.$$

Exercice 79 (Convergence en un nombre fini d'itérations)**.**

1 Soit α et β des réels. Soit $u^{(0)} \in \mathbb{R}$ et $(u^{(k)})_{k\in\mathbb{N}}$ la suite réelle définie par $u^{(k+1)} = \alpha u^{(k)} + \beta$.

1.a Donner les valeurs de α et β pour lesquelles la suite $(u^{(k)})_{k\in\mathbb{N}}$ converge.

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 $K \in \mathbb{N}$ tel que $u_K = \overline{u}$, alors $u^{(k)} = \overline{u}$ pour tout $k \in \mathbb{N}$.

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 $u^{(k+1)} = Bu^{(k)} + c$.

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 $u_0 \in \mathbb{R}^N$.

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 80 (SOR et Jacobi pour une matrice tridiagonale)**.**

Soit $A = (a_{i,j})_{i,j=1,\ldots,n} \in \mathcal{M}_n(\mathbb{R})$ une matrice carrée d'ordre n tridiagonale, c'est-à-dire telle que $a_{i,j} = 0$ si $|i-j| > 1$, et dont la matrice diagonale extraite $D = \mathrm{diag}(a_{i,i})_{i=1,\ldots,n}$ est inversible.
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 81 (Méthode de Jacobi et Gauss-Seidel pour des matrices triangulaires)**.** *Soit $A \in \mathcal{M}_n(\mathbb{R})$. On note D la partie diagonale de A, $-E$ la partie inférieure stricte et $-F$ la partie supérieure stricte, de sorte que $A = D - E - F$. On suppose que D est inversible et on note B_J et B_{GS} les matrices des itérations des méthodes de Jacobi et Gauss-Seidel. (On rappelle que $B_J = D^{-1}(E+F)$ et $B_{GS} = (D-E)^{-1}F$.)*

1. *On suppose dans cette questions que $F = 0$, calculer $\rho(B_J)$ et $\rho(B_{GS})$. Pour $b \in \mathbb{R}^n$, Les methodes de Jacobi et Gauss-Seidel donnent-elles la solution exacte du système $Ax = b$ après un nombre fini d'itérations ? si oui, combien faut-il au plus d'itérations ?*
   [Suggestion : Commencer par le cas $n = 2$.]
2. *On suppose dans cette questions $E = 0$, calculer $\rho(B_J)$ et $\rho(B_{GS})$. Pour $b \in \mathbb{R}^n$, Les methodes de Jacobi et Gauss-Seidel donnent-elles la solution exacte du système $Ax = b$ après un nombre fini d'itérations ? si oui, combien faut-il au plus d'itérations ?*

Exercice 82 (Méthode de Jacobi pour des matrices particulières)**.** *Suggestions en page 114, corrigé en page 121*

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 :

$$a_{i,j} \leq 0, \forall i, j = 1, \dots, n, \ i \neq j, \tag{1.118}$$

$$a_{i,i} > 0, \forall i = 1, \dots, n. \tag{1.119}$$

$$\sum_{i=1}^{n} a_{i,j} = 0, \forall j = 1, \dots, n. \tag{1.120}$$

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 $\lambda\mathrm{Id} + A$ est inversible.
3. On considère le système linéaire suivant :
$$(\lambda\mathrm{Id} + A)u = b \tag{1.121}$$
Montrer que la méthode de Jacobi pour la recherche de la solution de ce système définit une suite $(u^{(k)})_{k\in N}$ de $\mathbb{R}^n$.
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 $\alpha = \frac{\lambda}{\max_{i=1,\dots,n} a_{i,i}}$.
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 (1.121).

Exercice 83 (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 $Px^{(k+1)} = Nx^{(k)} + b$ avec $A = P - N$ et $N = \mathrm{diag}(\beta - \alpha_i)$ (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 $P^{-1}N$; montrer qu'il existe un vecteur $x \in \mathbb{C}^n$ non nul tel que $Nx \cdot \overline{x} = \lambda Px \cdot \overline{x}$ (où $\overline{x}$ désigne le conjugué de x). En déduire que toutes les valeurs propres de la matrice $P^{-1}N$ sont réelles.

2. Montrer que le rayon spectral $\rho(P^{-1}N)$ de la matrice vérifie : $\rho(P^{-1}N) \leq \max_{i=1,n} \dfrac{|\beta - \alpha_i|}{\beta}$

3. Déduire de la question 1. que si $\beta > \dfrac{\overline{\alpha}}{2}$, où $\overline{\alpha} = \max_{i=1,n} \alpha_i$, alors $\rho(P^{-1}N) < 1$, 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 $i = 1, \ldots, n$, avec $\underline{\alpha} = \min_{i=1,\ldots,n} \alpha_i$ et $\bar{\alpha} = \max_{i=1,\ldots,n} \alpha_i$ et en déduire que $\max_{i=1,n} |\beta - \alpha_i| = \max(\beta - \underline{\alpha}, \bar{\alpha} - \beta)$) .

Exercice 84 (Méthode des directions alternées)**.** Soit $n \in \mathbb{N}$, $n \geq 1$ et soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice carrée d'ordre n symétrique inversible et $\boldsymbol{b} \in \mathbb{R}^n$. On cherche à calculer $\boldsymbol{u} \in \mathbb{R}^n$, solution du système linéaire suivant :

$$A\boldsymbol{u} = \boldsymbol{b}, \tag{1.122}$$

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 $\boldsymbol{u}^{(0)} \in \mathbb{R}^n$, on propose la méthode itérative suivante pour résoudre (1.122) :

$$(X + \alpha\mathrm{Id})\boldsymbol{u}^{(k+1/2)} = -Y\boldsymbol{u}^{(k)} + \boldsymbol{b}, \tag{1.123a}$$

$$(Y + \alpha\mathrm{Id})u^{(k+1)} = -X\boldsymbol{u}^{(k+1/2)} + \boldsymbol{b}. \tag{1.123b}$$

1. Montrer que la méthode itérative (1.123) définit bien une suite $(\boldsymbol{u}^{(k)})_{k\in\mathbb{N}}$ et que cette suite converge vers la solution $\boldsymbol{u}$ de (1.1) si et seulement si

   $$\rho\big((Y + \alpha\mathrm{Id})^{-1}X(X + \alpha\mathrm{Id})^{-1}Y\big) < 1.$$

   (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 $(X+\frac{\alpha}{2}\mathrm{Id})$ et $(Y+\frac{\alpha}{2}\mathrm{Id})$ sont définies positives alors la méthode (1.123) converge. On pourra pour cela (mais ce n'est pas obligatoire) suivre la démarche suivante :
   (a) Montrer que

   $$\rho\big((Y + \alpha\mathrm{Id})^{-1}X(X + \alpha\mathrm{Id})^{-1}Y\big) = \rho\big(X(X + \alpha\mathrm{Id})^{-1}Y(Y + \alpha\mathrm{Id})^{-1}\big).$$

   (On pourra utiliser l'exercice 50 page 76).
   (b) Montrer que

   $$\rho\big(X(X + \alpha\mathrm{Id})^{-1}Y(Y + \alpha\mathrm{Id})^{-1}\big) \leq \rho\big(X(X + \alpha\mathrm{Id})^{-1}\big)\rho\big(Y(Y + \alpha\mathrm{Id})^{-1}\big).$$

   (c) Montrer que $\rho\big(X(X + \alpha\mathrm{Id})^{-1}\big) < 1$ si et seulement si la matrice $(X + \frac{\alpha}{2}\mathrm{Id})$ 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 (1.123) converge dans ce cas et qui justifient l'appellation "méthode des directions alternées" qui lui est donnée.

Exercice 85 (Systèmes linéaires, "mauvaise relaxation"). Soit $A = (a_{i,j})_{i,j=1,\dots,n} \in \mathcal{M}_n(\mathbb{R})$ une matrice s.d.p.. On note D la partie diagonale de A, $-E$ la partie triangulaire inférieure stricte de A et $-F$ la partie triangulaire supérieure stricte de A, c'est-à-dire :

$$\begin{array}{l} D = (d_{i,j})_{i,j=1,\dots,n},\ d_{i,j} = 0 \text{ si } i \neq j,\ d_{i,i} = a_{i,i}, \\ E = (e_{i,j})_{i,j=1,\dots,n},\ e_{i,j} = 0 \text{ si } i \leq j,\ e_{i,j} = -a_{i,j} \text{ si } i > j, \\ F = (f_{i,j})_{i,j=1,\dots,n},\ f_{i,j} = 0 \text{ si } i \geq j,\ f_{i,j} = -a_{i,j} \text{ si } i < j. \end{array}$$

Soit $b \in \mathbb{R}^n$. On cherche à calculer $x \in \mathbb{R}^n$ t.q. $Ax = b$. Pour $0 < \omega < 2$, on considère la méthode itérative suivante :

(a) Initialisation : On choisit $x^{(0)} \in \mathbb{R}^n$.

(b) Itérations : Pour $k \in \mathbb{N}$,
On calcule $\tilde{x}^{(k+1)}$ dans $\mathbb{R}^n$ solution de $(D - E)\tilde{x}^{(k+1)} = Fx^{(k)} + b$,
On pose $x^{(k+1)} = \omega\tilde{x}^{(k+1)} + (1 - \omega)x^{(k)}$.

Enfin, on pose $M = \frac{D-E}{\omega}$ et $N = M - A$.

1. Montrer que la méthode s'écrit aussi $Mx^{(k+1)} = Nx^{(k)} + b$ et que la suite $(x^{(k)})_{k\in\mathbb{N}}$ est bien définie (c'est-à-dire que M est inversible).
2. On suppose dans cette question que $0 < \omega \leq 1$. Montrer que $M^t + N$ est s.d.p..
   N.B. : Un lemme du polycopié (lemme 1.56) donne alors que la méthode est convergente.
3. On suppose, dans cette question, que $n = 2$. On pose $A = \begin{bmatrix} \alpha & \gamma \\ \gamma & \beta \end{bmatrix}$.
   (a) Montrer que $\alpha > 0$, $\beta > 0$ et $\gamma^2 < \alpha\beta$.
   (b) Montrer que la méthode est convergente pour $0 < \omega < 2$. [Indication : Soit μ une valeur propre de $M^{-1}N$, montrer que $\mu \in]-1, 1[$.]
   (c) Montrer, en donnant un exemple, que $M^t + N$ n'est pas toujours s.d.p.. [Prendre $\gamma \neq 0$.]
4. On suppose, dans cette question, que $n = 3$ et on prend $A = \begin{bmatrix} 1 & a & a \\ a & 1 & a \\ a & a & 1 \end{bmatrix}$, avec $-\frac{1}{2} < a < 1$.
   (a) Vérifier que A est bien s.d.p..
   (b) Montrer que si $a = -1/2$ et $\omega = 2$, la matrice M est toujours inversible (mais A n'est plus s.d.p.) et que $\rho(M^{-1}N) > 1$. En déduire qu'il existe $a \in]-1/2, 1[$ et $\omega \in]0, 2[$ tels que $\rho(M^{-1}N) > 1$.
5. On suppose $n \geq 3$. Donner un exemple de matrice A ($A \in M_n(\mathbb{R})$, A s.d.p) pour laquelle la méthode est non convergente pour certains $\omega \in]0, 2[$. [Utiliser la question précédente.]

Exercice 86 (Vitesse de convergence pour la méthode de Jacobi).

Soient A une matrice carrée d'ordre n, inversible, et $\boldsymbol{b} \in \mathbb{R}^n$, $n > 1$. On pose $\bar{\boldsymbol{x}} = A^{-1}\boldsymbol{b}$. On note D la partie diagonale de A, $-E$ la partie triangulaire inférieure stricte de A et $-F$ la partie triangulaire supérieure stricte de A. On suppose que D est inversible et on note B_J la matrice des itérations de la méthode de Jacobi, c'est-à-dire $B_J = D^{-1}(E + F)$. On munit $\mathbb{R}^n$ d'une norme notée $\|\cdot\|$. On note ρ le rayon spectral de B_J ; on choisit $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$, et on note $(\boldsymbol{x}^{(k)})_{k\geq 1}$ la suite des itérés par la méthode de Jacobi pour la résolution du système linéaire $A\boldsymbol{x} = \boldsymbol{b}$ à partir du choix initial $\boldsymbol{x}^{(0)}$.

1. On suppose que B_J est diagonalisable dans $\mathbb{R}$ (c'est-à-dire qu'il existe une base de $\mathbb{R}^n$ formée de vecteurs propres de B_J). Montrer qu'il existe $\beta > 0$, dépendant de A, $\boldsymbol{b}$, $\boldsymbol{x}^{(0)}$ et de la norme choisie sur $\mathbb{R}^n$, mais indépendant de k, telle que
$$\|\boldsymbol{x}^{(k)} - \bar{\boldsymbol{x}}\| \leq \beta\rho^k \text{ pour tout } k \geq 0.$$

2. On ne suppose plus que B_J est diagonalisable. Montrer que pour tout $\varepsilon > 0$, il existe $\beta_\varepsilon > 0$, dépendant de A, b, $\boldsymbol{x}^{(0)}$, ε et de la norme choisie sur $\mathbb{R}^n$, mais indépendant de k, telle que

$$\|\boldsymbol{x}^{(k)} - \bar{\boldsymbol{x}}\| \leq \beta_\varepsilon(\rho + \varepsilon)^k \text{ pour tout } k \geq 0. \tag{1.124}$$

Dans la suite de cet exercice on prend $n = 2$ et $A = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$.

3. Dans cette question, on choisit, pour norme dans $\mathbb{R}^2$, la norme euclidienne, c'est-à-dire $\|\boldsymbol{x}\|^2 = x_1^2 + x_2^2$ si $\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$. Montrer qu'il existe β (dépendant de $\boldsymbol{b}$ et $\boldsymbol{x}^{(0)}$, mais non de k) telle que

$$\|\boldsymbol{x}^{(k)} - \bar{\boldsymbol{x}}\| = \beta\rho^k \text{ pour tout } k \geq 0. \tag{1.125}$$

4. Montrer qu'il existe des normes dans $\mathbb{R}^2$ pour lesquelles la conclusion de la question 3 est fausse (c'est-à-dire pour lesquelles la suite $(\|\boldsymbol{x}^{(k)} - \bar{\boldsymbol{x}}\|/\rho^k)_{k\in\mathbb{N}}$ n'est pas une suite constante sauf éventuellement pour des valeurs particulières de $\boldsymbol{x}^{(0)}$).

Exercice 87 (Convergence d'une méthode itérative)**.**
Soit $A \in M_3(\mathbb{R})$ définie par $A = I - E - F$ avec

$$I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} E = -\begin{bmatrix} 0 & 2 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \text{ et } F = -\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix}.$$

1. Montrer que A est inversible.
2. Soit $0 < \omega < 2$. Montrer que $(\frac{I}{\omega} - E)$ est inversible si et seulement si $\omega \neq \sqrt{2}/2$.

Pour $0 < \omega < 2$, $\omega \neq \sqrt{2}/2$, on considère (pour trouver la solution de $Ax = b$) la méthode itérative suivante :

$$(\frac{I}{\omega} - E)x^{(n+1)} = (F + \frac{1-\omega}{\omega}I)x^{(n)} + b.$$

On note $B_\omega = (\frac{I}{\omega} - E)^{-1}(F + \frac{1-\omega}{\omega}I)$. (De sorte que $x^{(n+1)} = B_\omega x^{(n)} + (\frac{I}{\omega} - E)^{-1}b$.)

3. Calculer, en fonction de ω, les valeurs propres de B_ω.
4. Donner l'ensemble des valeurs de ω pour lesquelles la méthode est convergente (quelquesoit $x^{(0)}$).
5. Déterminer $\omega_0 \in]0, 2[$ t.q. $\rho(B_{\omega_0}) = \min\{\rho(B_\omega), \omega \in]0, 2[, \omega \neq \sqrt{2}/2\}$.
   Pour cette valeur ω_0, montrer que la méthode donne la solution exacte de $Ax = b$ après un nombre fini d'itérations (quelquesoit $x^{(0)}$).

Exercice 88 (Une méthode itérative à pas optimal)**.**
Soit $n > 1$ et $A \in \mathcal{M}_n(\mathbb{R})$. On suppose que A est inversible. Pour calculer la solution du système linéaire $Ax = b$, avec $b \in \mathbb{R}^n$ donné, on se donne une matrice M appartenant à $\mathcal{M}_n(\mathbb{R})$, inversible (et plus simple à "inverser" que A) et on considère la méthode itérative suivante :
Initialisation : $x^{(0)}$ vecteur donné de $\mathbb{R}^n$, on pose $r^{(0)} = b - Ax^{(0)}$.
Itérations : pour $k \geq 0$, on choisit un réel α_k, on résout $M(x^{(k+1)} - x^{(k)}) = \alpha_k r^{(k)}$ et on pose $r^{(k+1)} = b - Ax^{(k+1)}$.
Pour conclure la description de la méthode, il reste à donner le choix de α_k, ceci est fait à la question 3.

1. Montrer que la méthode considérée peut aussi s'écrire de la manière suivante :
   Initialisation : $x^{(0)}$ vecteur donné de $\mathbb{R}^n$, $r^{(0)} = b - Ax^{(0)}$, $My^{(0)} = r^{(0)}$.
   Itérations : pour $k \geq 0$, On choisit un réel α_k,
   $x^{(k+1)} = x^{(k)} + \alpha_k y^{(k)}$, $r^{(k+1)} = r^{(k)} - \alpha_k Ay^{(k)}$, $My^{(k+1)} = r^{(k+1)}$.

2. On suppose, dans cette question, que α_k ne dépend par de k (c'est-à-dire $\alpha_k = \alpha$ pour tout $k \geq 0$). Déterminer $B \in \mathcal{M}_n(\mathbb{R})$ et $c \in \mathbb{R}^n$ tels que

$$x^{(k+1)} = Bx^{(k)} + c \quad \text{pour tout } k \geq 0.$$

Montrer que si la suite $(x^{(k)})_{k\in\mathbb{N}}$ converge, sa limite est solution du système linéaire $Ax = b$.

Pour la suite de l'exercice, on suppose que M est une matrice symétrique définie positive et on note $\|\cdot\|_M$ la norme sur $\mathbb{R}^n$ induite par M, c'est-à-dire $\|z\|_M^2 = Mz \cdot z$ (où $y \cdot z$ désigne le produit scalaire usuel de y et z)

3. (Choix de α_k) Soit $k \geq 0$.
Pour $x^{(k)}$, $r^{(k)}$ et $y^{(k)}$ connus, on pose, pour $\alpha \in \mathbb{R}$, $f(\alpha) = \|M^{-1}(r^{(k)} - \alpha A y^{(k)})\|_M^2$.
Si $y^{(k)} \neq 0$, montrer qu'il existe un unique $\bar{\alpha} \in \mathbb{R}$ tel que $f(\bar{\alpha}) \leq f(\alpha)$ pour tout $\alpha \in \mathbb{R}$. Donner cette valeur de $\bar{\alpha}$.
Dans la suite de l'exercice, si $y^{(k)} \neq 0$, on choisit $\alpha_k = \bar{\alpha}$ lors de l'itération k (si $y^{(k)} = 0$, on prend, par exemple, $\alpha_k = 0$).

4. Soit $k \geq 0$. Si $y^{(k)} \neq 0$, montrer que

$$\|y^{(k)}\|_M^2 - \|y^{(k+1)}\|_M^2 = \frac{(Ay^{(k)} \cdot y^{(k)})^2}{M^{-1}Ay^{(k)} \cdot Ay^{(k)}}.$$

En déduire que la suite $(\|y^{(k)}\|_M)_{k\in\mathbb{N}}$ converge dans $\mathbb{R}$.

5. (question indépendante des précédentes) Montrer, en donnant un exemple avec $n = 2$, que la matrice $A + A^t$ est symétrique mais pas nécessairement inversible

On suppose dans la suite que $A + A^t$ est (symétrique) définie positive.

6. Montrer qu'il existe $\beta > 0$ tel que

$$Az \cdot z \geq \beta z \cdot z \quad \text{pour tout } z \in \mathbb{R}^n.$$

En déduire qu'il existe $\gamma > 0$ tel que

$$\frac{(Az \cdot z)^2}{M^{-1}Az \cdot z} \geq \gamma \|z\|_M^2.$$

7. Montrer que $y^{(k)} \to 0$ quand $k \to +\infty$ et en déduire que la suite $(x^{(k)})_{k\in\mathbb{N}}$ tend vers x, solution de $Ax = b$, quand $k \to +\infty$.

1.5.5 Exercices, suggestions

Exercice 68 page 103 (Méthode itérative du "gradient à pas fixe".)

1. Calculer le rayon spectral $\rho(B)$ de la matrice d'itération $B = \mathrm{Id} - \alpha A$. 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 69 page 104 (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 74 page 106 (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 75 page 106 (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 71 page 104 (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 13 page 19 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 $\rho(B_{\text{J}}) = \max(-\mu_1, \mu_n)$, 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 82 page 109 (Méthode de Jacobi et relaxation.)

2. Utiliser l'exercice 75 page 106

1.5.6 Exercices, corrigés

Exercice 67 page 103

(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 $\overline{\boldsymbol{x}} = (Id - B)^{-1}\boldsymbol{c} = \begin{bmatrix} 9 \\ 3 \end{bmatrix}$.

(b) Les valeurs propres sont $\frac{2}{3}$ et 2 et donc le rayon spectral est 2 qui est strictement supérieur à 1, donc la suite diverge (sauf si $x^{(0)} = Bx^{(0)} + c$).

Exercice 68 page 103 (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 (1.5) $\alpha_0 = \dfrac{2}{\lambda_1 + \lambda_n}$.

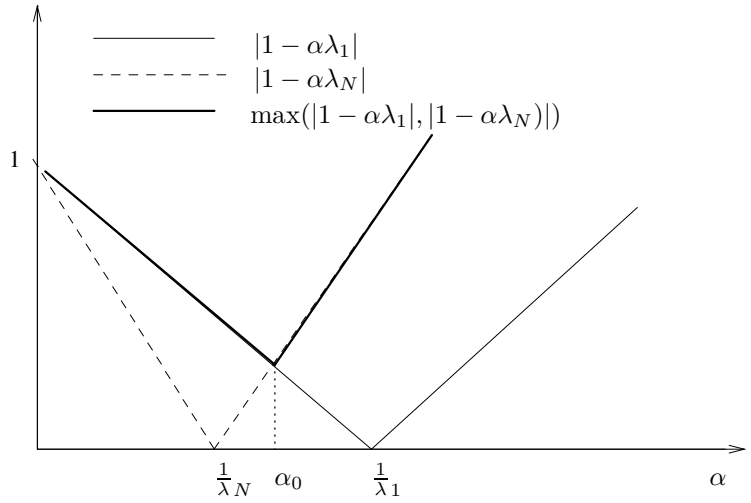


FIGURE 1.5: Graphes de $|1 - \alpha\lambda_1|$ et $|1 - \alpha\lambda_n|$ en fonction de α.

Exercice 69 page 104 (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 71 page 104 (Méthode de Jacobi et relaxation)

1. $J = D^{-1}(E+F)$ peut ne pas être symétrique, même si A est symétrique :
   En effet, prenons $A = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}$.
   Alors
   $$J = D^{-1}(E+F) = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{2} \\ 1 & 0 \end{pmatrix} \neq \begin{pmatrix} 0 & 1 \\ \frac{1}{2} & 0 \end{pmatrix}.$$
   donc J n'est pas symétrique.
2. On applique l'exercice 13 pour l'application linéaire T de matrice D, qui est, par hypothèse, définie positive (et évidemment symétrique puisque diagonale) et l'application S de matrice $E+F$, 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
   $$Jf_i = D^{-1}(E+F)f_i = \mu_i f_i, \quad \forall i = 1, \dots, n, \text{ et } Df_i \cdot f_j = \delta_{ij}.$$
3. La définition de J donne que tous les éléments diagonaux de J sont nuls et donc sa trace également. Or $Tr(J) = \sum_{i=1}^{n} \mu_i$. Si $\mu_i > 0 \ \forall i = 1, \dots, n$, alors $Tr(J) > 0$, donc $\exists i_0;\ \mu_{i_0} \leq 0$ 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 $\rho(J) < 1$ (théorème 1.53 page 94). Or, par la question précédente, $\rho(J) = \max(-\mu_1, \mu_n)$. 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 $(2D-A)f_1 \cdot f_1 = \beta \leq 0$, ce qui contredit le fait que $2D-A$ est définie positive. En conséquence, on a bien $\mu_1 > -1$.
   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 $Af_n \cdot f_n \leq 0$, ce qui contredit le fait que A est définie positive.
5. Par définition, on a :$D\tilde{x}^{(k+1)} = (E+F)x^{(k)} + b$ 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 $J_\omega = M_\omega^{-1}N_\omega$. En reprenant le raisonnement de la question 2 avec l'application linéaire T de matrice M_ω, qui est symétrique définie positive, et l'application S de matrice N_ω, qui est symétrique, il existe une base $(\tilde{f}_1, \dots, \tilde{f}_n)$ de $\mathbb{R}^n$ et $(\tilde{\mu}_1, \dots \tilde{\mu}_n) \subset \mathbb{R}$ 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 $\frac{2}{\omega}D - A$ 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

$$\left(\frac{2}{\omega}D - A\right) x \cdot x > 0, \quad \forall x \neq 0, \tag{1.126}$$

On va montrer que (1.126) est équivalent à

$$\left(\frac{2}{\omega}D - A\right) f_i \cdot f_i > 0, \quad \forall i = 1, \ldots, n, \tag{1.127}$$

où les $(f_i)_{i=1,n}$ sont les vecteurs propres de $D^{-1}(E+F)$ donnés à la question 2.
La famille $(f_i)_{i=1,\ldots,n}$ est une base de $\mathbb{R}^n$, et

$$\begin{aligned} \left(\frac{2}{\omega}D - A\right) f_i &= \left(\frac{2}{\omega}D - D + (E+F)\right) f_i \\ &= \left(\frac{2}{\omega} - 1\right) Df_i + \mu_i D f_i \\ &= \left(\frac{2}{\omega} - 1 + \mu_i\right) Df_i. \end{aligned} \tag{1.128}$$

On a donc en particulier $\left(\frac{2}{\omega}D - A\right) f_i \cdot f_j = 0$ si $i \neq j$, ce qui prouve que (1.126) est équivalent à (1.127).
De (1.127), on déduit aussi, grâce au fait que $Df_i \cdot f_i = 1$,

$$\left(\frac{2}{\omega}D - A\right) f_i \cdot f_i = \frac{2}{\omega} - 1 + \mu_i.$$

Une condition nécessaire et suffisante pour avoir (1.126) est donc $\frac{2}{\omega} - 1 + \mu_1 > 0$ car $\mu_1 = \inf \mu_i$, c'est–à–dire : $\frac{2}{\omega} > 1 - \mu_1$, 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 :

$$D^{-1}\left(\frac{1}{\omega}D - A\right) u = \lambda u, \text{ } \textit{i.e.} \text{ } \left(\frac{1}{\omega}D - A\right) u = \lambda D u.$$

On en déduit que

$$(D-A)u + \left(\frac{1}{\omega} - 1\right) Du = \lambda D u, \text{ soit encore}$$

$$D^{-1}(E+F)u = \left(1 - \frac{1}{\omega} + \lambda\right) u.$$

Ceci montre que u est un vecteur propre de J associé à la valeur propre $(1 - \frac{1}{\omega} + \lambda)$. Il existe donc $i \in \{1, \ldots, n\}$ tel que $(1 - \frac{1}{\omega} + \lambda) = \mu_i$. Les valeurs propres de I_ω sont donc les nombres $(\mu_i - 1 + \frac{1}{\omega})$ pour $i \in \{1, \ldots, n\}$. Finalement, les valeurs propres de J_ω sont donc les nombres $(\omega(\mu_i - 1) + 1)$ pour $i \in \{1, \ldots, n\}$.
On cherche maintenant à minimiser le rayon spectral

$$\rho(J_\omega) = \sup_i |\omega(\mu_i - 1) + 1)|$$

On a, pour tout i,

$$\omega(\mu_1 - 1) + 1) \leq \omega(\mu_i - 1) + 1) \leq \omega(\mu_n - 1) + 1),$$

et

$$-(\omega(\mu_n-1)+1) \leq -(\omega(\mu_i-1)+1) \leq -(\omega(\mu_1-1)+1),$$

donc

$$\rho(J_\omega) = \max\left(|\omega(\mu_1-1)+1)|, (|\omega(\mu_n-1)+1)|\right)$$

dont le minimum est atteint (voir Figure 1.6) pour

$$\omega(\mu_n-1)+1 = -\omega(\mu_1-1)-1 \text{ c'est–à–dire } \omega = \frac{2}{2-\mu_1-\mu_n}.$$

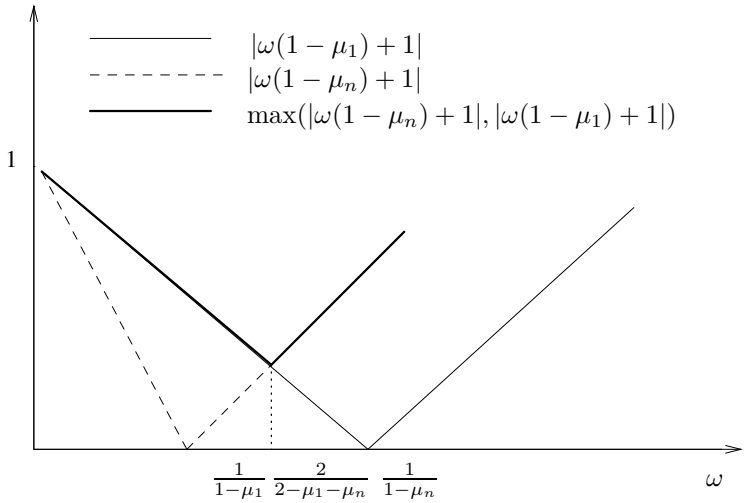


FIGURE 1.6: Détermination de la valeur de ω réalisant le minimum du rayon spectral.

Exercice 75 page 106 (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 $B_{\mathrm{J}}x = \lambda x$, 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 $\rho(B_{\mathrm{J}}) < 1$ ce qui prouve que la méthode de Jacobi converge.

Exercice 78 page 107 (Jacobi et Gauss-Seidel pour une matrice 3×3)

1.a La méthode de Jacobi s'écrit $D\boldsymbol{x}^{(k+1)} = (E+F)\boldsymbol{x}^{(k)} + \boldsymbol{b}$ 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 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 Le polynôme caractéristique de B_{J} est $P_J(\lambda) = \det(B_{\mathrm{J}} - \lambda \mathrm{Id}) = (-\lambda(-\lambda^2 + \frac{1}{2})$ et donc $\rho(B_{\mathrm{J}}) = \frac{\sqrt{2}}{2} < 1$. On en déduit que la méthode de Jacobi converge.

1.d 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.a 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 Il est facile de voir que $x \in \mathrm{Ker}(B_{GS})$ si et seulement si $x_2 = x_3 = 0$. Donc Ker$B_{GS} = \{t \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, t \in \mathbb{R}\}$.

2.c Le polynôme caractéristique de B_{GS} est $P_{GS}(\lambda) = \det(B_{GS} - \lambda \mathrm{Id})$. 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 On a bien $\rho(B_{\mathrm{GS}}) = \frac{1}{2} = \rho(B_{\mathrm{J}})^2$, ce qui est conforme au théorème 1.36 du cours.
2.e 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) : $x^{(1)} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, x^{(2)} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$.

Exercice 81 page 108 (Méthode de Jacobi et Gauss-Seidel pour des matrices triangulaires)

1. La matrice B_{GS} est nulle et la matrice B_J est triangulaire inférieure avec des 0 sur la diagonale. Pour ces deux matrices le rayon spectral est nul.
   Dans le cas de Gauss-Seidel, la première intération consiste à résoudre $Ax^{(1)} = b$ (car $A = D - E$). On obtient donc la solution après une itération au plus.
   Dans le cas de Jacobi, les intérations (avec les notations du cours) sont :
   $$a_{i,i}x_i^{(k+1)} = -\sum_{j<i} a_{i,j}x_j^{(k)} + b_i.$$
   La solution $x = A^{-1}b$ (noter que A est nécessairement inversible) vérifie
   $$a_{i,i}x_i = -\sum_{j<i} a_{i,j}x_j + b_i.$$
   On en déduit que $x_1^{(1)} = x_1$ (car il n'y pas de $j < 1$), puis que $x_2^{(2)} = x_2$ (car $x_1^{(1)} = x_1$), puis par récurrence que $x_k^{(k)} = x_k$. On obtient donc la solution après n itérations au plus.
2. Les matrices B_{GS} et B_J sont triangulaires supérieures avec des 0 sur la diagonale (et elles sont d'ailleurs égales). Pour ces deux matrices le rayon spectral est nul.
   Pour les deux méthodes, on obtient la solution après n itérations au plus. Le raisonnement est semblable à celui fait pour Jacobi à la question précédente. La différence ici est que $x_n^{(1)} = x_n$ puis par récurrence $x_{n-k+1}^{(k)} = x_{n-k+1}$.

Exercice 82 page 109 (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 $\tilde{A} = \lambda \mathrm{Id} + A$ et notons $\tilde{a}_{i,j}$ ses coefficients. Comme $\lambda \in \mathbb{R}_+^*$,et grâce aux hypothèses (1.118)–(1.120), ceux-ci vérifient :
   $$\tilde{a}_{i,j} \leq 0, \forall i, j = 1, \ldots, n, \ i \neq j, \tag{1.129}$$
   $$\tilde{a}_{i,i} > \sum_{\substack{i=1 \\ j\neq i}}^n a_{i,j}, \forall i = 1, \ldots, n. \tag{1.130}$$
   La matrice $\tilde{A}$ est donc à diagonale dominante stricte, et par l'exercice 75 page 106, elle est donc inversible.
3. La méthode de Jacobi pour la résolution du système (1.121) s'écrit :
   $$\tilde{D}u^{(k+1)} = (E + F)u^{(k)} + b, \tag{1.131}$$
   avec $\tilde{D} = \lambda \mathrm{Id} + D$, 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 :
   $$u_i^{(k+1)} = \frac{1}{a_{i,i} + \lambda}(-\sum_{\substack{j=1,n \\ j\neq i}} a_{i,j}u_j^{(k)} + b_i).$$

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 (1.131), on obtient que $\overline{u}$ est solution de (1.121).

Exercice 86 page 111 (Vitesse de convergence pour la méthode de Jacobi)

1. Soit $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n$ une base de $\mathbb{R}^n$ formée de vecteurs propres de B_J. On a donc, pour tout $i \in \{1, \ldots, n\}$, $B_J \boldsymbol{f}_i = \lambda_i \boldsymbol{f}_i$ avec $\lambda_i \in \mathbb{R}$ et $\rho = \max\{|\lambda_i|, i \in \{1, \ldots, n\}\}$.
La suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ donnée par la méthode de Jacobi vérifie, pour tout $k \geq 0$, $\boldsymbol{x}^{(k+1)} - \overline{\boldsymbol{x}} = B_J(\boldsymbol{x}^{(k)} - \overline{\boldsymbol{x}})$. En écrivant $\boldsymbol{x}^{(0)} - \bar{\boldsymbol{x}}$ dans la base $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n$ on a $x^{(0)} - \overline{\boldsymbol{x}} = \sum_{i=1}^{n} \alpha_i \boldsymbol{f}_i$. Par récurrence sur k, on en déduit, pour tout $k \geq 0$,

$$\boldsymbol{x}^{(k)} - \bar{\boldsymbol{x}} = \sum_{i=1}^{n} \lambda_i^k \alpha_i \boldsymbol{f}_i.$$

Comme $|\lambda_i| \leq \rho$, on en déduit $\|\boldsymbol{x}^{(k)} - \bar{\boldsymbol{x}}\| \leq \rho^k \sum_{i=1}^{n} |\alpha_i| \|\boldsymbol{f}_i\|$, ce qui donne le résultat demandé avec $\beta = \sum_{i=1}^{n} |\alpha_i| \|\boldsymbol{f}_i\|$.

2. Soit $\varepsilon > 0$; on sait qu'il existe une norme sur $\mathbb{R}^n$, notée $\|\cdot\|_\varepsilon$, telle que la norme induite sur $\mathcal{M}_n(\mathbb{R})$ (encore notée $\|\cdot\|_\varepsilon$) vérifie $\|B_J\|_\varepsilon \leq \rho + \varepsilon$. (On rappelle que cette norme dépend de ε et B_J.)
   Avec cette norme, on a donc
   $$\|\boldsymbol{x}^{(k)} - \bar{\boldsymbol{x}}\|_\varepsilon \leq \|B_J\|_\varepsilon^k \|\boldsymbol{x}^{(0)} - \bar{\boldsymbol{x}}\|_\varepsilon \leq \|\boldsymbol{x}^{(0)} - \bar{\boldsymbol{x}}\|_\varepsilon (\rho + \varepsilon)^k.$$
   D'autre part, comme sur $\mathbb{R}^n$ toutes les normes sont équivalentes, il existe $\gamma \in \mathbb{R}_+$ tel que $\|\cdot\| \leq \gamma \|\cdot\|_\varepsilon$. On a donc
   $$\|\boldsymbol{x}^{(k)} - \bar{\boldsymbol{x}}\| \leq \gamma \|\boldsymbol{x}^{(k)} - \bar{\boldsymbol{x}}\|_\varepsilon \leq \beta_\varepsilon (\rho + \varepsilon)^k \text{ avec } \beta_\varepsilon = \gamma \|\boldsymbol{x}^{(0)} - \bar{\boldsymbol{x}}\|_\varepsilon.$$

3. Comme $B_J = D^{-1}(E+F) = \begin{bmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{bmatrix}$, on a donc $\rho = \frac{1}{2}$.
   Soit $\boldsymbol{x}^{(0)} \in \mathbb{R}^2$, $\boldsymbol{x}^{(0)} - \bar{\boldsymbol{x}} = \alpha_1 \boldsymbol{e}_1 + \alpha_2 \boldsymbol{e}_2$. On a alors
   $$\boldsymbol{x}^{(k)} - \bar{\boldsymbol{x}} = (\frac{1}{2})^k (\alpha_1 \boldsymbol{e}_1 + \alpha_2 \boldsymbol{e}_2) \text{ si } k \text{ est pair,}$$
   $$\boldsymbol{x}^{(k)} - \bar{\boldsymbol{x}} = (\frac{1}{2})^k (\alpha_1 \boldsymbol{e}_2 + \alpha_2 \boldsymbol{e}_1) \text{ si } k \text{ est impair.}$$
   On en déduit que $\|\boldsymbol{x}^{(k)} - \bar{\boldsymbol{x}}\| = \rho^k \|\boldsymbol{x}^{(0)} - \bar{\boldsymbol{x}}\|$ pour tout $k \geq 0$.

4. On prend, par exemple, $\|\boldsymbol{x}\| = \sqrt{2x_1^2 + x_2^2}$ pour $\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$.
   Avec les notations de la question précédente, on a
   $$\|\boldsymbol{x}^{(k)} - \bar{\boldsymbol{x}}\| = \begin{cases} \rho^k \sqrt{2\alpha_1^2 + \alpha_2^2} \text{ si } k \text{ est pair,} \\ \\ \rho^k \sqrt{2\alpha_2^2 + \alpha_1^2} \text{ si } k \text{ est impair.} \end{cases}$$
   La suite $(\|\boldsymbol{x}^{(k)} - \bar{\boldsymbol{x}}\|/\rho^k)_{k \in \mathbb{N}}$ est une suite constante seulement si $\alpha_1^2 = \alpha_2^2$.

1.6 Valeurs propres et vecteurs propres

Les techniques de recherche des éléments propres, c.à.d. des valeurs et vecteurs propres (voir Définition 1.3 page 8) d'une matrice sont essentielles dans de nombreux domaines d'application, par exemple en dynamique des structures : 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 (voir par exemple le chapitre 5 du polycopié d'E. Hairer, cité dans l'introduction de ce cours, en ligne sur le web). 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

Pour expliquer l'algorithme de la puissance, commençons par un exemple simple. 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 en itérant les instructions suivantes :

```
-->x = A * x ; x = x/norm(x)
```

ce qui correspond à la construction de la suite

$$\boldsymbol{x}^{(0)} = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|},\ \boldsymbol{x}^{(1)} = \frac{A\boldsymbol{x}^{(0)}}{\|A\boldsymbol{x}^{(0)}\|}, \cdots, \boldsymbol{x}^{(k+1)} = \frac{A\boldsymbol{x}^{(k)}}{\|A\boldsymbol{x}^{(k)}\|} \tag{1.132}$$

où $\|\boldsymbol{x}\|$ désigne la norme euclidienne.
On obtient les résultats suivants :

0.8944272	0.7808688	0.7327935	0.7157819	0.7100107	0.7080761	0.7074300
- 0.4472136	- 0.6246950	- 0.6804511	- 0.6983239	- 0.7061361	- 0.7067834	- 0.706999

On voit clairement sur cet exemple que la suite $\boldsymbol{x}^{(k)}$ converge vers $\boldsymbol{f}_2 = \frac{\sqrt{2}}{2}\begin{bmatrix}-1\\1\end{bmatrix}$ lorsque $k \to +\infty$. Si maintenant on fait tourner Scilab en lui demandant de calculer ensuite le produit scalaire de $A\boldsymbol{x}$ avec $\boldsymbol{x}$:

```
-->x= A*x;x=x/norm(x);mu=(A*x)'*x
```

ce qui correspond au calcul de la suite $\mu_k = A\boldsymbol{x}^{(k)} \cdot \boldsymbol{x}^{(k)}, k \geq 0$, on obtient la suite :

$$2.8, 2.9756098, 2.9972603, 2.9996952, 2.9999661, ...$$

qui a tout l'air de converger vers 3 ! En fait on a le théorème suivant, qui montre que dans un certain nombre de cas, on a effectivement convergence de l'algorithme vers la valeur propre dite dominante (celle qui correspond au rayon spectral).

Théorème 1.64 (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 $\lambda_n \in \mathbb{R}$. Soit $\boldsymbol{x}^{(0)} \notin Vect(\boldsymbol{f}_1, \cdots, \boldsymbol{f}_{n-1})$. Alors, la suite de vecteurs $\boldsymbol{x}_{2k}$ définie par (1.132) *converge vers un vecteur unitaire qui est vecteur propre de A pour la valeur propre dominante λ_n.*
De plus, si la norme choisie dans l'algorithme (1.132) *est la norme 2, alors la suite $(A\boldsymbol{x}_k \cdot \boldsymbol{x}_k)_{k\in\mathbb{N}}$ converge vers λ_n lorsque $k \to +\infty$.*

Démonstration. La démonstration de ce résultat fait l'objet de l'exercice 90 dans le cas plus simple où A est une matrice symétrique, et donc diagonalisable dans $\mathbb{R}$. □

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 deux plus grandes valeurs propres est proche de 1, la méthode va converger trop lentement.

De manière assez miraculeuse, il existe un remède à chacun de ces maux :

1. Pour calculer plusieurs valeurs propres simultanément, on procède par blocs : on part de p vecteurs orthogonaux $\boldsymbol{x}_1^{(0)}, \ldots \boldsymbol{x}_p^{(0)}$ (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 $A^{-1}\boldsymbol{x}_k \cdot \boldsymbol{x}_k$ 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 $A\tilde{\boldsymbol{x}}_{k+1} = \boldsymbol{x}^{(k)}$ (et $\boldsymbol{x}_{k+1} = \tilde{\boldsymbol{x}}_{k+1}/\|\tilde{\boldsymbol{x}}_{k+1}\|$).
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 91.

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 1.65 (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 dans la question 1 de l'exercice 95. 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 $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n$ (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 93 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 : $A = A_0 = Q_0 R_0$ et on calcule $A_1 = R_0 Q_0$. 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, grâce à l'associativité du produit matriciel, on a :

$$A_1 = R_0 Q_0 = (Q_0^{-1} Q_0) R_0 Q_0 = Q_0^{-1} (Q_0 R_0) Q_0 = Q_0^{-1} A Q_0.$$

Les matrices A_0 et A_1 ont donc même valeurs propres.
On recommence alors l'opération : à l'itération k, on effectue la décomposition QR de A_k : $A_k = Q_k R_k$ et on calcule $A_{k+1} = R_k Q_k$.
Par miracle, pour la plupart des matrices, les coefficients diagonaux de la matrice A_k tendent vers les valeurs propres de la matrice A. Dans beaucoup de cas, on peut aussi obtenir des vecteurs propres associés (ils sont donnés par les colonnes de la matrice $\tilde{Q}_k$ de l'exercice 95). On sait démontrer cette convergence pour certaines matrices ; on pourra trouver par exemple dans les livres de Serre ou Hubbard-Hubert la démonstration sous une hypothèse assez technique et difficile à vérifier en pratique ; l'exercice 95 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 94 donne un exemple de l'application de la méthode QR avec shift.

1.6.3 Exercices (valeurs propres, vecteurs propres)

Exercice 89 (Multiplicités algébrique et géométrique)**.** *Soit $A \in \mathcal{M}_n(\mathbb{R})$, $n > 1$. On note $P(A)$ le polynôme caractéristique de A et $\{\lambda_i,\ i \in \{1, \ldots, p\}\}$ les valeurs propres de A (on a donc $1 \leq p \leq n$).*
On rappelle que $P(A)(x) = \prod_{i=1}^{p} (x - \lambda_i)^{m_i}$ et que $1 \leq n_i \leq m_i$ (pour tout i) où $n_i = \dim \ker(A - \lambda_i I)$ (I désigne la matrice identité de taille n).
Le nombre m_i est la multiplicité algébrique de λ_i alors que le nombre n_i est la multiplicité géométrique de λ_i.
On rappelle enfin que que $n_i < m_i$ si et seulement si $\ker(A - \lambda_i I)^2 \neq \ker(A - \lambda_i I)$.
Soit $\lambda \in \{\lambda_i,\ i \in \{1, \ldots, p\}\}$ et $x \in \ker(A - \lambda_i I)^2$.

1. *Montrer que $A^2 x - \lambda^2 x = 2\lambda(Ax - \lambda x)$, puis que, pour tout $k \in \mathbb{N}^\star$,*

$$A^k x - \lambda^k x = k\lambda^{k-1}(Ax - \lambda x). \tag{1.133}$$

   [Utiliser une récurrence sur k.]

On suppose dans les questions 2 et 3 que $|\lambda| = \rho(A)$ ($\rho(A)$ désigne le rayon spectral de A).

2. *On suppose dans cette question que $A \neq 0$ et qu'il existe une norme induite sur $\mathcal{M}_n(\mathbb{R})$, notée $\|\cdot\|_\star$ pour laquelle $\|A\|_\star = \rho(A)$.*
   (a) *Montrer que $\lambda \neq 0$.*
   (b) *Montrer que $Ax = \lambda x$. [Utiliser (1.133) et faire $k \to +\infty$. On pourra se limiter au cas $\lambda \in \mathbb{R}$ et $x \in \mathbb{R}^n$. Les courageux pourront, hors barème, faire le cas $\lambda \in \mathbb{C}$.]*
   (c) *Montrer que $m_i = n_i$.*
3. *Donner un exemple pour lequel $n_i < m_i$ (mais toujours avec $\lambda = \lambda_i$ et $|\lambda| = \rho(A)$). En déduire que, pour cet exemple, $\|A\|_\star > \rho(A)$ pour tout norme induite sur $\mathcal{M}_n(\mathbb{R})$.*

4. *On considère la méthode de la puissance pour la matrice* $A = \begin{pmatrix} \mu & 1 \\ 0 & \mu \end{pmatrix}$, *avec* $\mu \in \mathbb{R}$ *:*

$$x_{k+1} = \frac{Ax_k}{\|Ax_k\|},\ k \in \mathbb{N},\ x_0 = (\alpha, \beta)^t,$$

et $\|\cdot\|$ *désigne la norme euclidienne. Etudier la convergence de la suite* (x_k)*. [On pourra montrer que* x_k *est proportionnel à* $(\alpha\mu^k + k\beta\mu^{k-1}, \beta\mu^k)^t$*.]*

Exercice 90 (Méthode de la puissance)**.** *Suggestions en page 130, corrigé en page 131*

1. Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique (non nulle). Soit $\lambda_n \in \mathbb{R}$ valeur propre de A t.q. $|\lambda_n| = \rho(A)$ et soit $\boldsymbol{y}^{(0)} \in \mathbb{R}^n$. On suppose que $-\lambda_n$ n'est pas une valeur propre de A et que $\boldsymbol{y}^{(0)}$ n'est pas orthogonal à $\ker(A - \lambda_n \mathrm{Id})$, ce qui revient à dire que lorsqu'on écrit le $\boldsymbol{y}^{(0)}$ dans une base formée de vecteurs propres de A, la composante sur sous-espace propre associé à λ_n est non nulle. (L'espace $\mathbb{R}^n$ est muni de la norme euclidienne.) On définit la suite $(\boldsymbol{y}^{(k)})_{k\in\mathbb{N}}$ par $\boldsymbol{y}^{(k+1)} = A\boldsymbol{y}^{(k)}$ pour $k \in \mathbb{N}$. Montrer que

   (a) $\dfrac{\boldsymbol{y}^{(k)}}{(\lambda_n)^k} \to \boldsymbol{y}$, quand $k \to \infty$, avec $\boldsymbol{y} \neq 0$ et $A\boldsymbol{y} = \lambda_n \boldsymbol{y}$.

   (b) $\dfrac{\|\boldsymbol{y}^{(k+1)}\|}{\|\boldsymbol{y}^{(k)}\|} \to \rho(A)$ quand $k \to \infty$.

   (c) $\frac{1}{\|\boldsymbol{y}^{2k}\|}\boldsymbol{y}^{2k} \to \boldsymbol{x}$ quand $k \to \infty$ avec $A\boldsymbol{x} = \lambda_n \boldsymbol{x}$ et $\|\boldsymbol{x}\| = 1$.

   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 $\boldsymbol{b} \in \mathbb{R}^n$. Pour calculer $\boldsymbol{x}$ t.q. $A\boldsymbol{x} = \boldsymbol{b}$, on considère un méthode itérative : on se donne un choix initial $\boldsymbol{x}^{(0)}$, et on construit la suite $\boldsymbol{x}^{(k)}$ telle que $\boldsymbol{x}^{(k+1)} = B\boldsymbol{x}^{(k)} + \boldsymbol{c}$ avec $\boldsymbol{c} = (Id - B)A^{-1}\boldsymbol{b}$, et on suppose B symétrique. On rappelle que si $\rho(B) < 1$, la suite $(\boldsymbol{y}^{(k)})_{k\in\mathbb{N}}$ tend vers $\boldsymbol{x}$. Montrer que, sauf cas particuliers à préciser,

   (a) $\frac{\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}\|}{\|\boldsymbol{x}^{(k)} - \boldsymbol{x}\|} \to \rho(B)$ quand $k \to \infty$ (ceci donne une estimation de la vitesse de convergence de la méthode itérative).

   (b) $\frac{\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\|}{\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}\|} \to \rho(B)$ quand $k \to \infty$ (ceci permet d'estimer $\rho(B)$ au cours des itérations).

Exercice 91 (Méthode de la puissance inverse avec shift)**.** *Suggestions en page 130.*

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 $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$. On suppose que $\boldsymbol{x}^{(0)}$ n'est pas orthogonal à $\ker(A - \lambda_i \mathrm{Id})$. 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 $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ par $(A - \mu\mathrm{Id})\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)}$ pour $k \in \mathbb{N}$.

1. Vérifier que la construction de la suite revient à appliquer la méthode de la puissance à la matrice $(A - \mu\mathrm{Id})^{-1}$.
2. Montrer que $\boldsymbol{x}^{(k)}(\lambda_i - \mu)^k \to \boldsymbol{x}$, quand $k \to \infty$, où $\boldsymbol{x}$ est un vecteur propre associé à la valeur propre λ_i, c.à.d. $\boldsymbol{x} \neq 0$ et $A\boldsymbol{x} = \lambda_i \boldsymbol{x}$.
3. Montrer que $\frac{\|\boldsymbol{x}^{(k+1)}\|}{\|\boldsymbol{x}^{(k)}\|} \to \frac{1}{|\mu - \lambda_i|}$ quand $k \to \infty$.

Exercice 92 (Matrices antisymétriques)**.**
Soit $A \in \mathcal{M}_n(\mathbb{R})$ t.q. $A^t = -A$. On pose $B = A^tA$.

1. Montrer que toutes les valeurs propres de A sont imaginaires pures (c'est-à-dire de la forme $i\beta$ avec $\beta \in \mathbb{R}$).
   [On pourra montrer que $Az \cdot z = 0$ pour tout $z \in \mathbb{R}^n$. Puis si $A(x + iy) = (\alpha + i\beta)(x + iy)$, calculer $Ax \cdot x + Ay \cdot y$.]
   Montrer que toutes les valeurs propres de B sont dans $\mathbb{R}_+$.

2. Montrer que $\rho(A)^2 = \rho(B)$.
3. Soit $\mu = \rho(B)$. On note $(y^{(k)})_{k\in\mathbb{N}}$ la suite donnée par la méthode de la puissance pour le matrice B avec la norme euclidienne. On suppose que $y^{(0)} \notin (\ker(B-\mu I))^\perp$ Montrer que $By^{(k)}\cdot y^{(k)} \to \mu = \rho(A)^2$. [Suggestion : décomposer $y^{(0)}$ sur une base orthonormée de $\mathbb{R}^n$ formée de vecteurs propres de B.]
4. On suppose que n est impair.
   (a) Montrer que 0 est v.p. de A.
   (b) Pour cette question, $A = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix}$. Calculer B puis B^k pour tout $k \in \mathbb{N}$.
   Donner pour tout $k \in \mathbb{N}$, $y^{(k)}$ en fonction de $y^{(0)}$.
   A-t-on $\lim_{k\to+\infty} By^{(k)}\cdot y^{(k)} = \rho(A)^2$ si et seulement si $y^{(0)} \notin (\ker(B-\rho(B)I))^\perp$?

Exercice 93 (Orthogonalisation de Gram-Schmidt)**.** *Corrigé en page 132*

Soient $\boldsymbol{u}$ et $\boldsymbol{v}$ deux vecteurs de $\mathbb{R}^n$, $u \neq 0$. On rappelle que la projection orthogonale $\mathrm{proj}_{\boldsymbol{u}}(\boldsymbol{v})$ du vecteur $\boldsymbol{v}$ sur la droite vectorielle engendrée par $\boldsymbol{u}$ peut s'écrire de la manière suivante :

$$\mathrm{proj}_{\boldsymbol{u}}(\boldsymbol{v}) = \frac{\boldsymbol{v}\cdot\boldsymbol{u}}{\boldsymbol{u}\cdot\boldsymbol{u}}\boldsymbol{u},$$

où $\boldsymbol{u}\cdot\boldsymbol{v}$ désigne le produit scalaire des vecteurs $\boldsymbol{u}$ et $\boldsymbol{v}$. On note $\|\cdot\|$ la norme euclidienne sur $\mathbb{R}^n$.

1. Soient $(\boldsymbol{a_1},\ldots,\boldsymbol{a_n})$ une base de $\mathbb{R}^n$. On rappelle qu'à partir de cette base, on peut obtenir une base orthogonale $(\boldsymbol{v}_1,\ldots,\boldsymbol{v}_n)$ et une base orthonormale $(\boldsymbol{q}_1,\ldots,\boldsymbol{q}_n)$ par le procédé de Gram-Schmidt qui s'écrit :
Pour $k = 1,\ldots,n$,

$$\boldsymbol{v}_k = \boldsymbol{a}_k - \sum_{j=1}^{k-1} \frac{a_k\cdot v_j}{v_j\cdot v_j} v_j, \qquad \boldsymbol{q}_k = \frac{\boldsymbol{v}_k}{\|\boldsymbol{v}_k\|}. \tag{1.134}$$

1. Montrer par récurrence que la famille $(\boldsymbol{v}_1,\ldots,\boldsymbol{v}_n)$ est une base orthogonale de $\mathbb{R}^n$.

2. Soient A la matrice carrée d'ordre n dont les colonnes sont les vecteurs $\boldsymbol{a}_j$ et Q la matrice carrée d'ordre N dont les colonnes sont les vecteurs $\boldsymbol{q}_j$ définis par le procédé de Gram-Schmidt (1.134), ce qu'on note :

$$A = \begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \ldots & \boldsymbol{a}_n \end{bmatrix}, \qquad Q = \begin{bmatrix} \boldsymbol{q}_1 & \boldsymbol{q}_2 & \ldots & \boldsymbol{q}_n \end{bmatrix}.$$

Montrer que

$$\boldsymbol{a}_k = \|\boldsymbol{v}_k\|\boldsymbol{q}_k + \sum_{j=1}^{k-1} \frac{\boldsymbol{a}_k\cdot\boldsymbol{v}_j}{\|\boldsymbol{v}_j\|}\boldsymbol{q}_j.$$

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 $QQ^t = \mathrm{Id}$) 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 1.66 (Gram-Schmidt modifié)**.**

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

Calcul de la norme de $\boldsymbol{a}_k$

$r_{kk} := (\sum_{i=1}^n a_{ik}^2)^{\frac{1}{2}}$

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 $i = 1, \ldots, n$,
$a_{ij} := a_{ij} - a_{ik} r_{kj}$
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 94 (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 $\tilde{A}_1 = RQ - b\mathrm{Id}$ 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 95 (Méthode QR pour la recherche de valeurs propres)**.** *Corrigé en page 132*

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 :

$$A_1 = A = Q_1 R_1,\ R_1 Q_1 = A_2 = Q_2 R_2,\ \ldots,\ R_k Q_k = A_k = Q_{k+1} R_{k+1}. \tag{1.135}$$

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 et vérifiant l'hypothèse (1.137) (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 :

$$A = P\lambda P^t, \text{ avec } \lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n), \text{ et } P \text{ est une matrice orthogonale.} \tag{1.136}$$

(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

$$P^t \text{ admet une décomposition } LU \text{ et que les coefficients diagonaux de } U \text{ sont strictement positifs.} \tag{1.137}$$

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 (1.135).
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

$$A^k = \tilde{Q}_k\tilde{R}_k, \tag{1.138}$$

avec

$$\tilde{Q}_k = Q_1Q_2\dots Q_{k-1}Q_k \text{ et } \tilde{R}_k = R_kR_{k-1}\dots R_2R_1. \tag{1.139}$$

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 $b_1, b_2, \dots, b_n$ les colonnes de la matrice B et $b_1^{(k)}, b_2^{(k)}, \dots, b_n^{(k)}$ les colonnes de la matrice B_k, ou encore :

$$B = \begin{bmatrix} b_1 & b_2 & \dots & b_n \end{bmatrix}, \qquad B_k = \begin{bmatrix} b_1^{(k)} & b_2^{(k)} & \dots & b_n^{(k)} \end{bmatrix}.$$

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 90 page 127 (Méthode de la puissance pour calculer le rayon spectral de A.)

1. Décomposer $\boldsymbol{x}^{(0)}$ 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 $\boldsymbol{y}^{(k)} = \boldsymbol{x}^{(k)} - \boldsymbol{x}$ où $\boldsymbol{x}$ est la solution de $A\boldsymbol{x} = \boldsymbol{b}$ et appliquer la question 1.
b/ Raisonner avec $\boldsymbol{y}^{(k)} = \boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}$.

Exercice 91 page 127 (Méthode de la puissance inverse)

Appliquer l'exercice précédent à la matrice $B = (A - \mu\mathrm{Id})^{-1}$.

1.6.5 Corrigés

Exercice 90 page 127 (Méthode de la puissance pour calculer le rayon spectral de A)

1. Comme A est une matrice symétrique (non nulle), A est diagonalisable dans $\mathbb{R}$. Soit $(f_1, \ldots, f_n)$ une base orthonormée de $\mathbb{R}^n$ formée de vecteurs propres de A associée aux valeurs propres $\lambda_1, \ldots, \lambda_n$ (qui sont réelles). On décompose $y^{(0)}$ sur $(f_i)_{i=1,\ldots,n}$: $y^{(0)} = \sum_{i=1}^n \alpha_i f_i$. On a donc $Ay^{(0)} = \sum_{i=1}^n \lambda_i \alpha_i f_i$ et $A^k y^{(0)} = \sum_{i=1}^n \lambda_i^k \alpha_i f_i$.
   On en déduit :
   $$\frac{y^{(k)}}{\lambda_n^k} = \sum_{i=1}^n \left(\frac{\lambda_i}{\lambda_n}\right)^k \alpha_i f_i.$$
   Comme $-\lambda_n$ n'est pas valeur propre,
   $$\lim_{k\to+\infty} (\frac{\lambda_i}{\lambda_n})^k = 0 \text{ si } \lambda_i \neq \lambda_n. \tag{1.140}$$
   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
   $\lim_{n\to+\infty} \frac{y^{(k)}}{\lambda_n^k} = \sum_{i=p+1}^n \alpha_i f_i = y$, avec $Ay = \lambda_n y$.
   De plus, $y \neq 0$: en effet, $y^{(0)} \notin (Ker(A-\lambda_n \mathrm{Id}))^\perp = Vect\{f_1, \ldots, f_p\}$, et donc il existe $i \in \{p+1, \ldots, n\}$ tel que $\alpha_i \neq 0$.
   Pour montrer (b), remarquons que
   $$\frac{\|y^{(k+1)}\|}{\|y^{(k)}\|} = |\lambda_n| \frac{\|\frac{y^{(k+1)}}{\lambda_n^{k+1}}\|}{\|\frac{y^{(k)}}{\lambda_n^k}\|} \to |\lambda_n| \frac{\|y\|}{\|y\|} = |\lambda_n| \text{ lorsque } k \to +\infty.$$
   Enfin, $\frac{y^{(2k)}}{\|y^{(2k)}\|} = \frac{y^{(2k)}}{\lambda_n^{2k}} \frac{\lambda_n^{2k}}{\|y^{(2k)}\|}$ et $\lim_{k\to+\infty} \frac{\|y^{(2k)}\|}{\lambda_n^{2k}} = \|y\|$.
   On a donc $\lim_{k\to+\infty} \frac{y^{(2k)}}{\|y^{(2k)}\|} = x$, avec $x = \frac{y}{\|y\|}$.

2. a) La méthode I s'écrit à partir de $x^{(0)}$ connu : $x^{(k+1)} = Bx^{(k)} + c$ pour $k \geq 1$, avec $c = (I - B)A^{-1}b$.
   On a donc
   $$\begin{aligned} x^{(k+1)} - x &= Bx^{(k)} + (Id - B)x - x \\ &= B(x^{(k)} - x). \end{aligned} \tag{1.141}$$
   Si $y^{(k)} = x^{(k)} - x$, on a donc $y^{(k+1)} = By^{(k)}$, et d'après la question 1a) si $y^{(0)} \not\perp \ker(B - \mu_n \mathrm{Id})$ où μ_n est la plus grande valeur propre de B, (avec $|\mu_n| = \rho(B) et - \mu_n$ non valeur propre), alors
   $$\frac{\|y^{(k+1)}\|}{\|y^{(k)}\|} \longrightarrow \rho(B) \text{ lorsque } k \to +\infty,$$
   c'est–à–dire
   $$\frac{\|x^{(k+1)} - x\|}{\|x^{(k)} - x\|} \longrightarrow \rho(B) \text{ lorsque } k \to +\infty.$$
   b) On applique maintenant 1a) à $y^{(k)} = x^{(k+1)} - x^{(k)}$ avec
   $$y^{(0)} = x^{(1)} - x^{(0)} \text{ où } x^{(1)} = Ax^{(0)}.$$
   On demande que $x^{(1)} - x^{(0)} \notin \ker(B - \mu_n \mathrm{Id})^\perp$ comme en a), et on a bien $y^{(k+1)} = By^{(k)}$, donc $\frac{\|y^{(k+1)}\|}{\|y^{(k)}\|} \longrightarrow \rho(B)$ lorsque $k \to +\infty$.

Exercice 93 page 128 (Orthogonalisation par Gram-Schmidt)

1. Par définition de la projection orthogonale, on a $v_1 \cdot v_2 = a_1 \cdot (a_2 - \text{proj}_{a_1}(a_2)) = 0$.
Supposons la récurrence vraie au rang $k-1$ et montrons que v_k est orthogonal à tous les v_i pour $i = 1, \ldots, k-1$.
Par définition, $v_k = a_k - \sum_{j=1}^{k-1} \frac{a_k \cdot v_j}{v_j \cdot v_j} v_j$, et donc

$$v_k \cdot v_i = a_k \cdot v_i - \sum_{j=1}^{k-1} \frac{a_k \cdot v_j}{v_j \cdot v_j} v_j \cdot v_i = a_k \cdot v_i - a_k \cdot v_i$$

par hypothèse de récurrence. On en déduit que $v_k \cdot v_i = 0$ et donc que la famille $(v_1, \ldots v_n)$ est une base orthogonale.

2. De la relation (1.134), on déduit que :

$$a_k = v_k + \sum_{j=1}^{k-1} \frac{a_k \cdot v_j}{v_j \cdot v_j} v_j,$$

et comme $v_j = \|v_j\| q_j$, 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 $\frac{a_k \cdot v_j}{\|v_j\|}$. 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 (1.134) on construit la matrice $Q = \begin{bmatrix} q_1 & q_2 & \ldots & q_n \end{bmatrix}$, et par la question 2, 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 95 page 129 (Méthode QR pour la recherche de valeurs propres)

1.1 Par définition et associativité du produit des matrices,

$$A^2 = (Q_1 R_1)(Q_1 R_1) = Q_1 (R_1 Q_1) R_1 = Q_1 (R_1 Q_1) R_1 = Q_1 (Q_2 R_2) R_1 = (Q_1 Q_2)(R_2 R_1) = \tilde{Q}_2 \tilde{R}_2$$

avec $\tilde{Q}_2 = Q_1 Q_2$ et $\tilde{R}_2 = R_1 R_2$.

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, $PM_k = P\lambda^k L\lambda^{-k} = P\lambda^k P^t P^{-t} L\lambda^{-k} = A^k P^{-t} L\lambda^{-k}$.
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

$$(M_k)_{i,j} = (\lambda^k L\lambda^{-k})_{i,j} = \begin{cases} L_{i,i} & \text{si } i = j, \\ \frac{\lambda_j^k}{\lambda_i^k} Li, j & \text{si } i > j, \\ 0 & \text{sinon.} \end{cases}$$

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 $(M_k)_{i,i} \to 1$ 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 $\boldsymbol{b}_1^{(k)}$ de B_kC_k tend vers la première colonne de B, c'est -à-dire

$$c_{1,1}^{(k)}\boldsymbol{b}_1^{(k)} \to \boldsymbol{b}_1 \text{ lorsque } k \to \infty.$$

Comme les matrices B et B_k sont des matrices orthogonales, leurs vecteurs colonnes sont de norme 1, et donc

$$|c_{1,1}^{(k)}| = \|c_{1,1}^{(k)}\boldsymbol{b}_1^{(k)}\| \to \|\boldsymbol{b}_1\| = 1 \text{ lorsque } k \to \infty.$$

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 $c_{1,2}^{(k)}\boldsymbol{b}_1^{(k)} + c_{2,2}^{(k)}\boldsymbol{b}_2^{(k)}$.
On a donc

$$c_{1,2}^{(k)}\boldsymbol{b}_1^{(k)} + c_{2,2}^{(k)}\boldsymbol{b}_2^{(k)} \to \boldsymbol{b}_2 \text{ lorsque } k \to +\infty. \tag{1.142}$$

La matrice B_k est orthogonale, et donc $\boldsymbol{b}_1^{(k)} \cdot \boldsymbol{b}_1^{(k)} = 1$ et $\boldsymbol{b}_1^{(k)} \cdot \boldsymbol{b}_2^{(k)} = 0$. De plus, par la question précédente, $\boldsymbol{b}_1^{(k)} \to \boldsymbol{b}_1$ lorsque $k \to +\infty$, On a donc, en prenant le produit scalaire du membre de gauche de (1.142) avec $\boldsymbol{b}_1^{(k)}$,

$$c_{1,2}^{(k)} = \left(c_{1,2}^{(k)}\boldsymbol{b}_1^{(k)} + c_{2,2}^{(k)}\boldsymbol{b}_2^{(k)}\right) \cdot \boldsymbol{b}_1^{(k)} \to \boldsymbol{b}_2 \cdot \boldsymbol{b}_1 = 0 \text{ lorsque } k \to +\infty.$$

Comme $c_{1,2}^{(k)} \to 0$ et $\boldsymbol{b}_1^{(k)} \to \boldsymbol{b}_1$ on obtient par (1.142) que

$$c_{2,2}^{(k)}\boldsymbol{b}_2^{(k)} \to \boldsymbol{b}_2 \text{ lorsque } k \to +\infty.$$

Le même raisonnement que celui de la question précédente nous donne alors que $c_{2,2}^{(k)} \to 1$ et $\boldsymbol{b}_2^{(k)} \to \boldsymbol{b}_2$ 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 $\sum_{\ell=1}^{j-1} c_{\ell,j}^{(k)}\boldsymbol{b}_\ell^{(k)} + c_{j,j}^{(k)}\boldsymbol{b}_j^{(k)}$. On a donc

$$\sum_{\ell=1}^{j-1} c_{\ell,j}^{(k)}\boldsymbol{b}_\ell^{(k)} + c_{j,j}^{(k)}\boldsymbol{b}_j^{(k)} \to \boldsymbol{b}_j \text{ lorsque } k \to +\infty. \tag{1.143}$$

La matrice B_k est orthogonale, et donc $\boldsymbol{b}_i^{(k)} \cdot \boldsymbol{b}_j^{(k)} = \delta_{i,j}$. 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 (1.143) avec $\boldsymbol{b}_m^{(k)}$, pour $m < j$, on obtient

$$c_{m,j}^{(k)} = \left(\sum_{\ell=1}^{j-1} c_{\ell,j}^{(k)}\boldsymbol{b}_\ell^{(k)} + c_{j,j}^{(k)}\boldsymbol{b}_j^{(k)}\right) \cdot \boldsymbol{b}_m^{(k)} \to \boldsymbol{b}_m \cdot \boldsymbol{b}_j = 0 \text{ lorsque } k \to +\infty.$$

On déduit alors de (1.143) 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 :

$$\begin{cases} x \in \mathbb{R}^n \\ g(x) = 0. \end{cases} \tag{2.1}$$

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 Rappels et notations de calcul différentiel

Le premier chapitre faisait appel à vos connaissances en algèbre linéaire. Ce chapitre-ci, ainsi que le suivant (optimisation) s'appuieront sur vos connaissances en calcul différentiel, et nous allons donc réviser les quelques notions qui nous seront utiles.

2.1.1 Différentielle

Définition 2.1 (Application différentiable)**.** *Soient E et F des espaces vectoriels normés, f une application de E dans F et $x \in E$. On rappelle que f est différentiable en x s'il existe $T_x \in \mathcal{L}(E, F)$ (où $\mathcal{L}(E, F)$ est l'ensemble des applications linéaires continues de E dans F) telle que*

$$f(x + h) = f(x) + T_x(h) + \|h\|_E \varepsilon(h) \text{ avec } \varepsilon(h) \to 0 \text{ quand } h \to 0. \tag{2.2}$$

L'application T_x est alors unique [2] et on note $Df(x) = T_x \in \mathcal{L}(E, F)$ la différentielle de f au point x. Si f est différentiable en tout point de E, alors on appelle différentielle de f l'application $Df = E \to \mathcal{L}(E, F)$ qui à $x \in E$ associe l'application linéaire continue $Df(x)$ de E dans 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.

Remarquons tout de suite que si f est une application linéaire continue de E dans F, alors f est différentiable, et $Df = f$. En effet, si f est linéaire, $f(x+h) - f(x) = f(h)$, et donc l'égalité (2.2) est vérifiée avec $T_x = f$ et $\varepsilon = 0$.
Voyons maintenant quelques cas particuliers d'espaces E et F :

Cas où $E = \mathbb{R}$ et $F = \mathbb{R}$

Si f est une fonction de $\mathbb{R}$ dans $\mathbb{R}$, dire que f est différentiable en x revient à dire que f est dérivable en x. En effet, dire que f est dérivable en x revient à dire que

$$\lim_{h\to 0} \frac{f(x+h)-f(x)}{h} \text{ existe, et } \lim_{h\to 0} \frac{f(x+h)-f(x)}{h} = f'(x),$$

ce qui s'écrit encore

$$\frac{f(x+h)-f(x)}{h} = f'(x) + \varepsilon(h), \text{ avec } \varepsilon(h) \to 0 \text{ lorsque } h \to 0,$$

c'est-à-dire

$$f(x+h) - f(x) = T_x(h) + h\varepsilon(h), \text{ avec } T_x(h) = f'(x)h,$$

ce qui revient à dire que f est différentiable en x, et que sa différentielle en x est l'application linéaire $T_x : \mathbb{R} \to \mathbb{R}$, qui à h associe $f'(x)h$. On a ainsi vérifié que pour une fonction de $\mathbb{R}$ dans $\mathbb{R}$, la notion de différentielle coïncide avec celle de dérivée.

Exemple 2.2. *Prenons $f : \mathbb{R} \to \mathbb{R}$ définie par $f(x) = \sin x$. Alors f est dérivable en tout point et sa dérivée vaut $f'(x) = \cos x$. La fonction f est donc aussi différentiable en tout point. La différentielle de f au point x est l'application linéaire $Df(x)$ qui à $h \in \mathbb{R}$ associe $Df(x)(h) = \cos x\, h$. La différentielle de f est l'application de $\mathbb{R}$ dans $\mathcal{L}(\mathbb{R}, \mathbb{R})$, qui à x associe $Df(x)$ (qui est donc elle même une application linéaire).*

Cas où $E = \mathbb{R}^n$ et $F = \mathbb{R}^p$ Soit $f : \mathbb{R}^n \to \mathbb{R}^p$, $x \in \mathbb{R}^n$ et supposons que f est différentiable en x ; alors $Df(x) \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^p)$; par caractérisation d'une application linéaire de $\mathbb{R}^p$ dans $\mathbb{R}^n$, il existe une unique matrice $J_f(x) \in \mathcal{M}_{p,n}(\mathbb{R})$ telle que

$$\underbrace{Df(x)(y)}_{\in \mathbb{R}^p} = \underbrace{J_f(x)y}_{\in \mathbb{R}^p}, \ \forall y \in \mathbb{R}^n.$$

On confond alors souvent l'application linéaire $Df(x) \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^p)$ avec la matrice $J_f(x) \in \mathcal{M}_{p,n}(\mathbb{R})$ qui la représente, qu'on appelle **matrice jacobienne** de f au point x et qu'on note J_f. On écrit donc :

$$J_f(x) = Df(x) = (a_{i,j})_{1\le i\le p, 1\le j\le n} \text{ où } a_{i,j} = \partial_j f_i(x),$$

∂_j désignant la dérivée partielle par rapport à la j-ème variable.
Notons que si $n = p = 1$, la fonction f est de $\mathbb{R}$ dans $\mathbb{R}$ et la matrice jacobienne en x n'est autre que la dérivée en x : $J_f(x) = f'(x)$. On confond dans cette écriture la matrice $J_f(x)$ qui est de taille 1×1 avec le scalaire $f'(x)$.

Exemple 2.3. *Prenons $n = 3$ et $p = 2$; soit $f : \mathbb{R}^3 \to \mathbb{R}^2$ définie par :*

$$f(x) = \begin{pmatrix} x_1^2 + x_2^3 + x_3^4 \\ \\ 2x_1 - x_2 \end{pmatrix}, \ \forall x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

Soit $h \in \mathbb{R}^3$ de composantes (h_1, h_2, h_3). Pour calculer la différentielle de f (en x appliquée à h), on peut calculer $f(x+h) - f(x)$:

$$\begin{aligned} f(x+h) - f(x) &= \begin{bmatrix} (x_1+h_1)^2 - x_1^2 + (x_2+h_2)^3 - x_2^3 + (x_3+h_3)^4 - x_3^4 \\ \\ 2(x_1+h_1) - 2x_1 - (x_2+h_2) + x_2 \end{bmatrix} \\ &= \begin{bmatrix} 2x_1h_1 + h_1^2 + 3x_2^2h_2 + 3x_2h_2^2 + h_2^3 + 4x_3^3h_3 + 4x_3^2h_3^2 + h_3^4 \\ \\ 2h_1 - h_2 \end{bmatrix} \end{aligned}$$

et on peut ainsi vérifier l'égalité (2.2) *avec :*

$$Df(x)h = \begin{bmatrix} 2x_1h_1 + 3x_2^2h_2 + 4x_3^3h_3 \\ 2h_1 - h_2 \end{bmatrix}$$

et donc, avec les notations précédentes,

$$J_f(x) = \begin{bmatrix} 2x_1 & 3x_2^2 & 4x_3^3 \\ 2 & -1 & 0 \end{bmatrix}$$

Bien sûr, dans la pratique, on n'a pas besoin de calculer la différentielle en effectuant la différence $f(x+h)-f(x)$. On peut directement calculer les dériées partielles pour calculer la matrice jacobienne J_f.

Cas où $E = \mathbb{R}^n$, $F = \mathbb{R}$

C'est en fait un sous-cas du paragraphe précédent, puisqu'on est ici dans le cas $p = 1$. Soit $x \in \mathbb{R}^n$ et f une fonction de E dans F différentiable en x; on a donc $J_f(x) \in \mathcal{M}_{1,n}(\mathbb{R}) : J_f$ est une matrice ligne. On définit le **gradient** de f en x comme le vecteur de $\mathbb{R}^n$ dont les composantes sont les coefficients de la matrice colonne $(J_f(x))^t$, ce qu'on écrit, avec un abus de notation, $\nabla f(x) = (J_f(x))^t \in \mathbb{R}^n$. (L'abus de notation est dû au fait qu'à gauche, il s'agit d'un vecteur de $\mathbb{R}^n$, et à droite, une matrice $n \times 1$, qui sont des objets mathématiques différents, mais qu'on identifie pour alléger les notations). Pour $(x, y) \in (\mathbb{R}^n)^2$, on a donc

$$Df(x)(y) = J_f(x)y = \sum_{j=i}^{n} \partial_j f(x) y_j = \nabla f(x) \cdot y \text{ où } \nabla f(x) = \begin{bmatrix} \partial_1 f(x) \\ \vdots \\ \partial_n f(x) \end{bmatrix} \in \mathbb{R}^n.$$

Attention, lorsque l'on écrit $J_f(x)y$ il s'agit d'un *produit matrice vecteur*, alors que lorsqu'on écrit $\nabla f(x) \cdot y$, il s'agit du *produit scalaire entre les vecteurs* $\nabla f(x)$ et y, qu'on peut aussi écrire $\nabla(f(x))^t y$.

Cas où E est un espace de Hilbert et $F = \mathbb{R}$.

On généralise ici le cas présenté au paragraphe précédent. Soit $f : E \to \mathbb{R}$ différentiable en $x \in E$. Alors $Df(x) \in \mathcal{L}(E, \mathbb{R}) = E'$, où E' désigne le dual topologique de E, c.à.d. l'ensemble des formes linéaires continues sur E. Par le théorème de représentation de Riesz, il existe un unique $u \in E$ tel que $Df(x)(y) = (u|y)_E$ pour tout $y \in E$, où $(.|.)_E$ désigne le produit scalaire sur E. On appelle encore gradient de f en x ce vecteur u. On a donc $u = \nabla f(x) \in E$ et pour $y \in E$, $Df(x)(y) = (\nabla f(x)|y)_E$.

2.1.2 Différentielle d'ordre 2, matrice hessienne.

Revenons maintenant au cas général de deux espaces vectoriels normés E et F, et supposons maintenant que $f \in C^2(E, F)$. Le fait que $f \in C^2(E, F)$ signifie que $Df \in C^1(E, \mathcal{L}(E, F))$. Par définition, on a $D^2f(x) \in \mathcal{L}(E, \mathcal{L}(E, F))$ et donc pour $y \in E$, $D^2f(x)(y) \in \mathcal{L}(E, F)$, et pour $z \in E$, $D^2f(x)(y)(z) \in F$.
Considérons maintenant le cas particulier $E = \mathbb{R}^n$ et $F = \mathbb{R}$. On a :

$$f \in C^2(\mathbb{R}^n, \mathbb{R}) \Leftrightarrow \left[f \in C^1(\mathbb{R}^n, \mathbb{R}) \text{ et } \nabla f \in C^1(\mathbb{R}^n, \mathbb{R}^n) \right].$$

et

$$D^2f(x) \in \mathcal{L}(\mathbb{R}^n, \mathcal{L}(\mathbb{R}^n, \mathbb{R}))$$

Mais à toute application linéaire $\varphi \in \mathcal{L}(\mathbb{R}^n, \mathcal{L}(\mathbb{R}^n, \mathbb{R}))$, on peut associer de manière unique une forme bilinéaire ϕ sur $\mathbb{R}^n$ de la manière suivante :

$$\phi : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R} \tag{2.3}$$

$$(u, v) \mapsto \phi(u, v) = \underbrace{(\varphi(u))}_{\in \mathcal{L}(\mathbb{R}^n, \mathbb{R})} \underbrace{(v)}_{\in \mathbb{R}^n}. \tag{2.4}$$

On dit qu'il existe une isométrie canonique (un isomorphisme qui conserve la norme) entre l'espace vectoriel normé $\mathcal{L}(\mathbb{R}^n, \mathcal{L}(\mathbb{R}^n, \mathbb{R}))$ et l'espace des formes bilinéaires sur $\mathbb{R}^n$.
On appelle matrice hessienne de f et on note $H_f(x)$ la matrice de la forme bilinéaire ainsi associée à l'application linéaire $D^2f(x) \in \mathcal{L}(\mathbb{R}^n, \mathcal{L}(\mathbb{R}^n, \mathbb{R}))$.
On a donc $D^2f(x)(y)(z) = y^t H_f(x) z$. La matrice hessienne $H_f(x)$ peut se calculer à l'aide des dérivées partielles : $H_f(x) = (b_{i,j})_{i,j=1\ldots N} \in \mathcal{M}_n(\mathbb{R})$ où $b_{i,j} = \partial^2_{i,j} f(x)$ et $\partial^2_{i,j}$ désigne la dérivée partielle par rapport à la variable i de la dérivée partielle par rapport à la variable j. Notons que par définition (toujours avec l'abus de notation qui consiste à identifier les applications linéaires avec les matrices qui les représentent), $Dg(x)$ est la matrice jacobienne de $g = \nabla f$ en x.

Remarque 2.4 (Sur les différentielles, gradient et Hessienne)**.** *Pour définir la différentielle d'une fonction f d'un expace vectoriel de dimension finie E dans $\mathbb{R}$, on a besoin d'une norme sur E.*
Si f est différentiable en $x \in E$, pour définir le gradient de f en x, on a besoin d'un produit scalaire sur E pour pouvoir utiliser le théorème de representation de Riesz mentionné plus haut. Le gradient est défini de manière unique par le produit scalaire, mais ses composantes dépendent de la base choisie.
Enfin, si f est deux fois différentiable en $x \in E$, on a besoin d'une base de E pour définir la matrice hessienne en x, et cette matrice hessienne dépend de la base choisie.

2.1.3 Exercices (calcul différentiel)

Enoncés

Exercice 96 (Différentielle et gradient)**.** *Suggestions en page 139, corrigé détaillé en page 140*

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 97 (Calcul de différentielles)**.**

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 $f(x_1, x_2, x_3) = x_1^2 + x_1^2 x_2 + x_2 \sin(x_3)$.

Exercice 98 (Différentielle de l'inverse des matrices)**.** *Suggestions en page 139, corrigé en page 140*

1. Soit $\phi : \mathrm{GL}_n(\mathbb{R}) \to \mathrm{GL}_n(\mathbb{R})$ la fonction définie par $\phi(A) = A^{-1}$ pour $A \in \mathrm{GL}_n(\mathbb{R})$, où $\mathrm{GL}_n(\mathbb{R})$ désigne le groupe des matrices inversibles.
Donner l'expression de $D\phi(A)H$, différentielle de ϕ en A appliquée à H, pour $A \in \mathrm{GL}_n(\mathbb{R})$ et $H \in \mathcal{M}_n(\mathbb{R})$.
2. Soit $A \in C^1(\mathbb{R}^n, \mathrm{GL}_n(\mathbb{R}))$, et $\psi : \mathbb{R}^n \to \mathrm{GL}_n(\mathbb{R})$ définie par $\psi(\boldsymbol{x}) = A(\boldsymbol{x})^{-1}$. Montrer que ψ est différentiable et donner l'expression de $D\psi(\boldsymbol{x})(\boldsymbol{h})$ pour $\boldsymbol{h} \in \mathbb{R}^n$.

Suggestions

Exercice 96 page 139 (Différentielle et gradient)
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.

Exercice 98 page 139 (Différentielle de l'inverse des matrices) Calculer $\phi(A + H)$, pour $\|H\|$ suffisamment petit, à l'aide des séries de Neumann (voir exercice 53 page 77 et corollaire 1.39 page 70) et prendre garde à la non commutativité de la mmultiplication dans $\mathcal{M}_n(\mathbb{R})$.

Corrigés

Exercice 96 page 139 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 $\partial_i f \in C^1(\mathbb{R}^n, \mathbb{R})$, 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 $(A(x))_{i,j} = a_{i,j}(x) = \partial^2_{i,j} f(x)$.

Exercice 98 page 139 (Différentielle de l'inverse des matrices)
1. Soient $A \in \mathrm{GL}_n(\mathbb{R})$, $\|\cdot\|$ une norme matricielle sur $\mathcal{M}_n(\mathbb{R})$, et $H \in \mathcal{M}_n(\mathbb{R})$ telle que $\|A^{-1}H\| < 1$ On a $\phi(A + H) = (A + H)^{-1} = (\mathrm{Id} + A^{-1}H)^{-1}A^{-1}$. Comme $\|A^{-1}H\| < 1$, alors par le corollaire 1.39 page 70) $\rho(A^{-1}H) < 1$ et par l'exercice 53 page 77, on a

$$(\mathrm{Id} + A^{-1}H)^{-1} = \mathrm{Id} - A^{-1}H + R(H) \text{ avec } \|R(H)\| \leq \|H\|^2,$$

et donc

$$(A + H)^{-1} = (\mathrm{Id} - A^{-1}H + R(H))A^{-1} = A^{-1} - A^{-1}HA^{-1} + S(H) \text{ avec } \|S(H)\| \leq C\|H\|^2.$$

La fonction $\phi : A \mapsto A^{-1}$ est donc différentiable, et sa différentielle en A, notée $D\phi(A)$, est l'application linéaire de $\mathcal{M}_n(\mathbb{R})$ dans $\mathcal{M}_n(\mathbb{R})$, définie par $D\phi(A)H = -A^{-1}HA^{-1}$.

2. Soit $A \in C^1(\mathbb{R}^n, \mathrm{GL}_n(\mathbb{R}))$, et $\psi : \mathbb{R}^n \to \mathrm{GL}_n(\mathbb{R})$ définie par $\psi(\boldsymbol{x}) = A(\boldsymbol{x})^{-1}$.
Comme $A \in C^1(\mathbb{R}^n, \mathrm{GL}_n(\mathbb{R}))$, on peut écrire $A(\boldsymbol{x} + \boldsymbol{h}) = A(\boldsymbol{x}) + DA(\boldsymbol{x})(\boldsymbol{h}) + R(\boldsymbol{h})$, où $DA(\boldsymbol{x}) \in \mathcal{L}(\mathbb{R}^n, \mathrm{GL}_n(\mathbb{R}))$ est la différentielle de A en $\boldsymbol{x}$ et $R(\boldsymbol{h}) \in \mathcal{M}_n(\mathbb{R})$ est telle que $\|R(\boldsymbol{h})\| \leq \|\boldsymbol{h}\|^2$ pour une norme $\|\cdot\|$ sur $\mathbb{R}^n$ et sa norme induite $\|\cdot\|$ sur $\mathcal{M}_n(\mathbb{R})$.
On a donc, en supposant que $\rho(DA(\boldsymbol{x})(\boldsymbol{h}) + R(\boldsymbol{h})) < 1$, ce qui est vrai pour $\|\boldsymbol{h}\|$ suffisamment petit, et en raisonnant comme à la question 1,

$$\begin{aligned}\psi(\boldsymbol{x} + \boldsymbol{h}) &= (A(\boldsymbol{x} + \boldsymbol{h}))^{-1} \\ &= (A(\boldsymbol{x}) + DA(\boldsymbol{x})(\boldsymbol{h}) + R(\boldsymbol{h}))^{-1} \\ &= (A(\boldsymbol{x}))^{-1} - (A(\boldsymbol{x}))^{-1}DA(\boldsymbol{x})(\boldsymbol{h})(A(\boldsymbol{x}))^{-1} + S(\boldsymbol{h}))\end{aligned}$$

avec $S(\boldsymbol{h}) \in \mathcal{M}_n(\mathbb{R})$ telle que $\|S(\boldsymbol{h})\| \leq \|\boldsymbol{h}\|^2$. On a donc

$$\psi(\boldsymbol{x}+\boldsymbol{h}) - \psi(\boldsymbol{x}) = -(A(\boldsymbol{x}))^{-1} - DA(\boldsymbol{x})(\boldsymbol{h})(A(\boldsymbol{x}))^{-1} + S(\boldsymbol{h})),$$

ce qui montre que ψ est différentiable et que $D\psi(\boldsymbol{x}) \in \mathcal{L}(\mathbb{R}^n, \mathrm{GL}_n(\mathbb{R}))$ est définie par

$$D\psi(\boldsymbol{x})(\boldsymbol{h}) = -(A(\boldsymbol{x}))^{-1} DA(\boldsymbol{x})(\boldsymbol{h})(A(\boldsymbol{x}))^{-1},$$

2.2 Les méthodes de point fixe

2.2.1 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 2.5 (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 $\kappa \in]0,1[$ tel que $d(f(x), f(y)) \leq \kappa d(x,y)$ 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 $k \to +\infty$.*

DÉMONSTRATION – *Etape 1 : Existence de $\bar{x}$ et convergence de la suite*

Soit $x^{(0)} \in E$ et $(x^{(k)})_{k\in\mathbb{N}}$ la suite définie par $x^{(k+1)} = f(x^{(k)})$ pour $k \geq 0$. On va montrer que :

1. la suite $(x^{(k)})_{k\in\mathbb{N}}$ est de Cauchy (donc convergente car E est complet),
2. $\lim_{n\to+\infty} x^{(k)} = \bar{x}$ est point fixe de f.

Par hypothèse, on sait que pour tout $k \geq 1$,

$$d(x^{(k+1)}, x^{(k)}) = d(f(x^{(k)}), f(x^{(k-1)})) \leq \kappa d(x^{(k)}, x^{(k-1)}).$$

Par récurrence sur k, on obtient que

$$d(x^{(k+1)}, x^{(k)}) \leq \kappa^k d(x^{(1)}, x^{(0)}),\ \forall k \geq 0.$$

Soit $k \geq 0$ et $p \geq 1$, on a donc :

$$\begin{aligned} d(x^{(k+p)}, x^{(k)}) &\leq d(x^{(k+p)}, x^{(k+p-1)}) + \cdots + d(x^{(k+1)}, x^{(k)}) \\ &\leq \sum_{q=1}^{p} d(x^{(k+q)}, x^{(k+q-1)}) \\ &\leq \sum_{q=1}^{p} \kappa^{k+q-1} d(x^{(1)}, x^{(0)}) \\ &\leq d(x^{(1)}, x^{(0)}) \kappa^k (1 + \kappa + \ldots + \kappa^{p-1}) \\ &\leq d(x^{(1)}, x^{(0)}) \frac{\kappa^k}{1-\kappa} \longrightarrow 0 \text{ quand } k \to +\infty \text{ car } \kappa < 1. \end{aligned}$$

La suite $(x^{(k)})_{k\in\mathbb{N}}$ est donc de Cauchy, *i.e.* :

$$\forall \varepsilon > 0,\quad \exists k_\varepsilon \in \mathbb{N}\,;\quad \forall k \geq k_\varepsilon,\quad \forall\ p \geq 1\ \ d(x^{(k+p)}, x^{(k)}) \leq \varepsilon.$$

Comme E est complet, on a donc $x^{(k)} \longrightarrow \bar{x}$ dans E quand $k \to +\infty$. Comme la fonction f est strictement contractante, elle est continue, donc on a aussi $f(x^{(k)}) \longrightarrow f(\bar{x})$ dans E quand $k \to +\infty$. 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 $d(f(\bar{x}), f(\bar{y})) = d(\bar{x}, \bar{y}) \leq \kappa d(\bar{x}, \bar{y})$; comme $\kappa < 1$, 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 de la fonction $x \mapsto x^2$.

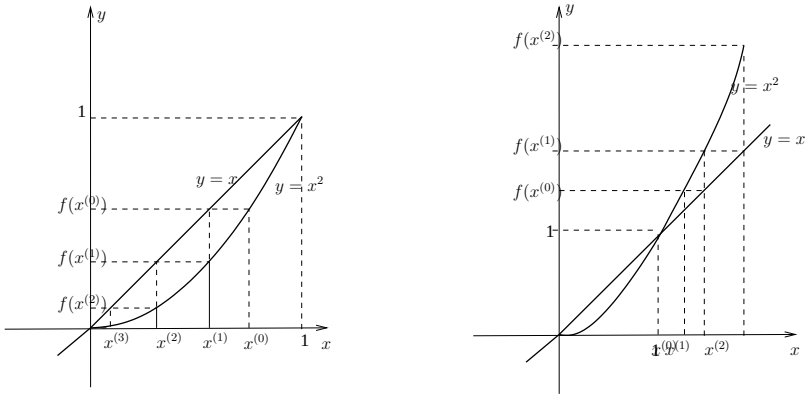


FIGURE 2.1: Comportement des itérés successifs du point fixe pour $x \mapsto x^2$– A gauche : $x^{(0)} < 1$, à droite : $x^{(0)} > 1$.

Pour la fonction $x \mapsto x^2$, on voit sur la figure 2.1, côté gauche, que si l'on part de $x = x^{(0)} < 1$, la méthode converge rapidement vers 0 ; or la fonction $x \mapsto x^2$ 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 $(x^{(k)})_{k \in \mathbb{N}}$ définie par le point fixe converge pour tout $x^{(0)} \in]-1, 1[$; ceci est très facile à voir car $x^{(k)} = (x^{(k)})^2$ et on a donc convergence vers 0 si $|x| < 1$.
Par contre si l'on part de $x^{(0)} > 1$ (à droite sur la figure 2.1), on diverge 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 la figure 2.2 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 $(x)_{k \in \mathbb{N}}$ 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.
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.

Remarque 2.6 (Vitesse de convergence)**.** *Sous les hypothèses du théorème 2.5, $d(x^{(k+1)}, \bar{x}) = d(f(x^{(k)}), f(\bar{x})) \leq k d(x^{(k)}, \bar{x})$; donc si $x^{(k)} \neq \bar{x}$ alors $\frac{d(x^{(k+1)}, \bar{x})}{d(x^{(k)}, \bar{x})} \leq \kappa$ (< 1), voir à ce sujet la définition 2.14. La convergence est donc au moins linéaire (même si de fait, cette méthode converge en général assez lentement).*

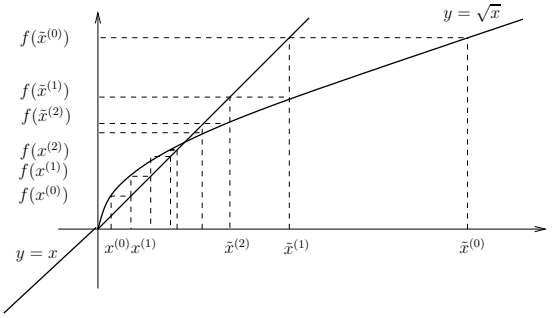


FIGURE 2.2: Comportement des itérés successifs du point fixe pour $x \mapsto \sqrt{x}$

Remarque 2.7 (Généralisation)**.** *Le théorème 2.5 se généralise en remplaçant l'hypothèse " f strictement contractante" par " il existe $k > 0$ tel que $f^{(k)} = \underbrace{f \circ f \circ \ldots \circ f}_{k \text{ fois}}$ 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

$$|f_\omega(x) - f_\omega(y)| = |x - y - \omega(g(x) - g(y))| \leq \kappa |x - y| \text{ pour } (x, y) \in \mathbb{R}^n \times \mathbb{R}^n, \text{ avec } \kappa < 1.$$

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 :

$$|g(x) - g(y)| \leq M|x - y|, \ \forall x, y \in \mathbb{R}^n. \tag{2.5}$$

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 $|x - y - \omega(g(x) - g(y))|^2 \leq \kappa |x - y|^2$, avec $\kappa < 1$. 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 :

$$\exists \alpha > 0 \text{ tel que } (g(x) - g(y)) \cdot (x - y) \geq \alpha |x - y|^2, \forall x, y \in \mathbb{R}^n, \tag{2.6}$$

On obtient alors :

$$|x - y - \omega(g(x) - g(y))|^2 \leq (1 + \omega^2 M^2 - 2\omega\alpha)|x - y|^2.$$

Et donc si $\omega \in]0, \frac{2\alpha}{M^2}[$, le polynôme $\omega^2 M^2 - 2\omega\alpha$ 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 2.8 (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* (2.6) *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 $k \to +\infty$ avec $x^{(k+1)} = f_\omega(x^{(k)}) = x^{(k)} - \omega g(x^{(k)})$.*

Remarque 2.9. *Le théorème 2.8 permet de montrer que sous les hypothèses* (2.6) *et* (2.5)*, et pour $\omega \in]0, \frac{2\alpha}{M^2}[$, on peut obtenir la solution de* (2.1) *en construisant la suite :*

$$\begin{cases} x^{(k+1)} = x^{(k)} - \omega g(x^{(k)}) \ \ n \geq 0, \\ x^{(0)} \in \mathbb{R}^n. \end{cases} \tag{2.7}$$

Or on peut aussi écrire cette suite de la manière suivante (avec $f(x) = x - g(x)$) :

$$\begin{cases} \tilde{x}^{(k+1)} &= f(x^{(k)}), \ \forall n \geq 0 \\ x^{(k+1)} &= \omega \tilde{x}^{(k+1)} + (1-\omega)x^{(k)}, \quad x^{(0)} \in \mathbb{R}^n. \end{cases} \tag{2.8}$$

En effet si $x^{(k+1)}$ est donné par la suite (2.8)*, 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 (2.8) *est l'algorithme de relaxation sur f.*

La proposition suivante donne une condition suffisante pour qu'une fonction vérifie les hypothèses (2.6) et (2.5).

Proposition 2.10. *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 α et M tels que*

$$\alpha \leq \lambda_i(x) \leq M, \ \forall i \in \{1 \dots n\}, \ \forall x \in \mathbb{R}^n.$$

(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 (2.6) *et* (2.5) *du théorème 2.8.*

DÉMONSTRATION – Montrons d'abord que l'hypothèse (2.6) est vérifiée. Soit $(x, y) \in (\mathbb{R}^n)^2$, on veut montrer que $(g(x) - g(y)) \cdot (x - y) \geq \alpha |x - y|^2$. 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 :

$$(g(y) - g(x)) \cdot (y - x) = \int_0^1 \left(Dg(x + t(y - x))(y - x) \cdot (y - x) \right) dt.$$

Comme $\lambda_i(x) \in [\alpha, M] \ \forall i \in \{1, \dots, n\}$, on a

$$\alpha |w|^2 \leq Dg(z)w \cdot w \leq M |w|^2 \text{ pour tout } w, z \in \mathbb{R}^n$$

On a donc :

$$(g(y) - g(x)) \cdot (y - x) \geq \int_0^1 \alpha |y - x|^2 dt = \alpha |y - x|^2$$

ce qui montre que l'hypothèse (2.6) est bien vérifiée.

Montrons maintenant que l'hypothèse (2.5) est vérifiée. On veut montrer que $|g(y) - g(x)| \leq M|y - x|$. 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 $\lambda_i(x) \in [\alpha, M]$ pour tout $i = 1, \ldots, n$, la matrice $Dg(x + t(y - x))$ est symétrique définie positive et donc, d'après la proposition 1.33 page 66, son rayon spectral est égal à sa norme, pour la norme induite par la norme euclidienne. On a donc :

$$|Dg(x + t(y - x)| = \rho(Dg(x + t(y - x)) \leq M.$$

On a donc ainsi montré que : $|g(y) - g(x)| \leq M|y - x|$, ce qui termine la démonstration. ∎

2.2.2 Point fixe de monotonie

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 :

$$\tilde{x}^{(k+1)} = A^{-1}R(x^{(k)}), \qquad x^{(k+1)} = \omega\tilde{x}^{(k+1)} + (1 - \omega)x^{(k)}.$$

Si la matrice A possède une propriété dite "de monotonie", on peut montrer la convergence de l'algorithme du point fixe ; c'est l'objet du théorème suivant.

Théorème 2.11 (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 inversible d'inverse à coefficients positis, ou ICP-matrice (voir la proposition 1.18 et l'exercice 14), c'est-à-dire 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'est-à-dire 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. $0 \leq x^{(k)} \leq \tilde{x},\ \forall k \in \mathbb{N}$,
2. $x^{(k+1)} \geq x^{(k)},\ \forall k \in \mathbb{N}$,
3. $x^{(k)} \to \bar{x}$ *quand* $k \to +\infty$ *et* $A\bar{x} = R(\bar{x})$.

DÉMONSTRATION – Comme A est inversible la suite $(x^{(k)})_{n\in\mathbb{N}}$ vérifiant

$$\begin{cases} x^{(0)} = 0, \\ Ax^{(k+1)} = R(x^{(k)}), \ \ k \geq 0 \end{cases}$$

est bien définie. On va montrer par récurrence sur k que $0 \leq x^{(k)} \leq \tilde{x}$ pour tout $k \geq 0$ et que $x^{(k)} \leq x^{(k+1)}$ pour tout $k \geq 0$.

1. Pour $k = 0$, 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 $p = k-1$, 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 $(x^{(k)})_{k\in\mathbb{N}}$, 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

$$0 \leq x^{(k)} \leq \tilde{x}, \quad \forall k \geq 0$$
$$x^{(k)} \leq x^{(k+1)}, \quad \forall k \geq 0.$$

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 $\forall k \geq 0$.

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 $\bar{x}_i = \lim_{k\to+\infty} x_i^{(k)}$. 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 $k \to +\infty$.

Enfin, comme $Ax^{(k+1)} = R(x^{(k)})$ et comme R est continue, on obtient par passage à la limite lorsque $k \to +\infty$ que $A\bar{x} = R(\bar{x})$ et que $0 \leq \bar{x} \leq \tilde{x}$. ∎

L'hypothèse 1 du théorème 2.11 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 12 et l'exercice 66) et Δu sur $]0,1[\times]0,1[$ (voir page 15).

Théorème 2.12 (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 il existe $\tilde{x} \geq 0$ 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 $A + \beta\mathrm{Id}$ au lieu de A et $R + \beta$ au lieu de R. ∎

Remarque 2.13 (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[3] *: 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.*

2.2.3 Vitesse de convergence

Définition 2.14 (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 $k \to +\infty$, que la suite est non stationnaire, c.à.d. que $x^{(k)} \neq \bar{x}$ pour tout $k \in \mathbb{N}$, et que*

$$\lim_{k\to+\infty} \frac{\|x^{(k+1)} - \bar{x}\|}{\|x^{(k)} - \bar{x}\|} = \beta \in [0,1]. \tag{2.9}$$

On s'intéresse à la "vitesse de convergence" de la suite $(x^{(k)})_{k\in\mathbb{N}}$. 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 $k \geq n_0$ alors $\|x^{(k+1)} - \bar{x}\| \leq \gamma\|x^{(k)} - \bar{x}\|^2$.*
   (b) *La convergence est* **quadratique** *si*

$$\lim_{k\to+\infty} \frac{\|x^{(k+1)} - \bar{x}\|}{\|x^{(k)} - \bar{x}\|^2} = \gamma > 0.$$

   Plus généralement, on dit que :
   (a) *La convergence est* **au moins d'ordre** *p s'il existe $\gamma \in \mathbb{R}_+$ et il existe $k_0 \in \mathbb{N}$ tels que si $k \geq k_0$ alors $\|x^{(k+1)} - \bar{x}\| \leq \gamma\|x^{(k)} - \bar{x}\|^p$.*
   (b) *La convergence est* **d'ordre** *p si*

$$\lim_{k\to+\infty} \frac{\|x^{(k+1)} - \bar{x}\|}{\|x^{(k)} - \bar{x}\|^p} = \gamma > 0.$$

Remarque 2.15 (Sur la vitesse de convergence des suites)**.**

— Remarquons d'abord que si une suite $(x^{(k)})_{k\in\mathbb{N}}$ de $\mathbb{R}^n$ converge vers $\bar{x}$ lorsque k tend vers l'infini, et qu'il existe β vérifiant (2.9), alors on a forcément $\beta \leq 1$. En effet, si la suite vérifie (2.9) avec $\beta > 1$, alors il existe $k_0 \in \mathbb{N}$ tel que si $k \geq k_0$, $|x_n - \bar{x}| \geq |x_{k_0} - \bar{x}|$ pour tout $k \geq k_0$, ce qui contredit la convergence.

— *Quelques exemples de suites qui convergent sous-linéairement* : $x_k = \frac{1}{\sqrt{k}}$, $x_k = \frac{1}{k}$, mais aussi, de manière moins intuitive : $x_k = \frac{1}{k^2}$. Toutes ces suites vérifient l'égalité (2.9) 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 $x_k = \beta^k$ avec $\beta \in]0,1[$ sont des suites qui convergent linéairement (vers 0), car elles verifient évidemment bien (2.9) avec $\beta \in]0,1[$.

— La convergence quadratique est encore plus rapide ! Par exemple la suite définie par $x_{k+1} = x_k^2$ converge de manière quadratique pour un choix initial $x_0 \in]-1,1[$. Mais si par malheur le choix initial est en dehors

3. Luitzen Egbertus Jan Brouwer (1881-1966), mathématicien néerlandais.

de cet intervalle, la suite diverge alors très vite... de manière exponentielle, en fait (pour $x_0 > 1$, on a $x_k = e^{2k \ln x_0}$).

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 2.16 (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 $k \to +\infty$. De plus si $x^{(k)} \neq \bar{x}$ pour tout $k \in \mathbb{N}$, 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 $k \to +\infty$, et si $x^{(k)} \neq \bar{x}$, $\forall k \in \mathbb{N}$ 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 $\xi_k \in]x^{(k)}, \bar{x}[$ ou $]\bar{x}, x^{(k)}[$, tel que $f(x^{(k)}) - f(\bar{x}) = f'(\xi_k)(x^{(k)} - \bar{x})$. On a donc

$$\frac{|x^{(k+1)} - \bar{x}|}{|x^{(k)} - \bar{x}|} = |f'(\xi_k)| \longrightarrow |f'(\bar{x})| \text{ car } x^{(k)} \to \bar{x} \text{ et } f' \text{ est continue.}$$

   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 $k \to +\infty$. On veut estimer la vitesse de convergence ; on suppose pour cela que $x^{(k)} \neq \bar{x}$ pour tout $k \in \mathbb{N}$. Comme $f \in C^2(\mathbb{R}, \mathbb{R})$, il existe $\xi_k \in]x^{(k)}, \bar{x}[$ tel que

$$f(x^{(k)}) - f(\bar{x}) = f'(\bar{x})(x^{(k)} - \bar{x}) + \frac{1}{2} f''(\xi_k)(x^{(k)} - \bar{x})^2.$$

On a donc : $x^{(k+1)} - \bar{x} = \frac{1}{2} f''(\xi_k)(x^{(k)} - \bar{x})^2$ ce qui entraîne, par continuité de f'', que

$$\frac{|x^{(k+1)} - \bar{x}|}{|x^{(k)} - \bar{x}|^2} = \frac{1}{2}|f''(\xi_k)| \longrightarrow \frac{1}{2}|f''(\bar{x})| \text{ quand } k \to +\infty.$$

La convergence est donc au moins quadratique.

■

2.2.4 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 [4] 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 $(x^{(k)})_{k\in\mathbb{N}} \subset \mathbb{R}^n$ 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 2.16) $(x^{(k)})_{n\in\mathbb{N}}$ 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 $g'(\overline{x}) \neq 0$ 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 2.16, 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.

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 :

$$g(x^{(k)}) + g'(x^{(k)})(x^{(k+1)} - x^{(k)}) = 0 \text{ ou encore } g'(x^{(k)})(x^{(k+1)} - x^{(k)}) = -g(x^{(k)}). \tag{2.10}$$

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

$$x^{(0)} \text{ donné ,}$$
$$x^{(k+1)} = x^{(k)} - (x^{(k)})^2 + 3.$$

4. 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 !

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

$$\begin{aligned} x^{(0)} \text{ donné ,} \\ x^{(k+1)} = -\frac{(x^{(k)})^2 - 3}{2x^{(k)}}. \end{aligned}$$

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 2.17 (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* $\varphi(2\pi) - \varphi(0) = 2\pi\varphi'(\xi)$.

2.2.5 Exercices (méthodes de point fixe)

Enoncés

Exercice 99 (Un point fixe dans $\mathbb{R}$)**.** *Corrigé en page 154*

1. Etudier la convergence de la suite $(x^{(k)})_{k\in\mathbb{N}}$, définie par $x^{(0)} \in [0,1]$ et $x^{(k+1)} = \cos\left(\dfrac{1}{1+x^{(k)}}\right)$.
2. 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 100 (Un autre point fixe dans $\mathbb{R}$)**.** *Corrigé détaillé en page 155.*

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 l'algorithme donné en (b).
2. On veut résoudre l'équation $x^2 - 2 = 0$, $x > 0$.
   (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 101 (Point fixe pour la résolution d'une équation non linéaire)**.** Soient $g \in C^3(\mathbb{R}, \mathbb{R})$ et $\bar{x} \in \mathbb{R}$ tels que $g(\bar{x}) = 0$, et soit $\omega \in \mathbb{R}^\star$. On définit f_ω par $f_\omega(x) = x - \omega g(x)$ et on cherche à calculer $\bar{x}$ en utilisant l'algorithme du point fixe pour f_ω, c'est-à-dire que l'on se donne $x_0 \in \mathbb{R}$ et on définit la suite $(x_k)_{k\in\mathbb{N}}$ par $x_{k+1} = f_\omega(x_k)$.

1. On suppose dans cette question que $|f'_\omega(\bar{x})| > 1$. Montrer qu'il existe $M > 1$ et $\alpha > 0$ tels que $|x - \bar{x}| \leq \alpha$ implique $|f'_\omega(x)| > M$. En déduire que la suite $(x_k)_{k\in\mathbb{N}}$ ne peut pas converger vers $\bar{x}$ si $x_0 \neq \bar{x}$.

On suppose maintenant que $|f'_\omega(\bar{x})| < 1$.

2. Montrer que pour x_0 bien choisi, autre que $\bar{x}$, on a $\lim_{k\to+\infty} x_k = \bar{x}$.
3. On suppose que la suite $(x_k)_{k\in\mathbb{N}}$ converge vers $\bar{x}$ lorsque $k \to +\infty$. Montrer que si on suppose $x_k \neq \bar{x}$ pour tout k, il existe $\beta < 1$ tel que pour k suffisamment grand,

$$\frac{|x_{k+1} - \bar{x}|}{|x_k - \bar{x}|} \leq \beta.$$

   Donner une valeur possible pour β (dépendant de f_ω et $\bar{x}$). En déduire que la convergence de x_k vers $\bar{x}$ est au moins linéaire.
4. On suppose dans cette question que $g'(\bar{x}) \neq 0$ et que $\omega = \dfrac{1}{g'(\bar{x})}$. Montrer que la convergence de x_k vers $\bar{x}$ est au moins quadratique, c'est-à-dire que si on suppose $x_k \neq \bar{x}$ pour tout k, il existe $\beta > 0$ tel que $\lim_{k\to+\infty} \dfrac{|x_{k+1} - \bar{x}|}{|x_k - \bar{x}|^2} \leq \beta$.

Exercice 102 (Point fixe pour la résolution d'un système non linéaire)**.** *Soient $G \in C^2(\mathbb{R}^n, \mathbb{R}^n)$ et $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ t.q. $G(\bar{\boldsymbol{x}}) = 0$. On suppose (par miracle) que l'on connaît la matrice jacobienne $J_G(\bar{\boldsymbol{x}})$ de G en $\bar{\boldsymbol{x}}$, et que cette matrice est inversible. On pose $A = J_G(\bar{\boldsymbol{x}})$ et on considère la méthode itérative suivante*
Initialisation. $\boldsymbol{x}_0 \in \mathbb{R}^n$
Itération. *Pour $k \geq 0$, $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - A^{-1}G(\boldsymbol{x}_k)$, c'est-à-dire $\boldsymbol{x}_{k+1} = F(\boldsymbol{x}_k)$ en posant $F(\boldsymbol{x}) = x - A^{-1}G(\boldsymbol{x})$. Pour $\alpha > 0$ on note $B_\alpha = \{\boldsymbol{x} \in \mathbb{R}^n,\ \|\boldsymbol{x} - \overline{\boldsymbol{x}}\| \leq \alpha\}$ (où $\|\cdot\|$ est une norme sur $\mathbb{R}^n$).*

1. *Montrer que $J_F(\bar{\boldsymbol{x}}) = 0$. Puis, montrer qu'il existe $\alpha_0 > 0$ t.q., pour tout $0 < \alpha \leq \alpha_0$, F envoie B_α dans lui même.*
2. *Montrer qu'il existe $\alpha > 0$ t.q. F est strictement contractante de B_α dans lui même.*
   En déduire que $\lim_{k\to+\infty} \boldsymbol{x}_k = \bar{\boldsymbol{x}}$ si $\boldsymbol{x}_0 \in B_\alpha$.
3. *On suppose que $\lim_{k\to+\infty} \boldsymbol{x}_k = \bar{\boldsymbol{x}}$, montrer que la convergence est au moins quadratique.*

Exercice 103 (Méthode itérative pour la résolution d'un système non linéaire)**.** *Soit $G \in C^1(\mathbb{R}^n, \mathbb{R}^n)$. Pour $x \in \mathbb{R}^n$, on note $J_G(x)$ la matrice jacobienne de G au point x. 0n suppose que pour toute valeur propre λ de $A^{-1}J_G(\overline{\boldsymbol{x}})$ on a $|\lambda - 1| < 1$.*
Notation : Si $\|\cdot\|$ est une norme sur $\mathbb{R}^n$, on note aussi $\|\cdot\|$ la norme induite sur $\mathcal{M}_n(\mathbb{R})$.

Soit $\overline{x} \in \mathbb{R}$ t.q. $G(\overline{x}) = 0$. On cherche à calculer $\overline{x}$ par une méthode itérative. On se donne une matrice $A \in \mathcal{M}^n(\mathbb{R})$ inversible et on considère l'algorithme suivant :

Initialisation : *On se donne $x_0 \in \mathbb{R}^n$.*
Itération : *Pour $k \in \mathbb{N}$, $A(x_{k+1} - x_k) = -G(x_k)$.*

1. *Ecrire cet algorithme comme un algorithme de point fixe pour une fonction $F \in C^1(\mathbb{R}^n, \mathbb{R}^n)$ qu'on explicitera.*
2. *Montrer que $\rho(I - A^{-1}J_G(\overline{x})) < 1$.*
3. *En déduire qu'il existe une norme sur $\mathbb{R}^n$, notée $\|\cdot\|$, pour laquelle $\|I - A^{-1}J_G(\overline{x})\| < 1$.*

Dans la suite, on utilise cette norme sur $\mathbb{R}^n$ et pour $\alpha > 0$, on note $B_\alpha = \{x \in \mathbb{R}^n,\ \|x - \overline{x}\| \leq \alpha\}$.

4. *Montrer qu'il existe $\alpha > 0$ tel que F est strictement contractante de B_α dans B_α.*
5. *En déduire que si $x_0 \in B_\alpha$, l'algorithme proposé converge vers $\overline{x}$ (c'est-à-dire $\lim_{k\to+\infty} x_k = \overline{x}$). Montrer que cette convergence est au moins linéaire.*

Dans la suite, on suppose que $n = 1$. La fonction G est notée g, la fonction F est notée f et la matrice A (qui est donc ici un scalaire) est notée a. L'itération de l'algorithme est donc $a(x_{k+1} - x_k) = -g(x_k)$.

6. *L'algorithme proposé correspond-t-il à un algorithme vu en cours ? Si oui, lequel ?*
7. *(Etude d'un exemple.) On suppose que $g(x) = x^2 - 3$ et $\overline{x} = \sqrt{3}$.*
   (a) *Donner explicitement les valeurs de a pour lesquelles il est certain que si x_0 est assez proche de $\bar{x}$, l'algorithme converge vers $\bar{x}$.*
   (b) *Pour une telle valeur de a, donner explicitement un intervalle B_α pour lequel cette convergence est assurée si $x_0 \in B_\alpha$.*

Exercice 104 (Méthode de monotonie)**.** *Suggestions en page 154, corrigé détaillé en page 156.*

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 $u_1, \ldots, u_n$) :

$$\begin{array}{l} (Au)_i = \alpha_i f(u_i) + \lambda b_i \ \forall i \in \{1, \ldots, n\}, \\ u = (u_1, \ldots, u_n)^t \in \mathbb{R}^n, \end{array} \tag{2.11}$$

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

$$u \in \mathbb{R}^n,\ Au \geq 0 \Rightarrow u \geq 0. \tag{2.12}$$

On suppose qu'il existe $\mu > 0$ t.q. (2.11) 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 $n \geq 0$,

$$(Av^{(k+1)})_i = \alpha_i f(v_i^{(k)}) + \lambda b_i \ \forall i \in \{1, \ldots, n\}. \tag{2.13}$$

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 (2.11) (et vérifie $0 \leq u^{(\lambda)} \leq u^{(\mu)}$).

Exercice 105 (Point fixe amélioré)**.** *Suggestions en page 154, Corrigé en page 156*

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 :

$$\left\{ \begin{array}{l} x_0 \in \mathbb{R}, \\ \\ x_{n+1} = h(x_n), n \geq 0. \end{array} \right. \tag{2.14}$$

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 (2.14) 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 (2.14) 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 (2.14) 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 (2.14) converge de manière cubique.

Exercice 106 (Un problème de rayonnement)**.** *Suggestions en page 154.*
On cherche à résoudre un modèle de diffusion thermique avec rayonnement (dans un matériau comme le verre, par exemple) avec une discrétisation par différences finies et une méthode de monotonie (voir aussi TP1).
Le modèle (simplifié) consiste à chercher la fonction u de $[0, 1]$ à valeurs dans $\mathbb{R}$ solution du problème suivant :

$$-\kappa u''(x) + \int_0^1 \frac{1}{\sqrt{|x-y|}}(u^4(x) - u^4(y))dy = 0 \text{ pour } x \in]0, 1[, \tag{2.15}$$

$$u(0) = c_1,\ u(1) = c_2. \tag{2.16}$$

Pour $n \geq 1$, on pose $h = 1/(n + 1)$.
La discrétisation de (2.15)-(2.16) par différences finies avec un pas uniforme $h = \frac{1}{n}$ consiste à chercher le vecteur u de $\mathbb{R}^n$ solution de

$$Au + R(u) = b, \tag{2.17}$$

où

$A \in \mathcal{M}_n(\mathbb{R})$, $A[i, i] = \frac{2\kappa}{h^2}$, $A[i, j] = -\frac{\kappa}{h^2}$ si $|i - j| = 1$ et $A[i, j] = 0$ si $|i - j| > 1$, $i, j \in \{1, \ldots, n\}$.

$R(u)_i = \sum_{j\in\{1,\ldots,n\},\, j\neq i} \frac{\sqrt{h}}{\sqrt{|i-j|}}(u_i^4 - u_j^4) + \frac{\sqrt{h}}{2\sqrt{i}}(u_i^4 - 1) + \frac{\sqrt{h}}{2\sqrt{n+1-i}}(u_i^4 - 16), \qquad i \in \{1, \ldots, n\}$

$b \in \mathbb{R}^n$, $b_1 = c_1\frac{\kappa}{h^2}$, $b_n = c_2\frac{\kappa}{h^2}$, $b_i = 0$ pour $1 < i < n$.

Pour trouver une solution de (2.17), on se donne $\beta \geq 0$ et on utilise la méthode itérative suivante :

$$\begin{aligned} &\textbf{Initialisation } u^{(0)} \in \mathbb{R}^n,\ u_i^{(0)} = 1 \text{ pour tout } i \in \{1, \ldots, n\}. \\ &\textbf{Itérations } Au^{(k+1)} + \beta u^{(k+1)} = -R(u^{(k)}) + \beta u^{(k)} + b. \end{aligned} \tag{2.18}$$

Dans toute la suite, on prendra $\kappa = 10$, $c_1 = 1$, $c_2 = 2$, et on note m l'élément de $\mathbb{R}^n$ dont toutes les composantes sont égales à 1 (c'est-à-dire $m = u^{(0)}$) et M l'élément de $\mathbb{R}^n$ dont toutes les composantes sont égales à 2.

1. (Propriété de A) Soient $\beta \geq 0$ et $u \in \mathbb{R}^n$. Montrer que $Au + \beta u \geq 0 \Rightarrow u \geq 0$.
   Suggestion : Considérer $i_0 = \min\{i \in \{1, \ldots, n\}$ tel que $u_i \leq u_j$ pour tout $j \in \{1, \ldots, n\}\}$ et montrer $u_{i_0} \geq 0$.
2. (Propriété de R)
   (a) Soient $u \in \mathbb{R}^n$, $m \leq u \leq M$, et $i, j \in \{1, \ldots, n\}$.
   Montrer que $\partial_j R_i(u) \leq 0$ pour $i \neq j$ et $\partial_i R_i(u) \leq 128$.
   Suggestion pour la minoration de $\partial_i R_i(u)$: calculer $\partial_i R_i(u)$ et remarquer que le terme en racine au dénominateur dans la somme obtenue est toujours du type $\sqrt{k}$ avec k entre 1 et n et pour un k donné, $\sqrt{k}$ ne peut apparaître qu'au plus deux fois.
   (b) Soient $u, v \in \mathbb{R}^n$, $m \leq u \leq v \leq M$, et $\beta \geq 128$. Montrer que $\beta u - R(u) \leq \beta v - R(v)$.

3. (Sous et sur solutions) Montrer que $Am + R(m) \leq b$ et $AM + R(M) \geq b$.
4. On choisit $\beta \geq 128$. Montrer, par récurrence sur k, que pour tout $k \geq 0$,

$$m \leq u^{(k)} \leq u^{(k+1)} \leq M.$$

En déduire qu'il existe $u \in \mathbb{R}^n$ tel que $u^{(k)} \to u$ quand $k \to +\infty$ et que $Au + R(u) = b$, $m \leq u \leq M$.

5. On initialise maintenant la méthode (2.18) par $u_i^{(0)} = 2$ au lieu de $u_i^0 = 1$ (pour tout $i = 1, \ldots, n$). La méthode converge-t'elle ?

Suggestions

Exercice 104 page 152 (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 105 page 152 (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

$$|x_{k+1} - \overline{x}| = (x_k - \overline{x})(1 - \frac{g(x_k) - g(\overline{x})}{(x_k - \overline{x})g'(\varphi(x_k))}. \tag{2.19}$$

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).

Exercice 106 page 153 (Rayonnement)

1. Considérer $i_0 = \min\{i \in \{1, \ldots, n\}$ tel que $u_i \leq u_j$ pour tout $j \in \{1, \ldots, n\}\}$ et montrer $u_{i_0} \geq 0$.
2. (Propriété de R)
   (a) Suggestion pour la minoration de $\partial_i R_i(u)$: calculer $\partial_i R_i(u)$ et remarquer que le terme en racine au dénominateur dans la somme obtenue est toujours du type $\sqrt{k}$ avec k entre 1 et n et pour un k donné, $\sqrt{k}$ ne peut apparaître qu'au plus deux fois.

Corrigés

Exercice 99 page 150 (Un point fixe dans $\mathbb{R}$)

1. On vérifie que l'application $f : x \mapsto \cos\left(\frac{1}{1+x}\right)$ est une application de $[0, 1]$ dans lui-même qui est contractante. En effet, $0 < \frac{1}{1+x} \leq 1 \leq \frac{\pi}{2}$ pour tout $x \in [0, 1]$, donc $f(x) \in [0, 1]$ pour tout $x \in [0, 1]$. De plus, $f'(x) = \frac{1}{(1+x)^2} \sin\left(\frac{1}{1+x}\right)$. On voit que $f'(x) \geq 0$ pour tout $x \in [0, 1]$ et $f'(x) \leq \sin(1) < 1$. On peut donc appliquer le théorème de point fixe de Banach pour déduire que f admet un unique point fixe dans l'intervalle $[0, 1]$ qui est limite de toutes les suites définies par $x^{(0)} \in [0, 1]$, $x^{(k+1)} = f(x^{(k)})$.
2. La suite des itérés de point fixe est définie par $x_0 \in [0, 1]$ et $x_{n+1} = (x_n)^4$.
   (a) Si $x_0 = 0$, la suite est stationnaire et égale à 0.
   (b) Si $x_0 = 1$, la suite est stationnaire et égale à 1.
   (c) Si $x_0 \in]0, 1[$, on montre par une récurrence facile que

i. $x_{n+1} < x_n$,

ii. $x_{n+1} \in]0,1[$.

On en déduit que la suite converge vers une limite ℓ, et en passant à la limite sur $x_{n+1} = (x_n)^4$, on obtient $\ell = 0$ ou 1. Comme $\ell \leq x_0 < 1$, on en dd́uit que $\ell = 0$.

Exercice 100 page 150 (Un autre point fixe dans $\mathbb{R}$)

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

$$x^{(0)} \text{ donné} \tag{2.20a}$$
$$x^{(k+1)} = f(x^{(k)}). \tag{2.20b}$$

   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 $|f'(x)| \leq \frac{1}{2}$ pour $x \in [0,1]$. De plus $f(x) \in [0,1]$ si $x \in [0,1]$. L'application $x \mapsto f(x) = \frac{1}{2}e^{-x}$ est donc strictement 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 $]0,4[$. 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 (2.20), 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.

Exercice 104 page 152 (Méthode de monotonie) Montrons que la suite $v^{(k)}$ est bien définie. Supposons $v^{(k)}$ connu ; alors $v^{(k+1)}$ est bien défini si le système

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

où $d^{(x)}$ est défini par : $d_i^{(k)} = \alpha_i f(v_i^{(k)}) + \lambda b_i$ pour $i = 1, \ldots, n$, admet une solution. Or, grâce au fait que $Av \geq 0 \Rightarrow v \geq 0$, la matrice A est inversible, ce qui prouve l'existence et l'unicité de $v^{(k+1)}$.
Montrons maintenant que les hypothèses du théorème de convergence du point fixe de monotonie sont bien satisfaites.
On pose $R_i^{(\lambda)}(u) = \alpha_i f(u_i) + \lambda b_i$. Le système à résoudre s'écrit donc :

$$Au = R^{(\lambda)}(u)$$

Or 0 est sous–solution car $0 \leq \alpha_i f(0) + \lambda b_i$ (grâce au fait que $f(0) = 0$, $\lambda > 0$ et $b_i \geq 0$).
Cherchons maintenant une sur–solution, c'est–à–dire $\tilde{u} \in \mathbb{R}^n$ tel que

$$\tilde{u} \geq R^{(\lambda)}(\tilde{u}).$$

Par hypothèse, il existe $\mu > 0$ et $u^{(\mu)} \geq 0$ tel que

$$(Au^{(\mu)})_i = \alpha f(u_i^{(\mu)}) + \mu b_i.$$

Comme $\lambda < \mu$ et $b_i \geq 0$, on a

$$(Au^{(\mu)})_i \geq \alpha_i f(u_i^{(\mu)}) + \lambda b_i = R_i^{(\lambda)}(u^{(\mu)}).$$

Donc $u^{(\mu)}$ est sur–solution. Les hypothèses du théorème dont bien vérifiées, et donc $v^{(k)} \to \bar{u}$ lorsque $n \to +\infty$, où $\bar{u}$ est tel que $A\bar{u} = R(\bar{u})$.

Exercice 105 page 152 (Point fixe amélioré)

1) La suite donnée par l'algorithme (2.14) est bien définie si pour tout $n \in \mathbb{N}$, $g' \circ \varphi(x_n) \neq 0$. Remarquons d'abord que $g' \circ \varphi(\overline{x}) \neq 0$. Or la fonction $g' \circ \varphi$ est continue ; pour $\varepsilon > 0$ fixé, il existe donc $\beta \in \mathbb{R}_+$ tel que $|g' \circ \varphi(x)| \geq \varepsilon$ pour tout $x \in [\overline{x} - \beta, \overline{x} + \beta] = I_\beta$. Remarquons ensuite que $h'(\overline{x}) = 1 - \frac{(g'(\bar{x}))^2}{(g'(\bar{x}))^2} = 0$. Or h' est aussi continue. On en déduit l'existence de $\gamma \in \mathbb{R}_+$ tel que $|h'(x)| < 1$ pour tout $x \in [\overline{x} - \gamma, \overline{x} + \gamma] = I_\gamma$.
Soit maintenant $\alpha = \min(\beta, \gamma)$; si $x_0 \in I_\alpha$, alors $g' \circ \varphi(x_0) \neq 0$. Comme h est strictement contractante sur I_α (et que $h(\overline{x}) = \overline{x}$), on en déduit que $x_1 \in I_\alpha$, et, par récurrence sur n, $x_n \in I_\alpha$ pour tout $n \in \mathbb{N}$ (et la suite est bien définie). De plus, comme h est strictement contractante sur I_α, le théorème du point fixe (théorème 2.5 page 141 donne la convergence de la suite $(x_n)_{n\in\mathbb{N}}$ vers $\bar{x}$.

2) Remarquons d'abord que si $\varphi \in C^2(\mathbb{R}, R)$, on peut directement appliquer la proposition 2.16 (item 2), car dans ce cas $h \in C^2(\mathbb{R}, \mathbb{R})$, puisqu'on a déjà vu que $h'(\overline{x}) = 0$. Effectuons maintenant le calcul dans le cas où l'on n'a que $\varphi \in C^1(\mathbb{R}, R)$. Calculons $|x_{k+1} - \overline{x}|$. Par définition de x_{k+1}, on a :

$$x_{k+1} - \overline{x} = x_k - \overline{x} - \frac{g(x_k)}{g'(\varphi(x_k))},$$

ce qui entraîne que

$$x_{n+1} - \overline{x} = (x_n - \overline{x}) \left(1 - \frac{g(x_n) - g(\overline{x})}{(x_n - \overline{x}) g'(\varphi(x_n))}\right). \tag{2.21}$$

Or il existe $\theta_n \in I(\overline{x}, x_n)$, où $I(\overline{x}, x_n)$ désigne l'intervalle d'extrémités $\overline{x}$ et x_n, tel que

$$\frac{g(x_n) - g(\overline{x})}{x_n - \overline{x}} = g'(\theta_n).$$

Mais comme $g \in C^3(\mathbb{R}, \mathbb{R})$ il existe $\zeta_n \in I(\theta_n, \varphi(x_n))$ tel que :

$$g'(\theta_n) = g'(\varphi(x_n)) + (\theta_n - \varphi(x_n))g''(\zeta_n).$$

On en déduit que

$$x_{n+1} - \overline{x} = (x_n - \overline{x})(\theta_n - \varphi(x_n))\frac{g''(\zeta_n)}{g'(\varphi(x_n))}. \tag{2.22}$$

Par inégalité triangulaire, on a :

$$|\theta_n - \varphi(x_n)| \leq |\theta_n - \overline{x}| + |\overline{x} - \varphi(x_n)| = |\theta_n - \overline{x}| + |\varphi(\overline{x}) - \varphi(x_n)|.$$

Comme $\theta_n \in I(\overline{x}, x_n)$, on a donc $|\theta_n - \overline{x}| \leq |x_n - \overline{x}|$; de plus : $|\varphi(\overline{x}) - \varphi(x_n)| \leq \sup_{x\in I_\alpha} |\varphi'(x)||x_n - \overline{x}|$. On en déduit que

$$|\theta_n - \varphi(x_n)| \leq |x_n - \overline{x}| \left(1 + \sup_{x\in I_\alpha} |\varphi'(x))|\right).$$

En reportant dans (2.22), on en déduit que :

$$|x_{n+1} - \overline{x}| \leq \frac{1}{\varepsilon}|x_n - \overline{x}|^2 \left(1 + \sup_{x\in I_\alpha} |\varphi'(x))|\right) \sup_{x\in I_\alpha} |g''(x)|,$$

où ε est donné à la question 1 par choix de α.
On a ainsi montré que la convergence de la suite $(x_n)_{n\in\mathbb{N}}$ définie par l'algorithme (2.14) est au moins quadratique.

3) Reprenons le calcul de la question précédente en montant en ordre sur les développements. Calculons $|x_{n+1} - \overline{x}|$. Ecrivons maintenant qu'il existe $\mu_n \in I(\overline{x}, x_n)$ tel que

$$g(x_n) = g(\overline{x}) + (x_n - \overline{x})g'(\overline{x}) + \frac{1}{2}(x_n - \overline{x})^2 g''(\mu_n).$$

De (2.21), on en déduit que

$$x_{n+1} - \overline{x} = (x_n - \overline{x}) \left(1 - (x_n - \overline{x})\frac{g'(\overline{x}) + \frac{1}{2}(x_n - \overline{x})g''(\mu_n)}{(x_n - \overline{x})g'(\varphi(x_n))}\right).$$

Or il existe $\nu_n \in I(\overline{x}, \varphi(x_n))$ tel que

$$g'(\varphi(x_n)) = g'(\overline{x}) + (\varphi(x_n) - \varphi(\overline{x}))g''(\nu_n).$$

On a donc :

$$x_{n+1} - \overline{x} = \frac{x_n - \overline{x}}{g'(\varphi(x_n))} \left((\varphi(x_n) - \varphi(\overline{x}))g''(\nu_n) - \frac{1}{2}(x_n - \overline{x})g''(\mu_n)\right).$$

Ecrivons maintenant que $\varphi(x_n) = \varphi(\overline{x}) + \varphi'(\xi_n)(x_n - \overline{x})$, où $\xi_n \in I(\overline{x}, x_n)$. Comme φ' est lipschitzienne, on a $\varphi'(\xi_n) = \varphi'(\overline{x}) + \epsilon_n = \frac{1}{2} + \epsilon_n$, avec $|\epsilon_n| \leq M|x_n - \overline{x}|$, où M est la constante de Lipschitz de φ'. On a donc :

$$x_{n+1} - \overline{x} = \frac{x_n - \overline{x}}{g'(\varphi(x_n))} \left((x_n - \overline{x})(\frac{1}{2} + \epsilon_n)g''(\nu_n) - \frac{1}{2}(x_n - \overline{x})g''(\mu_n)\right),$$

et donc (avec ε donné à la question 1 par choix de α) :

$$|x_{n+1} - \overline{x}| \leq \frac{1}{\varepsilon}|x_n - \overline{x}|^2 \left((\frac{1}{2}(g''(\nu_n) - g''(\mu_n)) + \epsilon_n g''(\nu_n)\right).$$

Mais de même, comme $g \in C^3(\mathbb{R}, \mathbb{R})$, et que μ_n et $\nu_n \in I(\overline{x}, x_n)$, on a

$$|g''(\mu_n) - g''(\nu_n)| \leq \sup_{x\in I_\alpha} |g'''(x)||x_n - \overline{x}|.$$

On en déduit finalement que :

$$|x_{n+1} - \overline{x}| \leq C|x_n - \overline{x}|^3, \text{ avec } C = \frac{1}{2\varepsilon} \sup_{x \in I_\alpha} |g'''(x)| + \frac{M}{\varepsilon} \sup_{x \in I_\alpha} |g''(x)|.$$

4) Pour montrer que la suite définie par l'algorithme (2.14) converge de manière cubique, il suffit de montrer que φ vérifie les hypothèses de la question 3). On a évidemment $\varphi(\bar{x}) = \bar{x}$. Comme $g \in C^3(\mathbb{R}, \mathbb{R})$ et que $g'(x) \neq 0, \quad \forall x \in I_\beta$, on en déduit que $\varphi \in C^2(\mathbb{R}, \mathbb{R})$. De plus

$$\varphi'(\overline{x}) = 1 - \frac{1}{2}\frac{g'(\overline{x})^2 - g''(\overline{x})g(\overline{x})}{g'(\overline{x})^2} = \frac{1}{2}.$$

La fonction φ vérifie donc bien les hypothèses de la question 3.

2.3 Méthode de Newton dans $\mathbb{R}^n$

2.3.1 Construction et convergence de la méthode

On a vu ci-dessus comment se construit la méthode de Newton à partir du point fixe de contraction en dimension $n = 1$. On va maintenant étudier cette méthode dans le cas n quelconque. Soient $g \in C^1(\mathbb{R}^n, \mathbb{R}^n)$ et $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ tels que $g(\bar{\boldsymbol{x}}) = 0$.
On généralise la méthode vue en 1D en remplaçant dans (2.10) la dérivée $g'(x^{(k)})$ par la matrice jacobienne de g au point $x^{(k)}$, qu'on note $Dg(\boldsymbol{x}^{(k)})$. La méthode s'écrit :

$$\left\{ \begin{array}{l} \boldsymbol{x}^{(0)} \in \mathbb{R}^n \\ Dg(\boldsymbol{x}^{(k)})(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}) = -g(\boldsymbol{x}^{(k)}), \ \forall k \geq 0. \end{array} \right. \tag{2.23}$$

Pour chaque $k \in \mathbb{N}$, il faut donc effectuer les opérations suivantes :

1. Calcul de $Dg(\boldsymbol{x}^{(k)})$,
2. Résolution du système linéaire $Dg(\boldsymbol{x}^{(k)})(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}) = -g(\boldsymbol{x}^{(k)})$.

Remarque 2.18. *Si la fonction g dont on cherche un zéro est linéaire, i.e. si g est définie par $g(\boldsymbol{x}) = A\boldsymbol{x} - b$ avec $A \in \mathcal{M}_n(\mathbb{R})$ et $b \in \mathbb{R}^n$, alors la méthode de Newton revient à résoudre le système linéaire $A\boldsymbol{x} = b$. En effet $Dg(\boldsymbol{x}^{(k)}) = A$ et donc* (2.23) *s'écrit $A\boldsymbol{x}^{(k+1)} = \boldsymbol{b}$.*

Pour assurer la convergence et la qualité de la méthode, on va chercher maintenant à répondre aux questions suivantes :

1. la suite $(\boldsymbol{x}^{(k)})_n$ est–elle bien définie ? A–t–on $Dg(\boldsymbol{x}^{(k)})$ inversible ?
2. A–t–on convergence $\boldsymbol{x}^{(k)} \to \bar{\boldsymbol{x}}$ quand $k \to +\infty$?
3. La convergence est-elle au moins quadratique ?

Théorème 2.19 (Convergence de la méthode de Newton, $g \in C^3$)**.** *Soient $g \in C^3(\mathbb{R}^n, \mathbb{R}^n)$ et $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ tels que $g(\bar{\boldsymbol{x}}) = 0$. On munit $\mathbb{R}^n$ d'une norme $\|\cdot\|$. On suppose que $Dg(\bar{\boldsymbol{x}})$ est inversible. Alors la méthode de Newton converge localement, et la convergence est au moins quadratique. Plus précisément, il existe $b > 0$, et $\beta > 0$ tels que*

1. *si $\boldsymbol{x}^{(0)} \in B(\bar{\boldsymbol{x}}, b) = \{\boldsymbol{x} \in \mathbb{R}^n, \|\boldsymbol{x} - \bar{\boldsymbol{x}}\| \leq b\}$ alors la suite $(\boldsymbol{x}^{(k)})_{k \in \mathbb{N}}$ est bien définie par* (2.23) *et $\boldsymbol{x}^{(k)} \in B(\bar{\boldsymbol{x}}, b)$ pour tout $n \in \mathbb{N}$,*
2. *si $\boldsymbol{x}^{(0)} \in B(\bar{\boldsymbol{x}}, b)$ et si la suite $(\boldsymbol{x}^{(k)})_{k \in \mathbb{N}}$ est définie par* (2.23) *alors $\boldsymbol{x}^{(k)} \to \bar{\boldsymbol{x}}$ quand $n \to +\infty$,*
3. *si $\boldsymbol{x}^{(0)} \in B(\bar{\boldsymbol{x}}, b)$ et si la suite $(\boldsymbol{x}^{(k)})_{k \in \mathbb{N}}$ est définie par* (2.23) *alors $\|\boldsymbol{x}^{(k+1)} - \bar{\boldsymbol{x}}\| \leq \beta \|\boldsymbol{x}^{(k)} - \bar{\boldsymbol{x}}\|^2$ $\forall k \in \mathbb{N}$.*

DÉMONSTRATION – Montrons d'abord que la suite converge si $\boldsymbol{x}^{(0)}$ est suffisamment proche de $\bar{\boldsymbol{x}}$. Pour cela on va utiliser le théorème du point fixe : Soit f la fonction définie sur un voisinage de $\bar{\boldsymbol{x}}$ (et à valeurs dans $\mathbb{R}^n$) par $\boldsymbol{x} \mapsto \boldsymbol{x} - (Dg(\boldsymbol{x}))^{-1} g(\boldsymbol{x})$. On a

$$Df(\bar{\boldsymbol{x}}) = \mathrm{Id} - (Dg(\bar{\boldsymbol{x}}))^{-1}(Dg(\bar{\boldsymbol{x}})) = 0.$$

Comme $g \in C^2(\mathbb{R}, \mathbb{R})$, la fonction f est de classe C^1 et donc par continuité de Df, il existe $b > 0$ tel que $\|Df(\boldsymbol{x})\| \leq \frac{1}{2}$ pour tout $\boldsymbol{x} \in B = B(\bar{\boldsymbol{x}}, b)$. Si on montre que $f(B) \subset B$, alors la fonction f est strictement contractante de B dans B, et donc par le théorème du point fixe, la suite définie par (2.23) converge. Soit $\boldsymbol{x} = \boldsymbol{x}^{(k)} \in B$, et soit $y = \boldsymbol{x}^{(k+1)} = f(\boldsymbol{x}^{(k)})$. Grâce au théorème des acroissements finis dans des espaces vectoriels normés [5], on a :

$$\|\boldsymbol{y} - \bar{\boldsymbol{x}}\| = \|f(\boldsymbol{x}) - f(\bar{\boldsymbol{x}})\| \leq \sup_{\boldsymbol{z} \in B} \|Df(\boldsymbol{z})\| \|\boldsymbol{x} - \bar{\boldsymbol{x}}\|, \tag{2.24}$$

et donc

$$\|\boldsymbol{y} - \bar{\boldsymbol{x}}\| \leq \frac{1}{2} \|\boldsymbol{x} - \bar{\boldsymbol{x}}\|.$$

On en déduit que $\boldsymbol{y} \in B$. La suite $(\boldsymbol{x}^{(k)})_{k \in \mathbb{N}}$ définie par (2.23) est donc bien convergente. Pour montrer le caractère quadratique de la convergence, on applique à nouveau l'inégalité des accroissements finis, cette fois-ci à $Df(\boldsymbol{z})$ dans (2.24). En effet, comme $Df \in C^1(\mathbb{R}^n, \mathbb{R}^n)$ (on utilise ici que g est de classe C^3), on a

$$\begin{aligned} \|Df(\boldsymbol{z})\| &= \|Df(\boldsymbol{z}) - Df(\bar{x})\| \\ &\leq \sup_{\xi \in B} \|D^2 f(\xi)\| \|\boldsymbol{z} - \bar{\boldsymbol{x}}\| \qquad (2.25) \\ &\leq \beta \|\boldsymbol{x} - \bar{\boldsymbol{x}}\|. \qquad (2.26) \end{aligned}$$

En reportant cette majoration de $\|Df(\boldsymbol{z})\|$ dans (2.24), on obtient alors (avec $\beta = \sup_{\xi \in B} \|D^2 f(\xi)\|$) :

$$\|y - \bar{\boldsymbol{x}}\| \leq \beta \|\boldsymbol{x} - \bar{\boldsymbol{x}}\|^2$$

ce qui donne la convergence locale au moins quadratique. ∎

La condition $g \in C^3(\mathbb{R}^n, \mathbb{R}^n)$ est une condition suffisante mais non nécessaire. Si $g \in C^1(\mathbb{R}^n, \mathbb{R}^n)$, on peut encore démontrer la convergence, mais sous des hypothèses pas très faciles à vérifier en pratique :

Théorème 2.20 (Convergence de la méthode de Newton, $g \in C^1$)**.**
Soient $g \in C^1(\mathbb{R}^n, \mathbb{R}^n)$ et $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ tels que $g(\bar{\boldsymbol{x}}) = 0$. On munit $\mathbb{R}^n$ d'une norme $\|\cdot\|$ et $\mathcal{M}_n(\mathbb{R})$ de la norme induite. On suppose que $Dg(\bar{\boldsymbol{x}})$ est inversible. On suppose de plus qu'il existe a, a_1, $a_2 \in \mathbb{R}_+^*$ tels que :

1. *si $\boldsymbol{x} \in B(\bar{\boldsymbol{x}}, a)$ alors $Dg(\boldsymbol{x})$ est inversible et $\|Dg(\boldsymbol{x}))^{-1}\| \leq a_1$;*
2. *si $\boldsymbol{x}, \boldsymbol{y} \in B(\bar{\boldsymbol{x}}, a)$ alors $\|g(\boldsymbol{y}) - g(\boldsymbol{x}) - Dg(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x})\| \leq a_2 \|\boldsymbol{y} - \boldsymbol{x}\|^2$.*

Alors, si on pose : $b = \min\left(a, \dfrac{1}{a_1 a_2}\right) > 0$, $\beta = a_1 a_2$ et si $\boldsymbol{x}^{(0)} \in B(\bar{x}, b)$, on a :

1. *$(\boldsymbol{x}^{(k)})_{k \in \mathbb{N}}$ est bien définie par* (2.23)*,*
2. *$\boldsymbol{x}^{(k)} \to \bar{\boldsymbol{x}}$ lorsque $n \to +\infty$,*
3. *$\|\boldsymbol{x}^{(k+1)} - \bar{\boldsymbol{x}}\| \leq \beta \|\boldsymbol{x}^{(k)} - \bar{\boldsymbol{x}}\|^2 \ \ \forall k \in \mathbb{N}$.*

DÉMONSTRATION – Soit $\boldsymbol{x}^{(0)} \in B(\bar{\boldsymbol{x}}, b) \subset B(\bar{\boldsymbol{x}}, a)$ où $b \leq a$. On va montrer par récurrence sur k que $\boldsymbol{x}^{(k)} \in B(\bar{\boldsymbol{x}}, b)$ $\forall k \in \mathbb{N}$ (et que $(\boldsymbol{x}^{(k)})_{k \in \mathbb{N}}$ est bien définie). L'hypothèse de récurrence est que $\boldsymbol{x}^{(k)}$ est bien défini, et que $\boldsymbol{x}^{(k)} \in B(\bar{\boldsymbol{x}}, b)$. On veut montrer que $\boldsymbol{x}^{(k+1)}$ est bien défini et $\boldsymbol{x}^{(k+1)} \in B(\bar{\boldsymbol{x}}, b)$. Comme $b \leq a$, la matrice $Dg(\boldsymbol{x}^{(k)})$ est inversible et $\boldsymbol{x}^{(k+1)}$ est donc bien défini ; on a :

$$\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)} = Dg(\boldsymbol{x}^{(k)})^{-1}(-g(\boldsymbol{x}^{(k)}))$$

5. **Théorème des accroissements finis :** Soient $(E, \|\cdot\|_E)$ et $(F, \|\cdot\|_F)$ des espaces vectoriels normés, soient $h \in C^1(E, F)$ et $(\boldsymbol{x}, y) \in E^2$. On définit $]x, y[= \{tx + (1-t)y, t \in]0, 1[\}$. Alors : $\|h(y) - h(x)\| \leq \|y - x\| \sup_{z \in]x, y[} \|Dh(z)\|_{\mathcal{L}(E,F)}$.
(On rappelle que si $T \in \mathcal{L}(E, F)$ alors $\|T\|_{[\mathcal{L}(E,F)} = \sup_{x \in E, \|x\|_E = 1} \|Tx\|_F|$.)
Attention piège ! ! : Si $\dim F > 1$, on ne peut pas dire, comme c'est le cas en dimension 1, que :$\exists \xi \in]x, y[$ t.q. $h(y) - h(x) = Dh(\xi)(y - x)$.

Pour montrer que $\boldsymbol{x}^{(k+1)} \in B(\bar{\boldsymbol{x}}, b)$ on va utiliser le fait que $b \leq \dfrac{1}{a_1 a_2}$. Par hypothèse, on sait que si $\boldsymbol{x}, \boldsymbol{y} \in B(\bar{\boldsymbol{x}}, a)$, on a

$$\|g(\boldsymbol{y}) - g(\boldsymbol{x}) - Dg(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x})\| \leq a_2 \|\boldsymbol{y} - \boldsymbol{x}\|^2.$$

Prenons $y = \bar{\boldsymbol{x}}$ et $\boldsymbol{x} = \boldsymbol{x}^{(k)} \in B(\bar{\boldsymbol{x}}, a)$ dans l'inégalité ci-dessus. On obtient alors :

$$\|g(\bar{\boldsymbol{x}}) - g(\boldsymbol{x}^{(k)}) - Dg(\boldsymbol{x}^{(k)})(\bar{\boldsymbol{x}} - \boldsymbol{x}^{(k)})\| \leq a_2 \|\bar{\boldsymbol{x}} - \boldsymbol{x}^{(k)}\|^2.$$

Comme $g(\bar{\boldsymbol{x}}) = 0$ et par définition de $\boldsymbol{x}^{(k+1)}$, on a donc :

$$\|Dg(\boldsymbol{x}^{(k)})(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}) - Dg(\boldsymbol{x}^{(k)})(\bar{\boldsymbol{x}} - \boldsymbol{x}^{(k)})\| \leq a_2 \|\bar{\boldsymbol{x}} - \boldsymbol{x}^{(k)}\|^2,$$

et donc

$$\|Dg(\boldsymbol{x}^{(k)})(\boldsymbol{x}^{(k+1)} - \bar{\boldsymbol{x}})\| \leq a_2 \|\bar{\boldsymbol{x}} - \boldsymbol{x}^{(k)}\|^2. \tag{2.27}$$

Or $\boldsymbol{x}^{(k+1)} - \bar{\boldsymbol{x}} = [Dg(\boldsymbol{x}^{(k)})]^{-1}(Dg(\boldsymbol{x}^{(k)}))(\boldsymbol{x}^{(k+1)} - \bar{\boldsymbol{x}})$, et donc

$$\|\boldsymbol{x}^{(k+1)} - \bar{\boldsymbol{x}}\| \leq \|Dg(\boldsymbol{x}^{(k)})^{-1}\| \ \|Dg(\boldsymbol{x}^{(k)})(\boldsymbol{x}^{(k+1)} - \bar{\boldsymbol{x}})\|.$$

En utilisant (2.27), les hypothèses 1 et 2 et le fait que $\boldsymbol{x}^{(k)} \in B(\bar{\boldsymbol{x}}, b)$, on a donc

$$\|\boldsymbol{x}^{(k+1)} - \bar{\boldsymbol{x}}\| \leq a_1 a_2 \|\boldsymbol{x}^{(k)} - \bar{\boldsymbol{x}}\|^2 < a_1 a_2 b^2. \tag{2.28}$$

Or $a_1 a_2 b^2 < b$ car $b \leq \dfrac{1}{a_1 a_2}$. Donc $\boldsymbol{x}^{(k+1)} \in B(\bar{\boldsymbol{x}}, b)$.

On a ainsi montré par récurrence que la suite $(\boldsymbol{x}^{(k)})_{k \in \mathbb{N}}$ est bien définie et que $\boldsymbol{x}^{(k)} \in B(\bar{\boldsymbol{x}}, b)$ pour tout $k \geq 0$.

Pour montrer la convergence de la suite $(\boldsymbol{x}^{(k)})_{k \in \mathbb{N}}$ vers $\bar{\boldsymbol{x}}$, on repart de l'inégalité (2.28) :

$$a_1 a_2 \|\boldsymbol{x}^{(k+1)} - \bar{\boldsymbol{x}}\| \leq (a_1 a_2)^2 \|\bar{\boldsymbol{x}} - \boldsymbol{x}^{(k)}\|^2 = (a_1 a_2 \|\boldsymbol{x}^{(k)} - \bar{\boldsymbol{x}}\|)^2, \ \forall k \in \mathbb{N},$$

et donc par récurrence

$$a_1 a_2 \|\boldsymbol{x}^{(k)} - \bar{\boldsymbol{x}}\| \leq (a_1 a_2 \|\boldsymbol{x}^{(0)} - \bar{\boldsymbol{x}}\|)^{2^k}, \quad \forall k \in \mathbb{N}$$

Comme $\boldsymbol{x}^{(0)} \in B(\bar{\boldsymbol{x}}, b)$ et $b \leq \frac{1}{a_1 a_2}$, on a $a_1 a_2 \|\boldsymbol{x}^{(0)} - \bar{\boldsymbol{x}}\| < 1$ et donc $\|\boldsymbol{x}^{(k)} - \bar{\boldsymbol{x}}\| \to 0$ quand $k \to +\infty$.

La convergence est au moins quadratique car l'inégalité (2.28) s'écrit :

$$\|\boldsymbol{x}^{(k+1)} - \bar{\boldsymbol{x}}\| \leq \beta \|\boldsymbol{x}^{(k)} - \boldsymbol{x}\|^2 \text{ avec } \beta = a_1 a_2.$$

■

Le théorème 2.19 peut aussi se démontrer comme corollaire du théorème 2.20. En effet, sous les hypothèses du théorème 2.19 (il est même suffisant de supposer g de classe C^2 au lieu de C^3), on peut démontrer qu'il existe $a, a_1, a_2 \in \mathbb{R}_+^*$ tels que

1. si $\boldsymbol{x} \in B(\bar{\boldsymbol{x}}, a)$ alors $Dg(\boldsymbol{x})$ est inversible et $\|(Dg(\boldsymbol{x}))^{-1}\| \leq a_1$,
2. si $\boldsymbol{x}, \boldsymbol{y} \in B(\bar{\boldsymbol{x}}, a)$ alors $\|g(\boldsymbol{y}) - g(\boldsymbol{x}) - Dg(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x})\| \leq a_2 \|\boldsymbol{y} - \boldsymbol{x}\|^2$.

et donc appliquer le théorème 2.20, voir exercice 125 page 169.

Remarque 2.21 (Choix de l'itéré initial)**.** *On ne sait pas bien estimer b dans le théorème 2.19, et ceci peut poser problème lors de l'implantation numérique : il faut choisir l'itéré initial $\boldsymbol{x}^{(0)}$ "suffisamment proche" de $\overline{\boldsymbol{x}}$ pour avoir convergence.*

2.3.2 Variantes de la méthode de Newton

L'avantage majeur de la méthode de Newton par rapport à une méthode de point fixe par exemple est sa vitesse de convergence d'ordre 2. On peut d'ailleurs remarquer que lorsque la méthode ne converge pas, par exemple si l'itéré initial $\boldsymbol{x}^{(0)}$ n'a pas été choisi "suffisamment proche" de $\overline{\boldsymbol{x}}$, alors la méthode diverge très vite...

L'inconvénient majeur de la méthode de Newton est son coût : on doit d'une part calculer la matrice jacobienne $Dg(\boldsymbol{x}^{(k)})$ à chaque itération, et d'autre part la factoriser pour résoudre le système linéaire $Dg(\boldsymbol{x}^{(k)})(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}) = -g(\boldsymbol{x}^{(k)})$. (On rappelle que pour résoudre un système linéaire, il ne faut pas calculer l'inverse de la matrice, mais plutôt la factoriser sous la forme LU par exemple, et on calcule ensuite les solutions des systèmes avec matrice triangulaires faciles à inverser, voir Chapitre 1.) Plusieurs variantes ont été proposées pour tenter de réduire ce coût.

"Faux quasi Newton"

Soient $g \in C^1(\mathbb{R}^n, \mathbb{R}^n)$ et $\bar{\boldsymbol{x}} \in \mathbb{R}$ tels que $g(\bar{\boldsymbol{x}}) = 0$. On cherche à calculer $\bar{\boldsymbol{x}}$. Si on le fait par la méthode de Newton, l'algorithme s'écrit :

$$\left\{ \begin{array}{l} \boldsymbol{x}^{(0)} \in \mathbb{R}^n, \\ Dg(\boldsymbol{x}^{(k)})(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}) = -g(\boldsymbol{x}^{(k)}), \ n \geq 0. \end{array} \right.$$

La méthode du "Faux quasi-Newton" (parfois appelée quasi-Newton) consiste à remplacer le calcul de la matrice jacobienne $Dg(\boldsymbol{x}^{(k)})$ à chaque itération par un calcul toutes les "quelques" itérations. On se donne une suite $(n_i)_{i\in\mathbb{N}}$, avec $n_0 = 0$ et $n_{i+1} > n_i \ \forall i \in \mathbb{N}$, et on calcule la suite $(\boldsymbol{x}^{(k)})_{n\in\mathbb{N}}$ de la manière suivante :

$$\left\{ \begin{array}{l} \boldsymbol{x}^{(0)} \in \mathbb{R}^n \\ Dg(\boldsymbol{x}^{(n_i)})(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}) = -g(\boldsymbol{x}^{(k)}) \text{ si } n_i \leq k < n_{i+1}. \end{array} \right. \tag{2.29}$$

Avec cette méthode, on a moins de calculs et de factorisations de la matrice jacobienne $Dg(\boldsymbol{x})$ à effectuer, mais on perd malheureusement la convergence quadratique : cette méthode n'est donc pas très utilisée en pratique.

Newton incomplet

On suppose que g s'écrit sous la forme :

$$g(\boldsymbol{x}) = A\boldsymbol{x} + F_1(\boldsymbol{x}) + F_2(\boldsymbol{x}), \text{ avec } A \in \mathcal{M}_n(\mathbb{R}) \text{ avec } F_1, F_2 \in C^1(\mathbb{R}^n, \mathbb{R}^n).$$

L'algorithme de Newton (2.23) s'écrit alors :

$$\left\{ \begin{array}{l} \boldsymbol{x}^{(0)} \in \mathbb{R}^n \\ \big(A + DF_1(\boldsymbol{x}^{(k)}) + DF_2(\boldsymbol{x}^{(k)})\big)\,(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}) = \\ \qquad\qquad -A\boldsymbol{x}^{(k)} - F_1(\boldsymbol{x}^{(k)}) - F_2(\boldsymbol{x}^{(k)}). \end{array} \right.$$

La méthode de Newton incomplet consiste à ne pas tenir compte de la jacobienne de F_2.

$$\left\{ \begin{array}{l} \boldsymbol{x}^{(0)} \in \mathbb{R}^n \\ (A + DF_1(\boldsymbol{x}^{(k)}))(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}) = -A\boldsymbol{x}^{(k)} - F_1(\boldsymbol{x}^{(k)}) - F_2(\boldsymbol{x}^{(k)}). \end{array} \right. \tag{2.30}$$

On dit qu'on fait du "Newton sur F_1" et du "point fixe sur F_2". Les avantages de cette procédure sont les suivants :

- La méthode ne nécessite pas le calcul de $DF_2(\boldsymbol{x})$, donc on peut l'employer si $F_2 \in C(\mathbb{R}^n, \mathbb{R}^n))$ n'est pas dérivable.
- On peut choisir F_1 et F_2 de manière à ce que la structure de la matrice $A + DF_1(\boldsymbol{x}^{(k)})$ soit "meilleure" que celle de la matrice $A + DF_1(\boldsymbol{x}^{(k)}) + DF_2(\boldsymbol{x}^{(k)})$; si par exemple A est la matrice issue de la discrétisation du Laplacien, c'est une matrice creuse. On peut vouloir conserver cette structure et choisir F_1 et F_2 de manière à ce que la matrice $A + DF_1(\boldsymbol{x}^{(k)})$ ait la même structure que A.
- Dans certains problèmes, on connaît a priori les couplages plus ou moins forts dans les non-linéarités : un couplage est dit fort si la variation d'une variable entraîne une variation forte du terme qui en dépend. Donnons un exemple : Soit f de $\mathbb{R}^2$ dans $\mathbb{R}^2$ définie par $f(x, y) = (x + \sin(10^{-5}y), \exp(x) + y)$, et considérons le système non linéaire $f(x, y) = (a, b)$ où $(a, b) \in \mathbb{R}^2$ est donné. Il est naturel de penser que pour ce système, le terme de couplage de la première équation en la variable y sera faible, alors que le couplage de deuxième équation en la variable x sera fort.

  On a alors intérêt à mettre en oeuvre la méthode de Newton sur la partie "couplage fort" et une méthode de point fixe sur la partie "couplage faible".

L'inconvénient majeur est la perte de la convergence quadratique. La méthode de Newton incomplet est cependant assez souvent employée en pratique en raison des avantages énumérés ci-dessus.

Remarque 2.22. *Si $F_2 = 0$, alors la méthode de Newton incomplet est exactement la méthode de Newton. Si $F_1 = 0$, la méthode de Newton incomplet s'écrit*

$$A(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}) = -A\boldsymbol{x}^{(k)} - F_2(\boldsymbol{x}^{(k)}),$$

En supposant A inversible, on a alors $\boldsymbol{x}^{(k+1)} = -A^{-1}F_2(\boldsymbol{x}^{(k)})$. C'est donc dans ce cas la méthode du point fixe sur la fonction $-A^{-1}F_2$.

Méthode de la sécante

La méthode de la sécante est une variante de la méthode de Newton dans le cas de la dimension 1 d'espace. On suppose ici $n = 1$ et $g \in C^1(\mathbb{R}, \mathbb{R})$. La méthode de Newton pour calculer $\overline{\boldsymbol{x}} \in \mathbb{R}$ tel que $g(\overline{\boldsymbol{x}}) = 0$ s'écrit :

$$\begin{cases} x^{(0)} \in \mathbb{R} \\ g'(x^{(k)})(x^{(k+1)} - x^{(k)}) = -g(x^{(k)}), \quad \forall n \geq 0. \end{cases}$$

On aimerait simplifier le calcul de $g'(x^{(k)})$, c'est–à–dire remplacer $g'(x^{(k)})$ par une quantité "proche" sans calculer g'. Pour cela, on remplace la dérivée par un quotient différentiel. On obtient la méthode de la sécante :

$$\begin{cases} x^{(0)}, x^{(1)} \in \mathbb{R} \\ \dfrac{g(x^{(k)}) - g(x^{(k-1)})}{x^{(k)} - x^{(k-1)}}(x^{(k+1)} - x^{(k)}) = -g(x^{(k)}) \quad n \geq 1. \end{cases} \tag{2.31}$$

Remarquons que dans la méthode de la sécante, $x^{(k+1)}$ dépend de $x^{(k)}$ et de $x^{(k-1)}$: on a une méthode à deux pas ; on a d'ailleurs besoin de deux itérés initiaux $x^{(0)}$ et $x^{(1)}$. L'avantage de cette méthode est qu'elle ne nécessite pas le calcul de g'. L'inconvénient est qu'on perd la convergence quadratique. On peut toutefois montrer (voir exercice 131 page 171) que si $g(\bar{x}) = 0$ et $g'(\bar{x}) \neq 0$, il existe $\alpha > 0$ tel que si $x^{(0)}, x^{(1)} \in [\bar{x} - \alpha.\bar{x} + \alpha] = I_\alpha$, $x^{(0)} \neq x^{(1)}$, la suite $(x^{(k)})_{n\in\mathbb{N}}$ construite par la méthode de la sécante (2.31) est bien définie, que $(x^{(k)})_{n\in\mathbb{N}} \subset I_\alpha$ et que $x^{(k)} \to \bar{x}$ quand $n \to +\infty$. De plus, la convergence est super linéaire, *i.e.* si $x^{(k)} \neq \bar{x}$ pour tout $n \in \mathbb{N}$, alors $\dfrac{x^{(k+1)} - \bar{x}}{x^{(k)} - \bar{x}} \to 0$ quand $n \to +\infty$. On peut même montrer (voir exercice 131 page 171) que la méthode de la sécante est convergente d'ordre d, où d est le nombre d'or.

Méthodes de type "Quasi Newton"

On veut généraliser la méthode de la sécante au cas $n > 1$. Soient donc $g \in C^1(\mathbb{R}^n, \mathbb{R}^n)$. Pour éviter de calculer $Dg(x^{(k)})$ dans la méthode de Newton (2.23), on va remplacer $Dg(x^{(k)})$ par $B^{(k)} \in \mathcal{M}_n(\mathbb{R})$ "proche de $Dg(x^{(k)})$". En s'inspirant de la méthode de la sécante en dimension 1, on cherche une matrice $B^{(k)}$ qui, $x^{(k)}$ et $x^{(k-1)}$ étant connus (et différents), vérifie la condition :

$$B^{(k)}(x^{(k)} - x^{(k-1)}) = g(x^{(k)}) - g(x^{(k-1)}) \tag{2.32}$$

Dans le cas où $n = 1$, cette condition détermine entièrement $B^{(k)}$;car on peut écrire : $B^{(k)} = \dfrac{g(x^{(k)}) - g(x^{(k-1)}}{x^{(k)} - x^{(k-1)}}$.
Si $n > 1$, la condition (2.32) ne permet pas de déterminer complètement $B^{(k)}$. Il y a plusieurs façons possibles de choisir $B^{(k)}$, nous en verrons en particulier dans le cadre des méthodes d'optimisation (voir chapitre 4, dans ce cas la fonction g est un gradient), nous donnons ici la méthode de Broyden[6]. Celle-ci consiste à choisir $B^{(k)}$ de la manière suivante : à $x^{(k)}$ et $x^{(k-1)}$ connus, on pose $\delta^{(k)} = x^{(k)} - x^{(k-1)}$ et $y^{(k)} = g(x^{(k)}) - g(x^{(k-1)})$; on suppose $B^{(k-1)} \in \mathcal{M}_n(\mathbb{R})$ connue (et $\delta^{(k)} \neq 0$), et on cherche $B^{(k)} \in \mathcal{M}_n(\mathbb{R})$ telle que

$$B^{(k)}\delta^{(k)} = y^{(k)} \tag{2.33}$$

(c'est la condition (2.32), qui ne suffit pas à déterminer $B^{(k)}$ de manière unique) et qui vérifie également :

$$B^{(k)}\xi = B^{(k-1)}\xi, \quad \forall \xi \in \mathbb{R}^n \text{ tel que } \xi \perp \delta^{(k)}. \tag{2.34}$$

6. C. G. Broyden, "A Class of Methods for Solving Nonlinear Simultaneous Equations." *Math. Comput.* 19, 577-593, 1965.

Proposition 2.23 (Existence et unicité de la matrice de Broyden)**.**
Soient $y^{(k)} \in \mathbb{R}^n$, $\delta^{(k)} \in \mathbb{R}^n$, $\delta^{(k)} \neq 0$, et $B^{(k-1)} \in \mathcal{M}_n(\mathbb{R})$. Il existe une unique matrice $B^{(k)} \in \mathcal{M}_n(\mathbb{R})$ vérifiant (2.33) *et* (2.34)*; la matrice $B^{(k)}$ s'exprime en fonction de $y^{(k)}$, $\delta^{(k)}$ et $B^{(k-1)}$ de la manière suivante :*

$$B^{(k)} = B^{(k-1)} + \frac{y^{(k)} - B^{(k-1)}\delta^{(k)}}{\delta^{(k)} \cdot \delta^{(k)}}(\delta^{(k)})^t. \tag{2.35}$$

DÉMONSTRATION – L'espace des vecteurs orthogonaux à $\delta^{(k)}$ est de dimension $n-1$. Soit $(\gamma_1, \ldots, \gamma_{n-1})$ une base de cet espace, alors $(\gamma_1, \ldots, \gamma_{n-1}, \delta^{(k)})$ est une base de $\mathbb{R}^n$ et si $B^{(k)}$ vérifie (2.33) et (2.34), les valeurs prises par l'application linéaire associée à $B^{(k)}$ sur chaque vecteur de base sont connues, ce qui détermine l'application linéaire et donc la matrice $B^{(k)}$ de manière unique. Soit $B^{(k)}$ définie par (2.35), on a :

$$B^{(k)}\delta^{(k)} = B^{(k-1)}\delta^{(k)} + \frac{y^{(k)} - B^{(k-1)}\delta^{(k)}}{\delta^{(k)} \cdot \delta^{(k)}}(\delta^{(k)})^t\delta^{(k)} = y^{(k)},$$

et donc $B^{(k)}$ vérifie (2.33). Soit $\xi \in \mathbb{R}^n$ tel que $\xi \perp \delta^{(k)}$, alors $\xi \cdot \delta^{(k)} = (\delta^{(k)})^t\xi = 0$ et donc

$$B^{(k)}\xi = B^{(k-1)}\xi + \frac{(y^{(k)} - B^{(k-1)}\delta^{(k)})}{\delta^{(k)} \cdot \delta^{(k)}}(\delta^{(k)})^t\xi = B^{(k-1)}\xi, \ \forall \xi \perp \delta^{(k)}.$$

■

L'algorithme de Broyden s'écrit donc :

$$\begin{cases} \text{Initialisation : } x^{(0)},\ x^{(1)} \in \mathbb{R}^n,\ x^{(0)}, \neq x^{(1)},\ B^{(0)} \in \mathcal{M}_n(\mathbb{R}) \\ \text{Itération k : } x^{(k)},\ x^{(k-1)} \text{ et } B^{(k-1)} \text{ connus, on pose} \\ \qquad\qquad \delta^{(k)} = x^{(k)} - x^{(k-1)} \text{ et } y^{(k)} = g(x^{(k)}) - g(x^{(k-1)}); \\ \text{Calcul de } B^{(k)} = B^{(k-1)} + \frac{y^{(k)} - B^{(k-1)}\delta^{(k)}}{\delta^{(k)} \cdot \delta^{(k)}}(\delta^{(k)})^t, \\ \text{résolution de } B^{(k)}(x^{(k+1)} - x^{(k)}) = -g(x^{(k)}). \end{cases}$$

Une fois de plus, l'avantage de cette méthode est de ne pas nécessiter le calcul de $Dg(x)$, mais l'inconvénient est la perte du caractère quadratique de la convergence .

2.3.3 Exercices (méthode de Newton)

Exercice 107 (Newton et logarithme)**.** *Suggestions en page 172*
Soit f la fonction de $\mathbb{R}_+^*$ dans $\mathbb{R}$ définie par $f(x) = \ln(x)$. Montrer que la méthode de Newton pour la recherche de $\overline{x}$ tel que $f(\overline{x}) = 0$ converge si et seulement si le choix initial $x^{(0)}$ est tel que $x^{(0)} \in]0, e[$.

Exercice 108 (Newton pour un système linéaire)**.** *Corrigé en page 174* Soit f l'application définie sur $\mathbb{R}^n$ par $f(x) = Ax - b$ où A est une matrice inversible et $b \in \mathbb{R}^n$. Ecrire l'algorithme de Newton pour la résolution de l'équation $f(x) = 0$ et montrer qu'il converge pour toute condition initiale $x^0 \in \mathbb{R}^n$.

Exercice 109 (Condition initiale et Newton)**.** *Corrigé en page 174* L'algorithme de Newton pour $F(x, y) = (\sin(x) + y, xy)^t$ est-il bien défini pour la condition initiale $(\frac{\pi}{2}, 0)$?

Exercice 110 (Newton dans $\mathbb{R}$ et $\mathbb{R}^2$)**.** Soit $a \in \mathbb{R}$ tel que $|a| < 1$ et $(x_0,\ y_0) \in \mathbb{R}^2$. On définit l'application

$$\begin{aligned} F : \mathbb{R}^{\ 2} &\to \mathbb{R}^2 \\ \begin{bmatrix} x \\ y \end{bmatrix} &\mapsto \begin{bmatrix} x - x_0 - ay \\ y - y_0 - a\sin x \end{bmatrix} \end{aligned}$$

.

1. Montrer qu'il existe une fonction $f : \mathbb{R} \to \mathbb{R}$, que l'on déterminera, telle que $F(x, y) = (0, 0)$ si et seulement si $x = x_0 + ay$ et $f(y) = 0$.
2. Montrer que pour tout triplet (a, x_0, y_0), il existe un unique couple $(\bar{x}, \bar{y}) \in \mathbb{R}^2$ tel que $F(\bar{x}, \bar{y}) = (0, 0)$.
3. Ecrire l'algorithme de Newton pour f et montrer que l'algorithme de Newton converge au voisinage de $\bar{y}$.

4. Ecrire l'algorithme de Newton pour la fonction F. Montrer que l'algorithme converge au voisinage de $(\bar{x}, \bar{y})$.

Exercice 111 (Méthode de Newton pour un système 2×2)**.**

1. Ecrire la méthode de Newton pour la résolution du système suivant :

$$-5x + 2\sin x + 2\cos y = 0, \tag{2.36}$$
$$2\cos x + 2\sin y - 5y = 0. \tag{2.37}$$

et montrer que la suite définie par cet algorithme est toujours bien définie.

2. Soit $(\overline{x}, \overline{y})$ une solution du problème (2.36)-(2.37). Montrer qu'il existe $\varepsilon > 0$ tel que si (x_0, y_0) est dans la boule B_ε de centre $(\overline{x}, \overline{y})$ et de rayon ε, alors la suite $(x_n, y_n)_{n\in\mathbb{N}}$ construite par la méthode de Newton converge vers $(\overline{x}, \overline{y})$ lorsque n tends vers $+\infty$.

3. Montrer qu'il existe au moins une solution $(\overline{x}, \overline{y})$ au problème (2.36)-(2.37).

Exercice 112 (Méthode de Newton pour un autre système 2×2)**.**

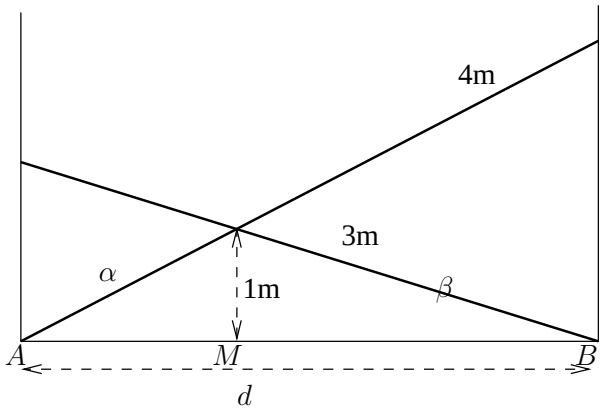
On considére le système non linéaire à deux inconnues suivant :

$$x^2 + 2xy = 0, \tag{2.38}$$
$$xy + 1 = 0, \tag{2.39}$$

1. Calculer les solutions du système (2.38)-(2.39).
2. Écrire l'algorithme de Newton pour la résolution du système (2.38)-(2.39) et donner les conditions sous lesquelles la suite définie par l'algorithme de Newton est bien définie.
3. Calculer les premiers itérés (x_1, y_1) construits par la méthode de Newton en partant de $(x_0, y_0) = (1, -1)$.

Exercice 113 (Newton et les échelles...)**.** *Corrigé en page 2.3.3 page 175*

Soient deux échelles de longueurs respectives 3 et 4 m, posées contre deux murs verticaux selon la figure ci-contre. On sait que les échelles se croisent à 1m du sol, et on cherche à connaître la distance d entre les deux murs.



1. Montrer que le problème revient à déterminer x et y tels que

$$16x^2 = (x^2 + 1)(x + y)^2 \tag{2.40}$$
$$9y^2 = (y^2 + 1)(x + y)^2. \tag{2.41}$$

2. Ecrire l'algorithme de Newton pour la résolution du système (2.40)-(2.41).

3. Calculer les premiers itérés $x^{(1)}$ et $y^{(1)}$ construits par la méthode de Newton en partant de $x^{(0)} = 1$ et $y^{(0)} = 1$.

Exercice 114 (Newton dans $\mathcal{M}_2(\mathbb{R})$)**.**

On considère l'application $f : \mathcal{M}_2(\mathbb{R}) \to \mathcal{M}_2(\mathbb{R})$ définie par $f(X) = X^2 - \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. L'objectif de cet exercice est de trouver les solutions de $f(X) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$.

1. Réécrire l'application f comme une application F de $\mathbb{R}^4$ dans $\mathbb{R}^4$.
2. Trouver l'ensemble des solutions de $f(X) = 0$.
3. Ecrire le premier itéré X_1 de l'algorithme de Newton pour l'application f partant de la donnée initiale $X_0 = \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}$ (On pourra passer par l'application F). Montrer que la suite $(X_k)_k$ définie par cet algorithme est définie par tout k et que l'on peut écrire sous la forme $X_k = \lambda_k \mathrm{Id}$ où $(\lambda_k)_k$ est une suite réelle dont on étudiera la convergence.
4. L'algorithme de Newton converge-t-il au voisinage de $X_* = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$?

Exercice 115 (Recherche d'un point fixe)**.** *Corrigé détaillé en page 175*
On définit la fonction f de $\mathbb{R}$ dans $\mathbb{R}$ par $f(x) = e^{(x^2)} - 4x^2$.

1. Montrer que f s'annule en 4 points de $\mathbb{R}$ et qu'un seul de ces points est entre 0 et 1.
2. On pose $g(x) = (1/2)\sqrt{e^{(x^2)}}$ (pour x dans $\mathbb{R}$).
   Montrer que la méthode du point fixe appliquée à g, initialisée avec un point de l'intervalle $]0, 1[$, est convergente et converge vers le point de $]0, 1[$ annulant f.
   Quel est l'ordre de convergence de cette méthode ?
3. Donner la méthode de Newton pour rechercher les points annulant f.
   Entre cette méthode et la méthode de la question précédente, quelle méthode vous semble *a priori* la plus efficace ?

Exercice 116 (Nombre d'itérations fini pour Newton)**.** *Corrigé détaillé en page 176*

1. Soit f la fonction de $\mathbb{R}$ dans $\mathbb{R}$ définie par : $f(x) = e^x - 1$. Pour $x^{(0)} \in \mathbb{R}$, on note $(x^{(k)})_{n\in\mathbb{N}}$ la suite des itérés construits par la méthode de Newton pour la recherche d'un point où f s'annule.

1.1 Montrer que pour tout $x^{(0)} \in \mathbb{R}$, la suite $(x^{(k)})_{n\in\mathbb{N}}$ est bien définie.

1.2 Montrer que si $x^{(0)} \neq 0$, alors $x^{(k+1)} \neq x^{(k)}$ pour tout $n \in \mathbb{N}$. En déduire que la méthode de Newton converge en un nombre fini d'opérations si et seulement si $f(x^{(0)}) = 0$.

1.3 Montrer que :
1.3 (a) si $x^{(0)} < 0$ alors $x^{(1)} > 0$.
1.3 (b) si $x^{(0)} > 0$ alors $0 < x^{(1)} < x^{(0)}$.

1.4 Montrer que la suite $(x^{(k)})_{n\in\mathbb{N}}$ converge lorsque n tend vers l'infini et donner sa limite.

2. Soit $F : \mathbb{R}^n \to \mathbb{R}$ une fonction continûment différentiable et strictement convexe $(n \geq 1)$ et dont la différentielle ne s'annule pas. Soit $x^{(0)} \in \mathbb{R}^n$ le choix initial (ou itéré 0) dans la méthode de Newton.
Montrer que la méthode de Newton converge en un nombre fini d'opérations si et seulement si $F(x^{(0)}) = 0$.

Exercice 117 (Méthode de Newton pour un système 2×2)**.** *Corrigé en page 176*

1. On considère l'application $f : \mathbb{R} \to \mathbb{R}$ définie par $f(x) = x^2$. Ecrire la méthode de Newton pour calculer la solution de $f(x) = 0$ et montrer qu'elle converge quel que soit le choix initial $x_0 \in \mathbb{R}$.
2. On considère maintenant l'application $F : \mathbb{R}^2 \to \mathbb{R}^2$ définie par

$$F(x, y) = \begin{bmatrix} x^2 - y \\ y^2 \end{bmatrix}$$

   (a) Déterminer l'ensemble des solutions de $F(x, y) = (0, 0)$.

(b) Ecrire l'algorithme de Newton pour la résolution, et montrer que l'algorithme est bien défini pour tous les couples (x_0, y_0) tels que $x_0 \neq 0$ et $y_0 > 0$.

(c) Soit $(x_0, y_0) = (1, 1)$. On note $(x_k, y_k), k \in \mathbb{N}$ les itérés de Newton.

i. Expliciter y_k et en déduire que la suite $(y_k)_{k\in\mathbb{N}}$ converge.

ii. Montrer que $x_k \geq 2^{-\frac{k}{2}}$ pour tout $k \in \mathbb{N}$ et en déduire que $x_{k+1} \leq x_k$ pour tout $k \in \mathbb{N}$.

iii. En déduire que la méthode de Newton converge vers une solution de $F(x, y) = (0, 0)$.

Exercice 118 (Méthode de Newton pour le calcul de l'inverse)**.** *Corrigé en page 177*

1. Soit $a > 0$. On cherche à calculer $\dfrac{1}{a}$ par l'algorithme de Newton.

(a) Montrer que l'algorithme de Newton appliqué à une fonction g (dont $\frac{1}{a}$ est un zéro) bien choisie s'écrit :

$$\begin{cases} x^{(0)} \text{ donné,} \\ x^{(k+1)} = x^{(k)}(2 - ax^{(k)}). \end{cases} \tag{2.42}$$

(b) Montrer que la suite $(x^{(k)})_{n\in\mathbb{N}}$ définie par (2.42) vérifie

$$\lim_{n\to+\infty} x^{(k)} = \begin{cases} \dfrac{1}{a} \text{ si } x^{(0)} \in]0, \dfrac{2}{a}[, \\ -\infty \text{ si } x^{(0)} \in]-\infty, 0[\cup]\dfrac{2}{a}, +\infty[\end{cases}$$

2. On cherche maintenant à calculer l'inverse d'une matrice par la méthode de Newton. Soit donc A une matrice carrée d'ordre n inversible, dont on cherche à calculer l'inverse.

(a) Montrer que l'ensemble $GL_n(\mathbb{R})$ des matrices carrées inversibles d'ordre n (où $n \geq 1$) est un ouvert de l'ensemble $\mathcal{M}_n(\mathbb{R})$ des matrices carrées d'ordre n.

(b) Soit T l'application définie de $GL_n(\mathbb{R})$ dans $GL_n(\mathbb{R})$ par $T(B) = B^{-1}$. Montrer que T est dérivable, et que $DT(B)H = -B^{-1}HB^{-1}$.

(c) Ecrire la méthode de Newton pour calculer A^{-1} en cherchant le zéro de la fonction g de $\mathcal{M}_n(\mathbb{R})$ dans $\mathcal{M}_n(\mathbb{R})$ définie par$g(B) = B^{-1} - A$. Soit $B^{(k)}$ la suite ainsi définie.

(d) Montrer que la suite $B^{(k)}$ définie dans la question précédente vérifie :

$$\mathrm{Id} - AB^{(k+1)} = (\mathrm{Id} - AB^{(k)})^2.$$

En déduire que la suite $(B^{(k)})_{n\in\mathbb{N}}$ converge vers A^{-1} si et seulement si $\rho(\mathrm{Id} - AB^{(0)}) < 1$.

Exercice 119 (Méthode de Newton pour le calcul de la racine)**.**

1. Soit $\lambda \in \mathbb{R}_+$ et f_λ la fonction de $\mathbb{R}$ dans $\mathbb{R}$ définie par $f_\lambda(x) = x^2 - \lambda$.

1.1 Soit $x^{(0)} \in \mathbb{R}$ fixé. Donner l'algorithme de Newton pour la résolution de l'équation $f_\lambda(x) = 0$.

1.2 On suppose $x^{(0)} > 0$.

(i) Montrer que la suite $(x^{(k)})_{k\geq 1}$ est minorée par $\sqrt{\lambda}$.

(ii) Montrer que la suite $(x^{(k)})_{k\geq 0}$ converge et donner sa limite.

Soit $n \in \mathbb{N}^*$ et soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice diagonalisable dans $\mathbb{R}$; on note $\lambda_i, i = 1, \ldots, n$ les valeurs propres de A. On suppose que $\lambda_i > 0$ pour tout $i = 1, \ldots, n$.

2. Montrer qu'il existe au moins une matrice $B \in \mathcal{M}_n(\mathbb{R})$ telle que $B^2 = A$. Calculer une telle matrice B dans le cas où $n = 2$ et $A = \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}$.

3. On suppose de plus que A est symétrique définie positive. Montrer qu'il existe une unique matrice symétrique définie positive B telle que $B^2 = A$. Montrer par un contre exemple que l'unicité n'est pas vérifiée si on ne demande pas que B soit symétrique définie positive.

Soit F l'application de $\mathcal{M}_n(\mathbb{R})$ dans $\mathcal{M}_n(\mathbb{R})$ définie par $F(X) = X^2 - A$.

4. Montrer que F est différentiable en tout $X \in \mathcal{M}_n(\mathbb{R})$, et déterminer $DF(X)H$ pour tout $H \in \mathcal{M}_n(\mathbb{R})$.

Dans la suite de l'exercice, on considère la méthode de Newton pour déterminer B.

5. On suppose maintenant $n \geq 1$. On note $(X^{(k)})_{k\in\mathbb{N}}$ la suite (si elle existe) donnée par l'algorithme de Newton à partir d'un choix initial $X^{(0)} = I$, où I est la matrice identité de $\mathcal{M}_n(\mathbb{R})$.
   5.1 Donner le procédé de construction de $X^{(k+1)}$ en fonction de $X^{(k)}$, pour $k \geq 0$.
   5.2 On note $\lambda_1 \leq \cdots \leq \lambda_n$ les valeurs propres de A (dont certaines peuvent être égales) et P la matrice orthogonale telle que $A = P\text{diag}(\lambda_1, \cdots, \lambda_n)P^{-1}$.
      (i) Montrer que pour tout $k \in \mathbb{N}$, $X^{(k)}$ est bien définie et est donné par

$$X^{(k)} = P\text{diag}(\mu_1^{(k)}, \cdots, \mu_n^{(k)})P^{-1},$$

      où $\mu_i^{(k)}$ est le $k^{ième}$ terme de la suite de Newton pour la résolution de $f_{\lambda_i}(x) = 0$, où $f_{\lambda_i}(x) = x^2 - \lambda_i$, avec comme choix initial $\mu_i^{(0)} = 1$.
      (ii) En déduire que la suite $X^{(k)}$ converge vers B quand $k \to +\infty$.

Exercice 120 (Valeurs propres et méthode de Newton)**.**

Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique. Soient $\overline{\lambda}$ une valeur propre simple de A et $\overline{x} \in \mathbb{R}^n$ un vecteur propre associé t.q. $\overline{x} \cdot \overline{x} = 1$. Pour calculer $(\overline{\lambda}, \overline{x})$ on applique la méthode de Newton au système non linéaire (de $\mathbb{R}^{n+1}$ dans $\mathbb{R}^{n+1}$) suivant :

$$\begin{aligned} Ax - \lambda x &= 0, \\ x \cdot x &= 1. \end{aligned}$$

Montrer que la méthode est localement convergente.

Exercice 121 (Problème aux valeurs propres généralisé)**.**

Soient $A, B \in \mathcal{M}_n(\mathbb{R})$, $n \geq 1$. Pour $\lambda \in \mathbb{R}$, on pose $P(\lambda) = \det(A + \lambda B)$.

1. Montrer que P est un polynôme de degré inférieur ou égal à n.
2. On suppose dans cette question que $\ker A \cap \ker B \neq \{0\}$. Montrer que P est le polynôme nul.

Dans toute la suite, on suppose que A et B sont des matrices symétriques et que $\ker A \cap \ker B = \{0\}$.
Soit $\lambda \in \mathbb{R}$ tel que $\dim\ker(A + \lambda B) = 1$. On suppose que $Bv \cdot v \neq 0$ pour $v \in \ker(A + \lambda B)$, $v \neq 0$. Soit $u \in \ker(A + \lambda B)$ tel que $Bu \cdot u = 1$. On cherche à calculer le couple (λ, u) par la méthode de Newton. Pour cela, on définit la fonction G de $\mathbb{R}^{n+1}$ dans $\mathbb{R}^{n+1}$ par

$$G(\begin{bmatrix} v \\ \mu \end{bmatrix}) = \begin{bmatrix} Av + \mu Bv \\ Bv \cdot v - 1 \end{bmatrix} \text{ pour } v \in \mathbb{R}^n,\ \mu \in \mathbb{R}.$$

3. Pour $v \in \mathbb{R}^n$ $\mu \in \mathbb{R}$, donner la matrice jacobienne de G au point $\begin{bmatrix} v \\ \mu \end{bmatrix}$ sous forme de blocs, avec quatre matrices appartenant à $M_n(\mathbb{R})$, $M_{n,1}(\mathbb{R})$, $M_{1,n}(\mathbb{R})$ et $M_{1,1}(\mathbb{R})$.
   Cette matrice jacobienne est notée ensuite $J_G(v, \mu)$ (elle appartient à $M_{n+1}(\mathbb{R})$)
4. La fonction J_G (de $\mathbb{R}^{n+1}$ dans $M_{n+1}(\mathbb{R})$) est-elle continue, de classe C^1, de classe C^2, ... de classe C^∞ ?

5. Soient $v \in \mathbb{R}^n$ et $\mu \in \mathbb{R}$ tels que $J_G(u,\lambda)\begin{bmatrix} v \\ \mu \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.
   Montrer que $Av + \lambda Bv + \mu Bu = 0$. En déduire que $\mu = 0$.[utiliser la symétrie de A et B.]
   Montrer que $v = 0$. [Utiliser $\dim \ker(A + \lambda B) = 1$.]
6. On considère l'algorithme de Newton pour calculer le couple (u, λ). Cet algorithme s'écrit
   Initialisation $u_0 \in \mathbb{R}^n$, $\lambda_0 \in \mathbb{R}$.
   Itérations pour $k \geq 0$, u_k et λ_k donnés, on calcule (si c'est possible) (u_{k+1}, λ_{k+1}) solution de
   $$J_G(u_k,\lambda_k)\begin{bmatrix} u_{k+1}-u_k \\ \lambda_{k+1}-\lambda_k \end{bmatrix} = -\begin{bmatrix} Au_k + \lambda_k Bu_k \\ Bu_k \cdot u_k - 1 \end{bmatrix}.$$
   Montrer qu'il existe $\varepsilon > 0$ tel que, si $\|u_0 - u\|_2 \leq \varepsilon$ et $|\lambda_0 - \lambda| \leq \varepsilon$, la suite $(u_k, \lambda_k)_{k\in\mathbb{N}}$ est bien définie (c'est-à-dire que la matrice $J_G(u_k, \lambda_k)$ est inversible pour tout $k \in \mathbb{N}$) et converge, quand $k \to +\infty$, vers (u, λ).
7. (Question indépendante des questions précédentes) On suppose que A ou B est s.d.p. Montrer que les racines de P sont nécessairement réelles.

Exercice 122 (Newton monotone)**.** Soit F une application de $\mathbb{R}^n$ dans $\mathbb{R}^n$ ($n \geq 1$) de classe C^3. On suppose que F est convexe, c'est-à-dire $F(tx + (1-t)y) \leq tF(x) + (1-t)F(y)$ pour tous $x, y \in \mathbb{R}^n$ et $t \in [0, 1]$. (On rappelle que $x \leq y$ signifie que $x_i \leq y_i$ pour toutes les composantes x_i, y_i de x et y.) On note $J_F(x)$ la matrice jacobienne de F au point x.

1. Soit $x, y \in \mathbb{R}^n$. Montrer que $F(x + t(y - x)) - F(x) \leq t(F(y) - F(x))$ pour tout $t \in]0, 1[$. En déduire que $F(y) \geq F(x) + J_F(x)(y - x)$.

Soit $\overline{x} \in \mathbb{R}^n$ tel que $F(\overline{x}) = 0$. On suppose que la matrice $J_F(\overline{x})$ est inversible et on s'intéresse à la méthode de Newton pour calculer $\overline{x}$. On rappelle que cette méthode s'écrit
Initialisation $x_0 \in \mathbb{R}^n$,
Itérations pour $k \geq 0$, x_k donné, on calcule (si c'est possible) x_{k+1} solution de

$$J_F(x_k)(x_{k+1} - x_k) = -F(x_k).$$

2. Montrer qu'il existe $\varepsilon > 0$ tel que, si $\|x_0 - \overline{x}\|_2 \leq \varepsilon$, la suite $(x_k)_{k\in\mathbb{N}}$ est bien définie (c'est-à-dire que la matrice $J_F(x_k)$ est inversible pour tout $k \in \mathbb{N}$) et que $x_k \to \overline{x}$ quand $k \to +\infty$.
3. On suppose maintenant que pour tout x tel que $\|x - \overline{x}\|_2 \leq \varepsilon$ (ε est donné à la question 2) la matrice $J_F(x)$ est inversible et la matrice inverse, notée $J_F(x)^{-1}$, a tous ses coefficients positifs. On suppose aussi que $F(x_0) \geq 0$.
   Montrer, par récurrence sur k, que $\bar{x} \leq x_{k+1} \leq x_k \leq x_0$.
   [Utiliser la convexité de F et le fait que, si $A \in \mathcal{M}_n(\mathbb{R})$ et A^{-1} a tous ses coefficients positifs, on a $(b \in \mathbb{R}^n,\, b \geq 0,\, Ax = b) \Rightarrow x \geq 0$.]

Exercice 123 (Modification de la méthode de Newton)**.** *Suggestions en page 173.*

Soient $f \in C^1(\mathbb{R}^n, \mathbb{R}^n)$ et $\overline{x} \in \mathbb{R}^n$ t.q. $f(\overline{x}) = 0$. On considère, pour $\lambda > 0$ donné, la méthode itérative suivante :

— Initialisation : $x^{(0)} \in \mathbb{R}^n$.
— Iterations : pour $k \geq 0$,

$$x^{(k+1)} = x^{(k)} - [Df(x^{(k)})^t Df(x^{(k)}) + \lambda \mathrm{Id}]^{-1} Df(x^{(k)})^t f(x^{(k)}).$$

[Noter que, pour $\lambda = 0$, on retrouve la méthode de Newton.]

1. Montrer que la suite $(x^{(k)})_{k\in\mathbb{N}}$ est bien définie.

2. On suppose, dans cette question, que $n = 1$ et que $f'(\overline{x}) \neq 0$. Montrer que la méthode est localement convergente en $\overline{x}$.
3. On suppose que le rang de $Df(\overline{x})$ est égal à n. Montrer que la méthode est localement convergente en $\overline{x}$. [Noter que cette question redonne la question précédente si $n = 1$.]

Exercice 124 (Méthode de Newton pour un système semi-linéaire)**.** *Suggestions en page 173.*
On suppose que $f \in C^2(\mathbb{R}, \mathbb{R})$ et que f est croissante. On s'intéresse au système non linéaire suivant de n équations à n inconnues (notées $u_1, \ldots, u_n$) :

$$\begin{array}{l} (Au)_i + \alpha_i f(u_i) = b_i \ \forall i \in \{1, \ldots, n\}, \\ u = (u_1, \ldots, u_n)^t \in \mathbb{R}^n, \end{array} \tag{2.43}$$

où $A \in \mathcal{M}_n(\mathbb{R})$ est une matrice symétrique définie positive, $\alpha_i > 0$ pour tout $i \in \{1, \ldots, n\}$ et $b_i \in \mathbb{R}$ pour tout $i \in \{1, \ldots, n\}$.
On admet que (2.43) admet au moins une solution (ceci peut être démontré mais est difficile).

1. Montrer que (2.43) admet une unique solution.
2. Soit u la solution de (2.43). Montrer qu'il existe $a > 0$ t.q. la méthode de Newton pour approcher la solution de (2.43) converge lorsque le point de départ de la méthode, noté $u^{(0)}$, vérifie $|u - u^{(0)}| < a$.

Exercice 125 (Autre démonstration de la convergence locale de Newton)**.** *Suggestions en page 173. Corrigé en page 179* On se place sous les sous les hypothèses du théorème 2.19 avec g de classe C^2 au lieu de C^3. Montrer qu'il existe $a, a_1, a_2 \in \mathbb{R}_+^*$ tels que

1. si $\boldsymbol{x} \in B(\bar{\boldsymbol{x}}, a)$ alors $Dg(\boldsymbol{x})$ est inversible et $\|(Dg(\boldsymbol{x}))^{-1}\| \leq a_1$,
2. si $\boldsymbol{x}, y \in B(\bar{\boldsymbol{x}}, a)$ alors $\|g(y) - g(\boldsymbol{x}) - Dg(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x})\| \leq a_2 \|\boldsymbol{y} - \boldsymbol{x}\|^2$.

et qu'on peut donc appliquer le théorème 2.20 pour obtenir le résultat de convergence locale du théorème 2.19.

Exercice 126 (Convergence de la méthode de Newton si $f'(\overline{x}) = 0$)**.** *Suggestions en page 173, corrigé détaillé en page 180*

Soient $f \in C^2(\mathbb{R}, \mathbb{R})$ et $\overline{x} \in \mathbb{R}$ t.q. $f(\overline{x}) = 0$.

1. Rappel du cours. Si $f'(\overline{x}) \neq 0$, la méthode de Newton est localement convergente en $\overline{x}$ et la convergence est au moins d'ordre 2.
2. On suppose maintenant que $f'(\overline{x}) = 0$ et $f''(\overline{x}) \neq 0$. Montrer que la méthode de Newton est localement convergente (en excluant le cas $x_0 = \overline{x} \ldots$) et que la convergence est d'ordre 1. Si on suppose f de classe C^3, donner une modification de la méthode de Newton donnant une convergence au moins d'ordre 2.

Exercice 127 (Point fixe et Newton)**.**

Soit $g \in C^3(\mathbb{R}, \mathbb{R})$ et $\overline{x} \in \mathbb{R}$ tels que $g(\overline{x}) = 0$ et $g'(\overline{x}) \neq 0$ et soit $f \in C^1(\mathbb{R}, \mathbb{R})$ telle que $f(\overline{x}) = \overline{x}$.
On considère l'algorithme suivant :

$$\left\{ \begin{array}{l} x_0 \in \mathbb{R}, \\ \\ x_{n+1} = h(x_n), n \geq 0. \end{array} \right. \tag{2.44}$$

avec $h(x) = x - \dfrac{g(x)}{g' \circ f(x))}$.

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 (2.44) est bien définie ; montrer que $x_n \to \overline{x}$ lorsque $n \to +\infty$ (on pourra montrer qu'on peut choisir α de manière à ce que $|h'(x)| < 1$ si $x \in I_\alpha$).
   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 (2.44) est au moins quadratique.

3. On suppose de plus que f est deux fois dérivable et que $f'(\overline{x}) = \dfrac{1}{2}$. Montrer que la convergence de la suite $(x_n)_{n\in\mathbb{N}}$ définie par (1) est au moins cubique, c'est-à-dire qu'il existe $c \in \mathbb{R}_+$ tel que

$$|x_{n+1} - \overline{x}| \leq c|x_n - \overline{x}|^3, \qquad \forall n \geq 1.$$

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 $f \in C^1(\mathbb{R}, \mathbb{R})$ telle que :

$$f(x) = x - \frac{g(x)}{2g'(x)} \qquad \text{si } x \in I_\beta,$$

alors la suite définie par l'algorithme (1) converge de manière cubique.

Exercice 128 (Variante de la méthode de Newton)**.**
Corrigé détaillé en page 181

Soit $f \in C^1(\mathbb{R}, \mathbb{R})$ et $\bar{x} \in \mathbb{R}$ tel que $f(\bar{x}) = 0$. Soient $x_0 \in \mathbb{R}$, $c \in \mathbb{R}_+^*$, $\lambda \in \mathbb{R}_+^*$. On suppose que les hypothèses suivantes sont vérifiées :

(i) $\bar{x} \in I = [x_0 - c, x_0 + c]$,
(ii) $|f(x_0)| \leq \frac{c}{2\lambda}$,
(iii) $|f'(x) - f'(y)| \leq \frac{1}{2\lambda}$, $\forall (x, y) \in I^2$
(iv) $|f'(x)| \geq \frac{1}{\lambda}\, \forall x \in I$.

On définit la suite $(x^{(k)})_{n\in\mathbb{N}}$ par :

$$\begin{aligned} x^{(0)} &= x_0, \\ x^{(k+1)} &= x^{(k)} - \frac{f(x^{(k)})}{f'(y)}, \end{aligned} \tag{2.45}$$

où $y \in I$ est choisi arbitrairement.

1. Montrer par récurrence que la suite définie par (2.45) satisfait $x^{(k)} \in I$ pour tout $n \in \mathbb{N}$.
(On pourra remarquer que si $x^{(k+1)}$ est donné par (2.45) *alors* $x^{(k+1)} - x_0 = x^{(k)} - x_0 - \frac{f(x^{(k)}) - f(x_0)}{f'(y)} - \frac{f(x_0)}{f'(y)}$.*)*
2. Montrer que la suite $(x^{(k)})_{n\in\mathbb{N}}$ définie par (2.45) vérifie $|x^{(k)} - \bar{x}| \leq \frac{c}{2^n}$ et qu'elle converge vers $\bar{x}$ de manière au moins linéaire.
3. On remplace l'algorithme (2.45) par

$$\begin{aligned} x^{(0)} &= x_0, \\ x^{(k+1)} &= x^{(k)} - \frac{f(x^{(k)})}{f'(y^{(k)})}, \end{aligned} \tag{2.46}$$

où la suite $(y^{(k)})_{n\in\mathbb{N}}$ est une suite donnée d'éléments de I. Montrer que la suite $(x^{(k)})_{n\in\mathbb{N}}$ converge vers $\bar{x}$ de manière au moins linéaire, et que cette convergence devient super-linéaire si $f'(y_n) \to f'(\bar{x})$ lorsque $n \to +\infty$.
4. On suppose maintenant que $n \geq 1$ et que $f \in C^1(\mathbb{R}^n, \mathbb{R}^n)$. La méthode définie par (2.45) ou (2.46) peut-elle se généraliser, avec d'éventuelles modifications des hypothèses, à la dimension n ?

Exercice 129 (Méthode de Steffensen)**.**
Suggestions en page 173, corrigé détaillé en page 183

Soient $f \in C^2(\mathbb{R}, \mathbb{R})$ et $\overline{x} \in \mathbb{R}$ t.q. $f(\overline{x}) = 0$ et $f'(\overline{x}) \neq 0$. On considère la méthode itérative suivante :

— Initialisation : $x^{(0)} \in \mathbb{R}^n$.

— Itérations : pour $n \geq 0$, si $f(x^{(k)} + f(x^{(k)})) \neq f(x^{(k)})$,

$$x^{(k+1)} = x^{(k)} - \frac{(f(x^{(k)}))^2}{f(x^{(k)} + f(x^{(k)})) - f(x^{(k)})}, \tag{2.47}$$

et si $f(x^{(k)} + f(x^{(k)})) = f(x^{(k)})$, $x^{(k+1)} = x^{(k)}$.

1. Montrer qu'il existe $\alpha > 0$ tel que si $x^{(k)} \in B(\overline{x}, \alpha)$, alors $f(x^{(k)} + f(x^{(k)})) \neq f(x^{(k)})$ si $x^{(k)} \neq \overline{x}$. En déduire que si $x_0 \in B(\overline{x}, \alpha)$, alors toute la suite $(x^{(k)})_{n\in\mathbb{N}}$ vérifie (2.47) pourvu que $x^{(k)} \neq \overline{x}$ pour tout $n \in \mathbb{N}$.
2. Montrer par des développements de Taylor avec reste intégral qu' il existe une fonction a continue sur un voisinage de $\overline{x}$ telle que si $x_0 \in B(\overline{x}, \alpha)$, alors

$$x^{(k+1)} - \overline{x} = a(x^{(k)})(x^{(k)} - \overline{x}), \text{ pour tout } n \in \mathbb{N} \text{ tel que } x^{(k)} \neq \overline{x}. \tag{2.48}$$

3. Montrer que la méthode est localement convergente en $\overline{x}$ et la convergence est au moins d'ordre 2.

Exercice 130 (Méthode de Newton-Tchebycheff)**.**

1. Soit $f \in C^3(\mathbb{R}, \mathbb{R})$ et soit $\bar{x} \in \mathbb{R}$ tel que $\bar{x} = f(\bar{x})$ et $f'(\bar{x}) = f''(\bar{x}) = 0$. Soit $(x_n)_{n\in\mathbb{N}}$ la suite définie par :

$$\left\{ \begin{array}{l} x_0 \in \mathbb{R}, \\ x_{n+1} = f(x_n). \end{array} \right. \tag{PF}$$

(a) Justifier l'appellation "(PF)" de l'algorithme.

(b) Montrer qu'il existe $a > 0$ tel que si $|y - \overline{x}| \leq a$ alors $|f'(y)| \leq \frac{1}{2}$.

(c) Montrer par récurrence sur n que si $x_0 \in B(\bar{x}, a)$ alors $x_n \in B(\bar{x}, \frac{a}{2^n})$.

(d) En déduire que la suite construite par (PF) converge localement, c'est–à–dire qu'il existe un voisinage V de $\bar{x}$ tel que si $x_0 \in V$ alors $x_n \to \bar{x}$ lorsque $n \to +\infty$. .

(e) Montrer que la vitesse de convergence de la suite construite par (PF) est au moins cubique (c'est-à-dire qu' il existe $\beta \in \mathbb{R}_+$ tel que $|x_{n+1} - \bar{x}| \leq \beta|x_n - \bar{x}|^3$) si la donnée initiale x_0 est choisie dans un certain voisinage de $\bar{x}$. (On pourra utiliser un développement de Taylor-Lagrange.)

2. Soit $g \in C^3(\mathbb{R}, \mathbb{R})$, et soit $\bar{x} \in \mathbb{R}$ tel que $g(\bar{x}) = 0$ et $g'(\bar{x}) \neq 0$. Pour une fonction $h \in C^3(\mathbb{R}, \mathbb{R})$ à déterminer, on définit $f \in C^3(\mathbb{R}, \mathbb{R})$ par $f(x) = x + h(x)g(x)$. Donner une expression de $h(\bar{x})$ et $h'(\bar{x})$ en fonction de $g'(\bar{x})$ et de $g''(\bar{x})$ telle que la méthode (PF) appliquée à la recherche d'un point fixe de f converge localement vers $\bar{x}$ avec une vitesse de convergence au moins cubique.

3. Soit $g \in C^5(\mathbb{R}, \mathbb{R})$, et soit $\bar{x} \in \mathbb{R}$ tel que $g(\bar{x}) = 0$ et $g'(\bar{x}) \neq 0$. On considère la modification suivante (dûe à Tchebychev) de la méthode de Newton :

$$x_{n+1} = x_n - \frac{g(x_n)}{g'(x_n)} - \frac{g''(x_n)[g(x_n)]^2}{2[g'(x_n)]^3}. \tag{2.49}$$

Montrer que la méthode (2.49) converge localement et que la vitesse de convergence est au moins cubique. [On pourra commencer par le cas où g' ne s'annule pas].

Exercice 131 (Méthode de la sécante)**.** *Corrigé en page 2.3.3 page 186*
Soient $f \in C^2(\mathbb{R}, \mathbb{R})$ et $\overline{x} \in \mathbb{R}$ t.q. $f(\overline{x}) = 0$ et $f'(\overline{x}) \neq 0$. Pour calculer $\overline{x}$, on considère la méthode itérative suivante (appelée "méthode de la sécante") :

— Initialisation : $x_0, x_1 \in \mathbb{R}$.

— Itérations : pour $n \geq 1$, $x_{n+1} = x_n - \dfrac{f(x_n)(x_n - x_{n-1})}{f(x_n) - f(x_{n-1})}$ si $f(x_n) \neq 0$ et $x_{n+1} = x_n$ si $f(x_n) = 0$.

Si la suite $(x_n)_{n\in\mathbb{N}}$ est bien définie par cette méthode, on pose $e_n = |x_n - \overline{x}|$ pour tout $n \in \mathbb{N}$.

1. Montrer qu'il existe $\varepsilon > 0$ t.q. pour $x_0, x_1 \in]\overline{x} - \varepsilon, \overline{x} + \varepsilon[$, $x_0 \neq x_1$, la méthode de la sécante définit bien une suite $(x_n)_{n\in\mathbb{N}}$ et l'on a $e_{n+1} \leq (1/2)e_n$ pour tout $n \geq 1$. [Raisonner par récurrence : on suppose $x_n, x_{n-1} \in]\overline{x} - \varepsilon, \overline{x} + \varepsilon[$, $x_n \neq x_{n-1}$ et $x_n \neq \overline{x}$. Montrer, grâce à un choix convenable de ε, que $f(x_n) \neq f(x_{n-1})$ et que $f(x_n) \neq 0$. En déduire que x_{n+1} est bien défini et $x_n \neq x_{n+1}$. Puis, toujours grâce à un choix convenable de ε, que $e_{n+1} \leq (1/2)e_n$. Conclure.]

   Dans les questions suivantes, on suppose que $x_0, x_1 \in]\overline{x} - \varepsilon, \overline{x} + \varepsilon[$, $x_0 \neq x_1$ (ε trouvé à la première question) et que la suite $(x_n)_{n\in\mathbb{N}}$ donnée par la méthode de la sécante vérifie $x_n \neq \overline{x}$ pour tout $n \in \mathbb{N}$. On pose $d = (1 + \sqrt{5})/2$ et on démontre que la convergence est en général d'ordre d.

2. Pour $x \neq \overline{x}$, on définit $\mu(x)$ comme la moyenne de f' sur l'intervalle dont les extrémités sont x et $\overline{x}$.

   (a) Montrer que $e_{n+1} = e_n e_{n-1} M_n$, pour tout $n \geq 1$, avec $M_n = |\frac{\mu(x_n)-\mu(x_{n-1})}{f(x_n)-f(x_{n-1})}|$.

   (b) Montrer que la fonction μ est dérivable sur $\mathbb{R} \setminus \{\overline{x}\}$. Calculer $\lim_{x\to\overline{x}} \mu'(x)$.

   (c) Calculer la limite de la suite $(M_n)_{n\geq 1}$ lorsque $n \to \infty$. En déduire que la suite $(M_n)_{n\geq 1}$ est bornée.

3. Soit $M > 0$, $M \geq M_n$ pour tout $n \geq 1$ (M_n donné à la question précédente). On pose $a_0 = Me_0$, $a_1 = Me_1$ et $a_{n+1} = a_n a_{n-1}$ pour $n \geq 1$.

   (a) Montrer que $Me_n \leq a_n$ pour tout $n \in \mathbb{N}$.

   (b) Montrer qu'il existe $\varepsilon_1 \in]0, \varepsilon]$ t.q. la suite $(a_n)_{n\geq 1}$ tend en décroissant vers 0 lorsque $n \to +\infty$, si $x_0, x_1 \in]\overline{x} - \varepsilon_1, \overline{x} + \varepsilon_1[$.

   (c) Dans cette question, on prend $x_0, x_1 \in]\overline{x} - \varepsilon_1, \overline{x} + \varepsilon_1[$. Montrer qu'il existe $\alpha > 0$ et $\beta \in]0, 1[$ t.q $Me_n \leq a_n \leq \alpha(\beta)^{d^n}$ pour tout $n \in \mathbb{N}$ (ceci correspond à une convergence d'ordre au moins d). [On pourra utiliser la relation de récurrence $\ln a_{n+1} = \ln a_n + \ln a_{n-1}$ pour $n \geq 1$].

   (d) (Question plus difficile) Si $f"(\overline{x}) \neq 0$, $e_{n+1} = e_n e_{n-1} M_n$, montrer que $M_n \to \overline{M}$, quand $n \to \infty$, avec $\overline{M} > 0$. En déduire qu'il existe $\gamma > 0$ t.q. $\frac{e_{n+1}}{(e_n)^d} \to \gamma$ quand $n \to \infty$ (ceci signifie que la convergence est exactement d'ordre d). [Considérer, par exemple, $\beta_n = \ln e_{n+1} - d \ln e_n$ et montrer que β_n converge dans $\mathbb{R}$ quand $n \to \infty$.]

   (e) Comparer l'ordre de convergence de la méthode de la sécante à celui de la méthode de Newton.

Exercice 132 (Algorithme de Newton pour calculer une racine cubique)**.**
Soient $n \geq 1$ et $A \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique définie positive. On note $\{e_1, \ldots, e_n\}$ une base orthonormée de $\mathbb{R}^n$ formée de vecteurs propres de A. On a donc $Ae_i = \lambda_i e_i$, avec $\lambda_i > 0$, $e_i \cdot e_j = 0$ pour $i \neq j$ et $e_i \cdot e_i = 1$. Pour $B \in \mathcal{M}_n(\mathbb{R})$, on pose $F(B) = B^3 - A$.

1. Montrer que la matrice X définie par $Xe_i = \mu_i e_i$ pour $i = 1, \ldots, n$, avec $\mu_i^3 = \lambda_i$, est une racine cubique de A (c'est-à-dire une solution de $F(X) = 0$). Montrer que X est symétrique.

Dans la suite de l'exercice, on s'intéresse à l'algorithme de Newton pour calculer X.

2. Soit $B \in \mathcal{M}_n(\mathbb{R})$. La différentielle de F au point B, notée $DF(B)$, est donc une application linéaire de $\mathcal{M}_n(\mathbb{R})$ dans $\mathcal{M}_n(\mathbb{R})$. Donner, pour tout $H \in \mathcal{M}_n(\mathbb{R})$, l'expression de $DF(B)H$.

3. Montrer que $DF(X)$ est une application inversible.
   [On pourra, par exemple, calculer $DF(X)He_i \cdot e_j$ et montrer que $DF(X)H = 0$ implique $H = 0$.]

4. Donner l'algorithme de Newton pour calculer X et montrer que l'algorithme de Newton donne une suite convergente vers X si l'initialisation de l'algorithme est faite avec une matrice suffisamment proche de X.

Suggestions

Exercice 107 page 163 (Newton et logarithme) Etudier les variations de la fonction φ définie par : $\varphi(x) = x - x \ln x$.

Exercice 119 page 166 (Méthode de Newton pour le calcul de la racine) 1.1 (ii) Montrer que la suite $(x^{(k)})_{k\geq 1}$ est décroissante.
4. Ecrire la définition de la différentiel en faisant attention que le produit matriciel n'est pas commutatif.
5. Ecrire l'algorithme de Newton dans la base des vecteurs propres associés aux valeurs propres de A.

Exercice 125 page 169 (Autre démonstration de la convergence locale de Newton)
Soit $S = Dg(\bar{x})^{-1}(Dg(x) - Dg(\bar{x}))$. Remarquer que

$$Dg(\boldsymbol{x}) = Dg(\bar{\boldsymbol{x}}) - Dg(\bar{\boldsymbol{x}}) + Dg(\boldsymbol{x}) = Dg(\bar{\boldsymbol{x}})(Id + S)$$

et démontrer que $\|S\| < 1$.
En déduire que $Dg(x) = Dg(\bar{x})(Id + S)$ est inversible, et montrer alors l'existence de a et de $a_1 = 2\|Dg(\bar{x})^{-1}\|$ tels que si $x \in B(\bar{x}, a)$ alors $Dg(x)$ est inversible et $\|Dg(x)^{-1}\| \leq a_1$.
Pour montrer l'existence de a_2, introduire la fonction $\varphi \in C^1(\mathbb{R}, \mathbb{R}^n)$ définie par

$$\varphi(t) = g(x + t(y - x)) - g(x) - tDg(x)(y - x).$$

Exercice 120 page 167 (Valeurs propres et méthode de Newton) Écrire le système sous la forme $F(x, \lambda) = 0$ où F est une fonction de $\mathbb{R}^{n+1}$ dans $\mathbb{R}^{n+1}$ et montrer que $DF(\overline{\lambda}, \overline{x})$ est inversible.

Exercice 123 page 168 (Modification de la méthode de Newton) 1. Remarquer que si $A \in \mathcal{M}_n(\mathbb{R})$ et $\lambda > 0$, alors$A^tA + \lambda\mathrm{Id}$ est symétrique définie positive.
2. En introduisant la fonction φ définie par $\varphi(t) = f(tx_n + (1 - t)\overline{x})$, montrer que $f(x_n) = (x_n - \overline{x})g(x_n)$, où $g(x) = \int_0^1 f'(tx + (1 - t)\overline{x})dt$. Montrer que g est continue.
Montrer que la suite $(x_n)_{n\in\mathbb{N}}$ vérifie $x_{n+1} - \overline{x} = a_n(x_n - \overline{x})$, où

$$a_n = 1 - \frac{f'(x_n)g(x_n)}{f'(x_n)^2 + \lambda},$$

et qu'il existe α tel que si $x_n \in B(\overline{x}, \alpha)$, alors $a_n \in]0, 1[$. Conclure.
3. Reprendre la même méthode que dans le cas $n = 1$ pour montrer que la suite $(x_n)_{n\in\mathbb{N}}$ vérifie $x_{n+1} - \overline{x} = D(x_n)(x_n - \overline{x})$, où $D \in \mathcal{C}(\mathbb{R}^n, \mathcal{M}_n(\mathbb{R}))$. Montrer que $D(\overline{x})$ est symétrique et montrer alors que $\|D(\overline{x})\|_2 < 1$ en calculant son rayon spectral. Conclure par continuité comme dans le cas précédent.

Exercice 126 page 169 (Convergence de la méthode de Newton si $f'(\overline{x}) = 0$) Supposer par exemple que $f''(\overline{x}) > 0$ et montrer que si x_0 est "assez proche" de $\overline{x}$ la suite $(x_n)_{n\in\mathbb{N}}$ est croissante majorée ou décroissante minorée et donc convergente. Pour montrer que l'ordre de la méthode est 1, montrer que

$$\frac{\|x_{n+1} - \overline{x}\|}{\|x_n - \overline{x}\|} \to \frac{1}{2} \text{ lorsque } n \to +\infty.$$

Exercice 124 page 169 (Méthode de Newton) 1. Pour montrer l'unicité, utiliser la croissance de f et le caractère s.d.p. de A.
2. Utiliser le théorème de convergence du cours.

Exercice 129 page 170 (Méthode de Steffensen)) 1. Utiliser la monotonie de f dans un voisinage de $\overline{x}$.
2. Développer le dénominateur dans l'expression de la suite en utilisant le fait que $f(x_n + f(x_n)) - f(x_n) = \int_0^1 \psi'(t)dt$ où $\psi(t) = f(x_n + tf(x_n))$, puis que $f'(x_n + tf(x_n)) = \int_0^t \xi'(s)ds$ où $\xi(t) = f'(x_n + tf(x_n))$. Développer ensuite le numérateur en utilisant le fait que $-f(x_n) = \int_0^1 \varphi'(t)dt$ où $\varphi(t) = f(t\overline{x} + (1 - t)x_n)$, et que $f'(t\overline{x} + (1 - t)x_n) = \int_0^1 \chi(s)ds + \chi(0)$, où $\chi(t) = f(\overline{x} + (1 - t)_n)$.
3. La convergence locale et l'ordre 2 se déduisent des résultats de la question 2.

Corrigés des exercices

Exercice 108 page 163 (Méthode de Newton pour un système linéaire) Pour tout $x \in \mathbb{R}^n$, la matrice $Df(x)$ est inversible ; donc l'algorithme de Newton est bien défini et s'écrit $Df(x)x^{(1)} = b$. Il converge donc en une itération, qui demande la résolution du sysème linéaire $Ax = b$...

Exercice 109 page 163 (Condition initiale et Newton) On vérifie que

$$DF(x,y) = \begin{bmatrix} \cos(x) & 1 \\ y & x \end{bmatrix}$$

et par conséquent

$$DF\left(\frac{\pi}{2}, 0\right) = \begin{bmatrix} 0 & 1 \\ 0 & \frac{\pi}{2} \end{bmatrix}$$

n'est pas inversible. La matrice $DF(x,y)$ est inversible pour $x = 0$ $y = 1$ par exemple.

Exercice 112 page 164 (Newton pour un autre système 2×2) ...)

1. Les solutions $(x,y) \in \mathbb{R}^2$ vérifient

$$\begin{aligned} x(x+2y) &= 0, \\ xy + 1 &= 0, \end{aligned}$$

   La première équation implique $x = 0$ ou $x = -2y$. Le choix $x = 0$ est impossible en raison de la seconde équation, on a donc forcément $x = -2y$ et donc $2y^2 = 1$. Les solutions sont donc $(\overline{x}_1, \overline{y}_1) = (\sqrt{2}, -\frac{\sqrt{2}}{2})$ et $(\overline{x}_2, \overline{y}_2) = (-\sqrt{2}, \frac{\sqrt{2}}{2})$. Notons que la jacobienne est bien définie en $(\overline{x}_1, \overline{y}_1)$ et $(\overline{x}_2, \overline{y}_2)$.

2. Soit F l'application de $\mathbb{R}^2$ dans $\mathbb{R}^2$ définie par $F(x,y) = (x^2 + 2xy, xy + 1)$. Calculons la matrice jacobienne de F au point (x,y) :

$$DF(x,y) = \begin{bmatrix} 2x+y & 2x \\ y & x \end{bmatrix}$$

   On a donc $Det(DF)(x,y) = (2x+y)x - 2xy = 2x(2x-y) \neq 0$ pour tout couple $(x,y) \in \mathbb{R}^2$ tel que $x \neq 0$ et $x \neq y$. L'algorithme de Newton s'écrit :

$$DF(x^{(k)}, y^{(k)}) \begin{bmatrix} x^{(k+1)} - x^{(k)} \\ y^{(k+1)} - y^{(k)} \end{bmatrix} = -F(x^{(k)}, y^{(k)})$$

   et la suite est donc bien définie si $2x^{(k)} \neq y^{(k)}$ et $x^{(k)} \neq 0$ pour tout $k \geq 0$.

3. On a $DF(1,-1) = \begin{bmatrix} 1 & 2 \\ -1 & 1 \end{bmatrix}$. Le système à résoudre est donc

$$\begin{aligned} \delta x + 2\delta y &= -1 \\ -\delta x + \delta y &= 0 \end{aligned}$$

   On en déduit $\delta x = \delta y = -\frac{1}{2}$, c.à.d. $x_1 = \frac{1}{2}$, $y_1 = -\frac{3}{2}$.

4. La fonction F est infiniment continûment différentiable. Pour appliquer le théorème de convergence du cours, il reste à vérifier que la matrice jacobienne DF est inversible dans un voisinage de $(\overline{x}, \overline{y})$ où $(\overline{x}, \overline{y})$ une solution de (2.38)-(2.39). Or $Det(DF(\overline{x}, \overline{y}))) = 5 \neq 0$. La matrice $DF(\overline{x}, \overline{y}))$ est donc inversible, et le théorème du cours s'applique : il existe donc $\varepsilon > 0$ tel que si (x_0, y_0) est dans la boule B_ε de centre $(\overline{x}, \overline{y})$ et de rayon ε, alors la suite $(x_n, y_n)_{n \in \mathbb{N}}$ construite par la méthode de Newton converge vers $(\overline{x}, \overline{y})$ lorsque n tends vers $+\infty$. De plus la convergence est quadratique.

Exercice 113 page 164 (Newton et les échelles...)
1. Notons A et B les deux pieds des murs, P le point de croisement des échelles et M sa projection sur le plan horizontal, comme indiqué sur la figure. Soient $x = d(A, M)$ la distance de A à M, $y = d(B, M)$, $\alpha = d(A, P)$ et $\beta = d(B, P)$.
Par le théorème de Pythagore, $x^2 + 1 = \alpha^2$ et $y^2 + 1 = \beta^2$. Par le théorème de Thalès, $\frac{x}{d} = \frac{\alpha}{4}$ et $\frac{y}{d} = \frac{\beta}{3}$. En éliminant α et β, on en déduit que x et y sont solutions du système non linéaire :

$$\begin{aligned} 16x^2 &= (x^2+1)(x+y)^2 \\ 9y^2 &= (y^2+1)(x+y)^2, \end{aligned}$$

qui est bien le système (2.40)-(2.41).

2. Le système précédent s'écrit sous la forme $F(X) = 0$, où F est la fonction de $\mathbb{R}^2$ dans $\mathbb{R}^2$ définie par

$$F(X) = F\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} (x^2+1)(x+y)^2 - 16x^2 \\ (y^2+1)(x+y)^2 - 9y^2 \end{bmatrix}.$$

On a donc

$$DF(X) = \begin{bmatrix} 4x^3 + 2xy^2 + 6x^2y + 2x + 2y - 32x & 2x^2y + 2x^3 + 2y + 2x \\ 2xy^2 + 2y^3 + 2x + 2y & 4y^3 + 2yx^2 + 6y^2x + 2y + 2x - 18y \end{bmatrix}$$

L'algorithme de Newton pour la résolution du système (2.40)-(2.41) s'écrit donc, pour $X_k = \begin{bmatrix} x_k \\ y_k \end{bmatrix}$ connu,

$$DF(X_k)(X_{k+1} - X_k) = -F(X_k).$$

A l'étape k, on doit donc résoudre le système linéaire

$$[DF(X_k) \begin{bmatrix} s \\ t \end{bmatrix} = - \begin{bmatrix} (x_k^2+1)(x_k+y_k)^2 - 16x_k^2 \\ (y_k^2+1)(x_k+y_k)^2 - 9y_k^2 \end{bmatrix} \tag{2.50}$$

avec

$$DF(X_k) = \begin{bmatrix} 4x_k^3 + 2x_k y_k^2 + 6x_k^2 y_k + 2x_k + 2y_k - 32x_k & 2x^2 y_k + 2x_k^3 + 2y_k + 2x_k \\ 2x_k y_k^2 + 2y_k^3 + 2x_k + 2y_k & 4y_k^3 + 2y_k x_k^2 + 6y_k^2 x + 2y_k + 2x_k - 18y_k \end{bmatrix}$$

et on obtient le nouvel itéré $X_{k+1} = \begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} x_k + s \\ y_k + t \end{bmatrix}$.

3. La première itération nécessite donc la résolution du système linéaire (2.50) pour $k = 0$, soit, pour $x_0 = 1$ et $y_0 = 1$,

$$\begin{bmatrix} -16 & 8 \\ 8 & -2 \end{bmatrix} \begin{bmatrix} s \\ t \end{bmatrix} = \begin{bmatrix} 8 \\ 1 \end{bmatrix}$$

dont la solution est $s = \frac{3}{4}$ et $t = \frac{5}{2}$. On en déduit que $x_1 = 1.75$ et $y_1 = 3.5$ construits par la méthode de Newton en partant de $x^{(0)} = 1$ et $y^{(0)} = 1$. La distance d à la première itération est donc $d = 5.25$.

Exercice 115 page 165 (Recherche d'un point fixe)
1. La fonction f est paire, il suffit de l'étudier sur $\mathbb{R}_+$. Comme $f'(x) = 2x(e^{(x^2)} - 4)$, la fonction f' ne s'annule qu'une fois sur $\mathbb{R}_+^\star$. Elle est négative pour x strictement positif et proche de 0. Comme $f(0) = 1$ et $f(1) = e - 4 < 0$, la fonction s'annule deux fois sur $\mathbb{R}_+$ et une seule fois entre 0 et 1.

2. Pour montrer que la méthode du point fixe appliquée à g, initialisée avec un point de l'intervalle $]0, 1[$, est convergente, il suffit de montrer de g est une application strictement contractante de $[0, 1]$ dans $[0, 1]$.

Pour $x \in [0,1]$, on a $0 < g(x) < (1/2)\sqrt{e} < 1$ et $0 \leq g'(x) = (1/2)x\sqrt{e^{(x^2)}} < (1/2)\sqrt{e} < 1$. Ceci prouve que g est une application strictement contractante de $[0,1]$ dans $[0,1]$. La méthode du point fixe appliquée à g, initialisée avec un point de l'intervalle $]0,1[$, est donc convergente. Elle converge vers le point fixe de g dans $[0,1]$. Ce point fixe, noté x, vérifie $x = (1/2)\sqrt{e^{(x^2)}}$, c'est-à-dire $f(x) = 0$. Ceci montre que x est le point de $]0,1[$ annulant f. Comme $g'(x) \neq 0$, la convergence est d'ordre 1.

3. La méthode de Newton pour rechercher les points annulant f consiste à construire une suite $(x_n)_{n\in\mathbb{N}}$ de la manière suivante :
Initialisation : $x_0 \in \mathbb{R}$.
Itérations : Pour $n \in \mathbb{N}$, $2x_n(e^{(x_n^2)} - 4)(x_{n+1} - x_n) = -e^{(x_n^2)} + 4x_n^2$.
Soit x le point annulant f dans $]0,1[$. Comme $f'(x) \neq 0$, il existe $\varepsilon > 0$ tel que cette méthode est bien définie si $|x_0 - x| \leq \varepsilon$ et on a alors $x_n \to x$ as $n \to +\infty$. Comme f est de classe C^3, la convergence est au moins d'ordre 2 et donc la méthode Newton est *a priori* plus efficace que celle de la question précédente.

Exercice 116 page 165 (Nombre d'itérations fini pour Newton) 1.1 Comme f' est définie sur tout $\mathbb{R}$ par $f'(x) = e^x$ et ne s'annule pas, on en déduit que la suite construite par la méthode de Newton, qui s'écrit :

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})} = x^{(k)} - \frac{e^{x^{(k)}} - 1}{e^{x^{(k)}}}$$

est bien définie.

1.2 Par définition de la suite, on a $x^{(k+1)} - x^{(k)} = -\frac{e^{x^{(k)}}-1}{e^{x^{(k)}}} = 0$ ssi $x^{(k)} = 0$. Donc par récurrence sur n, si $x^{(0)} \neq 0$, on a $x^{(k+1)} \neq x^{(k)}$ pour tout $n \in \mathbb{N}$. De plus, si $f(x^{(0)}) = 0$ (c.à.d. si $x^{(0)} = 0$), la suite est stationnaire. On en déduit que la méthode de Newton converge en un nombre fini d'opérations si et seulement si $f(x^{(0)}) = 0$.

1.3 Par définition, on a : $x^{(1)} = x^{(0)} - \frac{e^{x^{(0)}}-1}{e^{x^{(0)}}}$. Par le théorème de accroissements finis, on a donc : $x^{(1)} = x^{(0)}(1 - e^{\theta - x^{(0)}})$, avec $\theta \in]x^{(0)}, 0[$ si $x^{(0)} < 0$ et $\theta \in]0, x^{(0)}[$ si $x^{(0)} > 0$. Si $x^{(0)} < 0$, on a $e^{\theta - x^{(0)}} > 1$ et donc $x^{(1)} > 0$. En revanche, si $x^{(0)} > 0$, on a $e^{-x^{(0)}} < e^{\theta - x^{(0)}} < 1$ et donc $0 < x^{(1)} < x^{(0)}$.

1.4 On a vu à la question 1.2 que si $x^{(0)} = 0$ la suite est stationnaire et égale à 0. On a vu à la question 1.3 que si $x^{(0)} < 0$ alors $x^{(1)} > 0$. Il suffit donc d'étudier le cas $x^{(0)} > 0$. Or si $x^{(0)} > 0$, on a $0 < x^{(1)} < x^{(0)}$. Par récurrence sur n, on en déduit que si $x^{(0)} > 0$, la suite $(x^{(k)})_{n\in\mathbb{N}}$ est décroissante et minorée par 0, donc elle converge. La limite ℓ de la suite vérifie : $\ell = \ell - \frac{e^\ell - 1}{e^\ell}$, soit encore $\ell = 0$ (unique solution de l'équation $f(x) = 0$).

2. Soient $x^{(k)}$ et $x^{(k+1)}$ deux itérés successifs donnés par la méthode de Newton, tels que $F(x^{(k)}) \neq 0$. On a donc :

$$DF(x^{(k)})(x^{(k+1)} - x^{(k)}) = -F(x^{(k)}), \tag{2.51}$$

et en particulier, $x^{(k+1)} \neq x^{(k)}$. Or, la condition de stricte convexité pour une fonction continûment différentiable entraîne que :

$$DF(x^{(k)})(x^{(k+1)} - x^{(k)}) < F(x^{(k+1)}) - F(x^{(k)}),$$

et donc, avec (2.51), $F(x^{(k+1)}) > 0$. On montre ainsi, par récurrence sur n, que si $F(x^{(0)}) \neq 0$, alors $F(x^{(k)}) > 0$ pour tout $n > 0$, ce qui montre que la méthode de Newton converge en un nombre fini d'opérations si et seulement si $F(x^{(0)}) = 0$.

Exercice 117 page 165 ([Méthode de Newton pour un système 2×2)

1. Lla méthode de Newton pour calculer la solution de $f(x) = 0$ s'écrit : $x_{k+1} = \frac{1}{2}x_k$, et donc x_k converge vers 0 pour tout x_0.
2. (a) L'ensemble des solutions de $F(x,y) = (0,0)$.est $\{(0,0)\}$.

(b) La matrice jacobienne s'écrit : $DF(x,y) = \begin{bmatrix} 2x & -1 \\ 0 & 2y \end{bmatrix}$, et donc l'algorithme de Newton s'écrit :

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} + \begin{bmatrix} \delta x \\ \delta y \end{bmatrix} \text{ avec } \begin{bmatrix} 2x_k & -1 \\ 0 & 2y_k \end{bmatrix} \begin{bmatrix} \delta x \\ \delta y \end{bmatrix} = - \begin{bmatrix} x_k^2 - y_k \\ y_k^2 \end{bmatrix}$$

Cet algorithme est bien défini dès que la matrice jacobienne est inversible, c.à.d dès que $x_k y_k \neq 0$. C'est vrai pour $k = 0$ par hypothèse. Montrons que c'est vrai pour tout $k \in \mathbb{N}$ par récurrence sur k. Supposons qu'on a $x_p \neq 0$ et $y_p > 0$ pour tout $p \leq k$, et montrons que dans ce cas on a encore $x_{k+1} \neq 0$ et $y_{k+1} > 0$. Par définition de l'algorithme de Newton, on a $y_{k+1} = \frac{y_k}{2} > 0$ et $x_{k+1} = \frac{x_k}{2} + \frac{y_k}{4x_k} = \frac{1}{2x_k}(x_k^2 + \frac{y_k}{2}) \neq 0$, car $x_k^2 + \frac{y_k}{2} > 0$.

(c) i. Comme $y_0 = 1$ et que $y_{k+1} = \frac{y_k}{2}$, on a $y_k = 2^{-k}$

ii. On écrit que $x_{k+1} = g_k(x_k)$ avec $g_k(x) = \frac{x}{2} + 2^{-k-2}\frac{1}{x}$. L'étude de la fonction g_k montre que $\min g_k = 2^{-\frac{k+1}{2}}$.
Comme $x_k \geq 2^{-\frac{k}{2}} > 0$ et que $y_k = 2^{-k}$, on en déduit que $x_{k+1} - x_k = \frac{1}{2x_k}(-x_k^2 + \frac{y_k}{2} \leq 0$.

iii. La suite $(x_k)_{k\in\mathbb{N}}$ est décroissante minorée par 0, et donc elle converge. En passant à la limite sur l'expression de x_{k+1}, on en déduit que la suite x_k converge vers 0. Par l'expression de y_k, on sait qu'elle converge également vers 0. D'où la conclusion.

Exercice 118 page 166 (Méthode de Newton pour le calcul de l'inverse)

1. (a) Soit g la fonction définie de $\mathbb{R}^*$ dans $\mathbb{R}$ par $g(x) = \frac{1}{x} - a$. Cette fonction est continue et dérivable pour tout $x \neq 0$, et on a : $g'(x) = -\frac{1}{x^2}$. L'algorithme de Newton pour la recherche d'un zéro de cette fonction s'écrit donc bien :

$$\begin{cases} x^{(0)} \text{ donné,} \\ x^{(k+1)} = x^{(k)}(2 - ax^{(k)}). \end{cases} \tag{2.52}$$

(b) Soit $(x^{(k)})_{n\in\mathbb{N}}$ définie par (2.42). D'après le théorème du cours, on sait que la suite $(x^{(k)})_{n\in\mathbb{N}}$ converge de localement (de manière quadratique) dans un voisinage de $\frac{1}{a}$. On veut déterminer ici l'intervalle de convergence précisément. On a $x^{(k+1)} = \varphi(x^{(k)})$ où φ est la fonction définie par de $\mathbb{R}$ dans $\mathbb{R}$ par $\varphi(x) = x(2 - ax)$. Le tableau de variation de la fonction φ est le suivant :

x	$-\infty$		0		$\frac{1}{a}$		$\frac{2}{a}$	
$\varphi'(x)$		$+$	0	$+$	0	$-$		
$\varphi(x)$	$-\infty$	$\nearrow$	0	$\nearrow$	$\frac{1}{a}$	$\searrow$	$-\infty$	

(2.53)

Il est facile de remarquer que l'intervalle $]0, \frac{1}{a}[$ est stable par φ et que $\varphi(]\frac{1}{a}, \frac{2}{a}[) =]0, \frac{1}{a}[$ Donc si $x^{(0)} \in]\frac{1}{a}, \frac{2}{a}[$ alors $x^{(1)} \in]0, \frac{1}{a}[$, et on se ramène au cas $x^{(0)} \in]0, \frac{1}{a}[$.
On montre alors facilement que si $x^{(0)} \in]0, \frac{1}{a}[$, alors $x^{(k+1)} \geq x^{(k)}$ pour tout n, et donc la suite $(x^{(k)})_{n\in\mathbb{N}}$ est croissante. Comme elle est majorée (par $\frac{1}{a}$), elle est donc convergente. Soit ℓ sa limite, on a $\ell = \ell(2 - a\ell)$, et comme $\ell \geq x^{(0)} > 0$, on a $\ell = \frac{1}{a}$.

Il reste maintenant à montrer que si $x^{(0)} \in]-\infty, 0[\cup]\frac{2}{a}, +\infty[$ alors $\lim_{n\to+\infty} x^{(k)} = -\infty$. On montre d'abord facilement que si $x^{(0)} \in]-\infty, 0[$, la suite $(x_n)_{n\in\mathbb{N}}$ est décroissante. Elle admet donc une limite finie ou infinie. Appelons ℓ cette limite. Celle-ci vérifie : $\ell = \ell(2 - a\ell)$. Si ℓ est finie, alors $\ell = 0$ ou $\ell = \frac{1}{a}$ ce qui est impossible car $\ell \leq x^{(0)} < 0$. On en déduit que $\ell = -\infty$.

Enfin, l'étude des variations de la fonction φ montre que si $x^{(0)} \in]\frac{2}{a}, +\infty[$, alors $x^{(1)}] -\infty, 0[$, et on est donc ramené au cas pécédent.

2. (a) L'ensemble $GL_n(\mathbb{R})$ est ouvert car image réciproque de l'ouvert $\mathbb{R}^*$ par l'application continue qui a une matrice associe son déterminant.

   (b) L'application T est clairement définie de $GL_n(\mathbb{R})$ dans $GL_n(\mathbb{R})$. Montrons qu'elle est dérivable. Soit $H \in GL_n(\mathbb{R})$ telle que $B+H$ soit inversible. Ceci est vrai si $\|H\|\|B^{-1}\| < 1$, et on a alors, d'après le cours :

$$(B+H)^{-1} = \big(B(Id + B^{-1}H)\big)^{-1} = \sum_{k=0}^{+\infty}(-B^{-1}H)^k B^{-1}.$$

   On a donc :

$$\begin{aligned} T(B+H) - T(B) &= \sum_{k=0}^{+\infty}(B^{-1}H)^k B^{-1} - B^{-1} \\ &= (Id + \sum_{k=1}^{+\infty}(-B^{-1}H)^k - \mathrm{Id})B^{-1} \\ &= \sum_{k=1}^{+\infty}(-B^{-1}H)^k B^{-1}. \end{aligned}$$

   On en déduit que

$$T(B+H) - T(B) + B^{-1}HB^{-1} = \sum_{k=2}^{+\infty}(-B^{-1}H)^k B^{-1}.$$

   L'application qui à H associe $-B^{-1}HB^{-1}$ est clairement linéaire, et de plus,

$$\|T(B+H) - T(B) + B^{-1}HB^{-1}\| \leq \|B^{-1}\| \sum_{k=2}^{+\infty}(\|B^{-1}\|\|H\|)^k.$$

   Or $\|B^{-1}\|\|H\| < 1$ par hypothèse. On a donc

$$\begin{aligned} \frac{\|T(B+H) - T(B) - B^{-1}HB^{-1}}{\|H\|}\| &\leq \|B^{-1}\|^3\|H\| \sum_{k=0}^{+\infty}(\|B^{-1}\|\|H\|)^k \\ &\to 0 \text{ lorsque } \|H\| \to 0. \end{aligned}$$

   On en déduit que l'application T est différentiable et que $DT(B)(H) = -B^{-1}HB^{-1}$.

   (c) La méthode de Newton pour la recherche d'un zéro de la fonction g s'écrit :

$$\begin{cases} B^0 \in GL_n(\mathbb{R}), \\ Dg(B^n)(B^{n+1} - B^n) = -g(B^n). \end{cases}$$

   Or, d'apres la question précédente, $Dg(B^n)(H) = -(B^n)^{-1}H(B^n)^{-1}$. On a donc

$$Dg(B^n)(B^{n+1} - B^n) = -(B^n)^{-1}(B^{n+1} - B^n)(B^n)^{-1}.$$

   La méthode de Newton sécrit donc :

$$\begin{cases} B^0 \in GL_n(\mathbb{R}), \\ -(B^{n+1} - B^n) = (Id - B^n A)B^n. \end{cases} \tag{2.54}$$

   soit encore

$$\begin{cases} B^0 \in GL_n(\mathbb{R}), \\ B^{n+1} = 2B^n - B^n A B^n. \end{cases} \tag{2.55}$$

(d) Par définition, on a :

$$Id - AB^{n+1} = \mathrm{Id} - A(2B^n - B^nAB^n) = \mathrm{Id} - 2AB^n + AB^nAB^n.$$

Comme les matrices Id et AB^n commutent, on a donc :

$$Id - AB^{n+1} = (Id - AB^n)^2.$$

Une récurrence immédiate montre alors que $Id - AB^n = (Id - AB^0)^{2^n}$. On en déduit que la suite $Id - AB^n$ converge (vers la matrice nulle) lorsque $n \to +\infty$ ssi $\rho(Id - AB^0) < 1$, et ainsi que la suite B^n converge vers A^{-1} si et seulement si $\rho(Id - AB^0) < 1$.

Exercice 125 page 169 (Autre démonstration de la convergence locale de Newton) Remarquons d'abord que

$$Dg(\boldsymbol{x}) = Dg(\bar{\boldsymbol{x}}) - Dg(\bar{\boldsymbol{x}}) + Dg(\boldsymbol{x}) = Dg(\bar{\boldsymbol{x}})(Id + S)$$

où $S = Dg(\bar{x})^{-1}(Dg(x) - Dg(\bar{x}))$. Or si $\|S\| < 1$, la matrice $(Id + S)$ est inversible et

$$\|(Id + S)^{-1}\| \leq \frac{1}{1 - \|S\|}.$$

Nous allons donc essayer de majorer $\|S\|$. Par définition de S, on a :

$$\|S\| \leq \|Dg(\bar{x})^{-1}\| \ \ \|Dg(x) - Dg(\bar{x}).\|$$

Comme $g \in C^2(\mathbb{R}^n, \mathbb{R}^n)$, on a $Dg \in C^1(\mathbb{R}^n, \mathcal{M}_n(\mathbb{R}))$; donc par continuité de Dg, pour tout $\varepsilon \in \mathbb{R}_+^*$, il existe $a \in \mathbb{R}_+^*$ tel que si $\|x - \overline{x}\| \leq a$ alors $\|Dg(x) - Dg(\bar{x})\| \leq \varepsilon$. En prenant $\varepsilon = \frac{1}{2\|Dg(\bar{x})^{-1}\|}$, il existe donc $a > 0$ tel que si $x \in B(\bar{x}, a)$ alors

$$\|Dg(x) - Dg(\bar{x})\| \leq \frac{1}{2\|Dg(\bar{x})^{-1}\|}$$

et donc si $x \in B(\bar{x}, a)$, alors $\|S\| \leq \frac{1}{2}$. On en déduit que si $x \in B(\bar{x}, a)$ alors $Id + S$ est inversible et donc que $Dg(x) = Dg(\bar{x})(Id + S)$ est inversible (on rappelle que $Dg(\bar{x})$ est inversible par hypothèse). De plus, si $x \in B(\bar{x}, a)$ on a :

$$\|(Id + S)^{-1}\| \leq \frac{1}{1 - \|S\|} \leq 2,$$

et comme $(Id + S)^{-1} = (Dg(\bar{x}))^{-1}Dg(x)$, on a $\|Dg(x)^{-1}Dg(\bar{x})\| \leq 2$, et donc

$$\|Dg(x)^{-1}\| \leq \|(Dg(\overline{\boldsymbol{x}}))^{-1}\|\|(Dg(x))^{-1}Dg(\overline{\boldsymbol{x}})\| \leq 2\|(Dg(\overline{\boldsymbol{x}}))^{-1}\|.$$

En résumé, on a donc prouvé l'existence de a et de $a_1 = 2\|Dg(\bar{x})^{-1}\|$ tels que si $x \in B(\bar{x}, a)$ alors $Dg(x)$ est inversible et $\|Dg(x)^{-1}\| \leq a_1$.
Il reste maintenant à trouver a_2 tel que

$$x, y \in B(\bar{x}, a) \Longrightarrow \|g(y) - g(x) - Dg(x)(y - x)\| \leq a_2\|y - x\|^2.$$

Comme $g \in C^2(\mathbb{R}^n, \mathbb{R}^n)$, on a donc $Dg \in C^1(\mathbb{R}^n, \mathcal{M}_n(\mathbb{R}))$ (remarquons que jusqu'à présent on avait utilisé uniquement le caractère C^1 de g). On définit la fonction $\varphi \in C^1(\mathbb{R}, \mathbb{R}^n)$ par

$$\varphi(t) = g(x + t(y - x)) - g(x) - tDg(x)(y - x).$$

On a donc $\varphi(1) = g(y) - g(x) - Dg(x)(y - x)$ (c'est le terme dont on veut majorer la norme) et $\varphi(0) = 0$. On écrit maintenant que φ est l'intégrale de sa dérivée :

$$\varphi(1) - \varphi(0) = \int_0^1 \varphi'(t)dt = \int_0^1 Dg(x + t(y - x))(y - x) - Dg(x)(y - x)dt.$$

On a donc

$$\begin{aligned}\|\varphi(1)-\varphi(0)| &= \|g(y)-g(x)-Dg(x)(y-x)\| \\ &\leq \int_0^1 \|Dg(x+t(y-x))(y-x)-Dg(x)(y-x)\|dt \\ &\leq \|y-x\|\int_0^1 \|Dg(x+t(y-x))-Dg(x)\|dt.\end{aligned} \tag{2.56}$$

Pour majorer $\|Dg(x+t(y-x))-Dg(x))\|$, on utilise alors le théorème des accroissements finis (parfois aussi appelé "théorème de la moyenne") appliqué à Dg ; de l'inégalité (2.56), on tire donc que pour $x, y \in B(\bar{x}, a)$ et $t \in]0,1[$:

$$\|Dg(x+t(y-x))-Dg(x)\| \leq t\|y-x\| \sup_{c\in B(\bar{x},a)} \|D(Dg)(c)\|_{\mathcal{L}(\mathbb{R}^n,\mathcal{M}_n(\mathbb{R}))}. \tag{2.57}$$

Comme $D(Dg) = D^2g$ est continue par hypothèse, et comme $B(\bar{x}, a)$ est inclus dans un compact, on a

$$a_2 = \sup_{c\in B(\bar{x},a)} \|D(Dg)(c)\|_{\mathcal{L}(\mathbb{R}^n,\mathcal{M}_n(\mathbb{R}))} < +\infty.$$

De plus, $t < 1$ et on déduit de (2.57) que :

$$\|Dg(x+t(y-x)-Dg(x))\| \leq a_2\|y-x\|,$$

et de l'inégalité (2.56) on déduit ensuite que

$$\|g(y)-g(x)-Dg(x)(y-x)\| \leq \int_0^1 a_2\|y-x\|dt\,\|y-x\| = a_2\|y-x\|^2,$$

ce qui termine la preuve. On peut alors appliquer le théorème 2.20 pour obtenir le résultat de convergence locale du théorème 2.19.

Exercice 120 page 167 (Valeurs propres et méthode de Newton) On écrit le système sous la forme $F(x,\lambda) = 0$ où F est une fonction de $\mathbb{R}^{n+1}$ dans $\mathbb{R}^{n+1}$ définie par

$$F(y) = F(x,\lambda) = \begin{pmatrix} Ax-\lambda x \\ x\cdot x - 1 \end{pmatrix},$$

et on a donc

$$DF(\overline{\lambda},\overline{x})(z,\nu) = \begin{pmatrix} Az-\bar{\lambda}z-\nu\bar{x} \\ 2\bar{x}\cdot z \end{pmatrix},$$

Supposons que $DF(\overline{x},\overline{\lambda})(z,\nu) = 0$, on a alors $Az-\bar{\lambda}z-\nu\bar{x} = 0$ et $2\bar{x}\cdot z = 0$. En multipliant la première équation par $\bar{x}$ et en utilisant le fait que A est symétrique, on obtient :

$$z\cdot A\bar{x} - \bar{\lambda}z\cdot\bar{x} - \nu\bar{x}\cdot\bar{x} = 0, \tag{2.58}$$

et comme $A\bar{x} = \bar{\lambda}\bar{x}$ et $\bar{x}\cdot\bar{x} = 1$, ceci entraîne que $\nu = 0$. En revenant à (2.58) on obtient alors que $Ax-\bar{\lambda}x = 0$, c.à.d. que $x \in Ker(A-\bar{\lambda}\mathrm{Id}) = \mathbb{R}\bar{x}$ car $\bar{\lambda}$ est valeur propre simple. Or on a aussi $\bar{x}\cdot z = 0$, donc $z \perp \bar{x}$ ce qui n'est possible que si $z = 0$. On a ainsi montré que $Df(\bar{x},\bar{\lambda})$ est injective, et comme on est en dimension finie, $Df(\bar{x},\bar{\lambda})$ est bijective. Donc, d'après le théorème du cours, la méthode de Newton est localement convergente.

Exercice 126 page 169 (Convergence de la méthode de Newton si $f'(\overline{x}) = 0$) Comme $f''(\overline{x}) \neq 0$, on peut supposer par exemple $f''(\overline{x}) > 0$; par continuité de f'', il existe donc $\eta > 0$ tel que $f'(x) < 0$ si $x \in]\overline{x}-\eta,\overline{x}[$ et $f'(x) > 0$ si $x \in]\overline{x},\overline{x}+\eta[$, et donc f est décroissante sur $]\overline{x}-\eta,\overline{x}[$ (et croissante sur $]\overline{x},\overline{x}+\eta[$).
Supposons $x_0 \in]\overline{x},\overline{x}+\eta[$, alors $f'(x_0) > 0$ et $f''(x_0) > 0$.

On a par définition de la suite $(x_n)_{n\in\mathbb{N}}$,

$$\begin{aligned} f'(x_0)(x_1-x_0) &= -f(x_0) \\ &= f(\overline{x})-f(x_0) \\ &= f'(\xi_0)(\overline{x}-x_0),\ \text{où } \xi_0 \in]\overline{x}, x_0[\end{aligned}$$

Comme f' est strictement croissante sur $]\overline{x}, \overline{x}+\eta[$, on a $f'(\xi_0) < f'(x_0)$ et donc $x_1 \in]\overline{x}, x_0[$.
On montre ainsi par récurrence que la suite $(x_n)_{n\in\mathbb{N}}$ vérifie

$$x_0 > x_1 > x_2 \ldots > x_n > x_{n+1} > \ldots > \overline{x}.$$

La suite $(x_n)_{n\in\mathbb{N}}$ est donc décroissante et minorée, donc elle converge. Soit x sa limite ; comme

$$f'(x_n)(x_{n+1}-x_n) = -f(x_n) \text{ pour tout } n \in \mathbb{N},$$

on a en passant à la limite : $f(x) = 0$, donc $x = \overline{x}$.
Le cas $f''(\overline{x}) < 0$ se traite de la même manière.
Montrons maintenant que la méthode est d'ordre 1. Par définition, la méthode est d'ordre 1 si

$$\frac{\|x_{n+1}-\overline{x}\|}{\|x_n-\overline{x}\|} \to \beta \in \mathbb{R}_+^*.$$

Par définition de la suite $(x_n)_{n\in\mathbb{N}}$, on a :

$$f'(x_n)(x_{n+1}-x_n) = -f(x_n) \tag{2.59}$$

Comme $f \in \mathcal{C}^2(\mathbb{R})$ et $f'(\overline{x}) = 0$, il existe $\xi_n \in]\overline{x}, x_n[$ et $\eta_n \in]\overline{x}, x_n[$ tels que $f'(x_n) = f''(\xi_n)(x_n-\overline{x})$ et $-f(x_n) = -\frac{1}{2}f''(\eta_n)(\overline{x}-x_n)^2$. On déduit donc de (2.59) que

$$\begin{aligned} f''(\xi_n)(x_{n+1}-x_n) &= -\frac{1}{2}f''(\eta_n)(x_n-\overline{x}), \\ \text{soit } f''(\xi_n)(x_{n+1}-\overline{x}) &= (-\frac{1}{2}f''(\eta_n)+f''(\xi_n))(x_n-\overline{x}) \end{aligned}$$

On a donc

$$\frac{\|x_{n+1}-\overline{x}\|}{\|x_n-\overline{x}\|} = |1-\frac{1}{2}\frac{f''(\eta_n)}{f''(\xi_n)}| \to \frac{1}{2} \text{ lorsque } n \to +\infty.$$

La méthode est donc d'ordre 1.
On peut obtenir une méthode d'ordre 2 en appliquant la méthode de Newton à f'.

Exercice 128 page 170 (Variante de la méthode de Newton)

1. On a évidemment $x^{(0)} = x_0 \in I$. Supposons que $x^{(k)} \in I$ et montrons que $x^{(k+1)} \in I$. Par définition, on peut écrire :

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})-f(x_0)}{f'(y)} - \frac{f(x_0)}{f'(y)}.$$

Donc

$$x^{(k+1)} - x_0 = x^{(k)} - x_0 - \frac{f'(\xi_n)(x_n^{(k)}-x_0)-f(x_0)}{f'(y)},\ \text{où } \xi_n \in [x_0, x^{(k)}].$$

On en déduit que

$$x^{(k+1)} - x_0 = (1-\frac{f'(\xi_n)}{f'(y)})(x_n^{(k)}-x_0) - \frac{f(x_0)}{f'(y)}.$$

Ceci entraîne :

$$\begin{aligned} |x^{(k+1)} - x_0| &= \frac{1}{|f'(y)|}|f'(\xi_n) - f'(y)||x^{(k)} - x_0| + \frac{|f(x_0)|}{f'(y)} \\ &\leq \lambda\frac{1}{2\lambda}c + \frac{c}{2\lambda}\lambda = c. \end{aligned}$$

Donc $x^{(k+1)} \in I$.

2. On a :

$$\begin{aligned} x^{(k+1)} - \bar{x} &= x^{(k)} - \bar{x} - \frac{f(x^{(k)}) - f(\bar{x})}{f'(y)} - \frac{f(\bar{x})}{f'(y)}. \\ \text{Donc}|x^{(k+1)} - \bar{x}| &\leq |x^{(k)} - \bar{x}||f'(y) - f'(\eta_n)|\frac{1}{|f'(y)|} \text{ où } \eta_n \in [\bar{x}, x^{(k)}]; \end{aligned}$$

Par hypothèse, on a donc

$$\begin{aligned} |x^{(k+1)} - \bar{x}| &\leq |x^{(k)} - \bar{x}|\frac{1}{2\lambda}\lambda \\ &\leq \frac{c}{2}|x^{(k)} - \bar{x}|. \end{aligned}$$

On en déduit par récurrence que

$$|x^{(k)} - \bar{x}| \leq \frac{c}{2^n}|x^{(0)} - \bar{x}|.$$

Ceci entraîne en particulier que

$$\begin{aligned} x^{(k)} &\to \bar{x} \\ n &\to +\infty. \end{aligned}$$

Il reste à montrer que la convergence est au moins linéaire. On a :

$$\frac{|x^{(k+1)} - \bar{x}|}{|x^{(k)} - \bar{x}|} = |f'(y) - f'(x^{(k)})|\frac{1}{|f'(y)|}$$

$$\text{Donc} \quad \frac{|x^{(k+1)} - \bar{x}|}{|x^{(k)} - \bar{x}|} \underset{n \to +\infty}{\to} |1 - \frac{f'(\bar{x})}{f'(y)}| = \beta \geq 0$$

La convergence est donc au moins linéaire, elle est linéaire si $f'(\bar{x}) \neq f'(y)$ et super–linéaire si $f'(\bar{x}) = f'(y)$.

3. Le fait de remplacer y par $y^{(k)}$ ne change absolument rien à la preuve de la convergence de $x^{(k)}$ vers $\bar{x}$. Par contre, on a maintenant :

$$\begin{aligned} \frac{|x^{(k+1)} - \bar{x}|}{|x^{(k)} - \bar{x}|} &= |f'(y_n) - f'(\eta_n)|\frac{1}{|f'(y_n)|} \\ &= |1 - \frac{f'(\eta_n)}{f'(y_n)}| \end{aligned}$$

Or $f'(\eta_n) \underset{n\to+\infty}{\to} f'(\bar{x})$ et donc si $f'(y_n) \underset{n\to+\infty}{\to} f'(\bar{x})$ la convergence devient superlinéaire.

4. Pour $n \geq 1$, l'algorithme se généralise en :

$$\begin{cases} x^{(0)} = x_0 \\ x^{(k+1)} = x^{(k)} - (DF(y))^{-1}f(x^{(k)}). \end{cases}$$

On a donc

$$x^{(k+1)} - x_0 = x^{(k)} - x_0 - (DF(y))^{-1}(f(x^{(k)}) - f(x_0)) - (DF(y))^{-1}f(x_0). \tag{2.60}$$

On définit $\varphi : \mathbb{R} \to \mathbb{R}^n$ par $\varphi(t) = f(tx^{(k)} + (1-t)x^{(0)})$. On a

$$\varphi'(t) = Df(tx^{(k)} + (1-t)x^{(0)})(x^{(k)} - x^{(0)}).$$

et donc

$$\begin{aligned} f(x^{(k)}) - f(x^{(0)}) &= \varphi(1) - \varphi(0) \\ &= \int_0^1 \varphi'(t)dt \\ &= \int_0^1 Df(tx^{(k)} + (1-t)x^{(0)})(x^{(k)} - x^{(0)})dt. \end{aligned}$$

L'égalité (2.60) s'écrit donc

$$\begin{aligned} x^{(k+1)} - x^{(0)} &= \left(Id - (Df(y))^{-1} \int_0^1 Df(tx^{(k)} + (1-t)x^{(0)})dt \right)(x^{(k)} - x^{(0)}) - (Df(y))^{-1} f(x_0) \\ &= (Df(y))^{-1} \left(\int_0^1 \left(Df(y) - Df(tx^{(k)} + (1-t)x^{(0)} \right) dt \right)(x^{(k)} - x^{(0)}) - (Df(y))^{-1} f(x_0). \end{aligned}$$

On en déduit que :

$$\begin{aligned} \|x^{(k+1)} - x^{(0)}\| \leq \|(Df(y))^{-1}\| \int_0^1 \|Df(y) - Df(tx^{(k)} + (1-t)x^{(0)})\| dt \, \|x^{(k)} - x^{(0)}\| \\ + \|(Df(y))^{-1}\| \|f(x_0)\|. \quad (2.61) \end{aligned}$$

Si on suppose que $x^{(k)} \in I$, alors $tx^{(k)} + (1-t)x^{(0)} \in I$. L'hypothèse (iii) généralisée à la dimension n s'écrit :

$$\|Df(x) - Df(y)\| \leq \frac{1}{2\lambda} \ \forall (x,y) \in I^2,$$

si on suppose de plus que

(ii) $\|f(x_0)\| \leq \dfrac{c}{2\lambda}$ et

(iv) $\|(Df(x))^{-1}\| \leq \lambda \ \forall x \in I$, alors 2.61 donne que

$$\begin{aligned} \|x^{(k+1)} - x^{(0)}\| &\leq \|x^{(k)} - x^{(0)}\| \lambda \frac{1}{2\lambda} + \lambda \frac{c}{2\lambda} \\ &\leq c. \end{aligned}$$

ce qui prouve que $x^{(k+1)} \in I$.

On montre alors de la même manière que $x^{(k)}_{n \to \infty} \to \bar{x}$, (car $\|x^{(k+1)} - \bar{x}\| \leq \dfrac{1}{2}\|x^{(k)} - \bar{x}\|$).

Exercice 129 page 170 (Méthode de Steffensen)

1. Comme $f'(\overline{x}) \neq 0$, il existe $\bar{\alpha} > 0$ tel que f soit strictement monotone sur $B(\overline{x}, \bar{\alpha})$; donc si $f(x) = 0$ et $x \in B(\overline{x}, \bar{\alpha})$ alors $x = \overline{x}$. De plus, comme $x + f(x) \to \overline{x}$ lorsque $x \to \overline{x}$, il existe α tel que si $x \in B(\overline{x}, \alpha)$, alors $f(x + f(x) \in B(\overline{x}, \bar{\alpha})$. Or si $x \in B(x, \alpha)$, on a :$f(x) \neq 0$ si $x \neq \overline{x}$, donc $x + f(x) \neq x$ et comme $x + f(x) \in B(\overline{x}, \bar{\alpha})$ où f est strictement monotone, on a $f(x) \neq f(x + f(x))$ si $x \neq \overline{x}$. On en déduit que si $x_n \in B(\overline{x}, \alpha)$, alors $f(x_n + f(x_n)) \neq f(x_n)$ (si $x_n \neq \overline{x}$) et donc x_{n+1} est défini par $x_{n+1} = \dfrac{(f(x_n))^2}{f(x_n + f(x_n)) - f(x_n)}$. Ceci est une forme de stabilité du schéma).

2. Montrons maintenant que la suite $(x_n)_{n\in\mathbb{N}}$ vérifie :

$$x_{n+1}-\overline{x}=a(x_n)(x_n-\overline{x})^2 \quad \text{si } x_n\neq\overline{x} \ \text{ et } x_0\in B(\overline{x},\alpha),$$

où a est une fonction continue. Par définition de la suite $(x_n)_{n\in\mathbb{N}}$, on a :

$$x_{n+1}-\overline{x}=x_n-\overline{x}-\frac{(f(x_n))^2}{f(x_n+f(x_r))-f(x_n)}. \tag{2.62}$$

Soit $\psi_n:[0,1]\to\mathbb{R}$ la fonction définie par :

$$\psi_n(t)=f(x_n+tf(x_n))$$

On a $\psi_n\in\mathcal{C}^2(]0,1[,\mathbb{R})$, $\psi_n(0)=f(x_n)$ et $\psi_n(1)=f(x_n+f(x_n))$.
On peut donc écrire :

$$f(x_n+f(x_n))-f(x_n)=\psi_n(1)-\psi_n(0)=\int_0^1\psi_n'(t)dt$$

Ceci donne :

$$f(x_n+f(x_n))-f(x_n)=\int_0^1 f'(x_n+tf(x_n))f(x_n)dt$$

On pose maintenant $\xi_n(t)=f'(x_n+tf(x_n))$, et on écrit que $\xi_n(t)=\displaystyle\int_0^t\xi_n'(s)ds+\xi_n(0)$.
On obtient alors :

$$f(x_n+f(x_n))-f(x_n)=f(x_n)\left[f(x_n)\int_0^1\int_0^t f''(x_n+sf(x_n))ds+f'(x_n)\right]. \tag{2.63}$$

Soit $b\in\mathcal{C}(\mathbb{R},\mathbb{R})$ la fonction définie par :

$$b(x)=\int_0^1\left(\int_0^t f''(x+sf(x))ds\right)dt.$$

Comme $f\in\mathcal{C}(\mathbb{R},\mathbb{R})$, on a $b(x)\to\dfrac{1}{2}f''(\overline{x})$ lorsque $x\to\overline{x}$
L'égalité (2.63) s'écrit alors :

$$f(x_n+f(x_n))-f(x_n)=(f(x_n))^2b(x_n)+f(x_n)f'(x_n). \tag{2.64}$$

Comme $x_0\in B(\overline{x},\alpha)$, on a $x_n\in B(\overline{x},\alpha)$ et donc $f(x_n)\neq 0$ si $x_n\neq\overline{x}$.
Donc pour $x_n\neq\overline{x}$, on a grâce à (2.62) et (2.64) :

$$x_{n+1}-\overline{x}=x_n-\overline{x}-\frac{f(x_n)}{f(x_n)b(x_n)+f'(x_n))} \tag{2.65}$$

On a maintenant $-f(x_n)=f(\overline{x})-f(x_n)=\displaystyle\int_0^1\varphi'(t)dt$ où $\varphi\in\mathcal{C}^2(\mathbb{R},\mathbb{R})$ est définie par $\varphi(t)=f(t\overline{x}+(1-t)x_n)$.
Donc

$$-f(x_n)=\int_0^1 f'(t\overline{x}+(1-t)x_n)(\overline{x}-x_n)dt.$$

Soit $\chi\in\mathcal{C}^1(\mathbb{R},\mathbb{R})$ la fonction définie par $\chi(t)=f'(t\overline{x}+(1-t)x_n)$,

on a $\chi(0) = f'(x_n)$ et donc :

$$-f(x_n) = \int_0^1 \left[\int_0^t (f''(s\overline{x} + (1-s)x_n)(\overline{x} - x_n) + f'(x_n))\, ds(\overline{x} - x_n) \right] dt$$

Soit $c \in \mathcal{C}(\mathbb{R}, \mathbb{R})$ la fonction définie par

$$c(x) = \int_0^1 \left(\int_0^t f''(s\overline{x} + (1-s)x) ds \right) dt,$$

on a $c(x) \to \frac{1}{2} f''(\overline{x})$ lorsque $x \to \overline{x}$ et :

$$-f(x_n) = c(x)(\overline{x} - x_n)^2 + f'(x_n)(\overline{x} - x_n)$$

De cette égalité et de (2.65), on obtient :

$$\begin{aligned} x_{n+1} - \overline{x} &= (x_n - \overline{x}) \left[1 + \frac{c(x_n)(x_n - \overline{x}) - f'(x_n)}{f(x_n)b(x_n) + f'(x_n)} \right] \\ &= \frac{(x_n - \overline{x})}{f(x_n)b(x_n) + f'(x_n)} \left(-c(x_n)(\overline{x} - x_n)^2 b(x_n) - f'(x_n)(\overline{x} - x_n)b(x_n) + f'(x_n) \right. \\ &\qquad \left. + c(x_n)(x_n - \overline{x}) - f'(x_n) \right). \end{aligned}$$

On en déduit :

$$(x_{n+1} - \overline{x}) = (x_n - \overline{x})^2 a(x_n) \tag{2.66}$$

où

$$a(x) = \frac{c(x)b(x)(x - \overline{x}) + f'(x)b(x)b + c(x)}{f(x) + f'(x)}$$

La fonction a est continue en tout point x tel que

$$D(x) = f(x)b(x) + f'(x) \neq 0.$$

Elle est donc continue en $\overline{x}$ puisque $D(\overline{x}) = f(\overline{x})b(\overline{x}) + f'(\overline{x}) = f'(\overline{x}) \neq 0$.
De plus, comme f, f' et b sont continues, il existe un voisinage de $\overline{x}$ sur lequel D est non nulle et donc a continue.

3. Par continuité de a, pour tout $\varepsilon > 0$, il existe $\eta_\varepsilon > 0$ tel que si $x \in B(\overline{x}, \eta_\varepsilon)$ alors

$$|a(x) - a(\overline{x})| \leq \varepsilon. \tag{2.67}$$

Calculons

$$\begin{aligned} a(\overline{x}) &= \frac{f'(\overline{x})b(\overline{x}) + c(\overline{x})}{f'(\overline{x})} \\ &= \frac{1}{2} f''(\overline{x}) \frac{1 + f'(\overline{x})}{f'(\overline{x})} = \beta. \end{aligned}$$

Soit $\gamma = \min(\eta_1, \frac{1}{2(\beta+1)})$; si $x \in B(\overline{x}, \gamma)$, alors $|a(x)| \leq \beta + 1$ grâce à (2.67), et $|x - \overline{x}| \leq \frac{1}{2(\beta+1)}$.
On déduit alors de (2.66) que si $x_n \in B(\overline{x}, \gamma)$, alors

$$|x_{n+1} - \overline{x}| \leq \frac{1}{2} |x_n - \overline{x}|.$$

Ceci entraîne d'une part que $x_{n+1} \in B(\overline{x}, \gamma)$ et d'autre part, par récurrence, la convergence de la suite $(x_n)_{n\in\mathbb{N}}$ vers $\overline{x}$.
Il reste à montrer que la convergence est d'ordre 2. Grâce à (2.66), on a :

$$\frac{|x_{n+1} - \overline{x}|}{|x_n - \overline{x}|^2} = |a(x_n)|.$$

Or on a montré à l'étape 3 que a est continue et que $a(x) \to \beta \in \mathbb{R}$. On a donc une convergence d'ordre au moins 2.

Exercice 131 page 171 (Méthode de la sécante)

1. Supposons x_{n-1} et x_n connus.
Pour que x_{n+1} soit bien défini, il faut et il suffit que $f(x_n) \neq f(x_{n-1})$. Or par hypothèse, $f'(\bar{x}) \neq 0$. On en déduit qu'il existe un voisinage de $\bar{x}$ sur lequel f' est monotone, donc bijective. Donc il existe ε_1 tel que si $x_n, x_{n-1} \in]\overline{x} - \varepsilon_1, \overline{x} + \varepsilon_1[$, $x_n \neq x_{n-1}$ et $x_n \neq \overline{x}$, alors $f(x_n) \neq f(x_{n-1})$. De même, toujours par injectivité de f sur $]\overline{x} - \varepsilon_1, \overline{x} + \varepsilon_1[$, on a $f(x_n) \neq 0$.
En choisissant x_0 et x_1 dans l'intervalle $]\overline{x} - \varepsilon_1, \overline{x} + \varepsilon_1[$, on a par une récurrence immédiate que la suite $(x_n)_{n\in\mathbb{N}}$ est bien définie.
Par définition, si $f(x_n) \neq 0$, on a :

$$x_{n+1} - \bar{x} = x_n - \bar{x} - \frac{f(x_n) - f(\bar{x})}{x_n - \bar{x}}(x_n - \bar{x})\frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}.$$

En notant $I(a, b)$ l'intervalle d'extrémités a et b, il existe donc $\theta_n \in I(\bar{x}, x_n)$ et $\zeta_n \in I(x_{n-1}, x_n)$ tels que $x_{n+1} - \bar{x} = (x_n - \bar{x})(1 - \frac{f'(\theta_n)}{f'(\zeta_n)})$, , et donc : $e_{n+1} = |1 - \frac{f'(\theta_n)}{f'(\zeta_n)}|e_n$.
Or f' est continue, il existe ε_2 tel que x_n, $x_{n-1} \in]\overline{x} - \varepsilon_2, \overline{x} + \varepsilon_2[$, alors $1 - \frac{f'(\theta_n)}{f'(\zeta_n)} \leq 1/2$, et donc $e_{n+1} \leq \frac{1}{2}e_n$.
En posant $\varepsilon = \min(\varepsilon_1, \varepsilon_2)$, on a donc par récurrence le fait que si x_0 et x_1 appartiennent à l'intervalle $]\overline{x} - \varepsilon, \overline{x} + \varepsilon[$, la suite $(x_n)_{n\in\mathbb{N}}$ est bien définie et la méthode de la sécante est localement convergente.

2. (a) Par définition,

$$e_{n+1} = e_n - \frac{f(x_n) - f(\overline{x})}{f(x_n) - f(x_{n-1})}(x_n - x_{n-1}).$$

Donc :

$$\begin{aligned}(f(x_n) - f(x_{n-1})e_{n+1} &= e_n f(x_n) - e_n f(x_{n-1}) - f(x_n)e_n + f(x_n)e_{n-1} &(2.68)\\ &= -e_n f(x_{n-1}) + f(x_n)e_{n-1} &(2.69)\\ &= e_n e_{n-1}(\frac{f(x_n)}{e_n} - \frac{f(x_{n-1})}{e_{n-1}}). &(2.70)\end{aligned}$$

Or $\frac{f(x_n)}{e_n} = \frac{f(x_n) - f(\overline{x})}{e_n}$ (resp. $\frac{f(x_{n-1}}{e_{n-1}} = \frac{f(x_{n-1} - f(\overline{x})}{e_{n-1}}$)est la valeur moyenne de f' sur l'intervalle d'extrémités $\overline{x}$, x_n (resp. $\overline{x}$, x_{n-1}). On en déduit que $(f(x_n) - f(x_{n-1})e_{n+1} = e_n e_{n-1}(\mu_n - \mu_{n-1})$, d'où le résultat.

(b) Si $x > \overline{x}$, la fonction μ vérifie :

$$(x - \overline{x})\mu(x) = \int_{\overline{x}}^{x} f'(t)dt,$$

on en déduit que la fonction μ est continue et dérivable et sa dérivée μ' vérifie :

$$(x - \overline{x})\mu'(x) + \mu(x) = f'(x),\ \forall x > \overline{x}.$$

soit encore

$$\mu'(x) = \frac{f'(x) - \mu(x)}{x - \overline{x}}, \ \forall x > \overline{x}. \tag{2.71}$$

Or

$$\begin{aligned} \mu(x) &= \frac{1}{x-\overline{x}}(f(x) - f(\overline{x})) & (2.72)\\ &= \frac{1}{x-\overline{x}}(f(x) - (f(x) + (\overline{x}-x)f'(x) + \frac{1}{2}(\overline{x}-x)^2 f''(x)(\overline{x}-x)^3\varepsilon(x). & (2.73) \end{aligned}$$

On en déduit que

$$\mu(x) = f'(x) + \frac{1}{2}(x-\overline{x})f''(x) + (x-\overline{x})^2\varepsilon(x).$$

Et finalement, en reportant dans (2.71) :

$$\mu'(x) = \frac{1}{2}f''(x) + (x-\overline{x})\varepsilon(x), \ \forall x > \overline{x}. \tag{2.74}$$

On en déduit que μ' admet une limite lorsque x tend vers $\overline{x}$ par valeurs positives. Le même raisonnement pour $x < \overline{x}$ donne le même résultat.
Enfin, comme $f \in C^2(\mathbb{R}, \mathbb{R})$, on peut passer à la limite dans (2.74) et on obtient :

$$\lim_{x\to\overline{x}} \mu'(x) = \frac{1}{2}f''(\overline{x}). \tag{2.75}$$

(c) Par définition, on a

$$M_n = |\frac{\mu(x_n) - \mu(x_{n-1})}{x_n - x_{n-1}} \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}| = \frac{\mu'(\zeta_n)}{f'(\xi_n)},$$

où ζ_n et ξ_n sont compris entre x_{n-1} et x_n (par le théorème des accroissements finis). Comme la suite $(x_n)_{n\in\mathbb{N}}$ tend vers $\overline{x}$, comme f' est continue et grâce à (2.75), on a :

$$\lim_{n\to+\infty} M_n = \frac{1}{2}\frac{f''(\overline{x})}{f'(\overline{x})}.$$

Notons que cette limite est finie car $f'(\overline{x}) \neq 0$ par hypothèse. On en conclut que la suite $(M_n)_{n\in\mathbb{N}}$ est bornée.

3. (a) La relation à démontrer est vérifiée pour $n = 0$ et $n = 1$. Supposons-la vérifiée jusqu'au rang n. On a par définition : $a_{n+1} = a_n a_{n-1} \geq Me_n Me_{n-1}$. Or par la question 2a, $e_n e_{n-1} = M_n e_{n+1} \leq Me_{n+1}$. On en déduit que la relation est encore vérifiée au rang $n+1$.

(b) Par définition, $a_i = Me_i = M(x_i - \overline{x})$, pour $i = 0, 1$, donc si $x_0, x_1 \in]\overline{x}-\varepsilon_1, \overline{x}+\varepsilon_1[$ avec $\varepsilon_1 < 1/M$, alors $a_0 < 1$ et $a_1 < 1$. On en déduit alors facilement par récurrence que la suite $a_n < 1$, et donc que la suite $(a_n)_{n\in\mathbb{N}}$ est strictement décroissante. Elle converge donc vers une limite $\bar{a}$ qui vérifie $\bar{a} = \bar{a}^2$ et $\bar{a} < 1$. On en déduit que la limite est nulle.

(c) On pose $b_n = \ln a_n$ on a donc

$$b_{n+1} = b_n + b_{n-1}, \ \forall n \geq 1 \tag{2.76}$$

L'ensemble de suites $(b_n)_{n\in\mathbb{N}}$ vérifiant (2.76) est un espace vectoriel de dimension 2. Pour trouver une base de cet espace vectoriel, on cherche des éléments de cet espace sous la forme $b_n = r^n$, $n \geq 0$. Une telle suite vérifie (2.76) si et seulement si $r^2 = r + 1$, c.à.d. $r = \frac{1\pm\sqrt{5}}{2}$. Si la suite $(b_n)_{n\in\mathbb{N}}$ vérifie (2.76), il existe donc $C \in \mathbb{R}$ et $D \in \mathbb{R}$ tels que

$$b_n = \ln(a_n) = C(\frac{1+\sqrt{5}}{2})^n + D(\frac{1-\sqrt{5}}{2})^n.$$

On en déduit que $a_n \leq \alpha\beta^{d^n}$, avec $d = \frac{1+\sqrt{5}}{2}$, $\alpha = e^{|D|}$ et $\beta = e^C$. Notons qu'on a bien $0 < \beta < 1$ car $C < 0$ puisque $\ln(a_n) < 0$, pour tout $n \in \mathbb{N}$.

(d) Par la question 2(c) et l'hypothèse $f"(\overline{x}) \neq 0$, on déduit que $\overline{M} > 0$. Comme $e_{n+1} = M_n e_n e_{n-1}$, on a $\ln e_{n+1} = \ln M_n + \ln e_n + \ln e_{n-1}$; si on pose $\beta_n = \ln e_{n+1} - d \ln e_n$ (pour $n \geq 0$, on a donc

$$\begin{aligned} \beta_n &= (1-d)\ln e_n + \ln e_{n-1} + \ln M_n \\ &= (1-d)(\beta_{n-1} + d\ln e_{n-1}) + \ln e_{n-1} + \ln M_n \\ &= (1-d)(\beta_{n-1} + (1-d)d\ln e_{n-1}) + \ln e_{n-1} + \ln M_n. \end{aligned}$$

Or $(1-d)d = -1$ car d est racine de l'équation : $d^2 - d - 1 = 0$. On obtient donc finalement

$$\beta_n = (1-d)\beta_{n-1} + \ln M_n.$$

On pose maintenant $\beta_n = C_n(1-d)^n$ (obtenu par "variation de la constante" C pour la solution de l'équation homogène $\beta_n = (1-d)\beta_{n-1}$). On obtient alors

$$C_n(1-d)^n = (1-d)C_{n-1}(1-d)^{n-1} + \ln M_n.$$

Ceci entraîne :

$$C_n = C_{n-1} + \frac{\ln M_n}{(1-d)^n}.$$

Donc

$$C_n = C_0 + \sum_{p=1}^{n} \frac{\ln M_p}{(1-d)^p},$$

et comme la suite $(M_n)_{n\in\mathbb{N}}$ est bornée, la série de terme général $\frac{\ln M_p}{(1-d)^p}$ est convergente. Comme $(1-d)^n$ tend vers 0 lorsque n tend vers l'infini, on en déduit que $\beta_n \to 0$. On a donc $\ln e_{n+1} - d\ln e_n \to 0$, i.e. $\frac{e_{n+1}}{e_n^d} \to 1$ lorsque $n \to +\infty$.

(e) L'ordre de convergence de la méthode de la sécante est $d = \frac{1+\sqrt{5}}{2} < 2$, donc plus faible que l'ordre de convergence de la méthode de Newton.

Chapitre 3

Optimisation

3.1 Définitions et rappels

3.1.1 Extrema, points critiques et points selle.

L'objectif de ce chapitre est de rechercher des extrema, c'est-à-dire des minima ou des maxima d'une fonction $f \in C(\mathbb{R}^n, \mathbb{R})$ avec ou sans contrainte. Notons que la recherche d'un minimum ou d'un maximum implique que l'on ait une relation d'ordre, pour pouvoir comparer les valeurs prises par f. On insiste donc bien sur le fait que la fonction f est à valeurs dans $\mathbb{R}$ (et non pas $\mathbb{R}^n$, comme dans le chapitre précédent). Rappelons tout d'abord quelques définitions du cours de calcul différentiel.

Définition 3.1 (Extremum d'une fonction)**.** *Soit E un espace vectoriel normé et $f : E \to \mathbb{R}$.*
On dit que $\bar{x}$ est un minimum local de f s'il existe un voisinage V de $\bar{x}$ tel que

$$f(\bar{x}) \leq f(x), \forall x \in V.$$

De même, on dit que $\bar{x}$ est un maximum local de f s'il existe un voisinage V de $\bar{x}$ tel que

$$f(\bar{x}) \geq f(x), \forall x \in V.$$

On dit que $\bar{x}$ est un extremum local de f si c'est un minimum local ou un maximum local.
On dit que $\bar{x}$ est un minimum global de f si

$$f(\bar{x}) \leq f(x), \forall x \in E.$$

De même, on dit que $\bar{x}$ est un maximum global de f si

$$f(\bar{x}) \geq f(x), \forall x \in E.$$

On dit que $\bar{x}$ est un extremum global de f si c'est un minimum global ou un maximum global.

Le problème d'optimisation sans contrainte s'écrit :

$$\begin{cases} \text{Trouver } \bar{x} \in \mathbb{R}^n \text{ tel que :} \\ f(\bar{x}) \leq f(y), \quad \forall y \in \mathbb{R}^n. \end{cases} \tag{3.1}$$

Le problème d'optimisation avec contrainte s'écrit :

$$\begin{cases} \text{Trouver } \bar{x} \in K \text{tel que :} \\ f(\bar{x}) \leq f(y), \quad \forall y \in K. \end{cases} \tag{3.2}$$

où $K \subset \mathbb{R}^n$ et $K \neq \mathbb{R}^n$ L'ensemble K où l'on recherche la solution est donc l'ensemble qui représente les contraintes. Par exemple, si l'on cherche un miminum d'une fonction f de $\mathbb{R}$ dans $\mathbb{R}$ et que l'on demande que les points qui réalisent ce minimum soient positifs, on aura $K = \mathbb{R}_+$.
Si $\bar{x}$ est solution du problème (3.1), on dit que $\bar{x} \in \arg\min_{\mathbb{R}^n} f$, et si $\bar{x}$ est solution du problème (3.2), on dit que $\bar{x} \in \arg\min_K f$.
Vous savez déjà que si un point $\bar{x}$ réalise le minimum d'une fonction f dérivable de $\mathbb{R}$ dans $\mathbb{R}$, alors $f'(\overline{x}) = 0$. On dit que c'est un point critique (voir définition 3.2). La réciproque est évidemment fausse : la fonction $x \mapsto x^3$ est dérivable sur $\mathbb{R}$, et sa dérivée s'annule en 0 qui est donc un point critique, mais 0 n'est pas un extremum (c'est un point d'inflexion). Nous verrons plus loin que de manière générale, lorsque la fonctionnelle f est différentiable, les extrema sont des points critiques de f, au sens où ils annulent le gradient.

Définition 3.2 (Point critique)**.** *Soit E un espace vectoriel normé et $f : E \to \mathbb{R}$ différentiable. On dit que $x \in E$ est un point critique de f si $Df(x) = 0$.*

Pour illustrer un cas de point critique qui n'est pas un maximum ni un minimum, prenons un exemple en dimension 2, avec

$$f(x_1, x_2) = x_1^2 - x_2^2.$$

On a alors

$$Df(x_1, x_2)(h_1, h_2) = 2(x_1 h_1 - x_2 h_2) \text{ et } Df(0,0) = 0.$$

Le point $(0,0)$ est donc un point critique de f. Si on trace la surface $x \mapsto x_1^2 - x_2^2$, on se rend compte que le point $(0,0)$ est minimal dans une direction et maximal dans une direction indépendante de la première. C'est ce qu'on appelle un point selle

Définition 3.3 (Point selle)**.** *Soit E un espace vectoriel normé et $f : E \to \mathbb{R}$. On dit que $\bar{x}$ est un point selle de f s'il existe F et G des sous espaces vectoriels de E tels que $E = F \oplus G$ et un voisinage V de $\bar{x}$ tel que*

$$f(\bar{x} + z) \leq f(\bar{x}), \forall z \in F \,; \bar{x} + z \in V,$$
$$f(\bar{x} + z) \geq f(\bar{x}), \forall z \in G \,; \bar{x} + z \in V.$$

3.1.2 Convexité

Définition 3.4 (Convexité)**.** *Soit E un espace vectoriel (sur $\mathbb{R}$) et $f : E \to \mathbb{R}$. On dit que f est convexe si*

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) \text{ pour tout } (x,y) \in E^2 \text{ et } t \in [0,1].$$

On dit que f est strictement convexe si

$$f(tx + (1-t)y) < tf(x) + (1-t)f(y) \text{ pour tout } (x,y) \in E^2 \text{ t.q. } x \neq y \text{ et } t \in]0,1[.$$

Proposition 3.5 (Première caractérisation de la convexité)**.** *Soit E un espace vectoriel normé (sur $\mathbb{R}$) et $f \in C^1(E, \mathbb{R})$ alors :*

1. *la fonction f est convexe si et seulement si $f(y) \geq f(x) + Df(x)(y-x)$, pour tout couple $(x,y) \in E^2$,*
2. *la fonction f est strictement convexe si et seulement si $f(y) > f(x) + Df(x)(y-x)$ pour tout couple $(x,y) \in E^2$ tel que $x \neq y$.*

DÉMONSTRATION – *Démonstration de 1.*
($\Rightarrow$) Supposons que f est convexe : soit $(x,y) \in E^2$; on veut montrer que $f(y) \geq f(x) + Df(x)(y-x)$. Soit $t \in [0,1]$, alors $f(ty + (1-t)x) \leq tf(y) + (1-t)f(x)$ grâce au fait que f est convexe. On a donc :

$$f(x + t(y-x)) - f(x) \leq t(f(y) - f(x)). \tag{3.3}$$

Comme f est différentiable, $f(x + t(y-x)) = f(x) + Df(x)(t(y-x)) + t\varepsilon(t)$ où $\varepsilon(t)$ tend vers 0 lorsque t tend vers 0. Donc en reportant dans (3.3),

$$\varepsilon(t) + Df(x)(y-x) \leq f(y) - f(x), \quad \forall t \in]0,1[.$$

En faisant tendre t vers 0, on obtient alors :

$$f(y) \geq Df(x)(y-x) + f(x).$$

($\Leftarrow$) Montrons maintenant la réciproque : Soit $(x,y) \in E^2$, et $t \in]0,1[$ (pour $t = 0$ ou $= 1$ on n'a rien à démontrer). On veut montrer que $f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$. On pose $z = tx + (1-t)y$. On a alors par hypothèse :

$$\begin{array}{ll} f(y) & \geq f(z) + Df(z)(y-z), \\ \text{et } f(x) & \geq f(z) + Df(z)(x-z). \end{array}$$

En multipliant la première inégalité par $1-t$, la deuxième par t et en les additionnant, on obtient :

$$\begin{array}{ll} (1-t)f(y) + tf(x) & \geq f(z) + (1-t)Df(z)(y-z) + tDf(z)(x-z) \\ (1-t)f(y) + tf(x) & \geq f(z) + Df(z)((1-t)(y-z) + t(x-z)). \end{array}$$

Et comme $(1-t)(y-z) + t(x-z) = 0$, on a donc $(1-t)f(y) + tf(x) \geq f(z) = f(tx + (1-t)y)$.

Démonstration de 2
($\Rightarrow$) On suppose que f est strictement convexe, on veut montrer que $f(y) > f(x) + Df(x)(y-x)$ si $y \neq x$. Soit donc $(x,y) \in E^2$, $x \neq y$. On pose $z = \frac{1}{2}(y-x)$, et comme f est convexe, on peut appliquer la partie 1. du théorème et écrire que $f(x+z) \geq f(x) + Df(x)(z)$. On a donc $f(x) + Df(x)(\frac{y-x}{2}) \leq f(\frac{x+y}{2})$. Comme f est strictement convexe, ceci entraîne que $f(x) + Df(x)(\frac{y-x}{2}) < \frac{1}{2}(f(x) + f(y))$, d'où le résultat.
($\Leftarrow$) La méthode de démonstration est la même que pour le 1. ∎

Proposition 3.6 (Seconde caractérisation de la convexité)**.** *Soit $E = \mathbb{R}^n$ et $f \in C^2(E, \mathbb{R})$. Soit $H_f(x)$ la hessienne de f au point x, i.e. $(H_f(x))_{i,j} = \partial^2_{i,j} f(x)$. Alors*

1. *f est convexe si et seulement si $H_f(x)$ est symétrique et positive pour tout $x \in E$ (c.à.d. $H_f(x)^t = H_f(x)$ et $H_f(x)y \cdot y \geq 0$ pour tout $y \in \mathbb{R}^n$)*
2. *f est strictement convexe si $H_f(x)$ est symétrique définie positive pour tout $x \in E$. (Attention la réciproque est fausse.)*

DÉMONSTRATION – *Démonstration de 1.*
($\Rightarrow$) Soit f convexe, on veut montrer que $H_f(x)$ est symétrique positive. Il est clair que $H_f(x)$ est symétrique car $\partial^2_{i,j} f = \partial^2_{j,i} f$ car f est C^2. Par définition, $H_f(x) = D(\nabla f(x))$ et $\nabla f \in C^1(\mathbb{R}^n, \mathbb{R}^n)$. Soit $(x,y) \in E^2$, comme f est convexe et de classe C^1, on a, grâce à la proposition 3.5 :

$$f(y) \geq f(x) + \nabla f(x) \cdot (y-x). \tag{3.4}$$

Soit $\varphi \in C^2(\mathbb{R}, \mathbb{R})$ définie par $\varphi(t) = f(x + t(y-x))$. Alors :

$$f(y) - f(x) = \varphi(1) - \varphi(0) = \int_0^1 \varphi'(t)dt = [\varphi'(t)(t-1)]_0^1 - \int_0^1 \varphi''(t)(t-1)dt,$$

c'est–à dire : $f(y) - f(x) = \varphi'(0) + \int_0^1 \varphi''(t)(1-t)dt$. Or $\varphi'(t) = \nabla f(x + t(y-x)) \cdot (y-x)$, et

$$\varphi''(t) = D(\nabla f(x + t(y-x))(y-x) \cdot (y-x) = H_f(x + t(y-x))(y-x) \cdot (y-x).$$

On a donc :

$$f(y) - f(x) = \nabla f(x)(y-x) + \int_0^1 H_f(x + t(y-x))(y-x) \cdot (y-x)(1-t)dt. \tag{3.5}$$

Les inégalités (3.4) et (3.5) entraînent : $\int_0^1 H_f(x+t(y-x))(y-x)\cdot(y-x)(1-t)dt \geq 0\ \forall x,y\in E$. On a donc :

$$\int_0^1 H_f(x+tz)z\cdot z(1-t)dt \geq 0 \quad \forall x,\forall z\in E. \tag{3.6}$$

En fixant $x\in E$, on écrit (3.6) avec $z=\varepsilon y$, $\varepsilon>0$, $y\in\mathbb{R}^n$. On obtient :

$$\varepsilon^2\int_0^1 H_f(x+t\varepsilon y)y\cdot y(1-t)dt\geq 0 \quad \forall x,y\in E,\ \forall\varepsilon>0, \text{ et donc :}$$

$$\int_0^1 Hf(x+t\varepsilon y)y\cdot y(1-t)dt\geq 0 \quad \forall\varepsilon>0.$$

Pour $(x,y)\in E^2$ fixé, $H_f(x+t\varepsilon y)$ tend vers $H_f(x)$ uniformément lorsque $\varepsilon\to 0$, pour $t\in[0,1]$. On a donc :

$$\int_0^1 H_f(x)y\cdot y(1-t)dt\geq 0, \text{ c.à.d. } \frac{1}{2}H_f(x)y\cdot y\geq 0.$$

Donc pour tout $(x,y)\in(\mathbb{R}^n)^2$, $H_f(x)y\cdot y\geq 0$ donc $H_f(x)$ est positive.

($\Leftarrow$) Montrons maintenant la réciproque : On suppose que $H_f(x)$ est positive pour tout $x\in E$. On veut démontrer que f est convexe ; on va pour cela utiliser la proposition 3.5 et montrer que : $f(y)\geq f(x)+\nabla f(x)\cdot(y-x)$ pour tout $(x,y)\in E^2$. Grâce à (3.5), on a :

$$f(y)-f(x)=\nabla f(x)\cdot(y-x)+\int_0^1 H_f(x+t(y-x))(y-x)\cdot(y-x)(1-t)dt.$$

Or $H_f(x+t(y-x))(y-x)\cdot(y-x)\geq 0$ pour tout couple $(x,y)\in E^2$, et $1-t\geq 0$ sur $[0,1]$. On a donc $f(y)\geq f(x)+\nabla f(x)\cdot(y-x)$ pour tout couple $(x,y)\in E^2$. La fonction f est donc bien convexe.

Démonstration de 2.

($\Leftarrow$) On suppose que $H_f(x)$ est strictement positive pour tout $x\in E$, et on veut montrer que f est strictement convexe. On va encore utiliser la caractérisation de la proposition 3.5. Soit donc $(x,y)\in E^2$ tel que $y\neq x$. Alors :

$$f(y)=f(x)+\nabla f(x)\cdot(y-x)+\int_0^1 \underbrace{H_f(x+t(y-x))(y-x)\cdot(y-x)}_{>0 \text{ si } x\neq y}\ \underbrace{(1-t)}_{\neq 0 \text{ si } t\in]0,1[}\ dt.$$

Donc $f(y)>f(x)+\nabla f(x)(y-x)$ si $x\neq y$, ce qui prouve que f est strictement convexe. ∎

Contre-exemple Pour montrer que la réciproque de 2. est fausse, on propose le contre-exemple suivant : Soit $n=1$ et $f\in C^2(\mathbb{R},\mathbb{R})$, on a alors $H_f(x)=f''(x)$. Si f est la fonction définie par $f(x)=x^4$, alors f est strictement convexe mais $f''(0)=0$.

3.1.3 Exercices (extrema, convexité)

Exercice 133 (Vrai / faux)**.** *corrigé en page 194*

1. L'application $x\mapsto\|x\|_\infty$ est convexe sur $\mathbb{R}^2$.
2. L'application $x\mapsto\|x\|_\infty$ est strictement convexe sur $\mathbb{R}^2$.
3. L'application de $\mathbb{R}^2$ dans $\mathbb{R}$ définie par $F(x,y)=x^2-2xy+3y^2+y$ admet un unique minimum.
4. Soit $A\in\mathcal{M}_{n,m}(\mathbb{R})$, $b\in\mathbb{R}^n$, l'application $x\mapsto\|Ax-b\|_2$ admet un unique minimum.

Exercice 134 (Minimisation dans $\mathbb{R}$)**.** *Corrigé en page 194*

On considère les fonctions définies de $\mathbb{R}$ dans $\mathbb{R}$ par $f_0(x)=x^2$, $f_1(x)=x^2(x-1)^2$, $f_2(x)=|x|$, $f_3(x)=\cos x$, $f_4(x)=|\cos x|$, $f_5(x)=e^x$. On pose $K=[-1,1]$. Pour chacune de ces fonctions, répondre aux questions suivantes :

1. Etudier la différentiabilité et la (stricte) convexité éventuelles de la fonction, ; donner l'allure de son graphe.
2. La fonction admet elle un minimum global sur $\mathbb{R}$; ce minimum est-il unique ? Le cas échéant, calculer ce minimum.

3. La fonction admet elle un minimum sur K ; ce minimum est-il unique ? Le cas échéant, calculer ce minimum.

Exercice 135 (Fonctions quadratiques)**.**

1. Montrer que la fonction f de $\mathbb{R}^2$ dans $\mathbb{R}$ définie par $f(x,y) = x^2 + 4xy + 3y^2$ n'admet pas de minimum en $(0,0)$.
2. Trouver la matrice symétrique S telle que $f(x) = x^t Sx$, pour $f_1(x) = 2(x_1^2 + x_2^2 + x_3^2 - x_1x_2 - x_2x_3)$, puis pour $f_2(x) = 2(x_1^2 + x_2^2 + x_3^2 - x_1x_2 - x_1x_3 - x_2x_3)$ Etudier la convexité des fonctions f_1 et f_2.
3. Calculer les matrices hessiennes de g_1 et g_2 définies par : $g_1(x,y) = \frac{1}{4}x^4 + x^2y + y^2$ et $g_2(x,y) = x^3 + xy - x$ et étudier la convexité de ces deux fonctions.

Exercice 136 (Convexité et continuité)**.** *Suggestions en page 193.*

1. Soit $f : \mathbb{R} \to \mathbb{R}$ une fonction convexe.
   (a) Montrer que f est continue.
   (b) Montrer que f est localement lipschitzienne.
2. Soit $n \geq 1$ et $f : \mathbb{R}^n \to \mathbb{R}$. On suppose que f est convexe.
   (a) Montrer f est bornée supérieurement sur les bornés (c'est-à-dire : pour tout $R > 0$, il existe m_R t.q. $f(x) \leq m_R$ si la norme de x est inférieure ou égale à R).
   (b) Montrer que f est continue.
   (c) Montrer que f est localement lipschitzienne.
   (d) On remplace maintenant $\mathbb{R}^n$ par E, e.v.n. de dimension finie. Montrer que f est continue et que f est localement lipschitzienne.
3. Soient E un e.v.n. de dimension infinie et $f : E \to \mathbb{R}$. On suppose que f est convexe.
   (a) On suppose, dans cette question, que f est bornée supérieurement sur les bornés. Montrer que f est continue.
   (b) Donner un exemple d'e.v.n. (noté E) et de fonction convexe $f : E \to \mathbb{R}$ t.q. f soit non continue.

Suggestions pour les exercices

Exercice 136 page 193 (Convexité et continuité)

1. (a) Pour montrer la continuité en 0, soit $x \neq 0$, $|x| < 1$. On pose $a = \text{sgn}(x)$ $(= \frac{x}{|x|})$. Ecrire x comme une combinaison convexe de 0 et a et écrire 0 comme une combinaison convexe de x et $-a$. En déduire une majoration de $|f(x) - f(0)|$.
   (b) Utiliser la continuité de f et la majoration précédente.
2. (a) Faire une récurrence sur n et pour $x = (x_1, y)^t$ avec $-R < x_1 < R$ et $y \in \mathbb{R}^{n-1}$ $(n > 1)$, majorer $f(x)$ en utilisant $f(+R, y)$ et $f(-R, y)$.
   (b) Reprendre le raisonnement fait pour $n = 1$.
   (c) Se ramener à $E = \mathbb{R}^n$.
3. (a) reprendre le raisonnement fait pour $E = \mathbb{R}$.
   (b) On pourra, par exemple choisir $E = C([0,1], \mathbb{R})\ldots$

Corrigés des exercices

Exercice 133 page 192 (Vrai/faux)

1. Vrai.
2. Faux. L'application est convexe mais pas strictement convexe. Si on fixe $v_1 = (1,0)$ et $v_2 = (1,1)$, alors pour tout $t \in [0,1]$,
$$\|tv_1 + (1-t)v_2\|_\infty = \|(1, 1-t)\|_\infty = 1 = t\|v_1\|_\infty + (1-t)\|v_2\|_\infty.$$
3. Vrai. Posons $X = (x,y)^t$, on reconnait la fonctionnelle quadratique $F(x,y) = \frac{1}{2}(AX,X) - (b,X)$ avec $A = \begin{bmatrix} 1 & -1 \\ -1 & 3 \end{bmatrix}$ et $b = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. La matrice A une matrice symétrique définie positive. Le cours nous dit alors que F admet un unique minimum.
4. Contre-exemple. Soit $A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ et $b = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. Alors $\|Ax - b\|_2 = x_1^2$ et toute la droite $x_1 = 0$ réalise le minimum de f.

Exercice 134 page 192 (Minimisation dans $\mathbb{R}$)

1. La fonction f_0 est différentiable sur $\mathbb{R}$, et strictement convexe. Elle admet un minimum unique sur $\mathbb{R}$ et sur K et son minimum est réalisé en $\bar{x} = 0$, et on a $f_0(\bar{x}) = 0$.
2. La fonction f_1 est différentiable sur $\mathbb{R}$, et non convexe. La fonction f_1 admet un maximum local en $\bar{x} = \frac{1}{2}$, et on a $f(\bar{x}) = \frac{1}{16}$. Elle admet un minimum global non unique, réalisé en 0 et 1, et dont la valeur est 0.
3. La fonction f_2 est différentiable sur $\mathbb{R} \setminus \{0\}$, et convexe, mais pas strictement convexe. La fonction f_2 admet un minimum unique sur $\mathbb{R}$ et sur K et son minimum est réalisé en $\bar{x} = 0$, et on a $f_2(\bar{x}) = 0$, mais la fonction f_2 n'est pas différentiable en 0.
4. La fonction f_3 est différentiable sur $\mathbb{R}$, et non convexe. La fonction f_3 admet un minimum, qui est -1, et qui n'est pas unique car il est réalisé pour les points $(2k+1)\pi$, $k \in \mathbb{Z}$.
5. La fonction f_4 est différentiable sur $\mathbb{R}$, et non convexe. La fonction f_4 admet un minimum, qui est 0, et qui n'est pas unique car il est réalisé pour les points $(2k+1)\frac{\pi}{2}$, $k \in \mathbb{Z}$. La fonction f_4 n'est pas différentiable en ces points.
6. La fonction f_5 est différentiable et strictement convexe. Elle n'admet pas de minimum. On a $f_5(x) \to 0$ lorsque $x \to -\infty$ mais $f(x) > 0$ pour tout $x \in \mathbb{R}$.

3.2 Optimisation sans contrainte

3.2.1 Définition et condition d'optimalité

Soit $f \in C(E, \mathbb{R})$ et E un espace vectoriel normé. On cherche $\bar{x}$ minimum global de f, c.à.d. :

$$\bar{x} \in E \text{ tel que } f(\bar{x}) \leq f(y) \ \forall y \in E, \tag{3.7}$$

ou un minimum local, c.à.d. :

$$\bar{x} \text{ tel que } \exists \alpha > 0 \ \ f(\bar{x}) \leq f(y) \ \forall y \in B(\bar{x}, \alpha). \tag{3.8}$$

Proposition 3.7 (Condition nécessaire d'optimalité)**.**
Soit E un espace vectoriel normé, et soient $f \in C(E, \mathbb{R})$, et $\bar{x} \in E$ tel que f est différentiable en $\bar{x}$. Si $\bar{x}$ est solution de (3.8) alors $Df(\bar{x}) = 0$.

DÉMONSTRATION – Supposons qu'il existe $\alpha > 0$ tel que $f(\bar{x}) \leq f(y)$ pour tout $y \in B(\bar{x}, \alpha)$. Soit $z \in E \setminus \{0\}$; si $|t| < \frac{\alpha}{\|z\|}$, on a $\bar{x} + tz \in B(\bar{x}, \alpha)$ (où $B(\bar{x}, \alpha)$ désigne la boule ouverte de centre $\bar{x}$ et de rayon α) et on a donc $f(\bar{x}) \leq f(\bar{x} + tz)$. Comme f est différentiable en $\bar{x}$, on a :

$$f(\bar{x} + tz) = f(\bar{x}) + Df(\bar{x})(tz) + |t|\varepsilon_z(t),$$

où $\varepsilon_z(t) \to 0$ lorsque $t \to 0$. On a donc $f(\bar{x}) + tDf(\bar{x})(z) + |t|\varepsilon_z(t) \geq f(\bar{x})$. Et pour $\frac{\alpha}{\|z\|} > t > 0$, on a $Df(\bar{x})(z) + \varepsilon_z(t) \geq 0$. En faisant tendre t vers 0, on obtient que

$$Df(\bar{x})(z) \geq 0, \ \forall z \in E.$$

On a aussi $Df(\bar{x})(-z) \geq 0$ pour tout $z \in E$, et donc $-Df(\bar{x})(z) \geq 0$ pour tout $z \in E$. On en conclut que $Df(\bar{x}) = 0$. ■

Remarque 3.8 (Condition non suffisante)**.** Attention, la proposition précédente donne une condition nécessaire mais non suffisante. En effet, $Df(\bar{x}) = 0$ n'entraîne pas que f atteigne un minimum (ou un maximum) même local, en $\bar{x}$. Prendre par exemple $E = \mathbb{R}$, $\overline{x} = 0$ et la fonction f définie par : $f(x) = x^3$ pour s'en convaincre.

3.2.2 Résultats d'existence et d'unicité

Théorème 3.9 (Existence)**.** *Soit $E = \mathbb{R}^n$ et $f : E \to \mathbb{R}$ une application telle que*

(i) *f est continue,*

(ii) *$f(x) \to +\infty$ quand $\|x\| \to +\infty$.*

Alors il existe $\bar{x} \in \mathbb{R}^n$ tel que $f(\bar{x}) \leq f(y)$ pour tout $y \in \mathbb{R}^n$.

DÉMONSTRATION – La condition (ii) peut encore s'écrire

$$\forall A \in \mathbb{R}, \quad \exists R \in \mathbb{R}; \|x\| \geq R \Rightarrow f(x) \geq A. \tag{3.9}$$

On écrit (3.9) avec $A = f(0)$. On obtient alors :

$$\exists R \in \mathbb{R} \text{ tel que } \|x\| \geq R \Rightarrow f(x) \geq f(0).$$

On en déduit que $\inf_{\mathbb{R}^n} f = \inf_{B_R} f$, où $B_R = \{x \in \mathbb{R}^n; |x| \leq R\}$. Or, B_R est un compact de $\mathbb{R}^n$ et f est continue donc il existe $\bar{x} \in B_R$ tel que $f(\bar{x}) = \inf_{B_R} f$ et donc $f(\bar{x}) = \inf_{\mathbb{R}^n} f$. ■

Remarque 3.10.

1. Le théorème est faux si E est un espace de Banach (c'est-à-dire un espace vectoriel normé complet) de dimension infinie car, dans ce cas, la boule fermée B_R n'est pas compacte.
2. L'hypothèse (ii) du théorème peut être remplacée par

$$(ii)' \qquad \exists b \in \mathbb{R}^n, \exists R > 0 \text{ tel que } \|x\| \geq R \Rightarrow f(x) \geq f(b).$$

3. Sous les hypothèses du théorème il n'y a pas toujours unicité de $\bar{x}$ même dans le cas $n = 1$, prendre pour s'en convaincre la fonction f définie de $\mathbb{R}$ dans $\mathbb{R}$ par $f(x) = x^2(x-1)(x+1)$.

Théorème 3.11 (Condition suffisante d'unicité)**.** *Soit E un espace vectoriel normé et $f : E \to \mathbb{R}$ strictement convexe alors il existe au plus un $\bar{x} \in E$ tel que $f(\bar{x}) \leq f(y), \forall y \in E$.*

DÉMONSTRATION – Soit f strictement convexe, supposons qu'il existe $\bar{x}$ et $\bar{\bar{x}} \in E$ tels que $f(\bar{x}) = f(\bar{\bar{x}}) = \inf_{\mathbb{R}^n} f$. Comme f est strictement convexe, si $\bar{x} \neq \bar{\bar{x}}$ alors

$$f(\frac{1}{2}\bar{x} + \frac{1}{2}\bar{\bar{x}}) < \frac{1}{2}f(\bar{x}) + \frac{1}{2}f(\bar{\bar{x}}) = \inf_{\mathbb{R}^n} f,$$

ce qui est impossible ; donc $\bar{x} = \bar{\bar{x}}$. ∎

Ce théorème ne donne pas l'existence. Par exemple dans le cas $n = 1$ la fonction f définie par $f(x) = e^x$ n'atteint pas son minimumn ; en effet, $\inf_{\mathbb{R}^n} f = 0$ et $f(x) \neq 0$ pour tout $x \in \mathbb{R}$, et pourtant f est strictement convexe. Par contre, si on réunit les hypothèses des théorèmes 3.9 et 3.11, on obtient le résultat d'existence et unicité suivant :

Théorème 3.12 (Existence et unicité)**.** *Soit $E = \mathbb{R}^n$, et soit $f : E \to \mathbb{R}$. On suppose que :*

(i) f continue,

(ii) $f(x) \to +\infty$ quand $\|x\| \to +\infty$,

(iii) f est strictement convexe ;

alors il existe un unique $\bar{x} \in \mathbb{R}^n$ tel que $f(\bar{x}) = \inf_{\mathbb{R}^n} f$.

L'hypothèse (i) du théorème 3.12 est en fait inutile car une fonction convexe de $\mathbb{R}^n$ dans $\mathbb{R}$ est nécessairement continue.
Nous donnons maintenant des conditions suffisantes d'existence et d'unicité du minimum pour une fonction de classe C^1.

Proposition 3.13 (Condition suffisante d'existence et unicité)**.** *Soit $f \in C^1(\mathbb{R}^n, \mathbb{R})$. On suppose que :*

$$\exists \alpha > 0; (\nabla f(x) - \nabla f(y)) \cdot (x - y) \geq \alpha |x - y|^2, \ \forall (x, y) \in \mathbb{R}^n \times \mathbb{R}^n, \tag{3.10}$$

Alors :

1. *f est strictement convexe,*
2. *$f(x) \to +\infty$ quand $|x| \to +\infty$,*

et en conséquence, il existe un unique $\bar{x} \in \mathbb{R}^n$ tel que $f(\bar{x}) = \inf_{\mathbb{R}^n} f$.

DÉMONSTRATION –

1. Soit φ la fonction définie de $\mathbb{R}$ dans $\mathbb{R}^n$ par : $\varphi(t) = f(x + t(y - x))$. Alors

$$f(y) - f(x) = \varphi(1) - \varphi(0) = \int_0^1 \nabla f(x + t(y - x)) \cdot (y - x) dt,$$

On en déduit que

$$f(y) - f(x) - \nabla f(x) \cdot (y - x) = \int_0^1 (\nabla f(x + t(y - x)) \cdot (y - x) - \nabla f(x) \cdot (y - x)) dt,$$

c'est–à–dire :

$$f(y) - f(x) - \nabla f(x) \cdot (y - x) = \int_0^1 \underbrace{(\nabla f(x + t(y - x)) - \nabla f(x)) \cdot (y - x)}_{\geq \alpha t |y - x|^2} dt.$$

Grâce à l'hypothèse (3.10) sur f, ceci entraîne :

$$f(y) - f(x) - \nabla f(x) \cdot (y - x) \geq \alpha \int_0^1 t|y - x|^2 dt = \frac{\alpha}{2}|y - x|^2 > 0 \text{ si } y \neq x. \tag{3.11}$$

On a donc, pour tout $(x,y) \in E^2$, $f(y) > f(x) + \nabla f(x)\cdot(y-x)$; d'après la première caractérisation de la convexité, voir proposition 3.5, on en déduit que f est strictement convexe.

2. Montrons maintenant que $f(y) \to +\infty$ quand $|y| \to +\infty$. On écrit (3.11) pour $x = 0$: $f(y) \geq f(0) + \nabla f(0)\cdot y + \frac{\alpha}{2}|y|^2$. Comme $\nabla f(0)\cdot y \geq -|\nabla f(0)|(y)$, on a donc

$$f(y) \geq f(0) + |y|\left(\frac{\alpha}{2}|y| - |\nabla f(0)|\right) \to +\infty \text{ quand } |y| \to +\infty.$$

La fonction f vérifie donc bien les hypothèses du théorème 3.30, et on en déduit qu'il existe un unique $\bar{x}$ qui minimise f. ■

Remarque 3.14 (Généralisation à un espace de Hilbert)**.** Le théorème 3.12 reste vrai si E est un espace de Hilbert ; on a besoin dans ce cas pour la partie existence des hypothèses (i), (ii) et de la convexité de f.

Proposition 3.15 (Caractérisation des points tels que $f(\bar{x}) = \inf_E f$)**.** *Soit E espace vectoriel normé et f une fonction de E dans $\mathbb{R}$. On suppose que $f \in C^1(E,\mathbb{R})$ et que f est convexe. Soit $\bar{x} \in E$. Alors :*

$$f(\bar{x}) = \inf_E f \Leftrightarrow Df(\bar{x}) = 0.$$

En particulier si $E = \mathbb{R}^n$ alors $f(\bar{x}) = \inf_{x\in\mathbb{R}^n} f(x) \Leftrightarrow \nabla f(\bar{x}) = 0$.

Démonstration
($\Rightarrow$) Supposons que $f(\bar{x}) = \inf_E f$ alors on sait (voir Proposition 3.7) que $Df(\bar{x}) = 0$ (la convexité est inutile).
($\Leftarrow$) Si f est convexe et différentiable, d'après la proposition 3.5, on a : $f(y) \geq f(\bar{x}) + Df(\bar{x})(y-x)$ pour tout $y \in E$ et comme par hypothèse $Df(\bar{x}) = 0$, on en déduit que $f(y) \geq f(\bar{x})$ pour tout $y \in E$. Donc $f(\bar{x}) = \inf_E f$.

Cas d'une fonction quadratique On appelle fonction quadratique une fonction de $\mathbb{R}^n$ dans $\mathbb{R}$ définie par

$$\boldsymbol{x} \mapsto f(\boldsymbol{x}) = \frac{1}{2}A\boldsymbol{x}\cdot\boldsymbol{x} - \boldsymbol{b}\cdot\boldsymbol{x} + c, \tag{3.12}$$

où $A \in \mathcal{M}_n(\mathbb{R})$, $\boldsymbol{b} \in \mathbb{R}^n$ et $c \in \mathbb{R}$. On peut vérifier facilement que $f \in C^\infty(\mathbb{R}^n,\mathbb{R})$. Calculons le gradient de f et sa hessienne : on a

$$\begin{aligned} f(\boldsymbol{x}+\boldsymbol{h}) &= \frac{1}{2}A(\boldsymbol{x}+\boldsymbol{h})\cdot(\boldsymbol{x}+\boldsymbol{h}) - \boldsymbol{b}\cdot(\boldsymbol{x}+\boldsymbol{h}) + c \\ &= \frac{1}{2}A\boldsymbol{x}\cdot\boldsymbol{x} + \frac{1}{2}A\boldsymbol{x}\cdot\boldsymbol{h} + \frac{1}{2}A\boldsymbol{h}\cdot\boldsymbol{x} + \frac{1}{2}A\boldsymbol{h}\cdot\boldsymbol{h} - \boldsymbol{b}\cdot\boldsymbol{x} - \boldsymbol{b}\cdot\boldsymbol{h} + c \\ &= f(\boldsymbol{x}) + \frac{1}{2}(A\boldsymbol{x}\cdot\boldsymbol{h} + A\boldsymbol{h}\cdot\boldsymbol{x}) - \boldsymbol{b}\cdot\boldsymbol{h} + \frac{1}{2}A\boldsymbol{h}\cdot\boldsymbol{h} \\ &= f(\boldsymbol{x}) + \frac{1}{2}(A\boldsymbol{x} + A^t\boldsymbol{x})\cdot\boldsymbol{h} - \boldsymbol{b}\cdot\boldsymbol{h} + \frac{1}{2}A\boldsymbol{h}\cdot\boldsymbol{h}. \end{aligned}$$

Et comme $|A\boldsymbol{h}\cdot\boldsymbol{h}| \leq \|A\|_2\,|\boldsymbol{h}|^2$, on en déduit que :

$$\nabla f(\boldsymbol{x}) = \frac{1}{2}(A\boldsymbol{x} + A^t\boldsymbol{x}) - \boldsymbol{b}. \tag{3.13}$$

Si A est symétrique, on a donc $\nabla f(\boldsymbol{x}) = A\boldsymbol{x} - \boldsymbol{b}$. Calculons maintenant la hessienne de f. D'après (3.13), on a :

$$\nabla f(\boldsymbol{x}+\boldsymbol{h}) = \frac{1}{2}(A(\boldsymbol{x}+\boldsymbol{h}) + A^t(\boldsymbol{x}+\boldsymbol{h})) - \boldsymbol{b} = \nabla f(\boldsymbol{x}) + \frac{1}{2}(A\boldsymbol{h} + A^t\boldsymbol{h})$$

et donc $H_f(\boldsymbol{x}) = D(\nabla f(\boldsymbol{x})) = \frac{1}{2}(A + A^t)$. On en déduit que si A est symétrique, $H_f(\boldsymbol{x}) = A$. Dans le cas où A est symétrique définie positive, f est donc strictement convexe.

De plus on a $f(\boldsymbol{x}) \to +\infty$ quand $|\boldsymbol{x}| \to +\infty$. (On note comme d'habitude $|\cdot|$ la norme euclidienne de $\boldsymbol{x}$.) En effet,

$$A\boldsymbol{x}\cdot\boldsymbol{x} \geq \alpha|\boldsymbol{x}|^2 \text{ où } \alpha \text{ est la plus petite valeur propre de } A, \text{ et } \alpha > 0.$$

Donc

$$f(\boldsymbol{x}) \geq \frac{\alpha}{2}|\boldsymbol{x}|^2 - |\boldsymbol{b}\cdot\boldsymbol{x}| - |c|;$$

Mais comme $|\boldsymbol{b}\cdot\boldsymbol{x}| \leq |\boldsymbol{b}||\boldsymbol{x}|$, on a

$$f(\boldsymbol{x}) \geq |\boldsymbol{x}|\left(\frac{\alpha|\boldsymbol{x}|}{2} - |\boldsymbol{b}|\right) - |c| \longrightarrow +\infty \text{ quand } |\boldsymbol{x}| \to +\infty.$$

On en déduit l'existence et l'unicité de $\bar{\boldsymbol{x}}$ qui minimise f. On a aussi :

$$\nabla f(\bar{\boldsymbol{x}}) = 0 \Leftrightarrow f(\bar{\boldsymbol{x}}) = \inf_{\mathbb{R}^n} f$$

et donc $\bar{\boldsymbol{x}}$ est l'unique solution du système $A\boldsymbol{x} = \boldsymbol{b}$.
On en déduit le théorème suivant, très important, puisqu'il va nous permettre en particulier le lien entre certains algorithmes d'optimisation et les méthodes de résolution de systèmes linéaires vues au chapitre 1.

Théorème 3.16 (Minimisation d'une fonction quadratique)**.** *Soit f une fonction de $\mathbb{R}^n$ dans $\mathbb{R}$ définie par* (3.12) *où $A \in \mathcal{M}_n(\mathbb{R})$ est une matrice symétrique définie positive et $\boldsymbol{b} \in \mathbb{R}^n$. Alors il existe un unique $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ qui minimise f, et $\bar{\boldsymbol{x}}$ est l'unique solution du système linéaire $A\boldsymbol{x} = \boldsymbol{b}$.*

3.2.3 Exercices (optimisation sans contrainte)

Exercice 137 (Maximisation)**.** *Suggestions en page 200*
Soit E un espace vectoriel normé et $f : E \to \mathbb{R}$. En utilisant les résultats de la section 3.2.2, répondre aux questions suivantes :

1. Donner une condition suffisante d'existence de $\bar{x} \in E$ tel que $f(\bar{x}) = \sup_{x\in E} f(x)$.
2. Donner une condition suffisante d'unicité de $\bar{x} \in E$ tel que $f(\bar{x}) = \sup_{x\in E} f(x)$.
3. Donner une condition suffisante d'existence et unicité de $\bar{x} \in E$ tel que $f(\bar{x}) = \sup_{x\in E} f(x)$.

Exercice 138 (Complément de Schur)**.**

Soient n et p deux entiers naturels non nuls. Dans toute la suite, si u et v sont deux vecteurs de $\mathbb{R}^k$, $k \geq 1$, le produit scalaire de u et v est noté $u\cdot v$. Soient A une matrice carrée d'ordre n, inversible, soit B une matrice $n \times p$, C une matrice carrée d'ordre p, et soient $f \in \mathbb{R}^n$ et $g \in \mathbb{R}^p$. On considère le système linéaire suivant :

$$M\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix}, \text{ avec } M = \begin{bmatrix} A & B \\ B^t & C \end{bmatrix}. \tag{3.14}$$

1. On suppose dans cette question seulement que $n = p = 1$, et $A = \begin{bmatrix} a \end{bmatrix}$, $B = \begin{bmatrix} b \end{bmatrix}$, $C = \begin{bmatrix} c \end{bmatrix}$
   (a) Donner une condition nécessaire et suffisante sur a, b, et c pour que M soit inversible.
   (b) Donner une condition nécessaire et suffisante sur a, b, et c pour que M soit symétrique définie positive.

On définit la matrice $S = C - B^tA^{-1}B$, qu'on appelle "complément de Schur".
2. Calculer S dans le cas $A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$, $B = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$, $C = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.
3. Montrer qu'il existe une unique solution au problème (3.14) si et seulement si la matrice S est inversible. Est-ce le cas dans la question 2 ?

On suppose maintenant que A est symétrique définie positive.
4. On suppose dans cette question que C est symétrique.

(a) Vérifier que M est symétrique.

(b) Soient $x \in \mathbb{R}^n$, $y \in \mathbb{R}^p$ et $z = (x, y) \in \mathbb{R}^{n+p}$. Calculer $Mz \cdot z$ en fonction de A, B, C, x et y.

(c) On fixe maintenant $y \in \mathbb{R}^p$, et on définit la fonction F de $\mathbb{R}^n$ dans $\mathbb{R}$ par : $x \mapsto Ax \cdot x + 2By \cdot x + Cy \cdot y$. Calculer $\nabla F(x)$, et calculer $x_0 \in \mathbb{R}^n$ tel que $\nabla F(x_0) = 0$

(d) Montrer que la fonction F définie en 3(c) admet un unique minimum, et calculer la valeur de ce mimimum.

(e) En déduire que M est définie positive si et seulement si S est définie positive.

5. On suppose dans cette question que C est la matrice (carrée d'ordre p) nulle.

(a) Montrer que la matrice $\tilde{S} = -S$ est symétrique définie positive si et seulement si $p \leq n$ et rang$(B) = p$. On supposera que ces deux conditions sont vérifiées dans toute la suite de la question.

(b) En déduire que la matrice $P = \begin{bmatrix} A & 0 \\ 0 & \tilde{S} \end{bmatrix}$ est symétrique définie positive.

(c) Calculer les valeurs propres de la matrice $T = P^{-1}M$ (il peut être utile de distinguer les cas Ker$B^t = \{0\}$ et Ker$B^t \neq \{0\}$).

Exercice 139 (Approximation au sens des moindres carrés)**.** *Corrigé en page 200*

1. **Un premier exemple.** Dans le plan (s, t), on cherche la droite d'équation $t = \alpha + \beta s$ qui passe par les points $(0, 1)$, $(1, 9)$, $(3, 9)$, $(4, 21)$.

   (a) Montrer que si cette droite existait, le vecteur $x = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$ serait solution d'un système linéaire $Ax = b$; on donnera explicitement la matrice A et le vecteur b.

   (b) Montrer qu'une telle droite n'existe pas. Dans la suite du problème on va trouver la droite qui passe le "plus près" possible de ces quatre points, au sens de la norme euclidienne.

2. **Un second exemple.** On cherche maintenant à déterminer les coefficients α, β et γ d'une fonction linéaire T de $\mathbb{R}^3$ dans $\mathbb{R}$, dont on ne connaît la valeur qu'en deux points : $T(1, 1, 1) = 3$ et $T(0, 1, 1) = 2$.

   (a) Montrer que les coefficients α, β et γ s'ils existent, satisfont un système linéaire $Ax = b$; on donnera explicitement la matrice A et le vecteur b.

   (b) Montrer qu'il existe une infinité de solutions au système $Ax = b$. Dans la suite du problème on va trouver les coefficients α, β et γ qui donnent un vecteur x de norme euclidienne minimale.

On considère maintenant une matrice A d'ordre $n \times m$ et $b \in \mathbb{R}^n$, et on veut résoudre dans un sens aussi "satisfaisant" que possible le système linéaire

$$Ax = b,\ x \in \mathbb{R}^m, \tag{3.15}$$

lorsque $m \neq n$ ou lorsque $m = n$ mais que A n'est pas inversible. On note $\|y\| = \left(\sum_{i=1}^p y_i^2\right)^{\frac{1}{2}}$ la norme euclidienne sur $\mathbb{R}^p$, $p = n$ ou m suivant les cas et $(\cdot \mid \cdot)$ le produit scalaire associé. Soit f la fonction définie de $\mathbb{R}^m$ dans $\mathbb{R}$ par $f(x) = \|Ax - b\|^2$. On cherche à minimiser f, c.à.d. à trouver $\bar{x} \in \mathbb{R}^n$ tel que

$$f(\bar{x}) = \min\{f(x), x \in \mathbb{R}^n\}. \tag{3.16}$$

3. Soit E un sous espace vectoriel de $\mathbb{R}^m$ tel que $\mathbb{R}^m = E \bigoplus \mathrm{Ker}A$.

   (a) Montrer que $f(z) \to +\infty$ lorsque $\|z\| \to +\infty$ avec $z \in E$.

   (b) Montrer que f est strictement convexe de E dans $\mathbb{R}$.

   (c) En déduire qu'il existe un unique $\bar{z} \in E$ tel que

   $$f(\bar{z}) \leq f(z), \forall z \in E.$$

4. Soit $X_b = \{\bar{z} + y, y \in \mathrm{Ker}A\}$, où $\bar{z}$ est défini à la question précédente. Montrer que X_b est égal à l'ensemble des solutions du problème de minimisation (3.16).

5. Montrer que $x \in X_b \iff A^tAx = A^tb$, où A^t désigne la matrice transposée de A. On appelle système d'équations normales le système $A^tAx = A^tb$.
6. Ecrire les équations normales dans le cas de l'exemple de la question 1, et en déduire l'équation de la droite obtenue par moindres carrés, *i.e.* par résolution de (3.16). Tracer les quatre points donnés à la question 1 et la droite obtenue sur un graphique.
7. Ecrire les équations normales dans le cas de l'exemple de la question 2, et vérifier que le système obtenu n'est pas inversible.
8. Pour $y \in \mathrm{Ker}A$, on pose $g(y) = \|y + \bar{z}\|^2$, où $\bar{z}$ est définie à la question 3. Montrer qu'il existe un unique $\bar{y} \in \mathrm{Ker}A$ tel que $g(\bar{y}) \leq g(y)$ pour tout $y \in \mathrm{Ker}A$. En déduire qu'il existe un unique $\bar{x} \in X_b$ tel que $\|\bar{x}\|^2 \leq \|x\|^2$ pour tout $x \in X_b$. On appelle $\bar{x}$ pseudo-solution de (3.16).
9. Calculer $\bar{x}$ dans le cas des exemples des questions 1 et 2.

Dans la suite du problème, on considère, pour $\varepsilon > 0$ fixé, une version pénalisée du problème (3.16). On introduit la fonction f_ε de $\mathbb{R}^m$ dans $\mathbb{R}$, définie par $f_\varepsilon(x) = \|x\|^2 + \frac{1}{\varepsilon}\|A^tAx - A^tb\|^2$, et on cherche à trouver x_ε solution du problème de minimisation suivant :

$$f_\varepsilon(x_\varepsilon) \leq f_\varepsilon(x), \forall x \in \mathbb{R}^m. \tag{3.17}$$

10. Montrer que le problème (3.17) possède une unique solution x_ε.
11. Calculer $\nabla f_\varepsilon(x)$ et en déduire l'équation satisfaite par x_ε.
12. Montrer que x_ε converge vers $\bar{x}$ lorsque $\varepsilon \to 0$.

Suggestions pour les exercices

Exercice 137 page 198 (Maximisation) Appliquer les théorèmes du cours à $-f$.

Corrigés des exercices

Exercice 139 page 199 (Approximation au sens des moindres carrés)

1. (a) Une condition nécessaire pour que la droite existe est que α et β vérifie le système linéaire

$$\begin{aligned} \text{point } (0,1): &\quad \alpha = 1 \\ \text{point } (1,9): &\quad \alpha + \beta = 9 \\ \text{point } (3,9): &\quad \alpha + 3\beta = 9 \\ \text{point } (4,21): &\quad \alpha + 4\beta = 21 \end{aligned}$$

Autrement dit $x = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$ est une solution de $Ax = b$, avec

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} \text{ et } b = \begin{bmatrix} 1 \\ 9 \\ 9 \\ 21 \end{bmatrix}.$$

(b) Montrer qu'une telle droite n'existe pas.
Si l'on on retranche la ligne 2 à la ligne 3 du système, on obtient $\beta = 0$ et si l'on retranche la ligne 1 à la ligne 2, on obtient $\beta = 8$. Donc le système n'admet pas de solution.

2. (a) Une condition nécessaire pour que la droite existe est que α, β et γ vérifie le système

$$\begin{aligned} \alpha+\beta+\gamma &= 3 \\ \beta+\gamma &= 2 \end{aligned}$$

Autrement dit $x = \begin{bmatrix}\alpha\\ \beta\end{bmatrix}$ est une solution de $Ax = b$, avec

$$A = \begin{bmatrix}1 & 1 & 1\\ 0 & 1 & 1\end{bmatrix} \text{ et } b = \begin{bmatrix}3\\ 2\end{bmatrix}.$$

(b) Une solution particulière de ce système est $x = \begin{bmatrix}1\\ 2\\ 0\end{bmatrix}$, et le noyau de A est engendré par $\begin{bmatrix}0\\ -1\\ 1\end{bmatrix}$.

L'ensemble des solutions est de la forme $\{\begin{bmatrix}1\\ 2\\ 0\end{bmatrix} + \gamma \begin{bmatrix}0\\ -1\\ 1\end{bmatrix}, \gamma \in \mathbb{R}\}$, qui est infini.

3. (a) On a

$$\begin{aligned} f(z) &= (Az-b)\cdot(Az-b) \\ &= Az\cdot Az - 2Az\cdot b + b\cdot b \\ &\geq \|Az\|^2 - 2\|b\|\|Az\| + \|b\|^2 \qquad \text{d'après l'inégalité de Cauchy-Schwarz} \\ &\longrightarrow +\infty \text{ lorsque } \|Az\| \to \infty \end{aligned}$$

Il reste maintenant à montrer que $\|Az\| \longrightarrow +\infty$ lorsque $\|z\| \to \infty$. Pour cela, on remarque que

$$\|Az\| = \left\|A\frac{z}{\|z\|}\right\| \|z\| \geq \inf_{w\in E, \|w\|=1} \|Aw\|\|z\| = \|A\bar{w}\|\|z\|,$$

car l'ensemble $K = \{w \in E, \|w\| = 1\}$ est un compact de $\mathbb{R}^n$ et comme la fonction $\varphi : K \to \mathbb{R}$ définie par $\varphi(w) = \|Aw\|$ est continue, elle atteint son minimum en $\bar{w} \in K$:

$$\inf_{w\in E, \|w\|=1} \|Aw\| = \|A\bar{w}\|$$

Or $A\bar{w} \neq 0$, et donc $\|A\bar{w}\|\|z\| \to +\infty$ lorsque $\|z\| \to +\infty$.

(b) Calculons ∇f.

$$\forall x \in E, \nabla f(x) = 2(A^tAx - A^tb)$$

Par conséquent, pour $x, y \in E$,

$$\begin{aligned} f(y) - \nabla f(x)\cdot(y-x) &= (Ay-b)\cdot(Ay-b) - 2(Ax-b)\cdot A(y-x),\ \forall (x,y)\in E^2; x\neq y. \\ &= |Ay|^2 - 2Ay\cdot b + |b|^2 + 2|Ax|^2 - 2Ax\cdot Ay + 2b\cdot Ay - 2b\cdot Ax \\ &= |Ay|^2 + |b|^2 + 2|Ax|^2 - 2Ax\cdot Ay - 2b\cdot Ax \\ &= |Ay - Ax|^2 + |Ax - b|^2 \\ &> 0,\ \forall (x,y)\in E^2; x\neq y. \end{aligned}$$

On en déduit que f est strictement convexe par la proposition 3.5 (première caractérisation de la convexité).

(c) On applique le théorème 3.12 : f est une application continue de E dans $\mathbb{R}$, qui tend vers l'infini à l'infini et qui admet donc un minimum. L'unicité du minimum vient de la stricte convexité de cette application.

4. Soit $x \in \mathbb{R}^m$, x peut s'écrire $x = z + y$ avec $z \in E$ et $y \in \mathrm{Ker}A$, par suite

$$f(x) = \|A(z+y) - b\|^2 = \|Az - b\|^2 = f(z) \geq f(\bar{z}).$$

D'autre part,

$$f(\bar{z} + y) = f(\bar{z}) \forall y \in \mathrm{Ker}A.$$

Donc X_b est bien l'ensemble des solutions du problème de minimisation (3.16).

5. Condition nécessaire : On a déjà vu que f est différentiable et que $\nabla f(x) = 2(A^tAx - A^tb)$. Comme f est différentiable toute solution de (3.16) vérifie l'équation d'Euler $\nabla f(x) = 0$.
Condition suffisante : Soit x tel que $A^tAx = A^tb$, c'est à dire tel que $\nabla f(x) = 0$. Comme f est de classe C^1 et convexe sur $\mathbb{R}^m$, alors x est un minimum global de f. (Noter que la convexité de f peut se montrer comme à la question 3(b) en remplaçant E par $\mathbb{R}^m$.)

6. On a

$$A^tA = \begin{bmatrix} 4 & 8 \\ 8 & 26 \end{bmatrix} \text{ et } A^tb = \begin{bmatrix} 40 \\ 120 \end{bmatrix}$$

Les équations normales de ce problème s'écrivent donc

$$A^tA \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = A^tb.$$

La matrice A^tA est inversible, par conséquent il y a une unique solution à ces equations normales donnée par $\begin{bmatrix} 2 \\ 4 \end{bmatrix}$.

7. On a

$$A^tA = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 2 \end{bmatrix} \text{ et } A^tb = \begin{bmatrix} 3 \\ 5 \\ 5 \end{bmatrix}$$

Les deux dernières lignes de la matrice A^tA sont identiques donc la matrice n'est pas inversible. Comme les deux dernières lignes de A^tb sont elles aussi identiques, on en déduit que le système admet une infinité de solutions.
On peut échelonner le système :

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 3 \\ 1 & 2 & 2 & 5 \\ 1 & 2 & 2 & 5 \end{array}\right] \underset{T_{32}(-1), T_{21}(-1)}{\longrightarrow} \left[\begin{array}{ccc|c} 1 & 1 & 1 & 3 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{array}\right] \underset{T_{12}(-1)}{\longrightarrow} \left[\begin{array}{ccc|c} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{array}\right]$$

On retrouve les solutions obtenues à la question 2-b : $X_b = \underbrace{\begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}}_{\bar{z}} + \mathbb{R} \underbrace{\begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}}_{\bar{u}} = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix} + \mathrm{Ker}A.$

8. La fonction g est continue sur l'espace vectoriel $\mathrm{Ker}A$ et vérifie

$$g(y) \geq \|y\|^2 - 2\|y\|\|\bar{z}\| + \|\bar{z}\|^2 \longrightarrow +\infty \text{ lorsque } \|\bar{z}\| \longrightarrow +\infty;$$

par conséquent, g admet un minimum sur $\mathrm{Ker}A$. Ce minimum est unique car g est strictement convexe, car c'est le carré de la norme euclidienne. On peut le montrer directement, ou si l'on ne connaît pas ce résultat, on peut dire que c'est la composée d'une application convexe et d'une application strictement convexe et

croissante : $g = q(N(x))$ avec $N : x \mapsto \|x\|$ convexe et $q : s \mapsto s^2$. On pourrait également remarquer que l'application

$$\begin{aligned} &\mathrm{Ker}A \to \mathbb{R} \\ &v \mapsto D^2g(y)(v)(v) \end{aligned}$$

est une forme quadratique définie positive, car

$$Dg(y)(w) = 2(y + \bar{z}) \cdot w, \text{ et } D^2g(y)(h)(v) = 2h \cdot v$$

L'application $v \mapsto D^2g(y)(v)(v) = 2v \cdot v$ est clairement définie positive, ce qui montre une fois de plus que g est strictement convexe.
On en déduit alors qu'il existe un unique $\bar{x} \in X_b$ de norme minimale.

9. Dans le premier exemple, les équations normales admettent une seule solution $\bar{x} = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$, voir question 6.

   Pour le deuxième exemple, on calcule $\|x\|^2$ pour $x = \bar{z} + t\bar{u} \in X_b$:

$$\|x\|^2 = \|\bar{z} + t\bar{u}\|^2 = 1 + (2-t)^2 + t^2 = 5 - 4t + 2t^2 = 2\,(t-1)^2 + 3$$

   On voit $\|x\|^2$ est minimale pour $t = 1$, autrement dit $\bar{x} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$.

10. La fonction f_ε est une fonction continue, infinie à l'infini car

$$f_\varepsilon(x) \geq \|x\|^2 \longrightarrow +\infty \text{ lorsque } \|x\| \to \infty.$$

    On a donc existence d'un minimum pour f_ε. De plus, la fonction f_ε est de classe C^2 avec

$$\nabla f_\varepsilon(x) = 2(x + \frac{1}{\varepsilon}A^tA(A^tAx - A^tb)),\ D^2f_\varepsilon(x) = 2(Id + \frac{1}{\varepsilon}(A^tA)^2)$$

    La matrice A^tA est positive, donc la matrice $(A^tA)^2$ est positive par suite la matrice $D^2f(x)$ est définie positive. La fonction f_ε est donc strictement convexe. Par conséquent, f_ε admet un unique minimum.

11. On sait que le minimum x_ε de f_ε est un zéro de ∇f_ε, soit

$$x_\varepsilon + \frac{1}{\varepsilon}A^tA(A^tAx_\varepsilon - A^tb) = 0$$

    et donc $(Id + \frac{1}{\varepsilon}(A^tA)^2)x_\varepsilon = \frac{1}{\varepsilon}A^tAA^tb$, ce qui donne

$$x_\varepsilon = (Id + \frac{1}{\varepsilon}(A^tA)^2)^{-1}\frac{1}{\varepsilon}A^tAA^tb.$$

12. On commence par remarquer que $\|x_\varepsilon\|^2 \leq f_\varepsilon(x_\varepsilon) \leq f_\varepsilon(\bar{x}) = \|\bar{x}\|^2$. Par conséquent, la famille$\{x_\varepsilon,\, \varepsilon > 0\}$ est bornée dans $\mathbb{R}^m$ qui est de dimension finie. Pour montrer que $x_\varepsilon \to \bar{x}$ quand $\varepsilon \to 0$, il suffit donc de montrer que $\bar{x}$ est la seule valeur d'adhérence de la famille $\{x_\varepsilon,\, \varepsilon > 0\}$. Soit $\bar{y}$ une valeur d'adhérence de la famille $\{x_\varepsilon,\, \varepsilon > 0\}$. Il existe une suite $\varepsilon_n \to 0$ pour laquelle x_{ε_n} converge vers $\bar{y} \in \mathbb{R}^m$. Montrons que $A^tA\bar{y} = A^tb$. On rappelle que

$$\frac{1}{\varepsilon}\|A^tAx_\varepsilon - A^tb\|^2 \leq f_\varepsilon(x_\varepsilon) \leq \|\bar{x}\|^2.$$

    On en déduit que

$$\|A^tAx_{\varepsilon_n} - A^tb\|^2 \leq \varepsilon_n\|\bar{x}\|^2 \longrightarrow 0 \text{ lorsque } n \longrightarrow \infty$$

    et en passant à la limite on obtient $\|A^tA\bar{y} - A^tb\|^2 = 0$. On a également par un argument analogue $\|\bar{y}\|^2 \leq \|\bar{x}\|^2$. Donc $\bar{y} \in X_b$ et comme $\bar{x}$ est l'unique vecteur de X_b de norme minimale, on en déduit que $\bar{y} = \bar{x}$.

3.3 Algorithmes d'optimisation sans contrainte

Soit $f \in C(\mathbb{R}^n, \mathbb{R})$. On suppose qu'il existe $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ tel que $f(\bar{\boldsymbol{x}}) = \inf_{\mathbb{R}^n} f$.

On cherche à calculer $\bar{\boldsymbol{x}}$ (si f est de classe C^1, on a nécessairement $\nabla f(\bar{\boldsymbol{x}}) = 0$). On va donc maintenant développer des algorithmes (ou méthodes de calcul) du point $\bar{\boldsymbol{x}}$ qui réalise le minimum de f. Il existe deux grandes classes de méthodes :

— Les méthodes dites "directes" ou bien "de descente", qui cherchent à construire une suite minimisante, c.à.d. une suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ telle que :

$$f(\boldsymbol{x}^{(k+1)}) \leq f(\boldsymbol{x}^{(k)}),$$
$$\boldsymbol{x}^{(k)} \to \bar{\boldsymbol{x}} \text{ quand } k \to +\infty.$$

— Les méthodes basées sur l'équation d'Euler, qui consistent à chercher une solution de l'équation (dite d'Euler) $\nabla f(\boldsymbol{x}) = 0$ (ces méthodes nécessitent donc que f soit dérivable).

3.3.1 Méthodes de descente

Définition 3.17. *Soit $f \in C(\mathbb{R}^n, \mathbb{R})$.*

1. *Soit $\boldsymbol{x} \in \mathbb{R}^n$, on dit que $\boldsymbol{w} \in \mathbb{R}^n \setminus \{0\}$ est une direction de descente en $\boldsymbol{x}$ s'il existe $\alpha_0 > 0$ tel que*
$$f(\boldsymbol{x} + \alpha\boldsymbol{w}) \leq f(\boldsymbol{x}), \ \ \forall \alpha \in [0, \alpha_0]$$
2. *Soit $\boldsymbol{x} \in \mathbb{R}^n$, on dit que $w \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ est une direction de descente stricte en $\boldsymbol{x}$ si s'il existe $\alpha_0 > 0$ tel que*
$$f(\boldsymbol{x} + \alpha\boldsymbol{w}) < f(\boldsymbol{x}), \ \ \forall \alpha \in]0, \alpha_0].$$
3. *Une "méthode de descente" pour la recherche de $\bar{\boldsymbol{x}}$ tel que $f(\bar{\boldsymbol{x}}) = \inf_{\mathbb{R}^n} f$ consiste à construire une suite $(x_k)_{k\in\mathbb{N}}$ de la manière suivante :*
   (a) *Initialisation : $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$;*
   (b) *Itération k : on suppose $\boldsymbol{x}^{(0)}, \ldots, \boldsymbol{x}^{(k)}$ connus ($k \geq 0$) ;*
      i. *On cherche $\boldsymbol{w}^{(k)}$ direction de descente stricte en $\boldsymbol{x}^{(k)}$*
      ii. *On prend $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{w}^{(k)}$ avec $\alpha_k > 0$ "bien choisi".*

Proposition 3.18 (Caractérisation des directions de descente)**.** *Soient $f \in C^1(\mathbb{R}^n, \mathbb{R})$, $\boldsymbol{x} \in \mathbb{R}^n$ et $\boldsymbol{w} \in \mathbb{R}^n \setminus \{0\}$; alors*

1. *si $\boldsymbol{w}$ direction de descente en $\boldsymbol{x}$ alors $\boldsymbol{w} \cdot \nabla f(\boldsymbol{x}) \leq 0$,*
2. *si $\boldsymbol{w} \cdot \nabla f(\boldsymbol{x}) < 0$ alors $\boldsymbol{w}$ direction de descente stricte en $\boldsymbol{x}$,*
3. *si $\nabla f(\boldsymbol{x}) \neq 0$ alors $w = -\nabla f(\boldsymbol{x})$ est une direction de descente stricte en $\boldsymbol{x}$.*

DÉMONSTRATION –

1. Soit $\boldsymbol{w} \in \mathbb{R}^n \setminus \{0\}$ une direction de descente en $\boldsymbol{x}$: alors par définition,
$$\exists \alpha_0 > 0 \text{ tel que } f(\boldsymbol{x} + \alpha\boldsymbol{w}) \leq f(\boldsymbol{w}), \ \ \forall \alpha \in [0, \alpha_0].$$
Soit φ la fonction de $\mathbb{R}$ dans $\mathbb{R}$ définie par : $\varphi(\alpha) = f(\boldsymbol{x}+\alpha\boldsymbol{w})$. On a $\varphi \in C^1(\mathbb{R}, \mathbb{R})$ et $\varphi'(\alpha) = \nabla f(\boldsymbol{x}+\alpha\boldsymbol{w})\cdot\boldsymbol{w}$. Comme $\varphi(\alpha) \leq \varphi(0)$ pour tout $\alpha \in [0, \alpha_0]$ on a
$$\forall \alpha \in]0, \alpha_0[, \quad \frac{\varphi(\alpha) - \varphi(0)}{\alpha} \leq 0;$$
en passant à la limite lorsque α tend vers 0, on déduit que $\varphi'(0) \leq 0$, c.à.d. $\nabla f(\boldsymbol{x}) \cdot \boldsymbol{w} \leq 0$.

2. On reprend les notations précédentes. Si $\nabla f(\boldsymbol{x}) \cdot \boldsymbol{w} < 0$, on a $\varphi'(0) < 0$. Par continuité de ∇f, il existe $\alpha_0 > 0$ tel que $\varphi'(\alpha) < 0$ si $\alpha \in [0, \alpha_0]$. En utilisant le théorème des accroissements finis on en déduit que $\varphi(\alpha) < \varphi(0)$ si $\alpha \in]0, \alpha_0[$ et donc que w est une direction de descente stricte.
3. Si $\nabla f(\boldsymbol{x}) \neq 0$, $w = -\nabla f(\boldsymbol{x})$ est une direction de descente stricte en $\boldsymbol{x}$ car $\nabla f(\boldsymbol{x}) \cdot \boldsymbol{w} < 0 = -|\nabla f(\boldsymbol{x})|^2 < 0$.

■

Algorithme du gradient à pas fixe Soient $f \in C^1(E, \mathbb{R})$ et $E = \mathbb{R}^n$. On se donne $\alpha > 0$.

$$\left\{ \begin{array}{ll} \text{Initialisation :} & \boldsymbol{x}^{(0)} \in E, \\ \text{Itération } k : & \boldsymbol{x}^{(k)} \text{ connu, } (k \geq 0) \\ & \boldsymbol{w}^{(k)} = -\nabla f(\boldsymbol{x}^{(k)}), \\ & \boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha \boldsymbol{w}^{(k)}. \end{array} \right. \tag{3.18}$$

Théorème 3.19 (Convergence du gradient à pas fixe)**.** *Soient $E = \mathbb{R}^n$ et $f \in C^1(E, \mathbb{R})$ vérifiant les hypothèses*

$$\exists \omega > 0; (\nabla f(x) - \nabla f(y)) \cdot (x - y) \geq \omega |x - y|^2, \ \forall (x, y) \in \mathbb{R}^n \times \mathbb{R}^n, \tag{3.19a}$$

$$\exists M > 0; \|\nabla f(x) - \nabla f(y)\| \leq M |x - y|, \ \forall (x, y) \in \mathbb{R}^n \times \mathbb{R}^n. \tag{3.19b}$$

L'hypothese 3.19a est l'hypothese 3.10 de la proposition 3.13. La fonction f est donc strictement convexe et croissante à l'infini, et admet donc un unique minimum. De plus, si $0 < \alpha < \dfrac{2\omega}{M^2}$ alors la suite $(\boldsymbol{x}^{(k)})_{k \in \mathbb{N}}$ construite par (3.18) *converge vers $\bar{\boldsymbol{x}}$ lorsque $k \to +\infty$.*

DÉMONSTRATION –
Montrons la convergence de la suite construite par l'algorithme de gradient à pas fixe en nous ramenant à un algorithme de point fixe. On pose $h(\boldsymbol{x}) = \boldsymbol{x} - \alpha \nabla f(\boldsymbol{x})$. L'algorithme du gradient à pas fixe est alors un algorithme de point fixe pour h.

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

Grâce au théorème 2.8 page 144, on sait que h est strictement contractante si

$$0 < \alpha < \frac{2\omega}{M^2}.$$

Donc la suite $(x^{(k)})_{k \in \mathbb{N}}$ converge vers l'unique point fixe $\bar{\boldsymbol{x}}$ de h, caractérisé par

$$\bar{\boldsymbol{x}} = h(\bar{\boldsymbol{x}}) = \bar{\boldsymbol{x}} - \alpha \nabla f(\bar{\boldsymbol{x}})$$

On a donc $\nabla f(\bar{\boldsymbol{x}}) = 0$, et, comme f est strictement convexe, $f(\bar{\boldsymbol{x}}) = \inf_E f$. ■

Algorithme du gradient à pas optimal L'idée de l'algorithme du gradient à pas optimal est d'essayer de calculer à chaque itération le paramètre qui minimise la fonction dans la direction de descente donnée par le gradient. Soient $f \in C^1(E, \mathbb{R})$ et $E = \mathbb{R}^n$, cet algorithme s'écrit :

$$\left\{ \begin{array}{ll} \text{Initialisation :} & \boldsymbol{x}^{(0)} \in \mathbb{R}^n. \\ \text{Itération } n : & \boldsymbol{x}^{(k)} \text{ connu.} \\ & \text{On calcule } \boldsymbol{w}^{(k)} = -\nabla f(\boldsymbol{x}^{(k)}). \\ & \text{On choisit } \alpha_k \geq 0 \text{ tel que} \\ & \qquad f(\boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{w}^{(k)}) \leq f(\boldsymbol{x}^{(k)} + \alpha \boldsymbol{w}^{(k)}) \ \forall \alpha \geq 0. \\ & \text{On pose } \boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{w}^{(k)}. \end{array} \right. \tag{3.20}$$

Les questions auxquelles on doit répondre pour s'assurer du bien fondé de ce nouvel algorithme sont les suivantes :

1. Existe–t–il α_k tel que $f(\boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{w}^{(k)}) \leq f(\boldsymbol{x}^{(k)} + \alpha \boldsymbol{w}^{(k)}), \forall \alpha \geq 0$?

2. Comment calcule–t'on α_k ?
3. La suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ construite par l'algorithme converge–t–elle ?

La réponse aux questions 1. et 3. est apportée par le théorème suivant :

Théorème 3.20 (Convergence du gradient à pas optimal)**.**
Soit $f \in C^1(\mathbb{R}^n, \mathbb{R})$ telle que $f(\boldsymbol{x}) \to +\infty$ quand $|\boldsymbol{x}| \to +\infty$. Alors :

1. *La suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ est bien définie par (3.20). On choisit $\alpha_k > 0$ tel que $f(\boldsymbol{x}^{(k)}+\alpha_k\boldsymbol{w}^{(k)}) \leq f(x_k+\alpha\boldsymbol{w}^{(k)})$ $\forall\alpha \geq 0$ (α_k existe mais n'est pas nécessairement unique).*
2. *La suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ est bornée et si $(\boldsymbol{x}^{(k_\ell)})_{\ell\in\mathbb{N}}$ est une sous suite convergente, i.e. $x^{(k_\ell)} \to x$ lorsque $\ell \to +\infty$, on a nécessairement $\nabla f(\boldsymbol{x}) = 0$. De plus si f est convexe on a $f(x) = \inf_{\mathbb{R}^n} f$*
3. *Si f est strictement convexe on a alors $\boldsymbol{x}^{(k)} \to \bar{\boldsymbol{x}}$ quand $k \to +\infty$, avec $f(\bar{\boldsymbol{x}}) = \inf_{\mathbb{R}^n} f$*

La démonstration de ce théorème fait l'objet de l'exercice 142. On en donne ici les idées principales.

1. On utilise l'hypothèse $f(\boldsymbol{x}) \to +\infty$ quand $|\boldsymbol{x}| \to +\infty$ pour montrer que la suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ construite par (3.20) existe : en effet, à $\boldsymbol{x}^{(k)}$ connu,

   1er cas : si $\nabla f(\boldsymbol{x}^{(k)}) = 0$, alors $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)}$ et donc $\boldsymbol{x}^{(p)} = \boldsymbol{x}^{(k)}$ $\forall p \geq k$,

   2ème cas : si $\nabla f(\boldsymbol{x}^{(k)}) \neq 0$, alors $\boldsymbol{w}^{(k)} = \nabla f(\boldsymbol{x}^{(k)})$ est une direction de descente stricte.

   Dans ce deuxième cas, il existe donc α_0 tel que

   $$f(\boldsymbol{x}^{(k)} + \alpha\boldsymbol{w}^{(k)}) < f(\boldsymbol{x}^{(k)}),\ \forall\alpha \in]0, \alpha_0]. \tag{3.21}$$

   De plus, comme $\boldsymbol{w}^{(k)} \neq 0$, $|\boldsymbol{x}^{(k)} + \alpha\boldsymbol{w}^{(k)}| \to +\infty$ quand $\alpha \to +\infty$ et donc $f(\boldsymbol{x}^{(k)} + \alpha\boldsymbol{w}^{(k)}) \longrightarrow +\infty$ quand $\alpha \to +\infty$. Il existe donc $M > 0$ tel que si $\alpha > M$ alors $f(x_k + \alpha\boldsymbol{w}^{(k)}) \geq f(\boldsymbol{x}^{(k)})$. On a donc :

   $$\inf_{\alpha\in\mathbb{R}_+^*} f(\boldsymbol{x}^{(k)} + \alpha\boldsymbol{w}^{(k)}) = \inf_{\alpha\in[0,M]} f(\boldsymbol{x}^{(k)} + \alpha\boldsymbol{w}^{(k)}).$$

   Comme $[0, M]$ est compact, il existe $\alpha_k \in [0, M]$ tel que $f(x_k + \alpha_k\boldsymbol{w}^{(k)}) = \inf_{\alpha\in[0,M]} f(x_k + \alpha\boldsymbol{w}^{(k)})$. De plus on a grâce à (3.21) que $\alpha_k > 0$.
2. Le point 2. découle du fait que la suite $(f(\boldsymbol{x}^{(k)}))_{k\in\mathbb{N}}$ est décroissante, donc la suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ est bornée (car $f(\boldsymbol{x}) \to +\infty$ quand $|\boldsymbol{x}| \to +\infty$). On montre ensuite que si $\boldsymbol{x}^{(k_\ell)} \to \boldsymbol{x}$ lorsque $\ell \to +\infty$ alors $\nabla f(\boldsymbol{x}) = 0$ (ceci est plus difficile, les étapes sont détaillées dans l'exercice 142).

Reste la question du calcul de α_k, qui est le paramètre optimal dans la direction de descente $\boldsymbol{w}^{(k)}$, c.à.d. le nombre réel qui réalise le minimum de la fonction φ de $\mathbb{R}_+$ dans $\mathbb{R}$ définie par : $\varphi(\alpha) = f(\boldsymbol{x}^{(k)} + \alpha\boldsymbol{w}^{(k)})$. Comme $\alpha_k > 0$ et $\varphi(\alpha_k) \leq \varphi(\alpha)$ pour tout $\alpha \in \mathbb{R}_+$, on a nécessairement

$$\varphi'(\alpha_k) = \nabla f(\boldsymbol{x}^{(k)} + \alpha_k\boldsymbol{w}^{(k)}) \cdot \boldsymbol{w}^{(k)} = 0.$$

Cette équation donne en général le moyen de calculer α_k.
Considérons par exemple le cas (important) d'une fonctionnelle quadratique, *i.e.* $f(\boldsymbol{x}) = \frac{1}{2}A\boldsymbol{x}\cdot\boldsymbol{x} - \boldsymbol{b}\cdot\boldsymbol{x}$, A étant une matrice symétrique définie positive ; on a alors $\nabla f(\boldsymbol{x}^{(k)}) = A\boldsymbol{x}^{(k)} - \boldsymbol{b}$, et donc

$$\nabla f(\boldsymbol{x}^{(k)} + \alpha_k\, \boldsymbol{w}^{(k)}) \cdot \boldsymbol{w}^{(k)} = (A\boldsymbol{x}^{(k)} + \alpha_k A\boldsymbol{w}^{(k)} - \boldsymbol{b}) \cdot \boldsymbol{w}^{(k)} = 0.$$

On a ainsi dans ce cas une expression explicite de α_k, avec $\boldsymbol{r}^{(k)} = \boldsymbol{b} - A\boldsymbol{x}^{(k)}$,

$$\alpha_k = \frac{(\boldsymbol{b} - A\boldsymbol{x}^{(k)}) \cdot \boldsymbol{w}^{(k)}}{A\boldsymbol{w}^{(k)} \cdot \boldsymbol{w}^{(k)}} = \frac{\boldsymbol{r}^{(k)} \cdot \boldsymbol{w}^{(k)}}{A\boldsymbol{w}^{(k)} \cdot \boldsymbol{w}^{(k)}} \tag{3.22}$$

Remarquons que $A\boldsymbol{w}^{(k)} \cdot \boldsymbol{w}^{(k)} \neq 0$ (car A est symétrique définie positive).
Dans le cas d'une fonction f générale, on n'a pas en général de formule explicite pour α_k. On peut par exemple le calculer en cherchant le zéro de f' par la méthode de la sécante ou la méthode de Newton...

L'algorithme du gradient à pas optimal est donc une méthode de minimisation dont on a prouvé la convergence. Cependant, cette convergence est lente (en général linéaire), et de plus, l'algorithme nécessite le calcul du paramètre α_k optimal.

Algorithme du gradient à pas variable Dans ce nouvel algorithme, on ne prend pas forcément le paramètre optimal pour α, mais on lui permet d'être variable d'une itération à l'autre. L'algorithme s'écrit :

$$\left\{\begin{array}{ll} \text{Initialisation :} & x^{(0)} \in \mathbb{R}^n. \\ \text{Itération :} & \text{On suppose } x^{(k)} \text{ connu ; soit } \boldsymbol{w}^{(k)} = -\nabla f(\boldsymbol{x}^{(k)}) \text{ où : } \boldsymbol{w}^{(k)} \neq 0 \\ & (\text{si } \boldsymbol{w}^{(k)} = 0 \text{ l'algorithme s'arrête}). \\ & \text{On prend } \alpha_k > 0 \text{ tel que } f(\boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{w}^{(k)}) < f(x_k). \\ & \text{On pose } \boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{w}^{(k)}. \end{array}\right. \tag{3.23}$$

Théorème 3.21 (Convergence du gradient à pas variable)**.**
Soit $f \in C^1(\mathbb{R}^n, \mathbb{R})$ une fonction telle que $f(x) \to +\infty$ quand $|x| \to +\infty$, alors :

1. *On peut définir une suite $(x^{(k)})_{k\in\mathbb{N}}$ par (3.23).*
2. *La suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ est bornée. Si $\boldsymbol{x}^{(k_\ell)} \to \boldsymbol{x}$ quand $\ell \to +\infty$ et si $\nabla f(\boldsymbol{x}^{(k_\ell)}) \to 0$ quand $\ell \to +\infty$ alors $\nabla f(\boldsymbol{x}) = 0$. Si de plus f est convexe on a $f(\boldsymbol{x}) = \inf_{\mathbb{R}^n} f$*
3. *Si $\nabla f(\boldsymbol{x}^{(k)}) \to 0$ quand $k \to +\infty$ et si f est strictement convexe alors $\boldsymbol{x}^{(k)} \to \bar{\boldsymbol{x}}$ et $f(\bar{\boldsymbol{x}}) = \inf_{\mathbb{R}^n} f$.*

La démonstration s'effectue facilement à partir de la démonstration du théorème précédent : reprendre en l'adaptant l'exercice 142.

3.3.2 Algorithme du gradient conjugué

La méthode du gradient conjugué a été découverte en 1952 par Hestenes et Steifel pour la minimisation de fonctions quadratiques, c'est-à-dire de fonctions de la forme

$$f(\boldsymbol{x}) = \frac{1}{2} A\boldsymbol{x} \cdot \boldsymbol{x} - b \cdot \boldsymbol{x},$$

où $A \in \mathcal{M}_n(\mathbb{R})$ est une matrice symétrique définie positive et $b \in \mathbb{R}^n$. On rappelle (voir le paragraphe 3.2.2) que $f(\bar{\boldsymbol{x}}) = \inf_{\mathbb{R}^n} f \Leftrightarrow A\bar{\boldsymbol{x}} = b$.
L'idée de la méthode du gradient conjugué est basée sur la remarque suivante : supposons qu'on sache construire n vecteurs (les directions de descente) $\boldsymbol{w}^{(0)}, \boldsymbol{w}^{(1)}, \ldots, \boldsymbol{w}^{(n-1)}$ libres et tels que $\boldsymbol{r}^{(n)} \cdot \boldsymbol{w}^{(p)} = 0$ pour tout $p < n$. On a alors $\boldsymbol{r}^{(n)} = \boldsymbol{0}$: en effet la famille $(\boldsymbol{w}^{(0)}, \boldsymbol{w}^{(1)}, \ldots, \boldsymbol{w}^{(n-1)})$ engendre $\mathbb{R}^n$; le vecteur $\boldsymbol{r}^{(n)}$ est alors orthogonal à tous les vecteurs d'une $\mathbb{R}^n$, et il est donc nul.
Pour obtenir une famille libre de directions de descente stricte, on va construire les vecteurs $\boldsymbol{w}^{(0)}$, $\boldsymbol{w}^{(1)}$, ..., $\boldsymbol{w}^{(n-1)}$ de manière à ce qu'ils soient orthogonaux pour le produit scalaire induit par A. Nous allons voir que ce choix marche (presque) magnifiquement bien. Mais avant d'expliquer pourquoi, écrivons une méthode de descente à pas optimal pour la minimisation de f, en supposant les directions de descente $\boldsymbol{w}^{(0)}$ connues.
On part de $\boldsymbol{x}^{(0)}$ dans $\mathbb{R}^n$ donné ; à l'itération k, on suppose que $\boldsymbol{r}^{(k)} = \boldsymbol{b} - A\boldsymbol{x}^{(k)} \neq \boldsymbol{0}$ (sinon on a $\boldsymbol{x}^{(k)} = \bar{\boldsymbol{x}}$ et on a fini). On calcule le paramètre α_k optimal dans la direction $\boldsymbol{w}^{(k)}$ par la formule (3.22). Et on calcule ensuite le nouvel itéré :

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{w}^{(k)}.$$

Notons que $\boldsymbol{r}^{(k+1)} = \boldsymbol{b} - A\boldsymbol{x}^{(k+1)}$ et donc

$$\boldsymbol{r}^{(k+1)} = \boldsymbol{r}^{(k)} - \alpha_k A\boldsymbol{w}^{(k)}. \tag{3.24}$$

De plus, par définition du paramètre optimal α_k, on a $\nabla f(\boldsymbol{x}^{(k+1)}) \cdot \boldsymbol{w}^{(k)} = 0$ et donc

$$\boldsymbol{r}^{(k+1)} \cdot \boldsymbol{w}^{(k)} = 0 \tag{3.25}$$

Ces deux dernières propriétés sont importantes pour montrer la convergence de la méthode. Mais il nous faut maintenant choisir les vecteurs $\boldsymbol{w}^{(k)}$ qui soient des directions de descente strictes et qui forment une famille libre. A l'étape 0, il est naturel de choisir la direction opposée du gradient :

$$\boldsymbol{w}^{(0)} = -\nabla f(\boldsymbol{x}^{(0)}) = \boldsymbol{r}^{(0)}.$$

A l'étape $k \geq 1$, on choisit la direction de descente $\boldsymbol{w}^{(k)}$ comme combinaison linéaire de $\boldsymbol{r}^{(k)}$ et de $\boldsymbol{w}^{(k-1)}$, de manière à ce que $\boldsymbol{w}^{(k)}$ soit orthogonal à $\boldsymbol{w}^{(k-1)}$ pour le produit scalaire associé à la matrice A.

$$\boldsymbol{w}^{(0)} = \boldsymbol{r}^{(0)}, \tag{3.26a}$$

$$\boldsymbol{w}^{(k)} = \boldsymbol{r}^{(k)} + \lambda_k \boldsymbol{w}^{(k-1)}, \text{ avec } \boldsymbol{w}^{(k)} \cdot A\boldsymbol{w}^{(k-1)} = 0, \text{ pour } k \geq 1. \tag{3.26b}$$

La contrainte d'orthogonalité $A\boldsymbol{w}^{(k)} \cdot \boldsymbol{w}^{(k-1)} = 0$ impose le choix du paramètre λ_k suivant :

$$\lambda_k = -\frac{\boldsymbol{r}^{(k)} \cdot A\boldsymbol{w}^{(k-1)}}{\boldsymbol{w}^{(k-1)} \cdot A\boldsymbol{w}^{(k-1)}}.$$

Remarquons que si $\boldsymbol{r}^{(k)} \neq \boldsymbol{0}$ alors $\boldsymbol{w}^{(k)} \cdot \boldsymbol{r}^{(k)} > 0$ car $\boldsymbol{w}^{(k)} \cdot \boldsymbol{r}^{(k)} = \boldsymbol{r}^{(k)} \cdot \boldsymbol{r}^{(k)}$ en raison de la propriété (3.25). On a donc $\boldsymbol{w}^{(k)} \cdot \nabla f(\boldsymbol{x}^{(k)}) < 0$, ce qui montre que $\boldsymbol{w}^{(k)}$ est bien une direction de descente stricte.
On a donc (on a déjà fait ce calcul pour obtenir la formule (3.22) du paramètre optimal)

$$\alpha_k = \frac{\boldsymbol{r}^{(k)} \cdot \boldsymbol{w}^{(k)}}{A\boldsymbol{w}^{(k)} \cdot \boldsymbol{w}^{(k)}} = \frac{\boldsymbol{r}^{(k)} \cdot \boldsymbol{r}^{(k)}}{A\boldsymbol{w}^{(k)} \cdot \boldsymbol{w}^{(k)}}. \tag{3.27}$$

On suppose que $\boldsymbol{r}^{(k)} \neq \boldsymbol{0}$ pour tout $k \in \{0, \ldots, n-1\}$. Montrons alors par récurrence que pour $k = 1, \ldots, n-1$, on a :

$(i)_k$ $\quad \boldsymbol{r}^{(k)} \cdot \boldsymbol{w}^{(p)} = 0$ si $p < k$,

$(ii)_k$ $\quad \boldsymbol{r}^{(k)} \cdot \boldsymbol{r}^{(p)} = 0$ si $p < k$,

$(iii)_k$ $\quad A\boldsymbol{w}^{(k)} \cdot \boldsymbol{w}^{(p)} = 0$ si $p < k$,

Ces relations sont vérifiées pour $k = 1$. Supposons qu'elles le sont jusqu'au rang k, et montrons qu'elles le sont au rang $k+1$.

$(i)_{k+1}$: Pour $p = k$, la relation $(i)_{k+1}$ est verifiée au rang $k+1$ grâce à (3.25) ; pour $p < k$, on a

$$\boldsymbol{r}^{(k+1)} \cdot \boldsymbol{w}^{(p)} = \boldsymbol{r}^{(k)} \cdot \boldsymbol{w}^{(p)} - \alpha_k A\boldsymbol{w}^{(k)} \cdot \boldsymbol{w}^{(p)} = 0$$

par (3.24) et hypothèse de récurrence.

$(ii)_{k+1}$: Par les relations (3.26b) et $(i)_{k+1}$, on a, pour $p \leq k$,

$$\boldsymbol{r}^{(k+1)} \cdot \boldsymbol{r}^{(p)} = \boldsymbol{r}^{(k+1)} \cdot (\boldsymbol{w}^{(p)} - \lambda_p \boldsymbol{w}^{(p-1)}) = 0.$$

$(iii)_{k+1}$: Pour $p = k$ la relation $(iii)_{k+1}$ est vérifiée grâce au choix de λ_{k+1}.
Pour $p < k$, on remarque que, avec (3.26b) et $(iii)_k$

$$\boldsymbol{w}^{(k+1)} \cdot A\boldsymbol{w}^{(p)} = (\boldsymbol{r}^{(k+1)} + \lambda_{k+1}\boldsymbol{w}^{(k)}) \cdot A\boldsymbol{w}^{(p)} = \boldsymbol{r}^{(k+1)} \cdot A\boldsymbol{w}^{(p)}.$$

On utilise maintenant (3.24) et $(i)_{k+1}$ pour obtenir

$$\boldsymbol{w}^{(k+1)} \cdot A\boldsymbol{w}^{(p)} = \frac{1}{\alpha_p}\boldsymbol{r}^{(k+1)} \cdot (\boldsymbol{r}^{(p)} - \boldsymbol{r}^{(p+1)}) = 0.$$

On a ainsi démontré la convergence de la méthode du gradient conjugué.

Mettons sous forme algorithmique les opérations que nous avons exposées, pour obtenir l'algorithme du gradient conjugué.

Algorithme 3.22 (Méthode du gradient conjugué)**.**

1. **Initialisation**
   Soit $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$, *et soit* $\boldsymbol{r}^{(0)} = \boldsymbol{b} - A\boldsymbol{x}^{(0)} = -\nabla f(\boldsymbol{x}^{(0)})$.

   Si $\boldsymbol{r}^{(0)} = \boldsymbol{0}$, *alors* $A\boldsymbol{x}^{(0)} = \boldsymbol{b}$ *et donc* $\boldsymbol{x}^{(0)} = \bar{\boldsymbol{x}}$, *auquel cas l'algorithme s'arrête.*

   Sinon, on pose
   $$\boldsymbol{w}^{(0)} = \boldsymbol{r}^{(0)},$$
   et on choisit α_0 *optimal dans la direction* $\boldsymbol{w}^{(0)}$. *On pose alors*
   $$\boldsymbol{x}^{(1)} = \boldsymbol{x}^{(0)} + \alpha_0\boldsymbol{w}^{(0)}.$$

2. **Itération** k, $1 \leq k \leq n-1$*; on suppose* $\boldsymbol{x}^{(0)}, \ldots, \boldsymbol{x}^{(k)}$ *et* $\boldsymbol{w}^{(0)}, \ldots, \boldsymbol{w}^{(k-1)}$ *connus et on pose*
   $$\boldsymbol{r}^{(k)} = \boldsymbol{b} - A\boldsymbol{x}^{(k)}.$$

   Si $\boldsymbol{r}^{(k)} = \boldsymbol{0}$, *alors* $A\boldsymbol{x}^{(k)} = \boldsymbol{b}$ *et donc* $\boldsymbol{x}^{(k)} = \bar{\boldsymbol{x}}$, *auquel cas l'algorithme s'arrête.*

   Sinon on pose
   $$\boldsymbol{w}^{(k)} = \boldsymbol{r}^{(k)} + \lambda_{k-1}\boldsymbol{w}^{(k-1)},$$
   avec λ_{k-1} *tel que*
   $$\boldsymbol{w}^{(k)} \cdot A\boldsymbol{w}^{(k-1)} = 0,$$
   et on choisit α_k *optimal dans la direction* $\boldsymbol{w}^{(k)}$, *donné par* (3.22). *On pose alors*
   $$\boldsymbol{x}^{(k+1)} = x^{(k)} + \alpha_k\boldsymbol{w}^{(k)}.$$

Nous avons démontré plus haut la convergence de l'algorithme, résultat que nous énonçons dans le théorème suivant.

Théorème 3.23 (Convergence de l'algorithme du gradient conjugué)**.** *Soit* A *une symétrique définie positive,* $A \in \mathcal{M}_n(\mathbb{R})$, $\boldsymbol{b} \in \mathbb{R}^n$ *et* $f(\boldsymbol{x}) = \frac{1}{2}A\boldsymbol{x} \cdot \boldsymbol{x} - \boldsymbol{b} \cdot \boldsymbol{x}$. *L'algorithme* (3.22) *définit une suite* $(\boldsymbol{x}^{(k)})_{k=0,\ldots,p}$ *avec* $p \leq n$ *telle que* $\boldsymbol{x}^{(p)} = \bar{\boldsymbol{x}}$ *avec* $A\bar{\boldsymbol{x}} = \boldsymbol{b}$. *On obtient donc la solution exacte de la solution du système linéaire* $A\boldsymbol{x} = \boldsymbol{b}$ *en moins de* n *itérations.*

Efficacité de la méthode du gradient conjugué On peut calculer le nombre d'opérations nécessaires pour calculer $\bar{\boldsymbol{x}}$ (c.à.d. pour calculer $\boldsymbol{x}^{(n)}$, sauf dans le cas miraculeux où $\boldsymbol{x}^{(k)} = \bar{\boldsymbol{x}}$ pour $k < n$) et montrer (exercice) que :

$$N_{gc} = 2n^3 + \mathcal{O}(n^2).$$

On rappelle que le nombre d'opérations pour Choleski est $\dfrac{n^3}{6}$ donc la méthode du gradient conjugué n'est pas intéressante comme méthode directe car elle demande 12 fois plus d'opérations que Choleski.

On peut alors se demander si la méthode est intéressante comme méthode itérative, c.à.d. si on peut espérer que $\boldsymbol{x}^{(k)}$ soit "proche de $\bar{\boldsymbol{x}}$" pour "$k \ll n$". Malheureusement, si la dimension n du système est grande, ceci n'est pas le cas en raison de l'accumulation des erreurs d'arrondi. Il est même possible de devoir effectuer plus de n itérations pour se rapprocher de $\bar{\boldsymbol{x}}$. Cependant, dans les années 80, des chercheurs se sont rendus compte que ce défaut pouvait être corrigé à condition d'utiliser un "préconditionnement". Donnons par exemple le principe du préconditionnement dit de "Choleski incomplet".

Méthode du gradient conjugué préconditionné par Choleski incomplet On commence par calculer une "approximation" de la matrice de Choleski de A c.à.d. qu'on cherche L triangulaire inférieure inversible telle que A soit "proche" de LL^t, en un sens à définir. Si on pose $y = L^t\boldsymbol{x}$, alors le système $A\boldsymbol{x} = b$ peut aussi s'écrire $L^{-1}A(L^t)^{-1}y = L^{-1}b$, et le système $(L^t)^{-1}y = \boldsymbol{x}$ est facile à résoudre car L^t est triangulaire supérieure. Soit $B \in \mathcal{M}_n(\mathbb{R})$ définie par $B = L^{-1}A(L^t)^{-1}$, alors

$$B^t = ((L^t)^{-1})^t A^t (L^{-1})^t = L^{-1}A(L^t)^{-1} = B$$

et donc B est symétrique. De plus,

$$B\boldsymbol{x}\cdot\boldsymbol{x} = L^{-1}A(L^t)^{-1}\boldsymbol{x}\cdot\boldsymbol{x} = A(L^t)^{-1}\boldsymbol{x}\cdot(L^t)^{-1}\boldsymbol{x},$$

et donc $B\boldsymbol{x}\cdot\boldsymbol{x} > 0$ si $\boldsymbol{x} \neq 0$. La matrice B est donc symétrique définie positive. On peut donc appliquer l'algorithme du gradient conjugué à la recherche du minimum de la fonction f définie par

$$f(\boldsymbol{y}) = \frac{1}{2}B\boldsymbol{y}\cdot\boldsymbol{y} - L^{-1}b\cdot\boldsymbol{y}.$$

On en déduit l'expression de la suite $(\boldsymbol{y}^{(k)})_{k\in\mathbb{N}}$ et donc $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$.
On peut alors montrer (voir exercice 148) que l'algorithme du gradient conjugué préconditionné ainsi obtenu peut s'écrire directement pour la suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$, de la manière suivante :

Itération k On pose $\boldsymbol{r}^{(k)} = \boldsymbol{b} - A\boldsymbol{x}^{(k)}$,
on calcule $\boldsymbol{s}^{(k)}$ solution de $LL^t\boldsymbol{s}^{(k)} = \boldsymbol{r}^{(k)}$.
On pose alors $\lambda_{k-1} = \dfrac{\boldsymbol{s}^{(k)}\cdot\boldsymbol{r}^{(k)}}{\boldsymbol{s}^{(k-1)}\cdot\boldsymbol{r}^{(k-1)}}$ et $\boldsymbol{w}^{(k)} = \boldsymbol{s}^{(k)} + \lambda_{k-1}\boldsymbol{w}^{(k-1)}$.
Le paramètre optimal α_k a pour expression :

$$\alpha_k = \frac{\boldsymbol{s}^{(k)}\cdot\boldsymbol{r}^{(k)}}{A\boldsymbol{w}^{(k)}\cdot\boldsymbol{w}^{(k)}},$$

et on pose alors $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k\boldsymbol{w}^{(k)}$.
Le choix de la matrice L peut se faire par exemple dans le cas d'une matrice creuse, en effectuant une factorisation "LL^t" incomplète, qui consiste à ne remplir que certaines diagonales de la matrice L pendant la factorisation, et laisser les autres à 0.

Méthode du gradient conjugué pour une fonction non quadratique. On peut généraliser le principe de l'algorithme du gradient conjugué à une fonction f non quadratique. Pour cela, on reprend le même algorithme que (3.22), mais on adapte le calcul de λ_{k-1} et α_k.

Itération n :
A $\boldsymbol{x}^{(0)},\ldots,\boldsymbol{x}^{(k)}$ et $w^{(0)},\ldots,w^{(k-1)}$ connus, on calcule $\boldsymbol{r}^{(k)} = -\nabla f(\boldsymbol{x}^{(k)})$.
Si $\boldsymbol{r}^{(k)} = 0$ alors $\nabla f(\boldsymbol{x}^{(k)}) = 0$ auquel cas l'algorithme s'arrête (le point $\boldsymbol{x}^{(k)}$ est un point critique de f et il minimise f si f est convexe).
Si $\boldsymbol{r}^{(k)} \neq 0$, on pose $\boldsymbol{w}^{(k)} = \boldsymbol{r}^{(k)} + \lambda_{k-1}w^{(k-1)}$ où λ_{k-1} peut être choisi de différentes manières :

1ère méthode (Fletcher–Reeves)

$$\lambda_{k-1} = \frac{\boldsymbol{r}^{(k)}\cdot\boldsymbol{r}^{(k)}}{\boldsymbol{r}^{(k-1)}\cdot\boldsymbol{r}^{(k-1)}},$$

2ème méthode (Polak–Ribière)

$$\lambda_{k-1} = \frac{(\boldsymbol{r}^{(k)} - \boldsymbol{r}^{(k-1)}) \cdot \boldsymbol{r}^{(k)}}{\boldsymbol{r}^{(k-1)} \cdot \boldsymbol{r}^{(k-1)}}.$$

On pose alors $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{w}^{(k)}$, où α_k est choisi, si possible, optimal dans la direction $\boldsymbol{w}^{(k)}$.
La démonstration de la convergence de l'algorithme de Polak–Ribière fait l'objet de l'exercice 150 page 220.

En résumé, la méthode du gradient conjugué est très efficace dans le cas d'une fonction quadratique à condition de l'utiliser avec préconditionnement. Dans le cas d'une fonction non quadratique, le préconditionnement ne se trouve pas de manière naturelle et il vaut donc mieux réserver cette méthode dans le cas "n petit".

3.3.3 Méthodes de Newton et Quasi–Newton

Soit $f \in C^2(\mathbb{R}^n, \mathbb{R})$ et $g = \nabla f \in C^1(\mathbb{R}^n, \mathbb{R}^n)$. On a dans ce cas :

$$f(\boldsymbol{x}) = \inf_{\mathbb{R}^n} f \Rightarrow g(\boldsymbol{x}) = 0.$$

Si de plus f est convexe alors on a $g(\boldsymbol{x}) = 0 \Rightarrow f(\boldsymbol{x}) = \inf_{\mathbb{R}^n} f$. Dans ce cas d'équivalence, on peut employer la méthode de Newton pour minimiser f en appliquant l'algorithme de Newton pour chercher un zéro de $g = \nabla f$. On a $D(\nabla f) = H_f$ où $H_f(\boldsymbol{x})$ est la matrice hessienne de f en $\boldsymbol{x}$. La méthode de Newton s'écrit dans ce cas :

$$\left\{ \begin{array}{ll} \text{Initialisation} & \boldsymbol{x}^{(0)} \in \mathbb{R}^n, \\ \text{Itération } k & H_f(\boldsymbol{x}^{(k)})(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}) = -\nabla f(\boldsymbol{x}^{(k)}). \end{array} \right. \tag{3.29}$$

Remarque 3.24. La méthode de Newton pour minimiser une fonction f convexe est une méthode de descente. En effet, si $H_f(\boldsymbol{x}^{(k)})$ est inversible, on a $\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)} = [H_f(\boldsymbol{x}^{(k)})]^{-1}(-\nabla f(\boldsymbol{x}^{(k)}))$ soit encore $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{w}^{(k)}$ où $\alpha_k = 1$ et $w^{(k)} = [H_f(\boldsymbol{x}^{(k)})]^{-1}(-\nabla f(\boldsymbol{x}^{(k)}))$. Si f est convexe, H_f est une matrice symétrique positive (déjà vu). Comme on suppose $H_f(\boldsymbol{x}^{(k)})$ inversible par hypothèse, la matrice $H_f(\boldsymbol{x}^{(k)})$ est donc symétrique définie positive.
On en déduit que $w^{(k)} = 0$ si $\nabla f(\boldsymbol{x}^{(k)}) = 0$ et, si $\nabla f(\boldsymbol{x}^{(k)}) \neq 0$,

$$-\boldsymbol{w}^{(k)} \cdot \nabla f(\boldsymbol{x}^{(k)}) = [H_f(\boldsymbol{x}^{(k)})]^{-1} \nabla f(\boldsymbol{x}^{(k)}) \cdot \nabla f(\boldsymbol{x}^{(k)}) > 0,$$

ce qui est une condition suffisante pour que $\boldsymbol{w}^{(k)}$ soit une direction de descente stricte.
La méthode de Newton est donc une méthode de descente avec $\boldsymbol{w}^{(k)} = -H_f(\boldsymbol{x}^{(k)})^{-1}(\nabla f(\boldsymbol{x}^{(k)}))$ et $\alpha_k = 1$.

On peut aussi remarquer, en vertu du théorème 2.19 page 158, que si $f \in C^3(\mathbb{R}^n, \mathbb{R})$, si $\bar{\boldsymbol{x}}$ est tel que $\nabla f(\bar{\boldsymbol{x}}) = 0$ et si $H_f(\bar{\boldsymbol{x}}) = D(\nabla f)(\bar{\boldsymbol{x}})$ est inversible alors il existe $\varepsilon > 0$ tel que si $\boldsymbol{x}_0 \in B(\bar{\boldsymbol{x}}, \varepsilon)$, alors la suite $(\boldsymbol{x}^{(k)})_k$ est bien définie par (3.29) et $\boldsymbol{x}^{(k)} \to \bar{\boldsymbol{x}}$ lorsque $k \to +\infty$. De plus, d'après la proposition 2.16, il existe $\beta > 0$ tel que $|\boldsymbol{x}^{(k+1)} - \bar{\boldsymbol{x}}| \leq \beta |\boldsymbol{x}^{(k)} - \bar{\boldsymbol{x}}|^2$ pour tout $k \in \mathbb{N}$.

Remarque 3.25 (Sur l'implantation numérique)**.** La convergence de la méthode de Newton est très rapide, mais nécessite en revanche le calcul de $H_f(\boldsymbol{x})$, qui peut s'avérer impossible ou trop coûteux.

On va maintenant donner des variantes de la méthode de Newton qui évitent le calcul de la matrice hessienne.

Proposition 3.26. *Soient $f \in C^1(\mathbb{R}^n, \mathbb{R})$, $\boldsymbol{x} \in \mathbb{R}^n$ tel que $\nabla f(\boldsymbol{x}) \neq 0$, et soit $B \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique définie positive ; alors $w = -B\nabla f(\boldsymbol{x})$ est une direction de descente stricte en $\boldsymbol{x}$.*

DÉMONSTRATION – On a : $w \cdot \nabla f(\boldsymbol{x}) = -B\nabla f(\boldsymbol{x}) \cdot \nabla f(\boldsymbol{x}) < 0$ car B est symétrique définie positive et $\nabla f(\boldsymbol{x}) \neq 0$ donc $\boldsymbol{w}$ est une direction de descente stricte en $\boldsymbol{x}$. En effet, soit φ la fonction de $\mathbb{R}$ dans $\mathbb{R}$ définie par $\varphi(\alpha) = f(\boldsymbol{x} + \alpha \boldsymbol{w})$. Il est clair que $\varphi \in C^1(\mathbb{R}, \mathbb{R})$, $\varphi'(\alpha) = \nabla f(\boldsymbol{x} + \alpha \boldsymbol{w}) \cdot w$ et $\varphi'(0) = \nabla f(\boldsymbol{x}) \cdot w < 0$. Donc $\exists \alpha_0 > 0$ tel que $\varphi'(\alpha) < 0$ si $\alpha \in]0, \alpha_0[$. Par le théorème des accroissements finis, $\varphi(\alpha) < \varphi(0)$ $\forall \alpha \in]0, \alpha_0[$ donc w est une direction de descente stricte. ∎

Méthode de Broyden La première idée pour construire une méthode de type quasi Newton est de prendre comme direction de descente en $\boldsymbol{x}^{(k)}$ le vecteur $\boldsymbol{w}^{(k)} = -(B^{(k)})^{-1}(\nabla f(\boldsymbol{x}^{(k)}))$ où la matrice $B^{(k)}$ est censée approcher $H_f(\boldsymbol{x}^{(k)})$ (sans calculer la dérivée seconde de f). On suppose $\boldsymbol{x}^{(k)}, \boldsymbol{x}^{(k-1)}$ et $B^{(k-1)}$ connus. Voyons comment on peut déterminer $B^{(k)}$. On peut demander par exemple que la condition suivante soit satisfaite :

$$\nabla f(\boldsymbol{x}^{(k)}) - \nabla f(\boldsymbol{x}^{(k-1)}) = B^{(k)}(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}). \tag{3.30}$$

Ceci est un système à n équations et $n \times n$ inconnues, et ne permet donc pas de déterminer entièrement la matrice $B^{(k)}$ si $n > 1$. Voici un moyen possible pour déterminer entièrement $B^{(k)}$, dû à Broyden. On pose $\boldsymbol{s}^{(k)} = \boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}$, on suppose que $\boldsymbol{s}^{(k)} \neq 0$, et on pose $\boldsymbol{y}^{(k)} = \nabla f(\boldsymbol{x}^{(k)}) - \nabla f(\boldsymbol{x}^{(k-1)})$. On choisit alors $B^{(k)}$ telle que :

$$\begin{cases} B^{(k)}\boldsymbol{s}^{(k)} = \boldsymbol{y}^{(k)} \\ B^{(k)}\boldsymbol{s} = B^{(k-1)}\boldsymbol{s},\ \forall s \perp s^{(k)} \end{cases} \tag{3.31}$$

On a exactement le nombre de conditions qu'il faut avec (3.31) pour déterminer entièrement $B^{(k)}$. Ceci suggère la méthode suivante :

Initialisation Soient $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$ et $B^{(0)}$ une matrice symétrique définie positive. On pose

$$\boldsymbol{w}^{(0)} = (B^{(0)})^{-1}(-\nabla f(\boldsymbol{x}^{(0)}));$$

alors $\boldsymbol{w}^{(0)}$ est une direction de descente stricte sauf si $\nabla f(\boldsymbol{x}^{(0)}) = 0$.

On pose alors

$$\boldsymbol{x}^{(1)} = \boldsymbol{x}^{(0)} + \alpha^{(0)}\boldsymbol{w}^{(0)},$$

où $\alpha^{(0)}$ est optimal dans la direction $\boldsymbol{w}^{(0)}$.

Itération k On suppose $\boldsymbol{x}^{(k)}$, $\boldsymbol{x}^{(k-1)}$ et $B^{(k-1)}$ connus, $(k \geq 1)$, et on calcule $B^{(k)}$ par (3.31). On pose

$$\boldsymbol{w}^{(k)} = -(B^{(k)})^{-1}(\nabla f(\boldsymbol{x}^{(k)})).$$

On choisit $\alpha^{(k)}$ optimal en $\boldsymbol{x}^{(k)}$ dans la direction $\boldsymbol{w}^{(k)}$, et on pose $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha^{(k)}\boldsymbol{w}^{(k)}$.

Le problème avec cet algorithme est que si la matrice est $B^{(k-1)}$ symétrique définie positive, la matrice $B^{(k)}$ ne l'est pas forcément, et donc $\boldsymbol{w}^{(k)}$ n'est pas forcément une direction de descente stricte. On va donc modifier cet algorithme dans ce qui suit.

Méthode de BFGS La méthode BFGS (de Broyden [1], Fletcher [2], Goldfarb [3] et Shanno [4]) cherche à construire $B^{(k)}$ proche de $B^{(k-1)}$, telle que $B^{(k)}$ vérifie (3.30) et telle que si $B^{(k-1)}$ est symétrique définie positive alors $B^{(k)}$ est symétrique définie positive. On munit $\mathcal{M}_n(\mathbb{R})$ d'une norme induite par un produit scalaire, par exemple si $A \in \mathcal{M}_n(\mathbb{R})$ et $A = (a_{i,j})_{i,j=1,\ldots,n}$ on prend $\|A\| = \left(\sum_{i,j=1}^n a_{i,j}^2\right)^{1/2}$. $\mathcal{M}_n(\mathbb{R})$ est alors un espace de Hilbert.

On suppose $\boldsymbol{x}^{(k)}$, $\boldsymbol{x}^{(k-1)}$, $B^{(k-1)}$ connus, et on définit

$$\mathcal{C}_k = \{B \in \mathcal{M}_n(\mathbb{R}) | B \text{ symétrique, vérifiant (3.30)}\},$$

qui est une partie de $\mathcal{M}_n(\mathbb{R})$ convexe fermée non vide. On choisit alors $B^{(k)} = P_{\mathcal{C}_k}B^{(k-1)}$ où $P_{\mathcal{C}_k}$ désigne la projection orthogonale sur $\mathcal{C}_k$. La matrice $B^{(k)}$ ainsi définie existe et est unique ; elle est symétrique d'après le choix de $\mathcal{C}_k$. On peut aussi montrer que si $B^{(k-1)}$ symétrique définie positive alors $B^{(k)}$ est aussi symétrique définie positive.

1. Broyden, C. G., The Convergence of a Class of Double-rank Minimization Algorithms, *Journal of the Institute of Mathematics and Its Applications* 1970, 6, 76-90
2. Fletcher, R., A New Approach to Variable Metric Algorithms, Computer Journal 1970, 13, 317-322
3. Goldfarb, D., A Family of Variable Metric Updates Derived by Variational Means, Mathematics of Computation 1970, 24, 23-26
4. Shanno, D. F.,Conditioning of Quasi-Newton Methods for Function Minimization , Mathematics of Computation 1970, 24, 647-656

Avec un choix convenable de la norme sur $\mathcal{M}_n(\mathbb{R})$, on obtient le choix suivant de $B^{(k)}$ si $s^{(k)} \neq 0$ et $\nabla f(\boldsymbol{x}^{(k)}) \neq 0$ (sinon l'algorithme s'arrête) :

$$B^{(k)} = B^{(k-1)} + \frac{\boldsymbol{y}^{(k)}(\boldsymbol{y}^{(k)})^t}{(\boldsymbol{s}^{(k)})^t \cdot y^{(k)}} - \frac{B^{(k-1)}\boldsymbol{s}^{(k)}(\boldsymbol{s}^{(k)})^t B^{(k-1)}}{(\boldsymbol{s}^{(k)})^t B^{(k-1)}\boldsymbol{s}^{(k)}}. \tag{3.32}$$

L'algorithme obtenu est l'algorithme de BFGS.

Algorithme de BFGS

$$\left\{\begin{array}{ll}
\underline{\text{Initialisation}} & \text{On choisit } \boldsymbol{x}^{(0)} \in \mathbb{R}^n \text{ et} \\
& B^{(0)} \text{ symétrique définie positive} \\
& (\text{ par exemple } B^{(0)} = Id) \text{ et on pose} \\
& \boldsymbol{w}^{(0)} = -B^{(0)}\nabla f(\boldsymbol{x}^{(0)}) \\
& \text{si } \nabla f(\boldsymbol{x}^{(0)}) \neq 0, \text{ on choisit } \alpha^{(0)} \text{ optimal} \\
& \text{dans la direction } \boldsymbol{w}^{(0)}, \text{ et donc} \\
& \boldsymbol{w}^{(0)} \text{ est une direction de descente stricte.} \\
& \text{On pose } \boldsymbol{x}^{(1)} = \boldsymbol{x}^{(0)} + \alpha^{(0)}\boldsymbol{w}^{(0)}. \\
\underline{\text{Itération } k} & \text{A } \boldsymbol{x}^{(k)}, \boldsymbol{x}^{(k-1)} \text{ et } B_{k-1} \text{ connus } \quad (k \geq 1) \\
& \text{On pose} \\
& \boldsymbol{s}^{(k)} = \boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)} \;\; \boldsymbol{y}^{(k)} = \nabla f(\boldsymbol{x}^{(k)}) - \nabla f(\boldsymbol{x}^{(k-1)}) \\
& \text{si } \boldsymbol{s}^{(k)} \neq 0 \text{ et } \nabla f(\boldsymbol{x}^{(k)}) \neq 0, \\
& \text{on choisit } B^{(k)} \text{ vérifiant (3.32)} \\
& \text{On calcule } \boldsymbol{w}^{(k)} = -(B^{(k)})^{-1}(\nabla f(\boldsymbol{x}^{(k)})) \\
& (\text{direction de descente stricte en } \boldsymbol{x}^{(k)}). \\
& \text{On calcule } \alpha^{(k)} \text{ optimal dans la direction } \boldsymbol{w}^{(k)} \\
& \text{et on pose } \boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha^{(k)}\boldsymbol{w}^{(k)}.
\end{array}\right. \tag{3.33}$$

On donne ici sans démonstration le théorème de convergence suivant :

Théorème 3.27 (Fletcher, 1976)**.** *Soit $f \in C^2(\mathbb{R}^n, \mathbb{R})$ telle que $f(\boldsymbol{x}) \to +\infty$ quand $|\boldsymbol{x}| \to +\infty$. On suppose de plus que f est strictement convexe (donc il existe un unique $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ tel que $f(\bar{\boldsymbol{x}}) = \inf_{\mathbb{R}^n} f$) et on suppose que la matrice hessienne $H_f(\bar{\boldsymbol{x}})$ est symétrique définie positive.*
Alors si $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$ et si $B^{(0)}$ est symétrique définie positive, l'algorithme BFGS définit bien une suite $\boldsymbol{x}^{(k)}$ et on a $\boldsymbol{x}^{(k)} \to \bar{\boldsymbol{x}}$ quand $k \to +\infty$
De plus, si $\boldsymbol{x}^{(k)} \neq \bar{\boldsymbol{x}}$ pour tout k, la convergence est super linéaire i.e.

$$\left|\frac{\boldsymbol{x}^{(k+1)} - \bar{\boldsymbol{x}}}{\boldsymbol{x}^{(k)} - \bar{\boldsymbol{x}}}\right| \to 0 \text{ quand } k \to +\infty.$$

Pour éviter la résolution d'un système linéaire dans BFGS, on peut choisir de travailler sur $(B^{(k)})^{-1}$ au lieu de $B^{(k)}$.

$$\left\{\begin{array}{l}
\textbf{Initialisation Soit } \boldsymbol{x}^{(0)} \in \mathbb{R}^n \text{ et } K^{(0)} \text{ symétrique définie positive} \\
\text{telle que } \alpha_0 \text{ soit optimal dans la direction } -K^{(0)}\nabla f(\boldsymbol{x}^{(0)}) = \boldsymbol{w}^{(0)} \\
\boldsymbol{x}^{(1)} = \boldsymbol{x}^{(0)} + \alpha_0 w^{(0)} \\
\textbf{Itération } k : \text{A } \boldsymbol{x}^{(k)}, \boldsymbol{x}^{(k-1)}, K^{(k-1)} \text{ connus, } k \geq 1, \\
\text{on pose } \boldsymbol{s}^{(k)} = \boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}, \boldsymbol{y}^{(k)} = \nabla f(\boldsymbol{x}^{(k)}) - \nabla f(\boldsymbol{x}^{(k-1)}) \\
\text{et } K^{(k)} = P_{\mathcal{C}_k} K^{(k-1)}. \\
\text{On calcule } \boldsymbol{w}^{(k)} = -K^{(k)}\nabla f(\boldsymbol{x}^{(k)}) \text{ et on choisit } \alpha_k \\
\text{optimal dans la direction } \boldsymbol{w}^{(k)}. \\
\text{On pose alors } \boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{w}^{(k)}.
\end{array}\right. \tag{3.34}$$

Remarquons que le calcul de la projection de $P_{\mathcal{C}_k} K^{(k-1)}$ peut s'effectuer avec la formule (3.32) où on a remplacé $B^{(k-1)}$ par $K^{(k-1)}$. Malheureusement, on obtient expérimentalement une convergence nettement moins bonne pour l'algorithme de quasi-Newton modifié (3.34) que pour l'algorithme de BFGS (3.32).

3.3.4 Résumé sur les méthodes d'optimisation

Faisons le point sur les avantages et inconvénients des méthodes qu'on a vues sur l'optimisation sans contrainte.

Méthodes de gradient : Ces méthodes nécessitent le calcul de $\nabla f(\boldsymbol{x}^{(k)})$. Leur convergence est linéaire (donc lente).

Méthode de gradient conjugué : Si f est quadratique (c.à.d. $f(\boldsymbol{x}) = \frac{1}{2} A\boldsymbol{x} \cdot \boldsymbol{x} - b \cdot \boldsymbol{x}$ avec A symétrique définie positive), la méthode est excellente si elle est utilisée avec un préconditionnement (pour n grand). Dans le cas général, elle n'est efficace que si n n'est pas trop grand.

Méthode de Newton : La convergence de la méthode de Newton est excellente (convergence localement quadratique) mais nécessite le calcul de $H_f(\boldsymbol{x}^{(k)})$ (et de $\nabla f(\boldsymbol{x}^{(k)})$). Si on peut calculer $H_f(\boldsymbol{x}^{(k)})$, cette méthode est parfaite.

Méthode de quasi Newton : L'avantage de la méthode de quasi Newton est qu'on ne calcule que $\nabla f(x^{(k)})$ et pas $H_f(x^{(k)})$). La convergence est super linéaire. Par rapport à une méthode de gradient où on calcule $\boldsymbol{w}^{(k)} = -\nabla f(\boldsymbol{x}^{(k)})$, la méthode BFGS nécessite une résolution de système linéaire :

$$B^{(k)} \boldsymbol{w}^{(k)} = -\nabla f(\boldsymbol{x}^{(k)}).$$

Quasi–Newton modifié :

Pour éviter la résolution de système linéaire dans BFGS, on peut choisir de travailler sur $(B^{(k)})^{-1}$ au lieu de $B^{(k)}$, pour obtenir l'algorithme de quasi Newton (3.34). Cependant, on perd alors en vitesse de convergence.

Comment faire si on ne veut (ou peut) pas calculer $\nabla f(\boldsymbol{x}^{(k)})$? On peut utiliser des "méthodes sans gradient", c.à.d. qu'on choisit *a priori* les directions $\boldsymbol{w}^{(k)}$. Ceci peut se faire soit par un choix déterministe, soit par un choix stochastique.

Un choix déterministe possible est de calculer $\boldsymbol{x}^{(k)}$ en résolvant n problèmes de minimisation en une dimension d'espace. Pour chaque direction $i = 1, \ldots, n$, on prend $w^{(n,i)} = \boldsymbol{e}_i$, où $\boldsymbol{e}_i$ est le i-ème vecteur de la base canonique, et pour $i = 1, \ldots, n$, on cherche $\theta \in \mathbb{R}$ tel que :

$$f(x_1^{(k)}, x_2^{(k)}, \ldots, \theta, \ldots, x_n^{(k)}) \leq f(x_1^{(k)}, \boldsymbol{x}_2^{(k)}, \ldots, t, \ldots, \boldsymbol{x}_n^{(k)}), \forall t \in \mathbb{R}.$$

Remarquons que si f est quadratique, on retrouve la méthode de Gauss Seidel.

3.3.5 Exercices (algorithmes pour l'optimisation sans contraintes)

Exercice 140 (Minimisation d'une fonction quadratique)**.**

Soient $N \in \mathbb{N}$, $A \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique définie positive, $\boldsymbol{b} \in \mathbb{R}^n$ et $c \in \mathbb{R}$, et soit f une fonction de $\mathbb{R}^n$ dans $\mathbb{R}$ définie par (3.12), où $\boldsymbol{x} \cdot \boldsymbol{y}$ désigne le produit scalaire de $\boldsymbol{x} \in \mathbb{R}^{\mathbb{N}}$ et $\boldsymbol{y} \in \mathbb{R}^{\mathbb{N}}$; on note $|\cdot|_2$ la norme euclidienne sur $\mathbb{R}^{\mathbb{N}}$, et $\|\cdot\|_2$ la norme matricielle induite sur $\mathcal{M}_n(\mathbb{R})$.

1. Montrer que f est de classe $C^3(\mathbb{R}^n, \mathbb{R})$. Donner son gradient, sa hessienne et sa différentielle troisième.
2. Montrer que

$$(\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})) \cdot (\boldsymbol{x} - \boldsymbol{y}) \geq \alpha |\boldsymbol{x} - \boldsymbol{y}|_2^2, \ \forall (\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^n \times \mathbb{R}^n, \forall (\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^n, \tag{3.35}$$

$$|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}))|_2 \leq M |\boldsymbol{x} - \boldsymbol{y}|_2, \forall (\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^n. \tag{3.36}$$

   où α est la plus petite valeur propre de A et $M > 0$ son rayon spectral.
3. Montrer que f est strictement convexe et que $f(\boldsymbol{x}) \to +\infty$ quand $|\boldsymbol{x}| \to +\infty$.

4. En déduire que f admet un unique minimum $\overline{\boldsymbol{x}} \in \mathbb{R}^n$ et donner la valeur de ce minimum.
5. En utilisant la question 2, montrer que la fonction h définie par

$$\boldsymbol{x} \in \mathbb{R}^n \mapsto h(\boldsymbol{x}) = \boldsymbol{x} - \rho \nabla f(\boldsymbol{x}) \in \mathbb{R}^n$$

est strictement contractante pour $\rho \in]0, \frac{2\alpha}{M^2}[$.

6. En déduire que pour tout $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$, la suite définie par

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \rho \nabla f(\boldsymbol{x}^{(k)}), k \in \mathbb{N}, \tag{3.37}$$

converge vers $\overline{\boldsymbol{x}}$.

7. Montrer que

$$\boldsymbol{x}^{(k+1)} - \overline{\boldsymbol{x}} = (\mathrm{Id} - \rho A)(\boldsymbol{x}^{(k)} - \overline{\boldsymbol{x}}).$$

8. En déduire que la suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ converge vers $\overline{\boldsymbol{x}}$ dès que $\rho \in]0, \frac{2\alpha}{M^2}[$.
9. Quel est le nom de l'algorithme (3.37) ?
10. Proposer d'autres méthodes, directes ou itératives, pour trouver le minimum de f.

Exercice 141 (Mise en oeuvre de GPF, GPO)**.** *Corrigé en page 223.*
On considère la fonction $f : \mathbb{R}^2 \to \mathbb{R}$ définie par $f(x_1, x_2) = 2x_1^2 + x_2^2 - x_1x_2 - 3x_1 - x_2 + 4$.

1. Montrer qu'il existe un unique $\bar{x} \in \mathbb{R}^2$ tel que $\bar{x} = \min_{x\in\mathbb{R}^2} f(x)$ et le calculer.
2. Calculer le premier itéré donné par l'algorithme du gradient à pas fixe (GPF) et du gradient à pas optimal (GPO), en partant de $(x_1^{(0)}, x_2^{(0)}) = (0, 0)$, pour un pas de $\alpha = .5$ dans le cas de GPF.

Exercice 142 (Convergence de l'algorithme du gradient à pas optimal)**.** *Suggestions en page 222. Corrigé détaillé en page 224*
Soit $f \in C^2(\mathbb{R}^n, \mathbb{R})$ t.q. $f(x) \to \infty$ quand $|x| \to \infty$. Soit $x_0 \in \mathbb{R}^n$. On va démontrer dans cet exercice la convergence de l'algorithme du gradient à pas optimal.

1. Montrer qu'il existe $R > 0$ t.q. $f(x) > f(x_0)$ pour tout $x \notin B_R$, avec $B_R = \{x \in \mathbb{R}^n, |x| \leq R\}$.
2. Montrer qu'il existe $M > 0$ t.q. $|H(x)y \cdot y| \leq M|y|^2$ pour tout $y \in \mathbb{R}^n$ et tout $x \in B_{R+1}$ ($H(x)$ est la matrice hessienne de f au point x, R est donné à la question 1).
3. (Construction de "la" suite $(x_k)_{k\in\mathbb{N}}$ de l'algorithme du gradient à pas optimal.) On suppose x_k connu ($k \in \mathbb{N}$). On pose $w_k = -\nabla f(x_k)$. Si $w_k = 0$, on pose $x_{k+1} = x_k$. Si $w_k \neq 0$, montrer qu'il existe $\overline{\rho} > 0$ t.q. $f(x_k + \overline{\rho} w_k) \leq f(x_k + \rho w_k)$ pour tout $\rho \geq 0$. On <u>choisit</u> alors un $\rho_k > 0$ t.q. $f(x_k + \rho_k w_k) \leq f(x_k + \rho w_k)$ pour tout $\rho \geq 0$ et on pose $x_{k+1} = x_k + \rho_k w_k$.
   On considère, dans les questions suivantes, la suite $(x_k)_{k\in\mathbb{N}}$ ainsi construite.
4. Montrer que (avec R et M donnés aux questions précédentes)
   (a) la suite $(f(x_k))_{k\in\mathbb{N}}$ est une suite convergente,
   (b) $x_k \in B_R$ pour tout $k \in \mathbb{N}$,
   (c) $f(x_k + \rho w_k) \leq f(x_k) - \rho|w_k|^2 + (\rho^2/2)M|w_k|^2$ pour tout $\rho \in [0, 1/|w_k|]$.
   (d) $f(x_{k+1}) \leq f(x_k) - |w_k|^2/(2M)$, si $|w_k| \leq M$.
   (e) $-f(x_{k+1}) + f(x_k) \geq |w_k|^2/(2\overline{M})$, avec $\overline{M} = \sup(M, \tilde{M})$, $\tilde{M} = \sup\{|\nabla f(x)|, x \in B_R\}$.
5. Montrer que $\nabla f(x_k) \to 0$ (quand $k \to \infty$) et qu'il existe une sous suite $(n_k)_{k\in\mathbb{N}}$ t.q. $x_{n_k} \to x$ quand $k \to \infty$ et $\nabla f(x) = 0$.
6. On suppose qu'il existe un unique $\overline{x} \in \mathbb{R}^n$ t.q. $\nabla f(\overline{x}) = 0$. Montrer que $f(\overline{x}) \leq f(x)$ pour tout $x \in \mathbb{R}^n$ et que $x_k \to \overline{x}$ quand $k \to \infty$.

Exercice 143 (Fonction non croissante à l'infini)**.** *Suggestions en page 223.*

Soient $n \geq 1$, $f \in C^2(\mathbb{R}^n, \mathbb{R})$ et $a \in \mathbb{R}$. On suppose que $A = \{x \in \mathbb{R}^n;\ f(x) \leq f(a)\}$ est un ensemble borné de $\mathbb{R}^n$ et qu'il existe $M \in \mathbb{R}$ t.q. $|H(x)y \cdot y| \leq M|y|^2$ pour tout x, $y \in \mathbb{R}^n$ (où $H(x)$ désigne la matrice hessienne de f au point x).

1. Montrer qu'il existe $\overline{x} \in A$ t.q. $f(\overline{x}) = \min\{f(x),\ x \in \mathbb{R}^n\}$ (noter qu'il n'y a pas nécessairement unicité de $\overline{x}$).
2. Soit $x \in A$ t.q. $\nabla f(x) \neq 0$. On pose $T(x) = \sup\{\alpha \geq 0;\ [x, x - \alpha\nabla f(x)] \subset A\}$. Montrer que $0 < T(x) < +\infty$ et que $[x, x - T(x)\nabla f(x)] \subset A\}$ (où $[x, x - T(x)\nabla f(x)]$ désigne l'ensemble $\{tx + (1-t)(x - T(x)\nabla f(x)), t \in [0,1]\}$.
3. Pour calculer une valeur appochée de $\overline{x}$ (t.q. $f(\overline{x}) = \min\{f(x), x \in \mathbb{R}^n\}$), on propose l'algorithme suivant :
   Initialisation : $x_0 \in A$,

   Itérations : Soit $k \geq 0$.
   Si $\nabla f(x_k) = 0$, on pose $x_{k+1} = x_k$. Si $\nabla f(x_k) \neq 0$, on choisit $\alpha_k \in [0, T(x_k)]$ t.q. $f(x_k - \alpha_k \nabla f(x_k)) = \min\{f(x_k - \alpha\nabla f(x_k)),\ 0 \leq \alpha \leq T(x_k)\}$ (La fonction T est définie à la question 2) et on pose $x_{k+1} = x_k - \alpha_k \nabla f(x_k)$.

   (a) Montrer que, pour tout $x_0 \in A$, l'algorithme précédent définit une suite $(x_k)_{k\in\mathbb{N}} \subset A$ (c'est–à–dire que, pour $x_k \in A$, il existe bien au moins un élément de $[0, T(x_k)]$, noté α_k, t.q. $f(x_k - \alpha_k\nabla f(x_k) = \min\{f(x_k - \alpha\nabla f(x_k)),\ 0 \leq \alpha \leq T(x_k)\}$).

   (b) Montrer que cet algorithme n'est pas nécessairement l'algorithme du gradient à pas optimal. [on pourra chercher un exemple avec $n = 1$.]

   (c) Montrer que $f(x_k) - f(x_{k+1}) \geq \frac{|\nabla f(x_k)|^2}{2M}$, pour tout $k \in \mathbb{N}$.
4. On montre maintenant la convergence de la suite $(x_k)_{k\in\mathbb{N}}$ construite à la question précédente.
   (a) Montrer qu'il existe une sous suite $(x_{k_\ell})_{\ell\in\mathbb{N}}$ et $x \in A$ t.q. $x_{k_\ell} \to x$, quand $\ell \to \infty$, et $\nabla f(x) = 0$.

   (b) On suppose, dans cette question, qu'il existe un et un seul élément $z \in A$ t.q. $\nabla f(z) = 0$. Montrer que $x_k \to z$, quand $k \to \infty$, et que $f(z) = \min\{f(x), x \in A\}$.

Exercice 144 (Application du GPO)**.**

Soit $A \in \mathcal{M}_n(\mathbb{R})$ et J la fonction définie de $\mathbb{R}^n$ dans $\mathbb{R}$ par $J(x) = e^{\|Ax\|^2}$, où $\|\cdot\|$ désigne la norme euclidienne sur $\mathbb{R}^n$.

1. Montrer que J admet un minimum (on pourra le calculer...).
2. On suppose que la matrice A est inversible, montrer que ce minimum est unique.
3. Ecrire l'algorithme du gradient à pas optimal pour la recherche de ce minimum. [On demande de calculer le paramètre optimal α_k en fonction de A et de x_k.] A quelle condition suffisante cet algorithme converge-t-il ?

Exercice 145 (Méthode de relaxation)**.** *Corrigé détaillé en page 226*

Soit f une fonction continûment différentiable de $E = \mathbb{R}^n$ dans $\mathbb{R}$ vérifiant l'hypothèse (3.10) :
1. Justifier l'existence et l'unicité de $\overline{x} \in \mathbb{R}^n$ tel que $f(\overline{x}) = \inf_{x\in\mathbb{R}^n} f(x)$.

On propose l'algorithme de recherche de minimum de f suivant :

$$\left\{\begin{array}{ll}
\text{Initialisation :} & x^{(0)} \in E,\\
\text{Itération } n : & x^{(k)} \text{ connu, } (n \geq 0)\\
& \text{Calculer } x_1^{(k+1)} \text{ tel que, pour tout } \xi \in \mathbb{R},\\
& f(x_1^{(k+1)}, x_2^{(k)}, x_3^{(k)}, \ldots, x_n^{(k)}) \leq f(\xi, x_2^{(k)}, x_3^{(k)}, \ldots, x_n^{(k)}),\\
& \text{Calculer } x_2^{(k+1)} \text{ tel que, pour tout } \xi \in \mathbb{R},\\
& f(x_1^{(k+1)}, x_2^{(k+1)}, x_3^{(k)}, \ldots, x_n^{(k)}) \leq f(x_1^{(k+1)}, \xi, x_3^{(k)}, \ldots, x_n^{(k)}),\\
& \qquad \ldots\\
& \text{Calculer } x_k^{(k+1)} \text{ tel que, pour tout } \xi \in \mathbb{R},\\
& f(x_1^{(k+1)}, \ldots, x_{k-1}^{(k+1)}, x_k^{(k+1)}, x_{(k+1)}^{(k)}, \ldots, x_n^{(k)})\\
& \qquad \leq f(x_1^{(k+1)}, \ldots, x_{k-1}^{(k+1)}, \xi, x_{(k+1)}^{(k)}, \ldots, x_n^{(k)}),\\
& \qquad \ldots\\
& \text{Calculer } x_n^{(k+1)} \text{ tel que, pour tout } \xi \in \mathbb{R},\\
& f(x_1^{(k+1)}, x_2^{(k+1)}, \ldots, x_{n-1}^{(k+1)}, x_n^{(k+1)}) \leq f(x_1^{(k+1)}, \ldots, x_{n-1}^{(k+1)}, \xi).
\end{array}\right. \tag{3.38}$$

2. Pour $n \in \mathbb{N}$ et $1 \leq k \leq N$, soit $\varphi_k^{(k+1)}$ la fonction de $\mathbb{R}$ dans $\mathbb{R}$ définie par :

$$\varphi_k^{(k+1)}(s) = f(x_1^{(k+1)}, \ldots, x_{k-1}^{(k+1)}, s, x_{(k+1)}^{(k)}, \ldots, x_n^{(k)}).$$

Montrer qu'il existe un unique élément $\overline{s} \in \mathbb{R}$ tel que

$$\varphi_k^{(k+1)}(\overline{s}) = \inf_{s \in \mathbb{R}} \varphi_k^{(k+1)}(s).$$

En déduire que la suite $(x^{(k)})_{n\in\mathbb{N}}$ construite par l'algorithme (3.38) est bien définie.

Dans toute la suite, on note $\|\cdot\|$ la norme euclidienne sur $\mathbb{R}^n$ et $(\cdot|\cdot)$ le produit scalaire associé. Pour $i = 1, \ldots, n$, on désigne par $\partial_i f$ la dérivée partielle de f par rapport à la i-ème variable.

3. Soit $(x^{(k)})_{n\in\mathbb{N}}$ la suite définie par l'algorithme (3.38).
Pour $n \geq 0$, on définit $x^{(n+1,0)} = x^{(k)} = (x_1^{(k)}, \ldots, x_n^{(k)})^t$, et pour $1 \leq k \leq n$,

$$x^{(n+1,k)} = (x_1^{(k+1)}, \ldots, x_k^{(k+1)}, x_{k+1}^{(k)}, \ldots, x_n^{(k)})^t$$

(de sorte que $x^{(n+1,n)} = x^{(k+1)}$).

(a) Soit $n \in \mathbb{N}$. Pour $1 \leq k \leq n$, montrer que $\partial_k f(x^{(n+1,k)}) = 0$, pour $k = 1, \ldots, n$. En déduire que

$$f(x^{(n+1,k-1)}) - f(x^{(n+1,k)}) \geq \frac{\alpha}{2}\|x^{(n+1,k-1)} - x^{(n+1,k)}\|^2.$$

(b) Montrer que la suite $(x^{(k)})_{n\in\mathbb{N}}$ vérifie

$$f(x^{(k)}) - f(x^{(k+1)}) \geq \frac{\alpha}{2}\|x^{(k)} - x^{(k+1)}\|^2.$$

En déduire que $\lim_{n\to+\infty} \|x^{(k)} - x^{(k+1)}\| = 0$ et que, pour $1 \leq k \leq n$, $\lim_{n\to+\infty} \|x^{(n+1,k)} - x^{(k+1)}\| = 0$.

4. Montrer que

$$\|x^{(k+1)} - \overline{x}\| \leq \frac{1}{\alpha} \left(\sum_{k=1}^{n} |\partial_k f(x^{(k+1)})|^2 \right)^{\frac{1}{2}}.$$

5. Montrer que les suites $(x^{(k)})_{n\in\mathbb{N}}$, et $(x^{(n+1,k)})_{n\in\mathbb{N}}$, pour $k = 1, \ldots, n$, sont bornées.
Montrer que

$$|\partial_k f(x^{(k+1)})| \to 0 \text{ lorsque } n \to +\infty.$$

(On rappelle que $\partial_k f(x^{(n+1,k)}) = 0$.)
Conclure quant à la convergence de la suite$(x^{(k)})_{n\in\mathbb{N}}$ lorsque $n \to +\infty$.

6. On suppose dans cette question que $f(x) = \frac{1}{2}(Ax|x) - (b|x)$. Montrer que dans ce cas, l'algorithme (3.38) est équivalent à une méthode itérative de résolution de systèmes linéaires qu'on identifiera.

7. On suppose dans cette question que $n = 2$. Soit g la fonction définie de $\mathbb{R}^2$ dans $\mathbb{R}$ par : $g(x) = x_1^2 + x_2^2 - 2(x_1 + x_2) + 2|x_1 - x_2|$, avec $x = (x_1, x_2)^t$.

(a) Montrer qu'il existe un unique élément $\overline{x} = (\overline{x}_1, \overline{x}_2)^t$ de $\mathbb{R}^2$ tel que $g(\overline{x}) = \inf_{x\in\mathbb{R}^2} g(x)$.

(b) Montrer que $\overline{x} = (1, 1)^t$.

(c) Montrer que si $x^{(0)} = (0, 0)^t$, l'algorithme (3.38) appliqué à g ne converge pas vers $\overline{x}$. Quelle est l'hypothèse mise en défaut ici ?

Exercice 146 (Mise en oeuvre de GC)**.**

On considère la fonction $f : \mathbb{R}^2 \to \mathbb{R}$ définie par $f(x_1, x_2) = 2x_1^2 + x_2^2 - x_1x_2 - 3x_1 - x_2 + 4$.

1. Montrer qu'il existe un unique $\bar{x} \in \mathbb{R}^2$ tel que $\bar{x} = \min_{x\in\mathbb{R}^2} f(x)$ admet un unique minimum, et le calculer.
2. Calculer le premier itéré donné par l'algorithme du gradient conjugué, en partant de $(x_1^{(0)}, x_2^{(0)}) = (0, 0)$, pour un pas de $\alpha = .5$ dans le cas de GPF.

Exercice 147 (Gradient conjugué pour une matrice non symétrique)**.** *Corrigé détaillé en page 228*

Soit $n \in \mathbb{N}$, $n \geq 1$. On désigne par $\|\cdot\|$ la norme euclidienne sur $\mathbb{R}^n$, et on munit l'ensemble $\mathcal{M}_n(\mathbb{R})$ de la norme induite par la norme $\|\cdot\|$, $\|\cdot\|$. Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible. On définit $M \in \mathcal{M}_n(\mathbb{R})$ par $M = A^tA$. On se donne un vecteur $b \in \mathbb{R}^n$, et on s'intéresse à la résolution du système linéaire

$$Ax = b;. \tag{3.39}$$

1. Montrer que $x \in \mathbb{R}^n$ est solution de (1.121) si et seulement si x est solution de

$$Mx = A^tb;. \tag{3.40}$$

2. On rappelle que le conditionnement d'une matrice $C \in \mathcal{M}_n(\mathbb{R})$ inversible est défini par $\text{cond}(C) = \|C\|\|C^{-1}\|$ (et dépend donc de la norme considérée ; on rappelle qu'on a choisi ici la norme induite par la norme euclidienne).
   (a) Montrer que les valeurs propres de la matrice M sont toutes strictement positives.
   (b) Montrer que $\text{cond}(A) = \sqrt{\frac{\lambda_n}{\lambda_1}}$, où λ_n (resp. λ_1) est la plus grande (resp. plus petite) valeur propre de M.
3. Ecrire l'algorithme du gradient conjugué pour la résolution du système (3.40), en ne faisant intervenir que les matrices A et A^t (et pas la matrice M) et en essayant de minimiser le nombre de calculs. Donner une estimation du nombre d'opérations nécessaires et comparer par rapport à l'algorithme du gradient conjugué écrit dans le cas d'une matrice carré d'ordre n symétrique définie positive.

Exercice 148 (Gradient conjugué préconditionné par LL^t)**.**

Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique définie positive, et $b \in \mathbb{R}^n$. Soit L une matrice triangulaire inférieure inversible, soit $B = L^{-1}A(L^t)^{-1}$ et $\tilde{b} = L^{-1}b$.

1. Montrer que B est symétrique définie positive.
2. Justifier l'existence et l'unicité de $x \in \mathbb{R}^n$ tel que $Ax = b$, et de $y \in \mathbb{R}^n$ tel que $By = \tilde{b}$. Ecrire x en fonction de y.

Soit $y^{(0)} \in \mathbb{R}^n$ fixé. On pose $\tilde{r}^{(0)} = \tilde{w}^{(0)} = \tilde{b} - By^{(0)}$. Si $\tilde{r}^{(0)} \neq 0$, on pose alors $y^{(1)} = y^{(0)} + \rho_0 \tilde{w}^{(0)}$, avec $\rho_0 = \frac{\tilde{r}^{(0)} \cdot \tilde{r}^{(0)}}{\tilde{w}^{(0)} \cdot A\tilde{w}^{(0)}}$.
Pour $n > 1$, on suppose $y^{(0)}, \ldots, y^{(k)}$ et $\tilde{w}^{(0)}, \ldots, \tilde{w}^{(k-1)}$ connus, et on pose : $\tilde{\boldsymbol{r}}^{(k)} = \tilde{b} - By^{(k)}$. Si $\tilde{\boldsymbol{r}}^{(k)} \neq 0$, on calcule : $\tilde{w}^{(k)} = \tilde{\boldsymbol{r}}^{(k)} + \lambda_{k-1}\tilde{w}^{(k-1)}$ avec $\lambda_{k-1} = \frac{\tilde{\boldsymbol{r}}^{(k)} \cdot \tilde{\boldsymbol{r}}^{(k)}}{\tilde{r}^{(k-1)} \cdot \tilde{r}^{(k-1)}}$ et on pose alors : $y^{(k+1)} = y^{(k)} + \alpha_k \tilde{w}^{(k)}$ avec $\alpha_k = \frac{\tilde{\boldsymbol{r}}^{(k)} \cdot \tilde{\boldsymbol{r}}^{(k)}}{\tilde{w}^{(k)} \cdot B\tilde{w}^{(k)}}$,

3. En utilisant le cours, justifier que la famille $y^{(k)}$ ainsi construite est finie. A quoi est égale sa dernière valeur ?

Pour $n \in \mathbb{N}$, on pose : $x^{(k)} = L^{-t}y^{(k)}$ (avec $L^{-t} = (L^{-1})^t = (L^t)^{-1}$), $\boldsymbol{r}^{(k)} = b - Ax^{(k)}$, $w^{(k)} = L^{-t}\tilde{w}^{(k)}$ et $s^{(k)} = (LL^t)^{-1}\boldsymbol{r}^{(k)}$.

4. Soit $n > 0$ fixé. Montrer que :

$$(a) \quad \lambda_{k-1} = \frac{s^{(k)} \cdot \boldsymbol{r}^{(k)}}{s^{(k-1)} \cdot r^{(k-1)}}, \qquad (b) \quad \rho_n = \frac{s^{(k)} \cdot \boldsymbol{r}^{(k)}}{w^{(k)} \cdot Aw^{(k)}},$$

$$(c) \quad w^{(k)} = s^{(k)} + \lambda_n w^{(k-1)}, \qquad (d) \quad x^{(k+1)} = x^{(k)} + \alpha_k w^{(k)}.$$

5. On suppose que la matrice LL^t est une factorisation de Choleski incomplète de la matrice A. Ecrire l'algorithme du gradient conjugué préconditionné par cette factorisation, pour la résolution du système $Ax = b$.

Exercice 149 (Méthode de quasi-linéarisation)**.** Soit $f \in C^3(\mathbb{R}, \mathbb{R})$ une fonction croissante à l'infini, c. à.d. telle que $f(x) \to +\infty$ lorsque $|x| \to +\infty$; soit $d \in \mathbb{R}$ et soit J la fonction de $\mathbb{R}$ dans $\mathbb{R}$ par

$$J(x) = (f(x) - d)^2.$$

1. Montrer qu'il existe $\bar{x} \in \mathbb{R}$ tel que $J(\bar{x}) \leq J(x), \forall x \in \mathbb{R}$.
2. (Newton) On cherche ici à déterminer un minimum de J en appliquant la méthode de Newton pour trouver une solution de l'équation $J'(x) = 0$. Ecrire l'algorithme de Newton qui donne x_{k+1} en fonction de x_k et des données d, f, f' et f''.
3. L'algorithme dit de "quasi–linéarisation" consiste à remplacer, à chaque itération $k \in \mathbb{N}$, la minimisation de la fonctionnelle J par celle de la fonctionnelle J_k, définie de $\mathbb{R}$ dans $\mathbb{R}$ obtenue à partir de J en effectuant un développement limité au premier ordre de $f(x)$ en $x^{(k)}$, c.à.d.

$$J_k(x) = (f(x_k) + f'(x_k)(x - x_k) - d)^2$$

   Montrer que à k fixé, il existe un unique $\bar{x} \in \mathbb{R}$ qui minimise J_k et le calculer (on supposera que $f'(x_k) \neq 0$). On pose donc $x_{k+1} = \bar{x}$. Que vous rappelle l'expression de x_{k+1} ?
4. Ecrire l'algorithme du gradient à pas fixe pour la minimisation de J.
5. Dans cette question, on prend $f(x) = x^2$.
   (a) Donner l'ensemble des valeurs $\bar{x} \in \mathbb{R}$ qui minimisent J, selon la valeur de d. Y a t-il unicité de $\bar{x}$?
   (b) Montrer que quelque soit la valeur de d, l'algorithme de Newton converge si le choix initial x_0 est suffisamment proche de $\bar{x}$.
   (c) On suppose que $d > 0$; montrer que l'algorithme de quasi-linéarisation converge pour un choix initial x_0 dans un voisinage de 1.
   (d) On suppose maintenant que $d = -1$. Montrer que l'algorithme de quasi-linéarisation ne converge que pour un ensemble dénombrable de choix initiaux x_0.

Exercice 150 (Méthode de Polak-Ribière)**.** *Suggestions en page 223, corrigé en page 229*

Dans cet exercice, on démontre la convergence de la méthode de Polak-Ribière (méthode de gradient conjugué pour une fonctionnelle non quadratique) sous des hypothèses "simples" sur f.

Soit $f \in C^2(\mathbb{R}^n, \mathbb{R})$. On suppose qu'il existe $\alpha > 0$, $\beta \geq \alpha$ tel que $\alpha|y|^2 \leq H(x)y \cdot y \leq \beta|y|^2$ pour tout x, $y \in \mathbb{R}^n$. ($H(x)$ est la matrice hessienne de f au point x.)

1. Montrer que f est strictement convexe, que $f(x) \to \infty$ quand $|x| \to \infty$ et que le spectre $\mathcal{VP}(H(x))$ de $H(x)$ est inclus dans $[\alpha, \beta]$ pour tout $x \in \mathbb{R}^n$.

   On note $\overline{x}$ l'unique point de $\mathbb{R}^n$ t.q. $f(\overline{x}) \leq f(x)$ pour tout $x \in \mathbb{R}^n$ (l'existence et l'unicité de $\overline{x}$ est donné par la question précédente). On cherche une approximation de $\overline{x}$ en utilisant l'algorithme de Polak-Ribière :

   initialisation. $x^{(0)} \in \mathbb{R}^n$. On pose $g^{(0)} = -\nabla f(x^{(0)})$. Si $g^{(0)} = 0$, l'algorithme s'arrête (on a $x^{(0)} = \overline{x}$). Si $g^{(0)} \neq 0$, on pose $w^{(0)} = g^{(0)}$ et $x^{(1)} = x^{(0)} + \rho_0 w^{(0)}$ avec ρ_0 "optimal" dans la direction $w^{(0)}$.

   itération. $x^{(k)}$, $w^{(k-1)}$ connus ($k \geq 1$). On pose $g^{(k)} = -\nabla f(x^{(k)})$. Si $g^{(k)} = 0$, l'algorithme s'arrête (on a $x^{(k)} = \overline{x}$). Si $g^{(k)} \neq 0$, on pose $\lambda_{k-1} = [g^{(k)} \cdot (g^{(k)} - g^{(k-1)})]/[g^{(k-1)} \cdot g^{(k-1)}]$, $w^{(k)} = g^{(k)} + \lambda_{k-1} w^{(k-1)}$ et $x^{(k+1)} = x^{(k)} + \alpha_k w^{(k)}$ avec α_k "optimal" dans la direction w_k. (Noter que α_k existe bien.)

   On suppose dans la suite que $g^{(k)} \neq 0$ pour tout $k \in \mathbb{N}$.
2. Montrer (par récurrence sur k) que $g^{(k+1)} \cdot w^{(k)} = 0$ et $g^{(k)} \cdot g^{(k)} = g^{(k)} \cdot w^{(k)}$, pour tout $k \in \mathbb{N}$.
3. On pose
   $$J^{(k)} = \int_0^1 H(x^{(k)} + \theta \alpha_k w^{(k)}) d\theta.$$
   Montrer que $g^{(k+1)} = g^{(k)} + \alpha_k J^{(k)} w^{(k)}$ et que $\alpha_k = (-g^{(k)} \cdot w^{(k)})/(J^{(k)} w^{(k)} \cdot w^{(k)})$ (pour tout $k \in \mathbb{N}$).
4. Montrer que $|w^{(k)}| \leq (1 + \beta/\alpha)|g^{(k)}|$ pour tout $k \in \mathbb{N}$. [Utiliser, pour $k \geq 1$, la question précédente et la formule donnant λ_{k-1}.]
5. Montrer que $x^{(k)} \to \overline{x}$ quand $k \to \infty$.

Exercice 151 (Algorithme de quasi Newton)**.**
Corrigé détaillé en page 232

Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique définie positive et $b \in \mathbb{R}^n$. On pose $f(x) = (1/2)Ax \cdot x - b \cdot x$ pour $x \in \mathbb{R}^n$. On rappelle que $\nabla f(x) = Ax - b$. Pour calculer $\overline{x} \in \mathbb{R}^n$ t.q. $f(\overline{x}) \leq f(x)$ pour tout $x \in \mathbb{R}^n$, on va utiliser un algorithme de quasi Newton, c'est-à-dire :

initialisation. $x^{(0)} \in \mathbb{R}^n$.

itération. $x^{(k)}$ connu ($n \geq 0$. On pose $x^{(k+1)} = x^{(k)} - \alpha_k K^{(k)} g^{(k)}$ avec $g^{(k)} = \nabla f(x^{(k)})$, $K^{(k)}$ une matrice symétrique définie positive à déterminer et α_k "optimal" dans la direction $w^{(k)} = -K^{(k)} g^{(k)}$. (Noter que α_k existe bien.)

Partie 1. Calcul de α_k. On suppose que $g^{(k)} \neq 0$.

1. Montrer que $w^{(k)}$ est une direction de descente stricte en $x^{(k)}$ et calculer la valeur de α_k (en fonction de $K^{(k)}$ et $g^{(k)}$).
2. On suppose que, pour un certain $n \in \mathbb{N}$, on a $K^{(k)} = (H(x^{(k)}))^{-1}$ (où $H(x)$ est la matrice hessienne de f en x, on a donc ici $H(x) = A$ pour tout $x \in \mathbb{R}^n$). Montrer que $\alpha_k = 1$.
3. Montrer que la méthode de Newton pour calculer $\overline{x}$ converge en une itération (mais nécessite la résolution du système linéaire $A(x^{(1)} - x^{(0)}) = b - Ax^{(0)} \ldots$)

Partie 2. Méthode de Fletcher-Powell. On prend maintenant $K^{(0)} = \mathrm{Id}$ et

$$K^{(k+1)} = K^{(k)} + \frac{s^{(k)}(s^{(k)})^t}{s^{(k)} \cdot y^{(k)}} - \frac{(K^{(k)} y^{(k)})(K^{(k)}(y^{(k)})^t}{K^{(k)} y^{(k)} \cdot y^{(k)}},\ n \geq 0, \tag{3.41}$$

avec $s^{(k)} = x^{(k+1)} - x^{(k)}$ et $y^{(k)} = g^{(k+1)} - g^{(k)} = As^{(k)}$.

On va montrer que cet algorithme converge en au plus n itérations (c'est-à-dire qu'il existe $n \leq n+1$ t.q. $x_{N+1} = \overline{x}$.)

1. Soit $n \in \mathbb{N}$. On suppose, dans cette question, que $s^{(0)}, \dots, s^{(k-1)}$ sont des vecteurs A-conjugués et non-nuls et que $K^{(0)}, \dots, K^{(k)}$ sont des matrices symétriques définies positives t.q. $K^{(j)}As^{(i)} = s^{(i)}$ si $0 \leq i < j \leq n$ (pour $n = 0$ on demande seulement $K^{(0)}$ symétrique définie positive).

   (a) On suppose que $g^{(k)} \neq 0$. Montrer que $s^{(k)} \neq 0$ (cf. Partie I) et que, pour $i < n$,

   $$s^{(k)} \cdot As^{(i)} = 0 \Leftrightarrow g^{(k)} \cdot s^{(i)} = 0.$$

   Montrer que $g^{(k)} \cdot s^{(i)} = 0$ pour $i < n$. [On pourra remarquer que $g^{(i+1)} \cdot s^{(i)} = g^{(i+1)} \cdot w^{(i)} = 0$ et $(g^{(k)} - g^{(i+1)}) \cdot s^{(i)} = 0$ par l'hypothèse de conjugaison de $s^{(0)}, \dots, s^{(k-1)}$.] En déduire que $s^{(0)}, \dots, s^{(k)}$ sont des vecteurs A-conjugués et non-nuls.

   (b) Montrer que $K^{(k+1)}$ est symétrique.

   (c) Montrer que $K^{(k+1)}As^{(i)} = s^{(i)}$ si $0 \leq i \leq n$.

   (d) Montrer que, pour tout $x \in \mathbb{R}^n$, on a

   $$K^{(k+1)}x \cdot x = \frac{(K^{(k)}x \cdot x)(K^{(k)}y^{(k)} \cdot y^{(k)}) - (K^{(k)}y^{(k)} \cdot x)^2}{K^{(k)}y^{(k)} \cdot y^{(k)}} + \frac{(s^{(k)} \cdot x)^2}{As^{(k)} \cdot s^{(k)}}.$$

   En déduire que $K^{(k+1)}$ est symétrique définie positive. [On rappelle (inégalité de Cauchy-Schwarz) que, si K est symétrique définie positive, on a $(Kx \cdot y)^2 \leq (Kx \cdot x)(Ky \cdot y)$ et l'égalité a lieu si et seulement si x et y sont colinéaires.]

2. On suppose que $g^{(k)} \neq 0$ si $0 \leq n \leq n-1$. Montrer (par récurrence sur n, avec la question précédente) que $s^{(0)}, \dots, s^{(n-1)}$ sont des vecteurs A-conjugués et non-nuls et que $K^{(n)}As^{(i)} = s^{(i)}$ si $i < n$. En déduire que $K^{(n)} = A^{-1}$, $\alpha_n = 1$ et $x^{(n+1)} = A^{-1}b = \overline{x}$.

Exercice 152 (Méthodes de Gauss–Newton et de quasi–linéarisation)**.**

Soit $f \in C^2(\mathbb{R}^n, \mathbb{R}^p)$, avec $n, p \in \mathbb{N}^*$. Soit $C \in \mathcal{M}_p(\mathbb{R})$ une matrice réelle carrée d'ordre p, symétrique définie positive, et $\boldsymbol{d} \in \mathbb{R}^p$. Pour $\boldsymbol{x} \in \mathbb{R}^n$, on pose

$$J(\boldsymbol{x}) = (f(\boldsymbol{x}) - \boldsymbol{d}) \cdot C(f(\boldsymbol{x}) - \boldsymbol{d}).$$

On cherche à minimiser J.

I *Propriétés d'existence et d'unicité*

   (a) Montrer que J est bornée inférieurement.

   (b) Donner trois exemples de fonctions f pour lesquels les fonctionnelles J associées sont telles que l'on ait :

      i. existence et unicité de $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ qui réalise le minimum de J, pour le premier exemple.

      ii. existence et non unicité de $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ qui réalise le minimum de J, pour le second exemple.

      iii. non existence de $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ qui réalise le minimum de J, pour le troisième exemple.

   (On pourra prendre $n = p = 1$.)

II *Un peu de calcul différentiel*

   (a) On note Df et D_2f les différentielles d'ordre 1 et 2 de f. A quels espaces appartiennent $Df(\boldsymbol{x})$, $D_2f(\boldsymbol{x})$ (pour $\boldsymbol{x} \in \mathbb{R}^n$), ainsi que Df et D_2f ? Montrer que pour tout $\boldsymbol{x} \in \mathbb{R}^n$, il existe $M(\boldsymbol{x}) \in \mathcal{M}_{p,n}(\mathbb{R})$, où $\mathcal{M}_{p,n}(\mathbb{R})$ désigne l'ensemble des matrices réelles à p lignes et n colonnes, telle que $Df(\boldsymbol{x})(\boldsymbol{y}) = M(\boldsymbol{x})\boldsymbol{y}$ pour tout $\boldsymbol{y} \in \mathbb{R}^n$.

   (b) Pour $\boldsymbol{x} \in \mathbb{R}^n$, calculer $\nabla J(\boldsymbol{x})$.

   (c) Pour $\boldsymbol{x} \in \mathbb{R}^n$, calculer la matrice hessienne de J en $\boldsymbol{x}$ (qu'on notera $H(\boldsymbol{x})$). On suppose maintenant que M ne dépend pas de $\boldsymbol{x}$; montrer que dans ce cas $H(\boldsymbol{x}) = 2M(\boldsymbol{x})^t CM(\boldsymbol{x})$.

III *Algorithmes d'optimisation* Dans toute cette question, on suppose qu'il existe un unique $\bar{b}fx \in \mathbb{R}^n$ qui réalise le minimum de J, qu'on cherche à calculer de manière itérative. On se donne pour cela $x_0 \in \mathbb{R}^n$, et on cherche à construire une suite $(x_k)_{k\in\mathbb{N}}$ qui converge vers $\bar{x}$.

(a) On cherche à calculer $\bar{x}$ en utilisant la méthode de Newton pour annuler ∇J. Justifier brièvement cette procédure et écrire l'algorithme obtenu.

(b) L'algorithme dit de "Gauss-Newton" est une modification de la méthode précédente, qui consiste à approcher, à chaque itération n, la matrice jacobienne de ∇J en x_k par la matrice obtenue en négligeant les dérivées secondes de f. Ecrire l'algorithme ainsi obtenu.

(c) L'algorithme dit de "quasi–linéarisation" consiste à remplacer, à chaque itération $k \in \mathbb{N}$, la minimisation de la fonctionnelle J par celle de la fonctionnelle J_k, définie de $\mathbb{R}^n$ dans $\mathbb{R}$, et obtenue à partir de J en effectuant un développement limité au premier ordre de $f(\boldsymbol{x})$ en $\boldsymbol{x}^{(k)}$, c.à.d.

$$J_k(\boldsymbol{x}) = (f(\boldsymbol{x}^{(k)}) + Df(\boldsymbol{x}^{(k)})(\boldsymbol{x} - \boldsymbol{x}^{(k)}) - d) \cdot C(f(\boldsymbol{x}^{(k)}) + Df(\boldsymbol{x}^{(k)})(\boldsymbol{x} - \boldsymbol{x}^{(k)}) - \boldsymbol{d}).$$

i. Soit $k \geq 0$, $\boldsymbol{x}^{(k)} \in \mathbb{R}^n$ connu, $M_k = M(\boldsymbol{x}^{(k)}) \in \mathcal{M}_{p,n}(\mathbb{R})$, et $\boldsymbol{x} \in \mathbb{R}^n$. On pose $\boldsymbol{h} = \boldsymbol{x} - \boldsymbol{x}^{(k)}$. Montrer que

$$J_k(\boldsymbol{x}) = J(\boldsymbol{x}^{(k)}) + M_k^t C M_k \boldsymbol{h} \cdot \boldsymbol{h} + 2M_k^t C(f(\boldsymbol{x}^{(k)}) - \boldsymbol{d}) \cdot \boldsymbol{h}.$$

ii. Montrer que la recherche du minimum de J_k est équivalente à la résolution d'un système linéaire dont on donnera l'expression.

iii. Ecrire l'algorithme de quasi–linéarisation, et le comparer avec l'algorithme de Gauss-Newton.

Exercice 153 (Comparaison de la minimisation de deux fonctions)**.**

Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible et soit $\boldsymbol{b} \in \mathbb{R}^n$. On considère les deux fonctions suivantes de $\mathbb{R}^n$ dans $\mathbb{R}$:

$$J : \boldsymbol{u} \in \mathbb{R}^n \mapsto J(\boldsymbol{u}) = \frac{1}{2}\boldsymbol{u}^t A\boldsymbol{u} - \boldsymbol{b}^t\boldsymbol{u} \;\; \textit{et} \;\; \widetilde{J} : \boldsymbol{u} \in \mathbb{R}^n \mapsto \widetilde{J}(\boldsymbol{u}) = (A\boldsymbol{u} - \boldsymbol{b})^t(A\boldsymbol{u} - \boldsymbol{b})$$

1. On suppose dans cette question que A est symétrique définie positive. Montrer que la fonction J (resp. $\widetilde{J}$) admet un unique minimum et que ce minimum est l'unique solution de $\nabla J(\boldsymbol{u}) = 0$ (resp. $\nabla\widetilde{J}(\boldsymbol{u}) = 0$); on calculera ces gradients pour expliciter les équations correspondantes.
2. Donner un exemple d'une ICP-matrice 2×2 non symétrique.
3. Donner un exemple d'une ICP-matrice 2×2 symétrique qui ne soit pas symétrique définie positive.
4. On suppose dans cette question que A est une ICP-matrice qui est symétrique mais non définie positive;
   (a) Montrer que
   $$\inf_{\boldsymbol{u}\in\mathbb{R}^n} J(\boldsymbol{u}) = -\infty$$
   (b) Montrer que $\widetilde{J}$ admet un unique minimum que l'on caractérisera.

Suggestions pour les exercices

Exercice 142 page 215 (Algorithme du gradient à pas optimal)

2. Utiliser le fait que H est continue.

3. Etudier la fonction $\varphi : \mathbb{R}_+$ dans $\mathbb{R}$ définie par $\varphi(\rho) = f(x_k + \rho\boldsymbol{w}^{(k)})$.

4. a. Montrer que f est minorée et remarquer que la suite $(f(x_k))_{k\in\mathbb{N}}$ est décroissante.
4.b se déduit du 4.a
4.c. Utiliser la fonction φ définie plus haut, la question 4.b. et la question 2.

4.d. Utiliser le fait que le choix de α_k est optimal et le résultat de 4.c.
4.e. Etudier le polynôme du 2nd degré en ρ défini par : $P_k(\rho) = f(x_k) - \rho|\boldsymbol{w}^{(k)}|^2 + \frac{1}{2}M|\boldsymbol{w}^{(k)}|^2\rho^2$ dans les cas où $|\boldsymbol{w}^{(k)}| \leq M$ (fait la quesiton 4.c) puis dans le cas $|\boldsymbol{w}^{(k)}| \geq M$.

5. utiliser l'inégalité prouvée en 4.e. pour montrer que $|\boldsymbol{w}^{(k)}| \to 0$ lorsque $n \to +\infty$.

6. Pour montrer que toute la suite converge, utiliser l'argument d'unicité de la limite, en raisonnant par l'absurde (supposer que la suite ne converge pas et aboutir à une contradiction).

Exercice 143 page 216 (Cas où f n'est pas croissante à l'infini)

S'inspirer des techniques utilisées lors des démonstrations de la proposition 3.13 et du théorème 3.19 (il faut impérativement les avoir fait avant...).

Exercice 150 page 220 (Méthode de Polak-Ribière)

1. Utiliser la deuxième caractérisation de la convexité. Pour montrer le comportement à l'infini, introduire la fonction φ habituelle... ($\varphi(t) = f(x + ty)$).
2. Pour montrer la convergence, utiliser le fait que si $w_k \cdot \nabla f(x_k) < 0$ alors w_k est une direction de descente stricte de f en x_k, et que si α_k est optimal alors $\nabla f(x_k + \alpha_k \boldsymbol{w}^{(k)}) = 0$.
3. Utiliser la fonction φ définie par $\varphi(\theta) = \nabla f(x_k + \theta\alpha_k \boldsymbol{w}^{(k)})$.
4. C'est du calcul...
5. Montrer d'abord que $-g_k \boldsymbol{w}^{(k)} \leq -\gamma|\boldsymbol{w}^{(k)}||g_k|$. Montrer ensuite (en utilisant la bonne vieille fonction φ définie par $\varphi(t) = f(x_k + t\alpha_k)$, que $g_k \to 0$ lorsque $n \to +\infty$.

Exercice 156 page 241 (Fonctionnelle quadratique)

1. Pour montrer que K est non vide, remarquer que comme $d \neq 0$, il existe $\tilde{x} \in \mathbb{R}^n$ tel que $d \cdot \tilde{x} = \alpha \neq 0$. En déduire l'existence de $x \in \mathbb{R}^n$ tel que $d \cdot x = c$.
2. Montrer par le théorème de Lagrange que si $\bar{\boldsymbol{x}}$ est solution de (3.48), alors $y = (\bar{\boldsymbol{x}}, \lambda)^t$ est solution du système (3.57), et montrer ensuite que le système (3.57) admet une unique solution.

Corrigés des exercices

Exercice 141 page 215 (Mise en oeuvre de GPF et GPO)

1. On a

$$\nabla f(x) = \begin{bmatrix} 4x_1 - x_2 - 3 \\ 2x_2 - x_1 - 1 \end{bmatrix} \text{ et } H_f = \begin{bmatrix} 4 & -1 \\ -1 & 2 \end{bmatrix}$$

La fonction f vérifie les hypothèses du théorème 3.30 d'existence et d'unicité du minimum. En particulier la hessienne $H_f = \begin{bmatrix} 4 & -1 \\ -1 & 2 \end{bmatrix}$ est s.d.p.. Le minimum est obtenu pour

$$\begin{aligned} \partial_1 f(x_1, x_2) &= 4x_1 - x_2 - 3 = 0 \\ \partial_2 f(x_1, x_2) &= 2x_2 - x_1 - 1 = 0 \end{aligned}$$

c'est-à-dire $\bar{x}_1 = 1$ et $\bar{x}_2 = 1$. Ce minimum est $f(\bar{x}_1, \bar{x}_2) = 2$.

2. L'algorithme du gradient à pas fixe s'écrit :

$$\begin{cases} \text{Initialisation :} & x^{(0)} \in \mathbb{R}^2,\ \rho > 0 \\ \text{Itération } k : & x^{(k)} \text{ connu, } (k \geq 0) \\ & w^{(k)} = -\nabla f(x^{(k)}), \\ & x^{(k+1)} = x^{(k)} + \rho w^{(k)}. \end{cases}$$

A la première itération, on a $\nabla f(0,0) = (-3,-1)$ et donc $w^{(0)} = (3,1)$. On en déduit, pour $\rho = 0.5$, $x^{(1)} = (3\rho, \rho) = (3/2, 1/2)$ et $f(x^{(1)}) = 3$.

L'algorithme du gradient à pas optimal s'écrit :

$$\left\{\begin{array}{ll} \text{Initialisation :} & x^{(0)} \in \mathbb{R}^n. \\ \text{Itération } k : & x^{(k)} \text{ connu.} \\ & \text{On calcule } w^{(k)} = -\nabla f(x^{(k)}). \\ & \text{On choisit } \rho_k \geq 0 \text{ tel que} \\ & \qquad f(x^{(k)} + \rho_k w^{(k)}) \leq f(x^{(k)} + \rho w^{(k)}) \ \ \forall \rho \geq 0. \\ & \text{On pose } x^{(k+1)} = x^{(k)} + \rho_k w^{(k)}. \end{array}\right.$$

Calculons le ρ_0 optimal à l'itération 0. On a vu précédemment que $w^{(0)} = (3,1)$. Le ρ_0 optimal minimise la fonction $\rho \mapsto \varphi(\rho) = f(x^{(0)} + \rho w^{(0)}) = f(3\rho, \rho)$. On doit donc avoir $\varphi'(\rho_0) = 0$. Calculons $\varphi'(\rho)$. Par le théorème de dérivation des fonctions composées, on a :

$$\varphi'(\rho) = \nabla f(x^{(0)} + \rho w^{(0)}) \cdot w(0) = \begin{bmatrix} 11\rho - 3 \\ -\rho - 1 \end{bmatrix} \cdot \begin{bmatrix} 3 \\ 1 \end{bmatrix} = 3(11\rho - 3) + (-\rho - 1) = 32\rho - 10.$$

On en déduit que $\rho_0 = \frac{5}{16}$. On obtient alors $x^{(1)} = x^{(0)} + \rho_0 w^{(0)} = (\frac{15}{16}, \frac{5}{16})$, et $f(x^{(1)}) = 2.4375$, ce qui est, comme attendu, mieux qu'avec GPF.

Exercice 142 page 215 (Convergence de l'algorithme du gradient à pas optimal)

1. On sait que $f(x) \to +\infty$ lorsque $|x| \to +\infty$. Donc $\forall A > 0$, $\exists R \in \mathbb{R}_+ ; |x| > R \Rightarrow f(x) > A$. En particulier pour $A = f(x_0)$ ceci entraîne :

$$\exists R \in \mathbb{R}_+ ; \ \ x \in B_R \Rightarrow f(x) > f(x_0).$$

2. Comme $f \in C^2(\mathbb{R}^n, \mathbb{R})$, sa hessienne H est continue, donc $\|H\|_2$ atteint son $\max$ sur B_{R+1} qui est un fermé borné de $\mathbb{R}^n$. Soit $M = \max_{x \in B_{R+1}} \|H(x)\|_2$, on a $|H(x)y \cdot y| \leq My \cdot y \leq M|y|^2$.
3. Soit $w_k = -\nabla f(x_k)$.
   Si $w_k = 0$, on pose $x_{k+1} = x_k$.
   Si $w_k \neq 0$, montrons qu'il existe $\bar{\rho} > 0$ tel que

$$f(x_k + \bar{\rho} w_k) \leq f(x_k + \rho w_k) \ \ \forall \rho > 0.$$

   On sait que $f(x) \to +\infty$ lorsque $|x| \to +\infty$.
   Soit $\varphi : \mathbb{R}_+ \to \mathbb{R}$ définie par $\varphi(\rho) = f(x_k + \rho w_k)$. On a $\varphi(0) = f(x_k)$ et $\varphi(\rho) = f(x_k + \rho w_k) \to +\infty$ lorsque $\rho \to +\infty$.
   En effet si $\rho \to +\infty$, on a $|x_k + \rho w_k| \to +\infty$. Donc φ étant continue, φ admet un minimum, atteint en $\bar{\rho}$, et donc $\exists \bar{\rho} \in \mathbb{R}_+ ; f(x_k + \bar{\rho} w) \leq f(x_k + \rho w_k) \ \ \forall \rho > 0$.
4. a) Montrons que la suite $(f(x_k))_{n \in \mathbb{N}}$ est convergente. La suite $(f(x_k))_{n \in \mathbb{N}}$ vérifie

$$f(x_{k+1}) \leq f(x_k).$$

   De plus $f(x) \to +\infty$ lorsque $|x| \to +\infty$ donc f est bornée inférieurement. On en conclut que la suite $(f(x_k))_{n \in \mathbb{N}}$ est convergente.

   b) Montrons que $x_k \in B_R \ \ \forall k \in \mathbb{N}$. On sait que si $x \notin B_R$ alors $f(x) > f(x_0)$. Or la suite $(f(x_k))_{n \in \mathbb{R}}$ est décroissante donc $f(x_k) \leq f(x_0) \ \ \forall k$, donc $x_k \in B_R, \ \ \forall k \in \mathbb{N}$.

c) Montrons que $f(x_k+\rho w_k) \leq f(x_k) - \rho|w_k|^2 + \dfrac{\rho^2}{2}M|w_k|^2,\ \ \forall \rho \in [0, \dfrac{1}{|w_k|}]$. Soit φ définie de $\mathbb{R}_+$ dans $\mathbb{R}$ par $\varphi(\rho) = f(x_k+\rho w_k)$. On a

$$\varphi(\rho) = \varphi(0) + \rho\varphi'(0) + \frac{\rho^2}{2}\varphi''(\tilde{\rho}),\ \text{où}\quad \tilde{\rho} \in]0, \rho[.$$

Or $\varphi'(\rho) = \nabla f(x_k+\rho w_k)\cdot w_k$ et $\varphi''(\rho) = H(x_k+\rho w_k)w_k \cdot w_k$. Donc

$$\varphi(\rho) = \underbrace{\varphi(0)}_{f(x_k)} + \rho \underbrace{\nabla f(x_k)}_{-w_k}\cdot w_k + \frac{\rho^2}{2}H(x_k+\tilde{\rho}w_k)w_k\cdot w_k.$$

Si $\rho \in [0, \dfrac{1}{|w_k|}]$ on a

$$\begin{aligned} |x_k+\tilde{\rho}w_k| &\leq |x_k| + \frac{1}{|w_k|}|w_k| \\ &\leq R+1, \end{aligned}$$

donc $x_k+\tilde{\rho}w_k \in B_{R+1}$ et par la question 2,

$$H(x_k+\tilde{\rho}w_k)w_k\cdot w_k \leq M|w_k|^2.$$

On a donc bien

$$\varphi(\rho) = f(x_k+\rho w_k) \leq f(x_k) - \rho|w_k|^2 + \frac{\rho^2}{2}M|w_k|^2.$$

d) Montrons que $f(x_{k+1}) \leq f(x_k) - \dfrac{|w_k|^2}{2M}$ si $|w_k| \leq M$.

Comme le choix de α_k est optimal, on a

$$f(x_{k+1}) = f(x_k+\alpha_k w_k) \leq f(x_k+\rho w_k),\ \ \forall \rho \in \mathbb{R}_+.$$

donc en particulier

$$f(x_{k+1}) \leq f(x_k+\rho w_k),\ \ \forall \rho \in [0, \frac{1}{|w_k|}].$$

En utilisant la question précédente, on obtient

$$f(x_{k+1}) \leq f(x_k) - \rho|w_k|^2 + \frac{\rho^2}{2}M|w_k|^2 = \varphi(\rho),\ \ \forall \rho \in [0, \frac{1}{|w_k|}]. \tag{3.42}$$

Or la fonction φ atteint son minimum pour

$$-|w_k|^2 + \rho M|w_k|^2 = 0$$

c'est–à–dire $\rho M = 1$ ou encore $\rho = \frac{1}{M}$ ce qui est possible si $\dfrac{1}{|w_k|} \geq \dfrac{1}{M}$ (puisque 3.42 est vraie si $\rho \leq \dfrac{1}{|w_k|}$).

Comme on a supposé $|w_k| \leq M$, on a donc

$$f(x_{k+1}) \leq f(x_k) - \frac{|w_k|^2}{M} + \frac{|w_k|^2}{2M} = f(x_k) - \frac{|w_k|^2}{2M}.$$

e) Montrons que $-f(x_{k+1}) + f(x_k) \geq \dfrac{|w_k|^2}{2\bar{M}}$ où $\bar{M} = \sup(M, \tilde{M})$ avec $\tilde{M} = \sup\{|\nabla f(x)|,\ x \in B_R\}$.

On sait par la question précédente que si

$$|w_k| \leq M, \text{ on a } -f(x_{k+1}) - f(x_k) \geq \frac{|w_k|^2}{2M}.$$

Montrons que si $|w_k| \geq M$, alors $-f(x_{k+1}) + f(x_k) \geq \dfrac{|w_k|^2}{2\tilde{M}}$. On aura alors le résultat souhaité.

On a

$$f(x_{k+1}) \leq f(x_k) - \rho|w_k|^2 + \frac{\rho^2}{2}M|w_k|^2,\ \ \forall \rho \in [0, \frac{1}{|w_k|}].$$

Donc

$$f(x_{k+1}) \leq \min_{[0,\frac{1}{|w_k|}]} \underbrace{[f(x_k) - \rho|w_k|^2 + \frac{\rho^2}{2}M|w_k|^2]}_{P_k(\rho)}$$

— 1er cas si $|w_k| \leq M$, on a calculé ce $\min$ à la question c).

— si $|w_k| \geq M$, la fonction $P_k(\rho)$ est décroissante sur $[0, \dfrac{1}{|w_k|}]$ et le minimum est donc atteint pour $\rho = \dfrac{1}{|w_k|}$.

$$\begin{aligned} \text{Or } P_k\left(\frac{1}{|w_k|}\right) &= f(x_k) - |w_k| + \frac{M}{2} \leq f(x_k) - \frac{|w_k|}{2} \\ &\leq f(x_k) - \frac{|w_k|^2}{2\tilde{M}}, \end{aligned}$$

en remarquant que $|w_k| \leq \tilde{M}$.

5. Montrons que $\nabla f(x_k) \to 0$ lorsque $k \to +\infty$. On a montré que $\forall k$, $|w_k|^2 \leq 2\bar{M}(f(x_k) - f(x_{k+1}))$. Or la suite $(f(x_k))_{n\in\mathbb{N}}$ est convergente. Donc $|w_k| \to 0$ lorsque $n \to +\infty$ et $w_k = \nabla f(x_k)$ ce qui prouve le résultat.
La suite $(x_k)_{n\in\mathbb{N}}$ est bornée donc $\exists (n_k)_{k\in\mathbb{N}}$ et $\tilde{x} \in \mathbb{R}^n$; $x_{n_k} \to \tilde{x}$ lorsque $k \to +\infty$ et comme $\nabla f(x_{n_k}) \to 0$, on a, par continuité, $\nabla f(\tilde{x}) = 0$.
6. On suppose qu'il existe un unique $\bar{x} \in \mathbb{R}^n$ tel que $\nabla f(\bar{x}) = 0$. Comme f est croissante à l'infini, il existe un point qui réalise un minimum de f, et on sait qu'en ce point le gradient s'annule ; en utilisant l'hypothèse d'unicité, on en déduit que ce point est forcément $\bar{x}$. On remarque aussi que $\bar{x}$ est la seule valeur d'adhérence de la suite (bornée) $(x_k)_{k\in\mathbb{N}}$, et donc que $x_k \to \bar{x}$ quand $k \to +\infty$.

Exercice 145 page 216 (Méthode de relaxation)

1. On sait par la propostion 3.13 que si f vérifie l'hypothèse (3.10) alors f est strictement convexe et tend vers l'infini en l'infini, et donc il existe un unique $\bar{x} \in \mathbb{R}^n$ réalisant son minimum.

2. Ecrivons l'hypothèse (3.10) avec $x = se_k$ et $y = te_k$ où $(s,t) \in \mathbb{R}^2$ et e_k est le k-ième vecteur de la base canonique de $\mathbb{R}^n$; en notant $\partial_k f$ la dérivée partielle de f par rapport à la k-ième variable, il vient :

$$(\partial_k f(s) - \partial_k f(t))(s-t) \geq \alpha|s-t|^2.$$

En appliquant à nouveau la proposition 3.13 au cas $n = 1$, on en déduit l'existence et unicité de $\bar{s}$ tel que

$$\varphi_k^{(k+1)}(\bar{s}) = \inf_{s\in\mathbb{R}} \varphi_k^{(k+1)}(s).$$

Comme l'algorithme (3.38) procède à n minimisations de ce type à chaque itération, on en déduit que la suite $(x^{(k)})_{n\in\mathbb{N}}$ construite par cet algorithme est bien définie.

3.(a) Par définition, $x_k^{(k+1)}$ réalise le minimum de la fonction $\varphi_k^{(k+1)}$ sur $\mathbb{R}$. Comme de plus, $\varphi_k^{(k+1)} \in C^1(\mathbb{R},\mathbb{R})$, on a donc $(\varphi_k^{(k+1)})'(x_k^{(k+1)}) = 0$. Or $(\varphi_k^{(k+1)})'(x_k^{(k+1)}) = \partial_k f(x^{(n+1,k)})$, et donc $\partial_k f(x^{(n+1,k)}) = 0$.
D'après la démonstration de la proposition 3.13 (voir l'inégalité (3.11)), on a

$$\begin{aligned} f(x^{(n+1,k-1)}) - f(x^{(n+1,k)}) &\geq \nabla f(x^{(n+1,k)})\cdot(x^{(n+1,k-1)} - x^{(n+1,k)}) \\ &+ \frac{\alpha}{2}|x^{(n+1,k-1)} - x^{(n+1,k)}|^2. \end{aligned}$$

Or $x^{(n+1,k-1)} - x^{(n+1,k)} = -x_k^{(k+1)} e_k$ et $\nabla f(x^{(n+1,k)})\cdot e_k = \partial_k f(x^{(n+1,k)}) = 0$. On en déduit que :

$$f(x^{(n+1,k-1)}) - f(x^{(n+1,k)}) \geq \frac{\alpha}{2}|x^{(n+1,k-1)} - x^{(n+1,k)}|^2.$$

3.(b) Par définition de la suite $(x^{(k)})_{n\in\mathbb{N}}$, on a :

$$f(x^{(k)}) - f(x^{(k+1)}) = \sum_{k=1}^{n} f(x^{(n+1,k-1)}) - f(x^{(n+1,k)}).$$

Par la question précédente, on a donc :

$$f(x^{(k)}) - f(x^{(k+1)}) \geq \frac{\alpha}{2}\sum_{k=1}^{n} |x^{(n+1,k-1))} - x^{(n+1,k)}|^2.$$

Or $x^{(n+1,k-1))} - x^{(n+1,k)} = -x_k^{(k+1)} e_k$, et $(e_k)_{k\in n}$ est une base orthonormée. On peut donc écrire que

$$\begin{aligned} \sum_{k=1}^{n} |x^{(n+1,k-1))} - x^{(n+1,k)}|^2 &= \sum_{k=1}^{n} |(x_k^{(k)} - x_k^{(k+1)})e_k|^2 \\ &= |\sum_{k=1}^{n} (x_k^{(k)} - x_k^{(k+1)})e_k|^2 \\ &= |\sum_{k=1}^{n} (x^{(n+1,k-1))} - x^{(n+1,k)})|^2 \\ &= |x^{(k)} - x^{(k+1)}|^2. \end{aligned}$$

On en déduit que

$$f(x^{(k)}) - f(x^{(k+1)}) \geq \frac{\alpha}{2}|x^{(k)} - x^{(k+1)}|^2.$$

La suite $(f(x^{(k)}))_{k\in\mathbb{N}}$ est bornée inférieurement par $f(\bar{x})$; l'inégalité précédente montre qu'elle est décroissante, donc elle converge. On a donc $f(x^{(k)}) - f(x^{(k+1)} \to 0$ lorsque $n \to +\infty$, et donc par l'inégalité précédente,

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

De plus, pour $1 \leq k \leq n$,

$$\begin{aligned} |x^{(n+1,k)} - x^{(k+1)}|^2 &= \sum_{\ell=k}^{n} |(x_\ell^{(k)} - x_\ell^{(k+1)})e_\ell|^2 \\ &= |\sum_{\ell=k}^{n} (x_\ell^{(k)} - x_\ell^{(k+1)})e_\ell|^2 \\ &= |\sum_{\ell=k}^{n} (x^{(n+1,\ell-1))} - x^{(n+1,\ell)})|^2 \\ &\leq |x^{(k)} - x^{(k+1)}|^2. \end{aligned}$$

d'où l'on déduit que $\lim_{n\to+\infty} |x^{(n+1,k)} - x^{(k+1)}| = 0$.

4. En prenant $x = \bar{x}$ et $y = x^{(k+1)}$ dans l'hypothèse (3.10) et en remarquant que, puisque $\bar{x}$ réalise le minimum de f, on a $\nabla f(\bar{x}) = 0$, on obtient :

$$(-\nabla f(x^{(k+1)}) \cdot (\bar{x} - x^{(k+1)}) \geq \alpha |\bar{x} - x^{(k+1)}|^2,$$

et donc, par l'inégalité de Cauchy Schwarz :

$$|x^{(k+1)} - \overline{x}| \leq \frac{1}{\alpha} \left(\sum_{k=1}^{n} |\partial_k f(x^{(k+1)})|^2 \right)^{\frac{1}{2}}.$$

5. En vertu de la proposition 3.13, on sait que la fonction f est croissante à l'infini. Donc il existe $R > 0$ tel que si $|x| > R$ alors $f(x) > f(x_0)$. Or, la suite $(f(x_k))_{k\in\mathbb{N}}$ étant décroissante, on a $f(x_k)) \leq f(x_0)$ pour tout n, et donc $|x_k| \leq R$ pour tout n. Par la question 3(b), on sait que pour tout $k \geq 1$, $\lim_{n\to+\infty} |x^{(n+1,k)} - x^{(k+1)}| = 0$, ce qui prouve que les suites $(x^{(n+1,k)})_{n\in\mathbb{N}}$, pour $k = 1, \ldots, n$, sont également bornées.

Comme $\lim_{n\to+\infty} |x^{(n+1,k)} - x^{(k+1)}| = 0$, on a pour tout $\eta > 0$, l'existence de $N_\eta \in \mathbb{N}$ tel que $|x^{(n+1,k)} - x^{(k+1)}| < \eta$ si $n \geq N_\eta$. Comme $f \in C^1(\mathbb{R}, \mathbb{R})$, la fonction $\partial_k f$ est uniformément continue sur les bornés (théorème de Heine), et donc pour tout $\varepsilon > 0$, il existe $\eta > 0$ tel que si $|x - y| < \eta$ alors $|\partial_k f(x) - \partial_k f(y)| \leq \epsilon$. On a donc, pour $n \geq N_\eta$: $|\partial_k f(x^{(n+1,k)}) - \partial_k f(x^{(k+1)})| \leq \epsilon$, ce qui démontre que :

$$|\partial_k f(x^{(k+1)})| \to 0 \text{ lorsque } n \to +\infty.$$

On en conclut par le résultat de la question 4 que $x^{(k)} \to \bar{x}$ lorsque $n \to +\infty$.

6. On a vu au paragraphe 3.2.2 que dans ce cas, $\nabla f(x) = \frac{1}{2}(A + A^t)x - b$. L'algorithme 3.38 est donc la méthode de Gauss Seidel pour la résolution du système linéaire $\frac{1}{2}(A + A^t)x = b$.

7 (a) La fonction g est strictement convexe (car somme d'une fonction strictement convexe : $(x_1, x_2) \to x_1^2 + x_2^2$, d'une fonction linéaire par morceaux : $(x_1, x_2) \mapsto -2(x_1 + x_2) + 2|x_1 - x_2|$. et croissante à l'infini grâce aux termes en puissance 2. Il existe donc un unique élément $\overline{x} = (\overline{x}_1, \overline{x}_2)^t$ de $\mathbb{R}^2$ tel que $g(\overline{x}) = \inf_{x\in\mathbb{R}^2} g(x)$.
7 (b) Soit $\epsilon > 0$. On a, pour tout $x \in \mathbb{R}$, $\phi_x(\epsilon) = g(x, x + \epsilon) = x^2 + (x + \epsilon)^2 - 4x$, qui atteint (pour tout x) son minimum pour $\epsilon = 0$. Le minimum de g se situe donc sur l'axe $x = y$. Or $\psi(x) = g(x, x) = 2x^2 - 4x$ atteint son minimum en $x = 1$.
7 (c) Si $x^{(0)} = (0, 0)^t$, on vérifie facilement que l'algorithme (3.38) appliqué à g est stationnaire. La suite ne converge donc pas vers $\overline{x}$. La fonction g n'est pas différentiable sur la droite $x_1 = x_2$.

Exercice 147 page 218 (Gradient conjugué pour une matrice non symétrique)

1. Comme A est inversible, A^t l'est aussi et donc les systèmes (3.39) et (3.40) sont équivalents.

2 (a) La matrice M est symétrique définie positive, car A est inversible et $M = AA^t$ est symétrique. Donc ses valeurs propres sont strictement positives.

2 (b) On a $\text{cond}(A) = \|A\|\|A^{-1}\|$. Comme la norme est ici la norme euclidienne, on a : $\|A\| = (\rho(A^t A))^{\frac{1}{2}}$ et $\|A^{-1}\| = (\rho((A^{-1})^t A^{-1}))^{\frac{1}{2}} = (\rho(AA^t)^{-1}))^{\frac{1}{2}}$. On vérifie facilement que $M = A^t A$ et $A^t A$ ont mêmes valeurs propres et on en déduit le résultat.

3. Ecrivons l'algorithme du gradient conjugué pour la résolution du système (3.40)

Initialisation
Soit $x^{(0)} \in \mathbb{R}^n$, et soit $r^{(0)} = A^t b - A^t A x^{(0)} =$
1) Si $r^{(0)} = 0$, alors $Ax^{(0)} = b$ et donc $x^{(0)} = \bar{\boldsymbol{x}}$,
auquel cas l'algorithme s'arrête.
2) Si $r^{(0)} \neq 0$, alors on pose $w^{(0)} = r^{(0)}$, et on choisit $\rho_0 = \dfrac{r^{(0)} \cdot r^{(0)}}{A^t A w^{(0)} \cdot w^{(0)}}$.
On pose alors $x^{(1)} = x^{(0)} + \rho_0 w^{(0)}$.

Itération $1 \leq n \leq n-1$:
On suppose $x^{(0)}, \ldots, x^{(k)}$ et $w^{(0)}, \ldots, w^{(k-1)}$ connus et on pose
$\boldsymbol{r}^{(k)} = A^t b - A^t A x^{(k)}$.
1) Si $\boldsymbol{r}^{(k)} = 0$ on a $Ax^{(k)} = b$ donc $x^{(k)} = \bar{\boldsymbol{x}}$
auquel cas l'algorithme s'arrête.
2) Si $\boldsymbol{r}^{(k)} \neq 0$, alors on pose $w^{(k)} = \boldsymbol{r}^{(k)} + \lambda_{k-1} w^{(k-1)}$
avec $\lambda_{k-1} = \frac{\boldsymbol{r}^{(k)} \cdot \boldsymbol{r}^{(k)}}{r^{(k-1)} \cdot r^{(k-1)}}$. et on pose $\alpha_k = \dfrac{\boldsymbol{r}^{(k)} \cdot \boldsymbol{r}^{(k)}}{A^t A w^{(k)} \cdot w^{(k)}}$.
On pose alors $x^{(k+1)} = x^{(k)} + \alpha_k w^{(k)}$.

Si on implémente l'algorithme sous cette forme, on a intérêt à calculer d'abord $\tilde{b} = A^t b$ et $M = A^t A$ pour minimiser le nombre de mutliplications matrice matrice et matrice vecteur. Au lieu du coût de l'algorithme initial, qui est en $2n^3 + O(n^2)$, on a donc un coût en $3n^3 + O(n^2)$.
Maintenant si on est optimiste, on peut espérer converger en moins de n itérations (en fait, c'est malheureusement rarement le cas), et dans ce cas il est plus économique d'écrire l'algorithme précédent sous la forme suivante.

Initialisation
Soit $x^{(0)} \in \mathbb{R}^n$, et soit $s^{(0)} = b - Ax^{(0)}$ et soit $r^{(0)} = A^t s^{(0)}$
1) Si $r^{(0)} = 0$, alors $Ax^{(0)} = b$ et donc $x^{(0)} = \bar{\boldsymbol{x}}$,
auquel cas l'algorithme s'arrête.
2) Si $r^{(0)} \neq 0$, alors on pose $w^{(0)} = r^{(0)}, y^{(0)} = Aw^{(0)}$ et on choisit $\rho_0 = \dfrac{r^{(0)} \cdot r^{(0)}}{y^{(0)} \cdot y^{(0)}}$.
On pose alors $x^{(1)} = x^{(0)} + \rho_0 w^{(0)}$.

Itération $1 \leq n \leq n-1$:
On suppose $x^{(0)}, \ldots, x^{(k)}$ et $w^{(0)}, \ldots, w^{(k-1)}$ connus et on pose
$s^{(k)} = b - Ax^{(k)}$ et $\boldsymbol{r}^{(k)} = A^t s^{(k)}$.
1) Si $\boldsymbol{r}^{(k)} = 0$ on a $Ax^{(k)} = b$ donc $x^{(k)} = \bar{\boldsymbol{x}}$
auquel cas l'algorithme s'arrête.
2) Si $\boldsymbol{r}^{(k)} \neq 0$, alors on pose $w^{(k)} = \boldsymbol{r}^{(k)} + \lambda_{k-1} w^{(k-1)}$
avec $\lambda_{k-1} = \frac{\boldsymbol{r}^{(k)} \cdot \boldsymbol{r}^{(k)}}{r^{(k-1)} \cdot r^{(k-1)}}$. et on pose $\alpha_k = \dfrac{\boldsymbol{r}^{(k)} \cdot \boldsymbol{r}^{(k)}}{y^{(k)} \cdot y^{(k)}}$ avec $y^{(k)} = Aw^{(k)}$.
On pose alors $x^{(k+1)} = x^{(k)} + \alpha_k w^{(k)}$.

On peut facilement vérifier que dans cette version, on a un produit matrice vecteur en plus à chaque itération, donc le coût est le même pour n itérations, mais il est inférieur si on a moins de n itérations.
Remarque : Cette méthode s'appelle méthode du gradient conjugué appliquée aux équations normales. Elle est facile à comprendre et à programmer. Malheureusement, elle ne marche pas très bien dans la pratique, et on lui préfère des méthodes plus sophistiquées telles sue la méthode "BICGSTAB" ou "GMRES".

Exercice 150 page 220 (Méthode de Polak-Ribière)

1. Montrons que f est strictement convexe et croissante à l'infini. Soit φ la fonction de $\mathbb{R}$ dans $\mathbb{R}$ définie par

$$\varphi(t) = f(x + t(y - x)).$$

On a $\varphi \in C^2(\mathbb{R}, \mathbb{R})$, $\varphi(0) = f(x)$ et $\varphi(1) = f(y)$, et donc :

$$f(y) - f(x) = \varphi(1) - \varphi(0) = \int_0^1 \varphi'(t)dt.$$

En intégrant par parties, ceci entraîne :

$$f(y) - f(x) = \varphi'(0) + \int_0^1 (1 - t)\varphi''(t)dt. \tag{3.43}$$

Or $\varphi'(t) = \nabla(x + t(y - x)) \cdot (y - x)$ et donc $\varphi''(t) = H(x + t(y - x))(y - x) \cdot (y - x)$. On a donc par hypothèse $\varphi''(t) \geq \alpha|y - x|^2$. On déduit alors de 3.43 que

$$f(y) \geq f(x) + \nabla f(x) \cdot (y - x) + \frac{\alpha}{2}|y - x|^2. \tag{3.44}$$

L'inégalité 3.44 entraîne la stricte convexité de f et sa croissance à l'infini (voir la démonstration de la proposition 3.13).

Il reste à montrer que l'ensemble $\mathcal{VP}(H(x))$ des valeurs propres de $H(x)$ est inclus dans $[\alpha, \beta]$. Comme $f \in C^2(\mathbb{R}, \mathbb{R})$, $H(x)$ est symétrique pour tout $x \in \mathbb{R}$, et donc diagonalisable dans $\mathbb{R}$. Soit $\lambda \in \mathcal{VP}(H(x))$; il existe donc $y \in \mathbb{R}^n$, $y \neq 0$ tel que $H(x)y = \lambda y$, et donc $\alpha y \cdot y \leq \lambda y \cdot y \leq \beta y \cdot y$, $\forall \lambda \in \mathcal{VP}(H)(x))$. On en déduit que $\mathcal{VP}(H(x)) \subset [\alpha, \beta]$.

2. Montrons par récurrence sur n que $g^{(k+1)} \cdot w^{(k)} = 0$ et $g^{(k)} \cdot g^{(k)} = g^{(k)} \cdot w^{(k)}$ pour tout $k \in \mathbb{N}$.
Pour $k = 0$, on a $w^{(0)} = g^{(0)} = -\nabla f(x^{(0)})$.
Si $\nabla f(x^{(0)}) = 0$ l'algorithme s'arrête. Supposons donc que $\nabla f(x^{(0)}) \neq 0$. Alors $w^{(0)} = -\nabla f(x^{(0)})$ est une direction de descente stricte. Comme $x^{(1)} = x^{(0)} + \rho_0 w^{(0)}$ où ρ_0 est optimal dans la direction $w^{(0)}$, on a $g^{(1)} \cdot w^{(0)} = -\nabla f(x^{(1)}) \cdot w^{(0)} = 0$. De plus, on a évidemment $g^{(0)} \cdot w^{(0)} = g^{(0)} \cdot g^{(0)}$.
Supposons maintenant que $g^{(k)} \cdot w^{(k-1)} = 0$ et $g^{(k-1)} \cdot g^{(k-1)} = g^{(k-1)} \cdot w^{(k-1)}$, et montrons que $g^{(k+1)} \cdot w^{(k)} = 0$ et $g^{(k)} \cdot g^{(k)} = g^{(k)} \cdot w^{(k)}$.
Par définition, on a :

$$w^{(k)} = g^{(k)} + \lambda_{k-1} w^{(k-1)}, \text{ donc}$$
$$w^{(k)} \cdot g^{(k)} = g^{(k)} \cdot g^{(k)} + \lambda_{k-1} w^{(k-1)} \cdot g^{(k)} = g^{(k)} \cdot g^{(k)}$$

par hypothèse de récurrence. On déduit de cette égalité que $w^{(k)} \cdot g^{(k)} > 0$ (car $g^{(k)} \neq 0$) et donc $w^{(k)}$ est une direction de descente stricte en $x^{(k)}$. On a donc $\nabla f(x^{(k+1)}) \cdot w^{(k)} = 0$, et finalement $g^{(k+1)} \cdot w^{(k)} = 0$.

3. Par définition, $g^{(k)} = -\nabla f(x^{(k)})$; or on veut calculer $g^{(k+1)} - g^{(k)} = -\nabla f(x^{(k+1)}) + \nabla f(x^{(k)})$. Soit φ la fonction de $\mathbb{R}$ dans $\mathbb{R}$ définie par :

$$\varphi(t) = -\nabla f(x^{(k)} + t(x^{(k+1)} - x^{(k)})).$$

On a donc :

$$\begin{aligned} \varphi(1) - \varphi(0) &= g^{(k+1)} - g^{(k)} \\ &= \int_0^1 \varphi'(t)dt. \end{aligned}$$

Calculons φ' : $\varphi'(t) = H(x^{(k)} + t(x^{(k+1)} - x^{(k)}))(x^{(k+1)} - x^{(k)})$. Et comme $x^{(k+1)} = x^{(k)} + \alpha_k w^{(k)}$, on a donc :

$$g^{(k+1)} - g^{(k)} = \alpha_k J^{(k)} w^{(k)}. \tag{3.45}$$

De plus, comme $g^{(k+1)} \cdot w^{(k)} = 0$ (question 1), on obtient par (3.45) que

$$\alpha_k = \frac{g^{(k)} \cdot w^{(k)}}{J^{(k)} w^{(k)} \cdot w^{(k)}}$$

(car $J^{(k)} w^{(k)} \cdot w^{(k)} \neq 0$, puisque $J^{(k)}$ est symétrique définie positive).

4. Par définition, on a $w^{(k)} = g^{(k)} + \lambda_{k-1} w^{(k-1)}$, et donc

$$|w^{(k)}| \leq |g^{(k)}| + |\lambda_{k-1}||w^{(k-1)}|. \tag{3.46}$$

Toujours par définition, on a :

$$\lambda_{k-1} = \frac{g^{(k)} \cdot (g^{(k)} - g^{(k-1)})}{g^{(k-1)} \cdot g^{(k-1)}}.$$

Donc, par la question 3, on a :

$$\lambda_{k-1} = \frac{\alpha_{k-1} g^{(k)} \cdot J^{(k-1)} w^{(k-1)}}{g^{(k-1)} \cdot g^{(k-1)}}.$$

En utilisant la question 2 et à nouveau la question 3, on a donc :

$$\lambda_{k-1} = -\frac{J^{(k-1)} w^{(k-1)} \cdot g^{(k)}}{J^{(k-1)} w^{(k-1)} \cdot w^{(k-1)}},$$

et donc

$$|\lambda_{k-1}| = \frac{|J^{(k-1)} w^{(k-1)} \cdot g^{(k)}|}{J^{(k-1)} w^{(k-1)} \cdot w^{(k-1)}},$$

car $J^{(k-1)}$ est symétrique définie positive.
De plus, en utilisant les hypothèses sur H, on vérifie facilement que

$$\alpha |x|^2 \leq J^{(k)} x \cdot x \leq \beta |x|^2 \ \ \forall x \in \mathbb{R}^n.$$

On en déduit que

$$|\lambda_{k-1}| \leq \frac{|J^{(k-1)} w^{(k-1)} \cdot g^{(k)}|}{\alpha |w^{(k-1)}|^2}.$$

On utilise alors l'inégalité de Cauchy–Schwarz :

$$\begin{aligned} |J^{(k-1)} w^{(k-1)} \cdot g^{(k)}| &\leq \|J^{(k-1)}\|_2 \ |w^{(k-1)}| \ |g^{(k)}| \\ &\leq \beta |w^{(k-1)}| \ |g^{(k)}|. \end{aligned}$$

On obtient donc que

$$|\lambda_{k-1}| \leq \frac{\beta}{\alpha} \frac{|g^{(k)}|}{|w^{(k-1)|}},$$

ce qui donne bien grâce à (3.46) :

$$|w^{(k)}| \leq |g^{(k)}| (1 + \frac{\beta}{\alpha}).$$

5. • Montrons d'abord que la suite $(f(x^{(k)}))_{n \in \mathbb{N}}$ converge. Comme $f(x^{(k+1)}) = f(x^{(k)} + \alpha_k w^{(k)}) \leq f(x^{(k)} + \rho w^{(k)}) \ \ \forall \rho \geq 0$, on a donc en particulier $f(x^{(k+1)}) \leq f(x^{(k)})$. La suite $(f(x^{(k)}))_{n \in \mathbb{N}}$ est donc décroissante. De plus, elle est minorée par $f(\bar{\boldsymbol{x}})$. Donc elle converge, vers une certaine limite $\ell \in \mathbb{R}$, lorsque k tend vers $+\infty$.

   • La suite $(x^{(k)})_{k \in \mathbb{N}}$ est bornée : en effet, comme f est croissante à l'infini, il existe $R > 0$ tel que si $|x| > R$ alors $f(x) > f(x^{(0)})$. Or $f(x^{(k)}) \leq f(x^{(0)})$ pout tout $k \in \mathbb{N}$, et donc la suite $(x^{(k)})_{n \in \mathbb{N}}$ est incluse dans la boule de rayon R.

• Montrons que $\nabla f(x^{(k)}) \to 0$ lorsque $n \to +\infty$.
On a, par définition de $x^{(k+1)}$,

$$f(x^{(k+1)}) \leq f(x^{(k)} + \rho w^{(k)}), \quad \forall \rho \geq 0.$$

En introduisant la fonction φ définie de $\mathbb{R}$ dans $\mathbb{R}$ par $\varphi(t) = f(x^{(k)} + t\rho w^{(k)})$, on montre facilement (les calculs sont les mêmes que ceux de la question 1) que

$$f(x^{(k)} + \rho w^{(k)}) = f(x^{(k)}) + \rho \nabla f(x^{(k)}) \cdot w^{(k)} + \rho^2 \int_0^1 H(x^{(k)} + t\rho w^{(k)}) w^{(k)} \cdot w^{(k)} (1-t) dt,$$

pour tout $\rho \geq 0$. Grâce à l'hypothèse sur H, on en déduit que

$$f(x^{(k+1)}) \leq f(x^{(k)}) + \rho \nabla f(x^{(k)}) \cdot w^{(k)} + \frac{\beta}{2} \rho^2 |w^{(k)}|^2, \ \forall \rho \geq 0.$$

Comme $\nabla f(x^{(k)}) \cdot w^{(k)} = -g^{(k)} \cdot w^{(k)} = -|g^{(k)}|^2$ (question 2) et comme $|w^{(k)}| \leq |g^{(k)}|(1 + \frac{\beta}{\alpha})$ (question 4), on en déduit que :

$$f(x^{(k+1)}) \leq f(x^{(k)}) - \rho |g^{(k)}|^2 + \rho^2 \gamma |g^{(k)}|^2 = \psi_k(\rho), \ \forall \rho \geq 0,$$

où $\gamma = \frac{\beta^2}{2} + (1 + \frac{\beta}{\alpha})^2$. La fonction ψ_k est un polynôme de degré 2 en ρ, qui atteint son minimum lorsque $\psi_k'(\rho) = 0$, *i.e.* pour $\rho = \dfrac{1}{2\gamma}$. On a donc, pour $\rho = \dfrac{1}{2\gamma}$,

$$f(x^{(k+1)}) \leq f(x^{(k)}) - \frac{1}{4\gamma} |g^{(k)}|^2,$$

d'où on déduit que

$$|g^{(k)}|^2 \leq 4\gamma (f(x^{(k)}) - f(x^{(k+1)}) \underset{k \to +\infty}{\to} 0$$

On a donc $\nabla f(x^{(k)}) \to 0$ lorsque $k \to +\infty$.

• La suite $(x^{(k)})_{n \in \mathbb{N}}$ étant bornée, il existe une sous–suite qui converge vers $x \in \mathbb{R}^n$, comme $\nabla f(x^{(k)}) \to 0$ et comme ∇f est continue, on a $\nabla f(x) = 0$. Par unicité du minimum (f est croissante à l'infini et strictement convexe) on a donc $x = \bar{\boldsymbol{x}}$.
Enfin on conclut à la convergence de toute la suite par un argument classique (voir question 6 de l'exercice 142 page 215).

Exercice 151 page 220 (Algorithme de quasi Newton)

Partie 1

1. Par définition de $w^{(k)}$, on a :

$$w^{(k)} \cdot \nabla f(x^{(k)}) = -K^{(k)} \nabla f(x^{(k)}) \cdot \nabla f(x^{(k)}) < 0$$

car K est symétrique définie positive.
Comme α_k est le paramètre optimal dans la direction $w^{(k)}$, on a $\nabla f(x^{(k)} + \alpha_k w^{(k)}) \cdot w^{(k)} = 0$, et donc $Ax^{(k)} \cdot w^{(k)} + \alpha_k A w^{(k)} \cdot w^{(k)} = b \cdot w^{(k)}$; on en déduit que

$$\alpha_k = -\frac{g^{(k)} \cdot w^{(k)}}{A w^{(k)} \cdot w^{(k)}}.$$

Comme $w^{(k)} = -K^{(k)} g^{(k)}$, ceci s'écrit encore :

$$\alpha_k = \frac{g^{(k)} \cdot K^{(k)} g^{(k)}}{A K^{(k)} g^{(k)} \cdot K^{(k)} g^{(k)}}.$$

2. Si $K^{(k)} = A^{-1}$, la formule précédente donne immédiatement $\alpha_k = 1$.
3. La méthode de Newton consiste à chercher le zéro de ∇f par l'algorithme suivant (à l'itération 1) :

$$H_f(x^{(0)})(x^{(1)} - x^{(0)}) = -\nabla f(x^{(0)}),$$

(où $H_f(x)$ désigne la hessienne de f au point x) c'est–à–dire

$$A(x^{(1)} - x^{(0)}) = -Ax^{(0)} + b.$$

On a donc $Ax^{(k)} = b$, et comme la fonction f admet un unique minimum qui vérifie $Ax = b$, on a donc $x^{(1)} = x$, et la méthode converge en une itération.

Partie 2 Méthode de Fletcher–Powell.

1. Soit $n \in \mathbb{N}$, on suppose que $g^{(k)} \neq 0$. Par définition, on a $s^{(k)} = x^{(k+1)} - x^{(k)} = -\alpha_k K^{(k)} g^{(k)}$, avec $\alpha_k > 0$. Comme $K^{(k)}$ est symétrique définie positive elle est donc inversible ; donc comme $g^{(k)} \neq 0$, on a $K^{(k)} g^{(k)} \neq 0$ et donc $s^{(k)} \neq 0$.
   Soit $i < n$, par définition de $s^{(k)}$, on a :

$$s^{(k)} \cdot As^{(i)} = -\alpha_k K^{(k)} g^{(k)} \cdot As^{(i)}.$$

   Comme $K^{(k)}$ est symétrique,

$$s^{(k)} \cdot As^{(i)} = -\alpha_k g^{(k)} \cdot K^{(k)} As^{(i)}.$$

   Par hypothèse, on a $K^{(k)} As^{(i)} = s^{(i)}$ pour $i < n$, donc on a bien que si $i < n$

$$s^{(k)} \cdot As^{(i)} = 0 \Leftrightarrow g^{(k)} \cdot s^{(i)} = 0.$$

   Montrons maintenant que $g^{(k)} \cdot s^{(i)} = 0$ pour $i < n$.

   • On a

$$\begin{aligned} g^{(i+1)} \cdot s^{(i)} &= -\rho_i g^{(i+1)} \cdot K^{(i)} g^{(i)} \\ &= -\rho_i g^{(i+1)} \cdot w^{(i)}. \end{aligned}$$

   Or $g^{(i+1)} = \nabla f(x^{(i+1)})$ et ρ_i est optimal dans la direction $w^{(i)}$. Donc

$$g^{(i+1)} \cdot s^{(i)} = 0.$$

   • On a

$$\begin{aligned} (g^{(k)} - g^{(i+1)}) \cdot s^{(i)} &= (Ax^{(k)} - Ax^{(i+1)}) \cdot s^{(i)} \\ &= \sum_{k=i+1}^{n-1} (Ax^{(k+1)} - Ax^{(k)}) \cdot s^{(i)} \\ &= \sum_{k=i+1}^{n-1} As^{(k)} \cdot s^{(i)}, \\ &= 0 \end{aligned}$$

   Par hypothèse de A–conjugaison de la famille $(s^{(i)})_{i=1,k-1}$ on déduit alors facilement des deux égalités précédentes que $g^{(k)} \cdot s^{(i)} = 0$. Comme on a montré que $g^{(k)} \cdot s^{(i)} = 0$ si et seulement si $s^{(k)} \cdot As^{(i)} = 0$, on en conclut que la famille $(s^{(i)})_{i=1,\dots,n}$ est $A-$conjuguée, et que les vecteurs $s^{(i)}$ sont non nuls.
2. Montrons que $K^{(k+1)}$ est symétrique. On a :

$$(K^{(k+1)})^t = (K^{(k)})^t + \frac{(s^{(k)}(s^{(k)})^t)^t}{s^{(k)} \cdot y^{(k)}} - \frac{[(K^{(k)}y^{(k)})(K^{(k)}y^{(k)})^t]^t}{K^{(k)}y^{(k)} \cdot y^{(k)}} = K^{(k+1)},$$

   car $K^{(k)}$ est symétrique.

3. Montrons que $K^{(k+1)}As^{(i)} = s^{(i)}$ si $0 \leq i \leq n$. On a :

$$K^{(k+1)}As^{(i)} = K^{(k)}As^{(i)} + \frac{s^{(k)}(s^{(k)})^t}{s^{(k)} \cdot y^{(k)}}As^{(i)} - \frac{(K^{(k)}y^{(k)})(K^{(k)}(y^{(k)})^t}{K^{(k)}y^{(k)} \cdot y^{(k)}}As^{(i)}. \tag{3.47}$$

— Considérons d'abord le cas $i < n$. On a

$$s^{(k)}(s^{(k)})^tAs^{(i)} = s^{(k)}[(s^{(k)})^tAs^{(i)}] = s^{(k)}[s^{(k)} \cdot As^{(i)}] = 0$$

car $s^{(k)} \cdot As^{(i)} = 0$ si $i < n$. De plus, comme $K^{(k)}$ est symétrique, on a :

$$(K^{(k)}y^{(k)})(K^{(k)}y^{(k)})^tAs^{(i)} = K^{(k)}y^{(k)}(y^{(k)})^tK^{(k)}As^{(i)}.$$

Or par la question (c), on a $K^{(k)}As^{(i)} = s^{(i)}$ si $0 \leq i \leq n$. De plus, par définition, $y^{(k)} = As^{(k)}$. On en déduit que

$$(K^{(k)}y^{(k)})(K^{(k)}y^{(k)})^tAs^{(i)} = K^{(k)}y^{(k)}(As^{(k)})^ts^{(i)} = K^{(k)}y^{(k)}(s^{(k)})^tAs^{(i)} = 0$$

puisque on a montré en (a) que les vecteurs $s^{(0)}, \ldots, s^{(k)}$ sont A-conjugués. On déduit alors de (3.47) que

$$K^{(k+1)}As^{(i)} = K^{(k)}As^{(i)} = s^{(i)}.$$

— Considérons maintenant le cas $i = n$. On a

$$K^{(k+1)}As^{(k)} = K^{(k)}As^{(k)} + \frac{s^{(k)}(s^{(k)})^t}{s^{(k)} \cdot y^{(k)}}As^{(k)} - \frac{(K^{(k)}y^{(k)})(K^{(k)}(y^{(k)})^t}{K^{(k)}y^{(k)} \cdot y^{(k)}}As^{(k)},$$

et comme $y^{(k)} = As^{(k)}$,, ceci entraîne que

$$K^{(k+1)}As^{(k)} = K^{(k)}As^{(k)} + s^{(k)} - K^{(k)}y^{(k)} = s^{(k)}.$$

4. Pour $x \in \mathbb{R}^n$, calculons $K^{(k+1)}x \cdot x$:

$$K^{(k+1)}x \cdot x = K^{(k)}x \cdot x + \frac{s^{(k)}(s^{(k)})^t}{s^{(k)} \cdot y^{(k)}}x \cdot x - \frac{(K^{(k)}y^{(k)})(K^{(k)}y^{(k)})^t}{K^{(k)}y^{(k)} \cdot y^{(k)}}x \cdot x.$$

Or $s^{(k)}(s^{(k)})^tx \cdot x = s^{(k)}(s^{(k)} \cdot x) \cdot x = (s^{(k)} \cdot x)^2$, et de même, $(K^{(k)}y^{(k)})(K^{(k)}y^{(k)})^tx \cdot x = (K^{(k)}y^{(k)} \cdot x)^2$. On en déduit que

$$K^{(k+1)}x \cdot x = K^{(k)}x \cdot x + \frac{(s^{(k)} \cdot x)^2}{s^{(k)} \cdot y^{(k)}} - \frac{(K^{(k)}y^{(k)} \cdot x)^2}{K^{(k)}y^{(k)} \cdot y^{(k)}}.$$

En remarquant que $y^{(k)} = As^{(k)}$, et en réduisant au même dénominateur, on obtient alors que

$$K^{(k+1)}x \cdot x = \frac{(K^{(k)}x \cdot x)(K^{(k)}y^{(k)} \cdot y^{(k)}) - (K^{(k)}y^{(k)} \cdot x)^2}{(K^{(k)}y^{(k)} \cdot y^{(k)})} + \frac{(s^{(k)} \cdot x)^2}{As^{(k)} \cdot s^{(k)}}.$$

Montrons maintenant que $K^{(k+1)}$ est symétrique définie positive. Comme $K^{(k)}$ est symétrique définie positive, on a grâce à l'inégalité de Cauchy-Schwarz que $(K^{(k)}y^{(k)} \cdot x)^2 \leq (K^{(k)}x \cdot x)(K^{(k)}y^{(k)})$ avec égalité si et seulement si x et $y^{(k)}$ sont colinéaires. Si x n'est pas colinéaire à $y^{(k)}$, on a donc donc clairement

$$K^{(k+1)}x \cdot x > 0.$$

Si maintenant x est colinéaire à $y^{(k)}$, *i.e.* $x = \alpha y^{(k)}$ avec $\alpha \in \mathbb{R}_+^*$, on a, grâce au fait que $y^{(k)} = As^{(k)}$,

$$\frac{(s^{(k)} \cdot x)^2}{As^{(k)} \cdot s^{(k)}} = \alpha^2\frac{(s^{(k)} \cdot As^{(k)})^2}{As^{(k)} \cdot s^{(k)}} > 0, \text{ et donc } K^{(k+1)}x \cdot x > 0.$$

On en déduit que $K^{(k+1)}$ est symétrique définie positive.

5. On suppose que $g^{(k)} \neq 0$ si $0 \leq n \leq n-1$. On prend comme hypothèse de récurrence que les vecteurs $s^{(0)}, \ldots, s^{(k-1)}$ sont A-conjugués et non-nuls, que $K^{(j)} A s^{(i)} = s^{(i)}$ si $0 \leq i < j \leq n$ et que les matrices $K^{(j)}$ sont symétriques définies positives pour $j = 0, \ldots, n$.
   Cette hypothèse est vérifiée au rang $n = 1$ grâce à la question 1 en prenant $n = 0$ et $K^{(0)}$symétrique définie positive.
   On suppose qu'elle est vraie au rang n. La question 1 prouve qu'elle est vraie au rang $n+1$.
   Il reste maintenant à montrer que $x^{(n+1)} = A^{-1}b = \overline{x}$. On a en effet $K^{(n)} A s^{(i)} = s^{(i)}$ pour $i = 0$ à $n-1$. Or les vecteurs $s^{(0)}, \ldots, s^{(k-1)}$ sont A-conjugués et non-nuls : ils forment donc une base. On en déduit que $K^{(n)} A = \mathrm{Id}$, ce qui prouve que $K^{(n)} = A^{-1}$, et donc, par définition de $x^{(n+1)}$, que $x^{(n+1)} = A^{-1}b = \overline{x}$.

3.4 Optimisation sous contraintes

3.4.1 Définitions

Soit $E = \mathbb{R}^n$, soit $f \in C(E, \mathbb{R})$, et soit K un sous ensemble de E. On s'intéresse à la recherche de $\bar{u} \in K$ tel que :

$$\begin{cases} \bar{u} \in K \\ f(\bar{u}) = \inf\limits_{K} f \end{cases} \tag{3.48}$$

Ce problème est un problème de minimisation avec contrainte (ou "sous contrainte") au sens où l'on cherche u qui minimise f en restreignant l'étude de f aux éléments de K. Voyons quelques exemples de ces contraintes (définies par l'ensemble K), qu'on va expliciter à l'aide des p fonctions continues, $g_i \in C(E, \mathbb{R})$ $i = 1 \ldots p$.

1. **Contraintes égalités.** On pose $K = \{x \in E,\, g_i(x) = 0\ i = 1 \ldots p\}$. On verra plus loin que le problème de minimisation de f peut alors être résolu grâce au théorème des multiplicateurs de Lagrange (voir théorème 3.34).
2. **Contraintes inégalités.** On pose $K = \{x \in E,\, g_i(x) \leq 0\ i = 1 \ldots, p\}$. On verra plus loin que le problème de minimisation de f peut alors être résolu grâce au théorème de Kuhn–Tucker (voir théorème 3.38).
   — *Programmation linéaire.* Avec un tel ensemble de contraintes K, si de plus f est linéaire, c'est-à-dire qu'il existe $b \in \mathbb{R}^n$ tel que $f(x) = b \cdot x$, et les fonctions g_i sont affines, c'est-à-dire qu'il existe $b_i \in \mathbb{R}^n$ et $c_i \in \mathbb{R}$ tels que $g_i(x) = b_i \cdot x + c_i$, alors on dit qu'on a affaire à un problème de "programmation linéaire". Ces problèmes sont souvent résolus numériquement à l'aide de l'algorithme de Dantzig, inventé vers 1950.
   — *Programmation quadratique.* Avec le même ensemble de contraintes K, si de plus f est quadratique, c'est-à-dire si f est de la forme $f(x) = \dfrac{1}{2} Ax \cdot x - b \cdot x$, et les fonctions g_i sont affines, alors on dit qu'on a affaire à un problème de "programmation quadratique".
3. **Programmation convexe.** Dans le cas où f est convexe et K est convexe, on dit qu'on a affaire à un problème de "programmation convexe".

3.4.2 Existence – Unicité – Conditions d'optimalité simple

Théorème 3.28 (Existence). *Soit $E = \mathbb{R}^n$ et $f \in C(E, \mathbb{R})$.*

1. *Si K est un sous-ensemble fermé borné non vide de E, alors il existe $\bar{\boldsymbol{x}} \in K$ tel que $f(\bar{\boldsymbol{x}}) = \inf\limits_{K} f$.*
2. *Si K est un sous-ensemble fermé non vide de E, et si f est croissante à l'infini, c'est–à–dire que $f(x) \to +\infty$ quand $|x| \to +\infty$, alors $\exists \bar{\boldsymbol{x}} \in K$ tel que $f(\bar{\boldsymbol{x}}) = \inf\limits_{K} f$*

DÉMONSTRATION –

1. Si K est un sous-ensemble fermé borné non vide de E, comme f est continue, elle atteint ses bornes sur K, d'où l'existence de $\bar{\boldsymbol{x}}$.
2. Soit $x_0 \in K$. Si f est croissante à l'infini, alors il existe $R > 0$ tel que si $\|x - x_0\| > R$ alors $f(x) > f(x_0)$; donc $\inf_K f = \inf_{K \cap B(x_0,R)} f$, où $B(x_0, R)$ désigne la boule (fermé) de centre x_0 et de rayon R. L'ensemble $K \cap B(x_0, R)$ est compact, car intersection d'un fermé et d'un compact. Donc, par ce qui précède, il existe $\bar{\boldsymbol{x}} \in K$ tel que $f(\bar{\boldsymbol{x}}) = \inf_{K \cap B(x_0,R)} f = \inf_K f$.

■

Théorème 3.29 (Unicité)**.** *Soit $E = \mathbb{R}^n$ et $f \in C(E, \mathbb{R})$. On suppose que f est strictement convexe et que K est convexe. Alors il existe au plus un élément $\bar{\boldsymbol{x}}$ de K tel que $f(\bar{\boldsymbol{x}}) = \inf_K f$.*

DÉMONSTRATION – Supposons que $\bar{\boldsymbol{x}}$ et $\bar{\bar{\boldsymbol{x}}}$ soient deux solutions du problème (3.48), avec $\bar{\boldsymbol{x}} \neq \bar{\bar{\boldsymbol{x}}}$
Alors $f(\frac{1}{2}\bar{\boldsymbol{x}} + \frac{1}{2}\bar{\bar{\boldsymbol{x}}}) < \frac{1}{2}f(\bar{\boldsymbol{x}}) + \frac{1}{2}f(\bar{\bar{\boldsymbol{x}}}) = \inf_K f$. On aboutit donc à une contradiction. ■

Des théorèmes d'existence 3.28 et d'unicité 3.29 on déduit immédiatement le théorème d'existence et d'unicité suivant :

Théorème 3.30 (Existence et unicité)**.** *Soient $E = \mathbb{R}^n$, $f \in C(E, \mathbb{R}^n)$ une fonction strictement convexe et K un sous ensemble convexe fermé de E. Si K est borné ou si f est croissante à l'infini, c'est–à–dire si $f(x) \to +\infty$ quand $\|x\| \to +\infty$, alors il existe un unique élément $\bar{\boldsymbol{x}}$ de K solution du problème de minimisation* (3.48)*, i.e. tel que $f(\bar{\boldsymbol{x}}) = \inf_K f$*

Remarque 3.31. *On peut remplacer $E = \mathbb{R}^n$ par E espace de Hilbert de dimension infinie dans le dernier théorème, mais on a besoin dans ce cas de l'hypothèse de convexité de f pour assurer l'existence de la solution (voir cours de maîtrise).*

Proposition 3.32 (Condition simple d'optimalité)**.** *Soient $E = \mathbb{R}^n$, $f \in C(E, \mathbb{R})$ et $\bar{\boldsymbol{x}} \in K$ tel que $f(\bar{\boldsymbol{x}}) = \inf_K f$. On suppose que f est différentiable en $\bar{\boldsymbol{x}}$*

1. *Si $\bar{\boldsymbol{x}} \in \overset{\circ}{K}$ alors $\nabla f(\bar{\boldsymbol{x}}) = 0$.*
2. *Si K est convexe, alors $\nabla f(\bar{\boldsymbol{x}}) \cdot (x - \bar{\boldsymbol{x}}) \geq 0$ pour tout $x \in K$.*

DÉMONSTRATION –

1. Si $\bar{\boldsymbol{x}} \in \overset{\circ}{K}$, alors il existe $\varepsilon > 0$ tel que $B(\bar{\boldsymbol{x}}, \varepsilon) \subset K$ et $f(\bar{\boldsymbol{x}}) \leq f(x)\ \forall x \in B(\bar{\boldsymbol{x}}, \varepsilon)$. Alors on a déjà vu (voir preuve de la Proposition 3.7 page 194) que ceci implique $\nabla f(\bar{\boldsymbol{x}}) = 0$.
2. Soit $x \in K$. Comme $\bar{\boldsymbol{x}}$ réalise le minimum de f sur K, on a : $f(\bar{\boldsymbol{x}} + t(x - \bar{\boldsymbol{x}})) = f(tx + (1-t)\bar{\boldsymbol{x}}) \geq f(\bar{\boldsymbol{x}})$ pour tout $t \in]0, 1]$, par convexité de K. On en déduit que
$$\frac{f(\bar{\boldsymbol{x}} + t(x - \bar{\boldsymbol{x}})) - f(\bar{\boldsymbol{x}})}{t} \geq 0 \text{ pour tout } t \in]0, 1].$$
En passant à la limite lorsque t tend vers 0 dans cette dernière inégalité, on obtient : $\nabla f(\bar{\boldsymbol{x}}) \cdot (x - \bar{\boldsymbol{x}}) \geq 0$.

■

3.4.3 Conditions d'optimalité dans le cas de contraintes égalité

Dans tout ce paragraphe, on considèrera les hypothèses et notations suivantes :

$$\begin{array}{l} f \in C(\mathbb{R}^n, \mathbb{R}), \quad g_i \in C^1(\mathbb{R}^n, \mathbb{R}), \ \ i = 1 \dots p\,; \\ K = \{u \in \mathbb{R}^n, \ \ g_i(u) = 0 \ \ \forall i = 1 \dots p\}\,; \\ g = (g_1, \dots, g_p)^t \in C^1(\mathbb{R}^n, \mathbb{R}^p) \end{array} \tag{3.49}$$

Remarque 3.33 (Quelques rappels de calcul différentiel)**.**
Comme $g \in C^1(\mathbb{R}^n, \mathbb{R}^p)$, si $u \in \mathbb{R}^n$, alors $Dg(u) \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^p)$, ce qui revient à dire, en confondant l'application linéaire $Dg(u)$ avec sa matrice, que $Dg(u) \in \mathcal{M}_{p,n}(\mathbb{R})$. Par définition, $Im(Dg(u)) = \{Dg(u)z, \ \ z \in \mathbb{R}^n\} \subset \mathbb{R}^p$, et $\text{rang}(Dg(u)) = \dim(Im(Dg(u))) \leq p$. On rappelle de plus que

$$Dg(u) = \begin{pmatrix} \dfrac{\partial g_1}{\partial x_1}, & \dots, & \dfrac{\partial g_1}{\partial x_n} \\ \dots, & \ddots, & \dots \\ \dfrac{\partial g_p}{\partial x_1}, & \dots, & \dfrac{\partial g_p}{\partial x_n} \end{pmatrix},$$

et que $\text{rang}(Dg(u)) \leq \min(n, p)$. De plus, si $\text{rang}(Dg(u)) = p$, alors les vecteurs $(Dg_i(u))_{i=1\dots p}$ sont linéairement indépendants dans $\mathbb{R}^n$.

Théorème 3.34 (Multiplicateurs de Lagrange)**.** *Soit $\bar{u} \in K$ tel que $f(\bar{u}) = \inf_K f$. On suppose que f est différentiable en $\bar{u}$ et $\dim(Im(Dg(\bar{u})) = p$ (ou rang$(Dg(\bar{u})) = p$), alors :*

$$\textit{il existe } (\lambda_1, \dots, \lambda_p)^t \in \mathbb{R}^p \textit{ tels que } \nabla f(\bar{u}) + \sum_{i=1}^{p} \lambda_i \nabla g_i(\bar{u}) = 0.$$

(Cette dernière égalité a lieu dans $\mathbb{R}^n$)

DÉMONSTRATION – Pour plus de clarté, donnons d'abord une idée "géométrique" de la démonstration dans le cas $n = 2$ et $p = 1$. On a dans ce cas $f \in C^1(\mathbb{R}^2, \mathbb{R})$ et $K = \{(x, y) \in \mathbb{R}^2 g(x, y) = 0\}$, et on cherche $u \in K$ tel que $f(u) = \inf_K f$.
Traçons dans le repère (x, y) la courbe $g(x, y) = 0$, ainsi que les courbes de niveau de f. Si on se "promène" sur la courbe $g(x, y) = 0$, en partant du point P_0 vers la droite (voir figure 3.1), on rencontre les courbes de niveau successives de f et on se rend compte sur le dessin que la valeur minimale que prend f sur la courbe $g(x, y) = 0$ est atteinte lorsque cette courbe est tangente à la courbe de niveau de f : sur le dessin, ceci correspond au point P_1 où la courbe $g(x, y) = 0$ est tangente à la courbe $f(x, y) = 3$. Une fois qu'on a passé ce point de tangence, on peut remarquer que f augmente.
On utilise alors le fait que si φ est une fonction continûment différentiable de $\mathbb{R}^2$ dans $\mathbb{R}$, le gradient de φ est orthogonal à toute courbe de niveau de φ, c'est-à-dire toute courbe de la forme $\varphi(x, y) = c$, où $c \in \mathbb{R}$. (En effet, soit $(x(t), y(t))$, $t \in \mathbb{R}$ un paramétrage de la courbe $g(x, y) = c$, en dérivant par rapport à t, on obtient : $\nabla g(x(t), y(t)) \cdot (x'(t), y'(t))^t = 0$). En appliquant ceci à f et g, on en déduit qu'au point de tangence entre une courbe de niveau de f et la courbe $g(x, y) = 0$, les gradients de f et g sont colinéaires. Et donc si $\nabla g(u) \neq 0$, il existe $\lambda \neq 0$ tel que $\nabla f(u) = \lambda \nabla g(u)$.

Passons maintenant à la démonstration rigoureuse du théorème dans laquelle on utilise le théorème des fonctions implicites [5].
Par hypothèse, $Dg(\bar{u}) \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^p)$ et $Im(Dg(\bar{u})) = \mathbb{R}^p$. Donc il existe un sous espace vectoriel F de $\mathbb{R}^n$ de dimension p, tel que $Dg(\bar{u})$ soit bijective de F dans $\mathbb{R}^p$. En effet, soit $(e_1 \dots e_p)$ la base canonique de $\mathbb{R}^p$, alors pour tout $i \in \{1, \dots p\}$, il existe $y_i \in \mathbb{R}^n$ tel que $Dg(\bar{x})y_i = e_i$. Soit F le sous espace engendré par la famille $\{y_1 \dots y_p\}$; on

5. **Théorème des fonctions implicites** Soient p et q des entiers naturels, soit $h \in C^1(\mathbb{R}^q \times \mathbb{R}^p, \mathbb{R}^p)$, et soient $(\bar{x}, \bar{y}) \in \mathbb{R}^q \times \mathbb{R}^p$ et $c \in \mathbb{R}^p$ tels que $h(\bar{x}, \bar{y}) = c$. On suppose que la matrice de la différentielle $D_2h(\bar{x}, \bar{y})(\in \mathcal{M}_p(\mathbb{R}))$ est inversible. Alors il existe $\varepsilon > 0$ et $\nu > 0$ tels que pour tout $x \in B(\bar{x}, \varepsilon)$, il existe un unique $y \in B(\bar{y}, \nu)$ tel que $h(x, y) = c$. on peut ainsi définir une application ϕ de $B(\bar{x}, \varepsilon)$ dans $B(\bar{y}, \nu)$ par $\phi(x) = y$. On a $\phi(\bar{x}) = \bar{y}$, $\phi \in C^1(\mathbb{R}^p, \mathbb{R}^p)$ et $D\phi(x) = -[D_2h(x, \phi(x))]^{-1} \cdot D_1h(x, \phi(x))$.

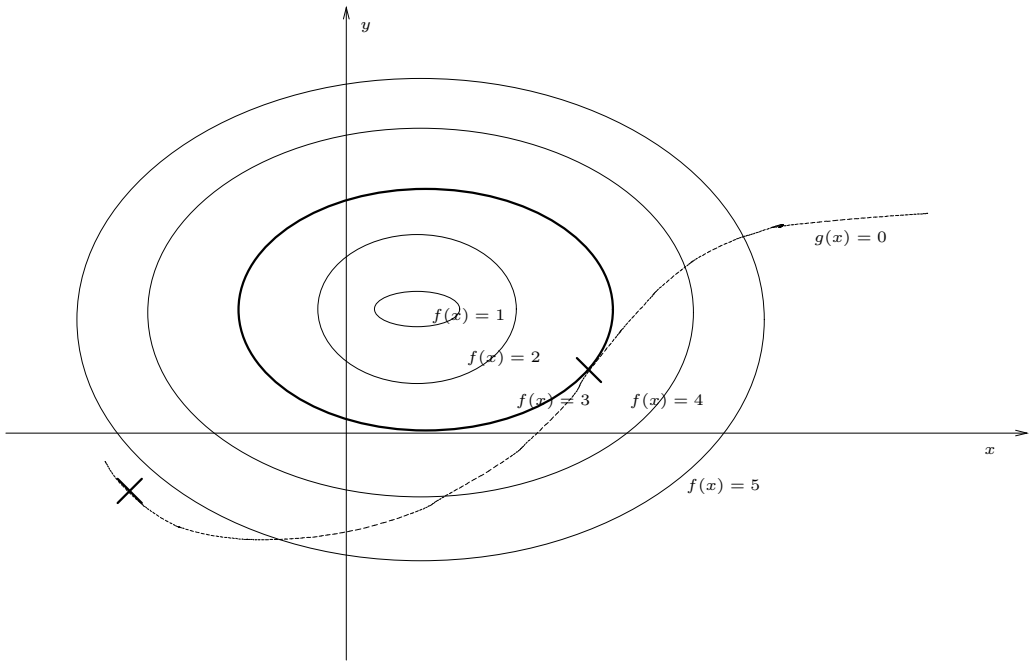


FIGURE 3.1: Interprétation géométrique des multiplicateurs de Lagrange

remarque que cette famille est libre, car si $\sum_{i=1}^{p} \lambda_i y_i = 0$, alors $\sum_{i=1}^{p} \lambda_i e_i = 0$, et donc $\lambda_i = 0$ pour tout $i = 1, \dots p$. On a ainsi montré l'existence d'un sous espace F de dimension p telle que $Dg(\bar{x})$ soit bijective (car surjective) de F dans $\mathbb{R}^p$.

Il existe un sous espace vectoriel G de $\mathbb{R}^n$, tel que $\mathbb{R}^n = F \bigoplus G$. Pour $v \in F$ et $w \in G$; on pose $\bar{g}(w, v) = g(v + w)$ et $\bar{f}(w, v) = f(v + w)$. On a donc $\bar{f} \in C(G \times F,\ \mathbb{R})$ et $\bar{g} \in C^1(G \times F,\ \mathbb{R})$. De plus, $D_2\bar{g}(w, v) \in \mathcal{L}(F,\ \mathbb{R}^p)$, et pour tout $z \in F$, on a $D_2\bar{g}(w, v)z = Dg(v + w)z$.

Soit $(\bar{v}, \bar{w}) \in F \times G$ tel que $\bar{u} = \bar{v} + \bar{w}$. Alors $D_2\bar{g}(\bar{w}, \bar{v})z = Dg(\bar{u})z$ pour tout $z \in F$. L'application $D_2\bar{g}(\bar{w}, \bar{v})$ est une bijection de F sur $\mathbb{R}^p$, car, par définition de F, $Dg(\bar{u})$ est bijective de F sur $\mathbb{R}^p$.

On rappelle que $K = \{u \in \mathbb{R}^n : g(u) = 0\}$ et on définit $\overline{K} = \{(w, v) \in G \times F,\ \ \bar{g}(w, v) = 0\}$. Par définition de $\bar{f}$ et de $\bar{g}$, on a

$$\begin{cases} (\bar{w}, \bar{v}) \in \overline{K} \\ \bar{f}(\bar{w}, \bar{v}) \leq f(w, v)\ \ \forall (w, v) \in \overline{K} \end{cases} \tag{3.50}$$

D'autre part, le théorème des fonctions implicites (voir note de bas de page 237) entraîne l'existence de $\varepsilon > 0$ et $\nu > 0$ tels que pour tout $w \in B_G(\bar{w}, \varepsilon)$ il existe un unique $v \in B_F(\bar{v}, \nu)$ tel que $\bar{g}(w, v) = 0$. On note $v = \phi(w)$ et on définit ainsi une application $\phi \in C^1(B_G(\bar{w}, \varepsilon), B_F(\bar{v}, \nu))$.

On déduit alors de (3.50) que :

$$\bar{f}(\bar{w}, \phi(\bar{w})) \leq \bar{f}(w, \phi(w))),\ \forall w \in B_G(\bar{w}, \varepsilon),$$

et donc

$$f(\bar{u}) = f(\bar{w} + \phi(\bar{w})) \leq f(w + \phi(w)),\ \forall w \in B_G(\bar{w}, \varepsilon).$$

En posant $\psi(w) = \bar{f}(w, \phi(w))$, on peut donc écrire

$$\psi(\bar{w}) = \bar{f}(\bar{w}, \phi(\bar{w})) \leq \psi(w),\ \forall w \in B_G(\bar{w}, \varepsilon).$$

On a donc, grâce à la proposition 3.32,

$$D\psi(\bar{w}) = 0. \tag{3.51}$$

Par définition de ψ, de $\bar{f}$ et de $\bar{g}$, on a :

$$D\psi(\bar{w}) = D_1\bar{f}(\bar{w}, \phi((\bar{w})) + D_2\bar{f}(\bar{w}, \phi(\bar{w}))D\phi(\bar{w}).$$

D'après le théorème des fonctions implicites,

$$D\phi(\bar{w}) = -[D_2\bar{g}(\bar{w}, \phi((\bar{w}))]^{-1} D_1\bar{g}(\bar{w}, \phi((\bar{w})).$$

On déduit donc de (3.51) que

$$D_1\bar{f}(\bar{w},\phi((\bar{w}))w - [D_2\bar{g}(\bar{w},\phi((\bar{w}))]^{-1}D_1\bar{g}(\bar{w},\phi((\bar{w}))w = 0, \text{ pour tout } w \in G. \tag{3.52}$$

De plus, comme $D_2\bar{g}(\bar{w},\phi((\bar{w}))]^{-1}D_2\bar{g}(\bar{w},\phi((\bar{w})) = \mathrm{Id}$, on a :

$$D_2\bar{f}(\bar{w},\phi((\bar{w}))z - D_2\bar{f}(\bar{w},\phi((\bar{w}))\left[D_2\bar{g}(\bar{w},\phi((\bar{w}))\right]^{-1}D_2\bar{g}(\bar{w},\phi((\bar{w}))z = 0,\ \forall z \in F. \tag{3.53}$$

Soit $x \in \mathbb{R}^n$, et $(z,w) \in F \times G$ tel que $x = z + w$. En additionant (3.52) et (3.53), et en notant

$$\lambda = -D_2\bar{f}(\bar{w},\phi((\bar{w}))\left[D_2\bar{g}(\bar{w},\phi((\bar{w}))\right]^{-1},$$

on obtient :

$$Df(\bar{u})x + \lambda Dg(\bar{u})x = 0,$$

ce qui donne, en transposant : $\nabla f(\bar{u}) + \sum_{i=1}^{p}\lambda_i\nabla g_i(\bar{u}) = 0$, avec $\lambda = (\lambda_1,\ldots,\lambda_p)$.

■

Remarque 3.35 (Utilisation pratique du théorème de Lagrange)**.** Soit $f \in C^1(\mathbb{R}^n,\mathbb{R})$, $g = (g_1,\ldots,g_p)^t$ avec $g_i \in C(\mathbb{R}^n,\mathbb{R})$ pour $i = 1,\ldots,p.$, et soit $K = \{u \in \mathbb{R}^n,\ \ g_i(u) = 0,\ \ i = 1,\ldots,p\}$.
Le problème qu'on cherche à résoudre est le problème de minimisation (3.48) qu'on rappelle ici :

$$\begin{cases} \bar{u} \in K \\ f(\bar{u}) = \inf\limits_{K} f \end{cases}$$

D'après le théorème des multiplicateurs de Lagrange, si $\bar{u}$ est solution de (3.48) et $Im(Dg(\bar{u})) = \mathbb{R}^p$, alors il existe $(\lambda_1,\ldots,\lambda_p) \in \mathbb{R}^p$ tel que $\bar{u}$ est solution du problème

$$\begin{cases} \dfrac{\partial f}{\partial x_j}(\bar{u}) + \displaystyle\sum_{i=1}^{p}\lambda_i\frac{\partial g_i}{\partial x_j} = 0, j = 1,\ldots n, \\ g_i(\bar{u}) = 0,\ \ i = 1,\ldots,p. \end{cases} \tag{3.54}$$

Le système (3.54) est un système non linéaire de de $(n+p)$ équations et à $(n+p)$ inconnues $(\bar{x},\ldots,\bar{x}_n,\lambda_i\ldots\lambda_p)$. Ce système sera résolu par une méthode de résolution de système non linéaire (Newton par exemple).

Remarque 3.36. *On vient de montrer que si $\bar{x}$ solution de* (3.48) *et $Im(Dg(\bar{x})) = \mathbb{R}^p$, alors $\bar{x}$ solution de* (3.54)*. Par contre, si $\bar{x}$ est solution de* (3.54)*, ceci n'entraîne pas que $\bar{x}$ est solution de* (3.48)*.*

Des exemples d'application du théorème des multiplicateurs de Lagrange sont donnés dans les exercices 155 page 240 et 156 page 241.

3.4.4 Contraintes inégalités

Soit $f \in C(\mathbb{R}^n,\mathbb{R})$ et $g_i \in C^1(\mathbb{R}^n,\mathbb{R})\ \ i = 1,\ldots,p$, on considère maintenant un ensemble K de la forme : $K = \{x \in \mathbb{R}^n,\ \ g_i(x) \leq 0\ \ \forall i = 1\ldots p\}$, et on cherche à résoudre le problème de minimisation (3.48) qui sécrit :

$$\begin{cases} \bar{x} \in K \\ f(\bar{x}) \leq f(x), \quad \forall x \in K. \end{cases}$$

Remarque 3.37. *Soit $\bar{x}$ une solution de* (3.48) *et supposons que $g_i(\bar{x}) < 0$, pour tout $i \in \{1,\ldots,p\}$. Il existe alors $\varepsilon > 0$ tel que si $x \in B(\bar{x},\varepsilon)$ alors $g_i(x) < 0$ pour tout $i = 1,\ldots,p$.*
On a donc $f(\bar{x}) \leq f(x)\ \ \forall x \in B(\bar{x},\varepsilon)$. On est alors ramené à un problème de minimisation sans contrainte, et si f est différentiable en $\bar{x}$, on a donc $\nabla f(\bar{x}) = 0$.

On donne maintenant sans démonstration le théorème de Kuhn-Tücker qui donne une caractérisation de la solution du problème (3.48).

Théorème 3.38 (Kuhn–Tucker)**.** *Soit $f \in C(\mathbb{R}^n, \mathbb{R})$, soit $g_i \in C^1(\mathbb{R}^n, \mathbb{R})$, pour $i = 1, \dots, p$, et soit $K = \{\boldsymbol{x} \in \mathbb{R}^n,\ g_i(\boldsymbol{x}) \leq 0\ \forall i = 1 \dots p\}$. On suppose qu'il existe $\bar{x}$ solution de* (3.48)*, et on pose $I(\bar{\boldsymbol{x}}) = \{i \in \{1, \dots, p\}; |g_i(\bar{\boldsymbol{x}}) = 0\}$. On suppose que f est différentiable en $\bar{\boldsymbol{x}}$ et que la famille (de $\mathbb{R}^n$) $\{\nabla g_i(\bar{\boldsymbol{x}}), i \in I(\bar{\boldsymbol{x}})\}$ est libre. Alors il existe une famille $(\lambda_i)_{i \in I(\bar{\boldsymbol{x}})} \subset \mathbb{R}_+$ telle que*

$$\nabla f(\bar{\boldsymbol{x}}) + \sum_{i \in I(\bar{\boldsymbol{x}})} \lambda_i \nabla g_i(\bar{\boldsymbol{x}}) = 0.$$

Remarque 3.39.

1. Le théorème de Kuhn-Tucker s'applique pour des ensembles de contrainte de type inégalité. Si on a une contraite de type égalité, on peut évidemment se ramener à deux contraintes de type inégalité en remarquant que $\{h(x) = 0\} = \{h(x) \leq 0)\} \cap \{-h(x) \leq 0\}$. Cependant, si on pose $g_1 = h$ et $g_2 = -h$, on remarque que la famille $\{\nabla g_1(\bar{x}), \nabla g_2(\bar{x})\} = \{\nabla h(\bar{x}), -\nabla h(\bar{x})\}$ n'est pas libre. On ne peut donc pas appliquer le théorème de Kuhn-Tucker sous la forme donnée précédemment dans ce cas (mais on peut il existe des versions du théorème de Kuhn-Tucker permettant de traiter ce cas, voir Bonans-Saguez).
2. Dans la pratique, on a intérêt à écrire la conclusion du théorème de Kuhn-Tucker (i.e. l'existence de la famille $(\lambda_i)_{i \in I(\bar{x})}$) sous la forme du système de $n + p$ équations et $2p$ inéquations à résoudre suivant :

$$\begin{cases} \nabla f(\bar{x}) + \displaystyle\sum_{i=1}^{p} \lambda_i \nabla g_i(\bar{x}) = 0, \\ \lambda_i g_i(\bar{x}) = 0, \quad \forall i = 1, \dots, p, \\ g_i(\bar{x}) \leq 0, \quad \forall i = 1, \dots, p, \\ \lambda_i \geq 0, \quad \forall i = 1, \dots, p. \end{cases}$$

3.4.5 Exercices (optimisation avec contraintes)

Exercice 154 (Sur l'existence et l'unicité)**.** *Corrigé en page 243*
Etudier l'existence et l'unicité des solutions du problème (3.48), avec les données suivantes : $E = \mathbb{R}$, $f : \mathbb{R} \to \mathbb{R}$ est définie par $f(x) = x^2$, et pour les quatre différents ensembles K suivants :

$$\begin{array}{llll} (i) & K = \{|x| \leq 1\}\,; & (ii) & K = \{|x| = 1\} \\ (iii) & K = \{|x| \geq 1\}\,; & (iv) & K = \{|x| > 1\}. \end{array} \tag{3.55}$$

Exercice 155 (Aire maximale d'un rectangle à périmètre donné)**.** *Corrigé en page 244*

1. On cherche à maximiser l'aire d'un rectangle de périmètre donné égal à 2. Montrer que ce problème peut se formuler comme un problème de minimisation de la forme (3.48), où K est de la forme $K = \{x \in \mathbb{R}^2; g(x) = 0\}$. On donnera f et g de manière explicite.
2. Montrer que le problème de minimisation ainsi obtenu est équivalent au problème

$$\begin{cases} \bar{x} = (\bar{x}_1, \bar{x}_2)^t \in \tilde{K} \\ f(\bar{x}_1, \bar{x}_2) \leq f(x_1, x_2), \quad \forall\, (x_1, x_2)^t \in \tilde{K}, \end{cases} \tag{3.56}$$

   où $\tilde{K} = K \cap [0, 1]^2$, K et f étant obtenus à la question 1. En déduire que le problème de minimisation de l'aire admet au moins une solution.
3. Calculer $Dg(x)$ pour $x \in K$ et en déduire que si x est solution de (3.56) alors $x = (1/2, 1/2)$. En déduire que le problème (3.56) admet une unique solution donnée par $\bar{x} = (1/2, 1/2)$.

Exercice 156 (Fonctionnelle quadratique)**.** *Suggestions en page 223, corrigé en page 244*
Soit f une fonction quadratique, *i.e.* $f(x) = \frac{1}{2}Ax \cdot x - b \cdot x$, où $A \in \mathcal{M}_n(\mathbb{R})$ est une matrice symétrique définie positive et $b \in \mathbb{R}^n$. On suppose que la contrainte g est une fonction linéaire de $\mathbb{R}^n$ dans $\mathbb{R}$, c'est-à-dire $g(x) = d \cdot x - c$ où $c \in \mathbb{R}$ et $d \in \mathbb{R}^n$, et que $d \neq 0$. On pose $K = \{x \in \mathbb{R}^n,\ g(x) = 0\}$ et on cherche à résoudre le problème de minimisation (3.48).

1. Montrer que l'ensemble K est non vide, fermé et convexe. En déduire que le problème (3.48) admet une unique solution.
2. Montrer que si $\bar{x}$ est solution de (3.48), alors il existe $\lambda \in \mathbb{R}$ tel que $y = (\bar{x}, \lambda)^t$ soit l'unique solution du système :
$$\left[\begin{array}{c|c} A & d \\ \hline d^t & 0 \end{array}\right] \left[\begin{array}{c} \bar{x} \\ \hline \lambda \end{array}\right] = \left[\begin{array}{c} b \\ \hline c \end{array}\right] \tag{3.57}$$

Exercice 157 (Boites de chocolats)**.** Un chocolatier veut concevoir une boîte cylindrique en carton, ouverte sur le dessus. Le coût du carton est proportionnel à la surface utilisée. Le carton doré qui constitue le cylindre est deux fois plus cher que le carton dont est fait le fond. Le rayon du cylindre est r et sa hauteur h. La boîte doit contenir un volume V de chocolats. Le chocolatier souhaite choisir r et h afin de minimiser le coût du carton.

1. Écrire le problème d'optimisation associé.
2. Résoudre le problème sans utiliser les multiplicateurs de Lagrange.
3. Résoudre le problème avec la méthode des multiplicateurs de Lagrange.

Exercice 158 (Minimisation sans dérivabilité)**.**
Soient $A \in \mathcal{M}_n(\mathbb{R})$ une matrice s.d.p., $b \in \mathbb{R}^n$, $j : \mathbb{R}^n \to \mathbb{R}$ une fonction continue et convexe, à valeurs positives ou nulles (mais non nécessairement dérivable, par exemple $j(v) = \sum_{j=1}^n \alpha_i |v_i|$, avec $\alpha_i \geq 0$ pour tout $i \in \{1, \ldots, n\}$). Soit U une partie non vide, fermée convexe de $\mathbb{R}^n$. Pour $v \in \mathbb{R}^n$, on pose $J(v) = (1/2)Av \cdot v - b \cdot v + j(v)$.

1. Montrer qu'il existe un et un seul u tel que :
$$u \in U,\ J(u) \leq J(v),\ \forall v \in U. \tag{3.58}$$
2. Soit $u \in U$, montrer que u est solution de (3.58) si et seulement si $(Au - b) \cdot (v - u) + j(v) - j(u) \geq 0$, pour tout $v \in U$.

Exercice 159 (Utilisation du théorème de Lagrange)**.**

1. Pour $(x, y) \in \mathbb{R}^2$, on pose : $f(x, y) = -y$, $g(x, y) = x^2 + y^2 - 1$. Chercher le(s) point(s) où f atteint son maximum ou son minimum sous la contrainte $g = 0$.
2. Soit $\boldsymbol{a} = (a_1, \ldots, a_n) \in \mathbb{R}^n$, $a \neq 0$; pour $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$, on pose : $f(\boldsymbol{x}) = \sum_{i=1}^n |x_i - a_i|^2$ et $g(\boldsymbol{x}) = \sum_{i=1}^n |x_i|^2$. Chercher le(s) point(s) où f atteint son maximum ou son minimum sous la contrainte $g = 1$.
3. Soient $A \in \mathcal{M}_n(\mathbb{R})$ symétrique, $B \in \mathcal{M}_n(\mathbb{R})$ s.d.p. et $\boldsymbol{b} \in \mathbb{R}^n$; pour $\boldsymbol{v} \in \mathbb{R}^n$, on pose
$$f(\boldsymbol{v}) = (1/2)A\boldsymbol{v} \cdot \boldsymbol{v} - \boldsymbol{b} \cdot \boldsymbol{v} \text{ et } g(\boldsymbol{v}) = B\boldsymbol{v} \cdot \boldsymbol{v}.$$
Peut-on appliquer le théorème de Lagrange et quelle condition donne-t-il sur $\boldsymbol{u}$ si $f(\boldsymbol{u}) = \min\{f(\boldsymbol{v}),\ \boldsymbol{v} \in K\}$ avec $K = \{\boldsymbol{v} \in \mathbb{R}^n ; g(\boldsymbol{v}) = 1\}$?

Exercice 160 (Distance d'un point à une ellipse)**.**
Soit $\boldsymbol{a} \in \mathbb{R}^2$. On note a_1, a_2 les deux composantes de $\boldsymbol{a}$.
Pour $\boldsymbol{x} = (x_1, x_2) \in \mathbb{R}^2$, on pose $J(\boldsymbol{x}) = |\boldsymbol{x} - \boldsymbol{a}|^2$, $f(\boldsymbol{x}) = x_1^2 + 2x_2^2 - 3$, et on définit $K = \{\boldsymbol{x} \in \mathbb{R}^2 ; f(\boldsymbol{x}) = 0\}$.
On s'intéresse au problème de la minimisation et de la maximisation de J sur K, c'est-à-dire aux deux problèmes suivants :
$$\boldsymbol{x} \in K,\ J(\boldsymbol{x}) \leq J(\boldsymbol{y}) \text{ pour tout } \boldsymbol{y} \in K. \tag{3.59}$$
$$\boldsymbol{x} \in K,\ J(\boldsymbol{x}) \geq J(\boldsymbol{y}) \text{ pour tout } \boldsymbol{y} \in K. \tag{3.60}$$

1. Représenter graphiquement l'ensemble K et expliquer en quoi l'un des problèmes d'optimisation ci-dessus revient à chercher la distance d'un point à une ellipse.
2. Montrer que les problèmes (3.59) et (3.60) admettent au moins une solution et que si $\boldsymbol{x}$ est solution de l'un de ces problèmes, il existe $\lambda \in \mathbb{R}$ tel que

$$\nabla J(\boldsymbol{x}) + \lambda \nabla f(\boldsymbol{x}) = 0, \tag{3.61}$$
$$f(\boldsymbol{x}) = 0. \tag{3.62}$$

Le système (non linéaire) (3.61)-(3.62) s'écrit $G(x_1, x_2, \lambda) = 0$. Soit (x_1, x_2, λ) une solution de (3.61)-(3.62). Dans les questions suivantes, on cherche à calculer (x_1, x_2, λ).

3. On suppose dans cette question que $a_1 \neq 0$, $a_2 \neq 0$ et $f(\boldsymbol{a}) > 0$.
   (a) Montrer que $\lambda \neq -1$ et $\lambda \neq -1/2$.
   (b) Calculer la matrice jacobienne $J_G(x_1, x_2, \lambda)$ et montrer que son déterminant est non nul. [*Il peut être opportun d'utiliser le fait que* $\lambda \notin [-1, -1/2]$.]
   (c) Ecrire la méthode de Newton pour calculer le point (x_1, x_2, λ). Y-a-t-il convergence de la suite donnée par la méthode de Newton vers (x_1, x_2, λ), et sous quelles conditions ?
4. On suppose dans cette question que $a_1 = 0$, $a_2 \neq 0$.
   (a) On suppose que $f(\boldsymbol{a}) \geq 0$ (et donc $|a_2| \geq \sqrt{3}/2$). Montrer que $x_1 = 0$ et $x_2 = \pm\sqrt{3}/2$.
   Quel est l'unique point solution de (3.59) et l'unique point solution de (3.60) ?
   (b) On suppose maintenant que $f(\boldsymbol{a}) < 0$. Montrer que le système (3.61)-(3.62) a quatre solutions : deux solutions correspondant à $x_1 = 0$ et deux solutions à $\lambda = -1$. Pour chacune de ces solutions, indiquer, sans démonstration, si elle est solution de (3.59), de (3.60) ou de aucun de ces problèmes. [Il est peut-être utile de faire un dessin.]

Exercice 161 (Calcul d'un coût marginal)**.**
Soient $\psi, F \in C^2(\mathbb{R}^2, \mathbb{R})$. On suppose que $F(\boldsymbol{x}) \to +\infty$ quand $|\boldsymbol{x}| \to +\infty$.
Pour $s \in \mathbb{R}$, on pose $K_s = \{\boldsymbol{x} \in \mathbb{R}^2; \psi(\boldsymbol{x}) = s\}$. On suppose $K_s \neq \emptyset$ pour tout s et on s'intéresse au problème

$$\boldsymbol{x} \in K_s, \tag{3.63}$$
$$F(\boldsymbol{x}) \leq F(\boldsymbol{y}) \text{ pour tout } \boldsymbol{y} \in K_s. \tag{3.64}$$

1. Montrer que pour tout $s \in \mathbb{R}$, le problème (3.63)-(3.64) admet au moins une solution.

Pour tout $s \in \mathbb{R}$, on note $\bar{\boldsymbol{x}}(s)$ une solution de (3.63)-(3.64) et $p(s) = F(\bar{\boldsymbol{x}}(s))$. On suppose que $\nabla\psi(\bar{\boldsymbol{x}}(0)) \neq 0$.

2. Montrer qu'il existe $\lambda \in \mathbb{R}$ tel que $\nabla F(\bar{\boldsymbol{x}}(0)) + \lambda \nabla\psi(\bar{\boldsymbol{x}}(0)) = 0$.
3. Soit $s \in \mathbb{R}$. On pose $\boldsymbol{z}(s) = \bar{\boldsymbol{x}}(s) - \bar{\boldsymbol{x}}(0)$. Montrer que

$$p(s) - p(0) = -\lambda s + \int_0^1 (1-t)\big(\lambda H_\psi(\bar{\boldsymbol{x}}(0) + t\boldsymbol{z}(s))\boldsymbol{z}(s) \cdot \boldsymbol{z}(s) + H_F(\bar{\boldsymbol{x}}(0) + t\boldsymbol{z}(s))\boldsymbol{z}(s) \cdot \boldsymbol{z}(s)\big)dt,$$

   où λ est donné par la question 2.
4. (Question indépendante de la précédente et de la suivante) On suppose dans cette question que F et ψ sont de classe C^4. Soit λ donné par la question 2. Pour $\boldsymbol{x} \in \mathbb{R}^2$, on pose $G(\boldsymbol{x}) = F(\boldsymbol{x}) + \lambda\psi(\boldsymbol{x})$, et on suppose que la matrice hessienne de G est s.d.p au point $\bar{\boldsymbol{x}}(0)$ (c'est-à-dire $H_G(\bar{\boldsymbol{x}}(0))$ s.d.p.).
   La question 2 donne que le couple $(\bar{\boldsymbol{x}}(0), \lambda)$ est solution du système de 3 équations à 3 inconnues (notées $\boldsymbol{x}$, μ avec $\boldsymbol{x} \in \mathbb{R}^2$, $\mu \in \mathbb{R}$) :

$$\nabla F(\boldsymbol{x}) + \mu \nabla\psi(\boldsymbol{x}) = 0, \tag{3.65}$$
$$\psi(\boldsymbol{x}) = 0. \tag{3.66}$$

   Montrer que l'algorithme de Newton pour calculer une solution de (3.65)-(3.66) converge vers $(\bar{\boldsymbol{x}}(0), \lambda)$ à condition que l'algorithme soit initialisé avec un point suffisamment proche de $(\bar{\boldsymbol{x}}(0), \lambda)$.
   La condition "$H_G(\bar{\boldsymbol{x}}(0))$ s.d.p." est elle assurée si F est strictement convexe et ψ est affine ?

5. (Exemple) Dans cette question, on note x_1 et x_2 les deux composantes de $\boldsymbol{x} \in \mathbb{R}^2$ et on choisit $F(\boldsymbol{x}) = 2(x_1+1)^2 + (x_2+2)^2$, $\psi(\boldsymbol{x}) = x_1 + x_2 - a$, où $a > 0$ est donné.
   (a) Montrer que le problème (3.63)-(3.64) admet une unique solution (notée $\bar{\boldsymbol{x}}(s)$) pour tout $s \in \mathbb{R}$, sans la calculer.
   Montrer que $\nabla\psi(\bar{\boldsymbol{x}}(0)) \neq 0$.
   (b) Calculer $\bar{\boldsymbol{x}}(0)$.
   (c) Comparer $\lim_{s\to 0} \dfrac{p(s)-p(0)}{s}$, $\lim_{s\to 0} \dfrac{F(a+s,0)-F(a,0)}{s}$ et $\lim_{s\to 0} \dfrac{F(0,a+s)-F(0,a)}{s}$.
   (d) Cette application est une modélisation très simplifiée du problème suivant : une entreprise possède 2 usines. La solution du modèle lui indique comment répartir (de manière optimale) cette production entre les deux usines. Puis, le modèle lui indique le coût d'une "petite" augmentation de production (c'est le coût marginal). Essayez d'indiquer comment (3.63)-(3.64) modélise ce problème et de donner le sens (pour l'entreprise) du multiplicateur de Lagrange (noté ici λ).

Exercice 162 (Contre exemple aux multiplicateurs de Lagrange)**.**
Soient f et $g : \mathbb{R}^2 \to \mathbb{R}$, définies par : $f(x,y) = y$, et $g(x,y) = y^3 - x^2$. On pose $K = \{(x,y) \in \mathbb{R}^2; g(x,y) = 0\}$.

1. Calculer le minimum de f sur K et le point $(\overline{x},\overline{y})$ où ce minimum est atteint.
2. Existe-t-il λ tel que $Df(\overline{x},\overline{y}) = \lambda Dg(\overline{x},\overline{y})$?
3. Pourquoi ne peut-on pas appliquer le théorème des multiplicateurs de Lagrange ?
4. Que trouve-t-on lorsqu'on applique la méthode dite "de Lagrange" pour trouver $(\overline{x},\overline{y})$?

Exercice 163 (Application simple du théorème de Kuhn-Tucker)**.** *Corrigé en page 245*
Soit f la fonction définie de $E = \mathbb{R}^2$ dans $\mathbb{R}$ par $f(x) = x^2 + y^2$ et $K = \{(x,y) \in \mathbb{R}^2; x + y \geq 1\}$. Justifier l'existence et l'unicité de la solution du problème (3.48) et appliquer le théorème de Kuhn-Tucker pour la détermination de cette solution.

Exercice 164 (Exemple d'opérateur de projection)**.** *Correction en page 245*

1. Soit $K = C^+ = \{x \in \mathbb{R}^n,\ x = (x_1,\ldots,x_k)^t,\ x_i \geq 0,\ \forall i = 1,\ldots,N\}$.
   (a) Montrer que K est un convexe fermé non vide.
   (b) Montrer que pour tout $y \in \mathbb{R}^n$, on a : $(p_K(y))_i = \max(y_i, 0)$.
2. Soit $(\alpha_i)_{i=1,\ldots,n} \subset \mathbb{R}^n$ et $(\beta_i)_{i=1,\ldots,n} \subset \mathbb{R}^n$ tels que $\alpha_i \leq \beta_i$ pour tout $i = 1,\ldots,n$. Soit $K = \{x = (x_1,\ldots,x_n)^t; \alpha_i \leq \beta_i, i = 1,\ldots,n\}$.
   (a) Montrer que K est un convexe fermé non vide.
   (b) Soit p_K l'opérateur de projection définie à la proposition 3.40 page 246. Montrer que pour tout $y \in \mathbb{R}^n$, on a :
   $$(p_K(y))_i = \max(\alpha_i, \min(y_i, \beta_i)), \quad \forall i = 1,\ldots,n.$$

Corrigés des exercices d'optimisation sous contraintes

Exercice 154 page 240 (Sur l'existence et l'unicité)

La fonction $f : \mathbb{R} \to \mathbb{R}$ définie par $f(x) = x^2$ est continue, strictement convexe, et croissante à l'infini. Etudions maintenant les propriétés de K dans les quatre cas proposés :

(i) L'ensemble $K = \{|x| \leq 1\}$ est fermé borné et convexe. On peut donc appliquer le théorème d'existence et d'unicité 3.30 page 236. En remarquant que $f(x) \geq 0$ pour tout $x \in \mathbb{R}$ et que $f(0) = 0$, on en déduit que l'unique solution du problème (3.48) est donc $\bar{x} = 0$.

(ii) L'ensemble $K = \{|x| = 1\}$ est fermé borné mais non convexe. Le théorème d'existence 3.28 page 235 s'applique donc, mais pas le théorème d'unicité 3.29 page 236. De fait, on peut remarquer que $K = \{-1, 1\}$, et donc $\{f(x), x \in K\} = \{1\}$. Il existe donc deux solutions du problème (3.48) : $\bar{x}_1 = 1$ et $\bar{x}_1 = -1$.

(iii) L'ensemble $K = \{|x| \geq 1\}$ est fermé, non borné et non convexe. Cependant, on peut écrire $K = K_1 \cup K_2$ où $K_1 = [1, +\infty[$ et $K_2 =]-\infty, -1]$ sont des ensembles convexes fermés. On peut donc appliquer le théorème 3.30 page 236 : il existe un unique $\bar{x}_1 \in \mathbb{R}$ et un unique $\bar{x}_2 \in \mathbb{R}$ solution de (3.48) pour $K = K_1$ et $K = K_2$ respectivement. Il suffit ensuite de comparer $\bar{x}_1$ et $\bar{x}_2$. Comme $\bar{x}_1 = -1$ et $\bar{x}_2 = 1$, on a existence mais pas unicité.

(iv) L'ensemble $K = \{|x| > 1\}$ n'est pas fermé, donc le théorème 3.28 page 235 ne s'applique pas. De fait, il n'existe pas de solution dans ce cas, car on a $lim_{x \to 1^+} f(x) = 1$, et donc $\inf_K f = 1$, mais cet infimum n'est pas atteint.

Exercice 155 page 240 (Maximisation de l'aire d'un rectangle à périmètre donné)

1. On peut se ramener sans perte de généralité au cas du rectangle $[0, x_1] \times [0, x_2]$, dont l'aire est égale à $x_1 x_2$ et de périmètre $2(x_1 + x_2)$. On veut donc maximiser $x_1 x_2$, ou encore minimiser $-x_1 x_2$. Pour$x = (x_1, x_2)^t \in \mathbb{R}^2$, posons $f(x_1, x_2) = -x_1 x_2$ et $g(x_1, x_2) = x_1 + x_2$. Définissons

$$K = \left\{ x = (x_1, x_2)^t \in (\mathbb{R}_+)^2 \text{ tel que } x_1 + x_2 = 1 \right\}.$$

Le problème de minimisation de l'aire du rectangle de périmètre donné et égal à 2 s'écrit alors :

$$\left\{ \begin{array}{l} \left[\begin{array}{c} x_1 \\ x_2 \end{array} \right] \in K \\ f(\bar{x}_1, \bar{x}_2) \leq f(x_1, x_2) \quad \forall (x_1, x_2) \in K \end{array} \right. \tag{3.67}$$

2. Comme x_1 et x_2 sont tous deux positifs, puisque leur somme doit être égale à 1, ils sont forcément tous deux inférieurs à 1. Il est donc équivalent de résoudre (3.67) ou (3.56). L'ensemble $\tilde{K}$ est un convexe ferme borné, la fonction f est continue, et donc par le théorème 3.28 page 235, il existe au moins une solution du problème (3.56) (ou (3.67)).

3. Calculons ∇g : $\nabla g(x) = (1, 1)^t$, donc $\text{rang}(Dg(x, y)) = 1$. Par le théorème de Lagrange, si $x = (x_1, x_2)^t$ est solution de (3.67), il existe $\lambda \in \mathbb{R}$ tel que

$$\left\{ \begin{array}{l} \nabla f(\bar{x}, \bar{y}) + \lambda \nabla g(\bar{x}, \bar{y}) = 0, \\ \bar{x} + \bar{y} = 1. \end{array} \right.$$

Or $\nabla f(\bar{x}, \bar{y}) = (-\bar{x}, -\bar{y})^t$, et $\nabla g(\bar{x}, \bar{y}) = (1, 1)^t$. Le système précédent s'écrit donc :

$$\begin{aligned} -\bar{y} + \lambda &= 0 \\ -\bar{x} + \lambda &= 0 \\ \bar{x} + \bar{y} &= 1. \end{aligned}$$

On a donc

$$\bar{x} = \bar{y} = \frac{1}{2}.$$

Exercice 156 page 241 (Fonctionnelle quadratique)

1. Comme $d \neq 0$, il existe $\tilde{x} \in \mathbb{R}^n$ tel que $d \cdot \tilde{x} = \alpha \neq 0$. Soit $x = \frac{c}{\alpha} \tilde{x}$ alors $d \cdot x = c$. Donc l'ensemble K est non vide. L'ensemble K est fermé car noyau d'une forme linéaire continue de $\mathbb{R}^n$ dans $\mathbb{R}$, et K est évidemment convexe. La fonction f est strictement convexe et $f(x) \to +\infty$ quand $|x| \to +\infty$, et donc par les théorèmes 3.28 et 3.29 il existe un unique $\bar{x}$ solution de (3.48).

2. On veut calculer $\bar{x}$. On a : $Dg(x)z = d \cdot z$, et donc $Dg(x) = d^t$. Comme $d \neq 0$ on a rang$(Dg(x)) = 1$, ou encore $Im(Dg(x)) = \mathbb{R}$ pour tout x. Donc le théorème de Lagrange s'applique. Il existe donc $\lambda \in \mathbb{R}$ tel que $\nabla f(\bar{x}) + \lambda \nabla g(\bar{x}) = 0$, c'est-à-dire $A\bar{x} - b + \lambda d = 0$. Le couple $(\bar{x}, \lambda)$ est donc solution du problème suivant' :

$$\left\{ \begin{array}{l} A\bar{x} - b + \lambda d = 0, \\ d \cdot \bar{x} = c \end{array} \right. , \tag{3.68}$$

qui s'écrit sous forme matricielle : $By = e$, avec $B = \left[\begin{array}{c|c} A & d \\ \hline d^t & 0 \end{array} \right] \in \mathcal{M}_{n+1}(\mathbb{R})$, $y = \left[\begin{array}{c} \bar{x} \\ \lambda \end{array} \right] \in \mathbb{R}^{n+1}$ et $e = \left[\begin{array}{c} b \\ \hline \\ c \end{array} \right] \in \mathbb{R}^{n+1}$. Montrons maintenant que B est inversible. En effet, soit $z \left[\begin{array}{c} x \\ \mu \end{array} \right] \in \mathbb{R}^{n+1}$, avec $x \in \mathbb{R}^n$ et $\mu \in \mathbb{R}$ tel que $Bz = 0$. Alors

$$\left[\begin{array}{c|c} A & d \\ \hline d^t & 0 \end{array} \right] \left[\begin{array}{c} x \\ \mu \end{array} \right] = 0.$$

Ceci entraîne $Ax - d\mu = 0$ et $d^t x = d \cdot x = 0$. On a donc $Ax \cdot x - (d \cdot x)\mu = 0$. On en déduit que $x = 0$, et comme $d \neq 0$, que $\mu = 0$. On a donc finalement $z = 0$.
On en conclut que B est inversible, et qu'il existe un unique $(x, \lambda)^t \in \mathbb{R}^{n+1}$ solution de (3.68) et que $\bar{x}$ est solution de (3.48).

Exercice 163 page 243 (Application simple du théorème de Kuhn-Tucker

La fonction f définie de $E = \mathbb{R}^2$ dans $\mathbb{R}$ par $f(x) = x^2 + y^2$ est continue, strictement convexe et croissante à l'infini. L'ensemble K qui peut aussi être défini par : $K = \{(x, y) \in \mathbb{R}^2; g(x, y) \leq 0\}$, avec $g(x, y) = 1 - x - y$ est convexe et fermé. Par le théorème 3.30 page 236, il y a donc existence et unicité de la solution du problème (3.48). Appliquons le théorème de Kuhn-Tucker pour la détermination de cette solution. On a :

$$\nabla g(x, y) = \left(\begin{array}{c} -1 \\ -1 \end{array} \right) \text{ et } \nabla f(x, y) = \left(\begin{array}{c} 2x \\ 2y \end{array} \right).$$

Il existe donc $\lambda \in \mathbb{R}_+$ tel que :

$$\left\{ \begin{array}{r} 2x - \lambda = 0, \\ 2y - \lambda = 0, \\ \lambda(1 - x - y) = 0, \\ 1 - x - y \leq 0, \\ \lambda \geq 0. \end{array} \right.$$

Par la troisième équation de ce système, on déduit que $\lambda = 0$ ou $1 - x - y = 0$. Or si $\lambda = 0$, on a $x = y = 0$ par les première et deuxième équations, ce qui est impossible en raison de la quatrième. On en déduit que $1 - x - y = 0$, et donc, par les première et deuxième équations, $x = y = \frac{1}{2}$.

Exercice 164 page 243 (Exemple d'opérateur de projection)

2. Soit p_K l'opérateur de projection définie à la proposition 3.40 page 246, il est facile de montrer que, pour tout $i = 1, \ldots, n$, :

$(p_K(y))_i = y_i$ si $y_i \in [\alpha_i, \beta_i]$,
$(p_K(y))_i = \alpha_i$ si $y_i < \alpha_i$, ce qui entraîne
$(p_K(y))_i = \beta_i$ si $y_i > \beta_i$,

$$(p_K(y))_i = \max(\alpha_i, \min(y_i, \beta_i)) \text{ pour tout } i = 1, \ldots, n.$$

3.5 Algorithmes d'optimisation sous contraintes

3.5.1 Méthodes de gradient avec projection

On rappelle le résultat suivant de projection sur un convexe fermé :

Proposition 3.40 (Projection sur un convexe fermé)**.** *Soit E un espace de Hilbert, muni d'une norme $\|.\|$ induite par un produit scalaire $(.,.)$, et soit K un convexe fermé non vide de E. Alors, tout $x \in E$, il existe un unique $x_0 \in K$ tel que $\|x - x_0\| \leq \|x - y\|$ pour tout $y \in K$. On note $x_0 = p_K(x)$ la projection orthogonale de x sur K. Soient $x \in E$ et $x_0 \in K$. On a également :*

$$x_0 = p_K(x) \text{ si et seulement si } (x - x_0, x_0 - y) \geq 0, \quad \forall y \in K.$$

Dans le cadre des algorithmes de minimisation avec contraintes que nous allons développer maintenant, nous considèrerons $E = \mathbb{R}^n$, $f \in C^1(\mathbb{R}^n, \mathbb{R})$ une fonction convexe, et K fermé convexe non vide. On cherche à calculer une solution approchée de $\bar{x}$, solution du problème (3.48).

Algorithme du gradient à pas fixe avec projection sur K (GPFK) Soit $\rho > 0$ donné, on considère l'algorithme suivant :

Algorithme (GPFK)

Initialisation : $x_0 \in K$

Itération :

x_k connu $\qquad\qquad x_{k+1} = p_K(x_k - \rho \nabla f(x_k))$

où p_K est la projection sur K définie par la proposition 3.40.

Lemme 3.41. *Soit $(x_k)_k$ construite par l'algorithme (GPFK). On suppose que $x_k \to x$ quand $n + \infty$. Alors x est solution de* (3.48)*.*

DÉMONSTRATION – Soit $p_K : \mathbb{R}^n \to K \subset \mathbb{R}^n$ la projection sur K définie par la proposition 3.40. Alors p_K est continue. Donc si

$x_k \to x$ quand $n \to +\infty$ alors $x = p_K(x - \rho \nabla f(x))$ et $x \in K$ (car $x_k \in K$ et K est fermé).

La caractérisation de $p_K(x - \rho \nabla f(x))$ donnée dans la proposition 3.40 donne alors :

$(x - \rho \nabla f(x) - x/x - y) \geq 0$ pour tout $y \in K$, et comme $\rho > 0$, ceci entraîne $(\nabla f(x)/x - y) \leq 0$ pour tout $y \in K$. Or f est convexe donc $f(y) \geq f(x) + \nabla f(x)(y - x)$ pour tout $y \in K$, et donc $f(y) \geq f(x)$ pour tout $y \in K$, ce qui termine la démonstration.

■

Théorème 3.42 (Convergence de l'algorithme GPFK)**.**
Soit $f \in C^1(\mathbb{R}^n, \mathbb{R})$, et K convexe fermé non vide. On suppose que :

1. *il existe $\alpha > 0$ tel que $(\nabla f(x) - \nabla f(y)|x - y) \geq \alpha |x - y|^2$, pour tout $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$,*
2. *il existe $M > 0$ tel que $|\nabla f(x) - \nabla f(y)| \leq M|x - y|$ pour tout $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$,*

alors :

1. *il existe un unique élément $\bar{x} \in K$ solution de* (3.48)*,*
2. *si $0 < \rho < \dfrac{2\alpha}{M^2}$, la suite (x_k) définie par l'algorithme $(GPFK)$ converge vers $\bar{x}$ lorsque $n \to +\infty$.*

DÉMONSTRATION –

1. La condition 1. donne que f est strictement convexe et que $f(x) \to +\infty$ quand $|x| \to +\infty$. Comme K est convexe fermé non vide, il existe donc un unique $\bar{x}$ solution de (3.48).
2. On pose, pour $x \in \mathbb{R}^n$, $h(x) = p_K(x - \rho\nabla f(x))$. On a donc $x_{k+1} = h(x_k)$. Soit $0 < \rho < \frac{2\alpha}{M^2}$. Pour montrer que la suite $(x_k)_{k\in\mathbb{N}}$ converge, il suffit donc de montrer que h est strictement contractante. Grâce au lemme 3.43, on sait que p_K est contractante. Or h est définie par :
$$h(x) = p_K(\bar{h}(x)) \quad \text{où } \bar{h}(x) = x - \rho\nabla f(x).$$
On a déjà vu que $\bar{h}$ est strictement contractante (car $0 < \rho < \frac{2\alpha}{M^2}$, voir le théorème 3.19 page 205). Plus précisément, on a :
$$|\bar{h}(x) - \bar{h}(y)| \leq (1 - 2\alpha\rho + M^2\rho^2)|x - y|^2.$$
On en déduit que :
$$|h(x) - h(y)|^2 \leq |p_K(\bar{h}(x)) - p_K(\bar{h}(y))|^2 \leq |\bar{h}(x) - \bar{h}(y))|^2 \leq (1 - 2\alpha\rho + \rho^2M^2)|x - y|^2.$$
L'application h est donc strictement contractante. La suite $(x_k)_{k\in\mathbb{N}}$ est donc convergente. on note $\tilde{x}$ sa limite. il reste à montrer que $\tilde{x} = \bar{x}$. On remarque tout d'abord que $\tilde{x} \in K$ (car K est fermé). Puis, comme $\tilde{x}$ est un point fixe de h, on a $\tilde{x} = p_K(\tilde{x} - \rho\nabla f(\tilde{x}))$. La caractérisation de p_K donnée dans la proposition 3.40 donne alors
$$(\tilde{x} - \rho\nabla f(\tilde{x}) - \tilde{x}) \cdot (\tilde{x} - y) \geq 0 \text{ pour tout } y \in K,$$
ce qui donne $\nabla f(\tilde{x}) \cdot (y - \tilde{x}) \geq 0$ pour tout $y \in K$ et donc, comme f est convexe, $f(y) \geq f(\tilde{x}) + \nabla f(\tilde{x}) \cdot (y - \tilde{x}) \geq f(\tilde{x})$ pour tout $y \in K$. Ceci montre bien que $\tilde{x} = \bar{x}$.

■

Lemme 3.43 (Propriété de contraction de la projection orthogonale)**.** *Soit E un espace de Hilbert, $\|\cdot\|$ la norme et $(\cdot,\cdot)$ le produit scalaire, K un convexe fermé non vide de E et p_K la projection orthogonale sur K définie par la proposition 3.40, alors $\|p_K(x) - p_K(y)\| \leq \|x - y\|$ pour tout $(x, y) \in E^2$.*

DÉMONSTRATION – Comme E est un espace de Hilbert,
$$\|p_K(x) - p_K(y)\|^2 = (p_K(x) - p_K(y)|p_K(x) - p_K(y)).$$
On a donc
$$\begin{aligned}\|p_K(x) - p_K(y)\|^2 = &\quad (p_K(x) - x + x - y + y - p_K(y)|p_K(x) - p_K(y)) \\ = &\quad (p_K(x) - x|p_K(x) - p_K(y))_E + (x - y|p_K(x) - p_K(y)) + \\ &\quad (y - p_K(y)|p_K(x) - p_K(y)).\end{aligned}$$
Or $(p_K(x) - x|p_K(x) - p_K(y)) \leq 0$ et $(y - p_K(y)|p_K(x) - p_K(y)) \leq 0$, d'où :
$$\|p_K(x) - p_K(y)\|^2 \leq (x - y|p_K(x) - p_K(y)),$$
et donc, grâce à l'inégalité de Cauchy-Schwarz,
$$\|p_K(x) - p_K(y)\|^2 \leq \|x - y\| \, \|p_K(x) - p_K(y)\|,$$
ce qui permet de conclure. ■

Algorithme du gradient à pas optimal avec projection sur K (GPOK)
L'algorithme du gradient à pas optimal avec projection sur K s'écrit :

Initialisation	$x_0 \in K$
Itération	x_k connu
	$w_k = -\nabla f(x_k)$; calculer α_k optimal dans la direction w_k
	$x_{k+1} = p_K(x_k + \alpha_k \boldsymbol{w}^{(k)})$

La démonstration de convergence de cet algorithme se déduit de celle de l'algorithme à pas fixe.

Remarque 3.44. *On pourrait aussi utiliser un algorithme de type Quasi–Newton avec projection sur K.*

Les algorithmes de projection sont simples à décrire, mais ils soulèvent deux questions :

1. Comment calcule-t-on p_K ?
2. Que faire si K n'est pas convexe ?

On peut donner une réponse à la première question dans les cas simples :

Cas 1. On suppose ici que $K = C^+ = \{x \in \mathbb{R}^n,\ x = (x_1, \ldots, x_k)^t\ x_i \geq 0\ \forall i\}$.

Si $y \in \mathbb{R}^n\ y = (y_1 \ldots y_n)^t$, on peut montrer (exercice 164 page 243) que

$$(p_K(y))_i = y_i^+ = \max(y_i, 0), \quad \forall i \in \{1, \ldots, n\}$$

Cas 2. Soit $(\alpha_i)_{i=1,\ldots,n} \subset \mathbb{R}^n$ et $(\beta_i)_{i=1,\ldots,n} \subset \mathbb{R}^n$ tels que $\alpha_i \leq \beta_i$ pour tout $i = 1, \ldots, n$. Si

$$K = \prod_{i=1,n} [\alpha_i, \beta_i],$$

alors

$$(p_K(y))_i = \max(\alpha_i, \min(y_i, \beta_i)), \quad \forall i = 1, \ldots, n$$

Dans le cas d'un convexe K plus "compliqué", ou dans le cas où K n'est pas convexe, on peut utiliser des méthodes de dualité introduites dans le paragraphe suivant.

3.5.2 Méthodes de dualité

Supposons que les hypothèses suivantes sont vérifiées :

$$\left\{ \begin{array}{l} f \in C^1(\mathbb{R}^n, \mathbb{R}), \\ g_i \in C^1(\mathbb{R}^n, \mathbb{R}), \\ K = \{x \in \mathbb{R}^n,\ g_i(x) \leq 0\ i = 1, \ldots, p\},\ \text{et } K \text{ est non vide.} \end{array} \right. \tag{3.69}$$

On définit un problème "primal" comme étant le problème de minimisation d'origine, c'est–à–dire

$$\left\{ \begin{array}{l} \bar{x} \in K, \\ f(\bar{x}) \leq f(x),\ \text{pour tout } x \in K, \end{array} \right. \tag{3.70}$$

On définit le "lagrangien" comme étant la fonction L définie de $\mathbb{R}^n \times \mathbb{R}^p$ dans $\mathbb{R}$ par :

$$L(x, \lambda) = f(x) + \lambda \cdot g(x) = f(x) + \sum_{i=1}^{p} \lambda_i g_i(x), \tag{3.71}$$

avec $g(x) = (g_1(x), \ldots, g_p(x))^t$ et $\lambda = (\lambda_1, \ldots, \lambda_p)^t$.
On note C^+ l'ensemble défini par

$$C^+ = \{\lambda \in \mathbb{R}^p,\ \lambda = (\lambda_1, \ldots, \lambda_p)^t, \lambda_i \geq 0 \text{ pour tout } i = 1, \ldots, p\}.$$

Remarque 3.45. *Sous les hypothèses du théorème de Kuhn-Tucker, si $\bar{x}$ est solution du problème primal* (3.70) *alors il existe $\lambda \in C^+$ tel que $D_1 L(\bar{x}, \lambda) = 0$ (c'est–à–dire $Df(\bar{x}) + \lambda \cdot Dg(\bar{x}) = 0$) et $\lambda \cdot g(\bar{x}) = 0$.*

On définit alors l'application M de $\mathbb{R}^p$ dans $\mathbb{R}$ par :

$$M(\lambda) = \inf_{x \in \mathbb{R}^n} L(x, \lambda),\ \text{pour tout } \lambda \in \mathbb{R}^p. \tag{3.72}$$

On peut donc remarquer que $M(\lambda)$ réalise le minimum (en x) du problème sans contrainte, qui s'écrit, pour $\lambda \in \mathbb{R}^p$ fixé :

$$\left\{ \begin{array}{l} x \in \mathbb{R}^n \\ L(x, \lambda) \leq L(y, \lambda) \text{ pour tout } x \in \mathbb{R}^n, \end{array} \right. \tag{3.73}$$

Lemme 3.46. *L'application M de $\mathbb{R}^p$ dans $\mathbb{R}$ définie par* (3.72) *est concave (ou encore l'application -M est convexe), c'est–à–dire que pour tous $\lambda, \mu \in \mathbb{R}^p$ et pour tout $t \in]0,1[$ on a $M(t\lambda + (1-t)\mu) \geq tM(\lambda) + (1-t)\mathcal{M}(u)$*

DÉMONSTRATION – Soit $\lambda, \mu \in \mathbb{R}^p$ et $t \in]0,1[$; on veut montrer que $M(t\lambda + (1-t)\mu) \geq tM(\lambda) + (1-t)M(\mu)$. Soit $x \in \mathbb{R}^n$, alors :

$$\begin{aligned} L(x, t\lambda + (1-t)\mu) &= f(x) + (t\lambda + (1-t)\mu)g(x) \\ &= tf(x) + (1-t)f(x) + (t\lambda + (1-t)\mu)g(x). \end{aligned}$$

On a donc $L(x, t\lambda + (1-t)\mu) = tL(x,\lambda) + (1-t)L(x,\mu)$. Par définition de M, on en déduit que pour tout $x \in \mathbb{R}^n$,

$$L(x, t\lambda + (1-t)\mu) \geq tM(\lambda) + (1-t)M(\mu)$$

Or, toujours par définition de M,

$$M(t\lambda + (1-t)\mu) = \inf_{x \in \mathbb{R}^n} L(x, t\lambda + (1-t)\mu) \geq tM(\lambda) + (1-t)M(\mu).$$

■

On considère maintenant le problème d'optimisation dit "dual" suivant :

$$\begin{cases} \mu \in C^+, \\ M(\mu) \geq M(\lambda) \quad \forall \lambda \in C^+. \end{cases} \tag{3.74}$$

Définition 3.47. *Soit $L : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ et $(x, \mu) \in \mathbb{R}^n \times C^+$. On dit que (x, μ) est un point selle de L sur $\mathbb{R}^n \times C^+$ si*

$$L(x, \lambda) \leq L(x, \mu) \leq L(y, \mu) \text{ pour tout } y \in \mathbb{R}^n \text{ et pour tout } \lambda \in C^+.$$

Proposition 3.48. *Sous les hypothèses* (3.69)*, soit L définie par $L(x,\lambda) = f(x) + \lambda g(x)$ et $(\bar{x}, \mu) \in \mathbb{R}^n \times C^+$ un point selle de L sur $\mathbb{R}^n \times C^+$.*
alors

1. *$\bar{x}$ est solution du problème* (3.70)*,*
2. *μ est solution de* (3.74)*,*
3. *$\bar{x}$ est solution du problème* (3.73) *avec $\lambda = \mu$.*

On admettra cette proposition.

Réciproquement, on peut montrer que (sous des hypothèses convenables sur f et g), si μ est solution de (3.74), et si $\bar{x}$ solution de (3.73) avec $\lambda = \mu$, alors $(\bar{x}, \mu)$ est un point selle de L, et donc $\bar{x}$ est solution de (3.70).

De ces résultats découle l'idée de base des méthodes de dualité : on cherche μ solution de (3.74). On obtient ensuite une solution $\bar{x}$ du problème (3.70), en cherchant $\bar{x}$ comme solution du problème (3.73) avec $\lambda = \mu$ (qui est un problème de minimisation sans contraintes). La recherche de la solution μ du problème dual (3.74) peut se faire par exemple par l'algorithme très classique d'Uzawa, que nous décrivons maintenant.

Algorithme d'Uzawa L'algorithme d'Uzawa consiste à utiliser l'algorithme du gradient à pas fixe avec projection (qu'on a appelé "GPFK", voir page 246) pour résoudre de manière itérative le problème dual (3.74). On cherche donc $\mu \in C^+$ tel que $M(\mu) \geq M(\lambda)$ pour tout $\lambda \in C^+$. On se donne $\rho > 0$, et on note p_{C^+} la projection sur le convexe C^+ (voir proposition 3.40 page 246). L'algorithme (GPFK) pour la recherche de μ s'écrit donc :

Initialisation : $\mu_0 \in C_+$
Itération : $\mu_{k+1} = p_{C_+}(\mu_k + \rho \nabla M(\mu_k))$

Pour définir complètement l'algorithme d'Uzawa, il reste à préciser les points suivants :

1. Calcul de $\nabla M(\mu_k)$,
2. calcul de $p_{C^+}(\lambda)$ pour λ dans $\mathbb{R}^n$.

On peut également s'intéresser aux propriétés de convergence de l'algorithme.

La réponse au point 2 est simple (voir exercice 164 page 243) : pour $\lambda \in \mathbb{R}^p$, on calcule $p_{C_+}(\lambda) = \gamma$ avec $\gamma = (\gamma_1, \ldots, \gamma_p)^t$ en posant $\gamma_i = \max(0, \lambda_i)$ pour $i = 1, \ldots, p$, où $\lambda = (\lambda_1, \ldots, \lambda_p)^t$.

La réponse au point 1. est une conséquence de la proposition suivante (qu'on admettra ici) :

Proposition 3.49. *Sous les hypothèses* (3.69)*, on suppose que pour tout $\lambda \in \mathbb{R}^n$, le problème* (3.73) *admet une solution unique, notée x_λ et on suppose que l'application définie de $\mathbb{R}^p$ dans $\mathbb{R}^n$ par $\lambda \mapsto x_\lambda$ est différentiable. Alors $M(\lambda) = L(x_\lambda, \lambda)$, M est différentiable en λ pour tout λ, et $\nabla M(\lambda) = g(x_\lambda)$.*

En conséquence, pour calculer $\nabla M(\lambda)$, on est ramené à chercher x_λ solution du problème de minimisation sans contrainte (3.73). On peut dont maintenant donner le détail de l'itération générale de l'algorithme d'Uzawa :

Itération de l'algorithme d'Uzawa. Soit $\mu_k \in C^+$ connu ;

1. On cherche $x_k \in \mathbb{R}^n$ solution de $\left\{ \begin{array}{l} x_k \in \mathbb{R}^n, \\ L(x_k, \mu_k) \leq L(x, \mu_k), \ \forall x \in \mathbb{R}^n \end{array} \right.$ (On a donc $x_k = x_{\mu_k}$)
2. On calcule $\nabla M(\mu_k) = g(x_k)$
3. $\overline{\mu}_{k+1} = \mu_k + \rho \nabla M(\mu_k) = \mu_k + \rho g(x_k) = ((\overline{\mu}_{k+1})_1, \ldots, (\overline{\mu}_{k+1})_p)^t$
4. $\mu_{k+1} = p_{C^+}(\overline{\mu}_{k+1})$, c'est–à-dire $\mu_{k+1} = ((\mu_{k+1})_1, \ldots, (\mu_{k+1})_p)^t$ avec $(\mu_{k+1})_i = \max(0, (\overline{\mu}_{k+1})_i)$ pour tout $i = 1, \ldots, p$.

L'exercice 166 donne un résulat de convergence contenant en particulier le cas très intéressant d'une fonctionnelle quadratique avec des contraintes affines "suffisamment" indépendantes (pour pouvoir appliquer le théorème de Kuhn-Tucker).

Remarque 3.50 (Sur l'algorithme d'Uzawa)**.**

1. *L'algorithme est très efficace si les contraintes sont affines : (i.e. si $g_i(x) = \alpha_i \cdot x + \beta_i$ pour tout $i = 1, \ldots, p$, avec $\alpha_i \in \mathbb{R}^n$ et $\beta_i \in \mathbb{R}$).*
2. *Pour avoir l'hypothèse 3 du théorème, il suffit que les fonctions g_i soient convexes. (On a dans ce cas existence et unicité de la solution x_λ du problème* (3.73) *et existence et unicité de la solution $\bar{x}$ du problème* (3.70)*.)*

3.5.3 Exercices (algorithmes pour l'optimisation avec contraintes)

Exercice 165 (Méthode de pénalisation)**.**
Soit f une fonction continue et strictement convexe de $\mathbb{R}^n$ dans $\mathbb{R}$, satisfaisant de plus :

$$\lim_{|x| \to +\infty} f(x) = +\infty.$$

Soit K un sous ensemble non vide, convexe (c'est–à–dire tel que $\forall (x, y) \in K^2$, $tx + (1-t)y \in K$, $\forall t \in]0, 1[$), et fermé de $\mathbb{R}^n$.
Soit ψ une fonction continue de $\mathbb{R}^n$ dans $[0, +\infty[$ telle que $\psi(x) = 0$ si et seulement si $x \in K$. Pour $k \in \mathbb{N}$, on définit la fonction f_k par $f_k(x) = f(x) + k\psi(x)$.

1. Montrer qu'il existe au moins un élément $\bar{x}_k \in \mathbb{R}^n$ tel que $f_k(\bar{x}_k) = \inf_{x \in \mathbb{R}^n} f_k(x)$, et qu'il existe un unique élément $\bar{x}_K \in K$ tel que $f(\bar{x}_K) = \inf_{x \in K} f(x)$.
2. Montrer que pour tout $k \in \mathbb{N}$,
$$f(\bar{x}_k) \leq f_k(\bar{x}_k) \leq f(\bar{x}_K).$$

3. En déduire qu'il existe une sous-suite $(\bar{x}_{k_m})_{m\in\mathbb{N}}$ et $y \in K$ tels que $\bar{x}_{k_m} \to y$ lorsque $m \to +\infty$.
4. Montrer que $y = \bar{x}_K$. En déduire que toute la suite $(\bar{x}_k)_{k\in\mathbb{N}}$ converge vers $\bar{x}_K$ lorsque $k \to +\infty$.
5. Déduire de ces questions un algorithme (dit "de pénalisation") de résolution du problème de minimisation suivant :
$$\begin{cases} \text{Trouver } \bar{x}_K \in K; \\ f(\bar{x}_K) \leq f(x),\ \forall x \in K, \end{cases}$$
en donnant un exemple de fonction ψ.

Exercice 166 (Convergence de l'algorithme d'Uzawa)**.** *Corrigé en page 253*
Soient $n \geq 1$ $p \in \mathbb{N}^\star$. Soit $f \in C^1(\mathbb{R}^n, \mathbb{R})$ une fonction telle que

$$\exists \alpha > 0,\ (\nabla f(x) - \nabla f(y)) \cdot (x - y) \geq \alpha |x - y|^2,\ \forall x, y \in \mathbb{R}^n.$$

Soit $C \in M_{p,n}(\mathbb{R})$ (C est donc une matrice, à éléments réels, ayant p lignes et n colonnes) et $d \in \mathbb{R}^p$. On note $D = \{x \in \mathbb{R}^n, Cx \leq d\}$ et $\mathcal{C}^+ = \{u \in \mathbb{R}^p, u \geq 0\}$.
On suppose $D \neq \emptyset$ et on s'intéresse au problème suivant :

$$x \in D,\ \ f(x) \leq f(y),\ \forall y \in D. \tag{3.75}$$

1. Montrer que $f(y) \geq f(x) + \nabla f(x) \cdot (y - x) + \frac{\alpha}{2}|x - y|^2$ pour tout $x, y \in \mathbb{R}^n$.
2. Montrer que f est strictement convexe et que $f(x) \to \infty$ quand $|x| \to \infty$. En déduire qu'il existe une et une seule solution au problème (3.75).

Dans la suite, on note $\overline{x}$ cette solution.
Pour $u \in \mathbb{R}^p$ et $x \in \mathbb{R}^n$, on pose $L(x, u) = f(x) + u \cdot (Cx - d)$.

3. Soit $u \in \mathbb{R}^p$ (dans cette question, u est fixé). Montrer que l'application $x \to L(x, u)$ est strictement convexe (de $\mathbb{R}^n$ dans $\mathbb{R}$) et que $L(x, u) \to \infty$ quand $|x| \to \infty$ [Utiliser la question 1]. En déduire qu'il existe une et une seule solution au problème suivant :
$$x \in \mathbb{R}^n,\ \ L(x, u) \leq L(y, u),\ \forall y \in \mathbb{R}^n. \tag{3.76}$$
Dans la suite, on note x_u cette solution. Montrer que x_u est aussi l'unique élément de $\mathbb{R}^n$ t.q. $\nabla f(x_u) + C^t u = 0$.
4. On admet que le théorème de Kuhn-Tucker s'applique ici (cf. cours). Il existe donc $\overline{u} \in \mathcal{C}^+$ t.q. $\nabla f(\overline{x}) + C^t\overline{u} = 0$ et $\overline{u} \cdot (C\overline{x} - d) = 0$. Montrer que $(\overline{x}, \overline{u})$ est un point selle de L sur $\mathbb{R}^n \times \mathcal{C}^+$, c'est-à-dire :
$$L(\overline{x}, v) \leq L(\overline{x}, \overline{u}) \leq L(y, \overline{u}),\ \forall (y, v) \in \mathbb{R}^n \times \mathcal{C}^+. \tag{3.77}$$

Pour $u \in \mathbb{R}^p$, on pose $M(u) = L(x_u, u)$ (de sorte que $M(u) = \inf\{L(x, u),\ x \in \mathbb{R}^n\}$). On considère alors le problème suivant :

$$u \in \mathcal{C}^+,\ \ M(u) \geq M(v),\ \forall v \in \mathcal{C}^+. \tag{3.78}$$

5. Soit $(x, u) \in \mathbb{R}^n \times \mathcal{C}^+$ un point selle de L sur $\mathbb{R}^n \times \mathcal{C}^+$ (c'est-à-dire $L(x, v) \leq L(x, u) \leq L(y, u)$, pour tout $(y, v) \in \mathbb{R}^n \times \mathcal{C}^+$). Montrer que $x = \overline{x} = x_u$ (on rappelle que $\overline{x}$ est l'unique solution de (3.75) et x_u est l'unique solution de (3.76)) et que u est solution de (3.78). [On pourra commencer par montrer, en utilisant la première inégalité, que $x \in D$ et $u \cdot (Cx - d) = 0$.]
Montrer que $\nabla f(\bar{x}) + C^t u = 0$ et que $u = P_{\mathcal{C}^+}(u + \rho(C\overline{x} - d))$, pour tout $\rho > 0$, où $P_{\mathcal{C}^+}$ désigne l'opérateur de projection orthogonale sur $\mathcal{C}^+$. [on rappelle que si $v \in \mathbb{R}^p$ et $w \in \mathcal{C}^+$, on a $w = P_{\mathcal{C}^+}v \iff ((v - w) \cdot (w - z) \geq 0, \forall z \in \mathcal{C}^+)$.]
6. Déduire des questions 2, 4 et 5 que le problème (3.78) admet au moins une solution.

7. On admet que l'application $u \mapsto x_u$ est dérivable. Montrer que l'algorithme du gradient à pas fixe avec projection pour trouver la solution de (3.78) s'écrit (on désigne par $\rho > 0$ le pas de l'algorithme) :

   Initialisation. $u_0 \in \mathcal{C}^+$.

   Itérations. Pour $u_k \in \mathcal{C}^+$ connu ($k \geq 0$). On calcule $x_k \in \mathbb{R}^n$ t.q. $\nabla f(x_k) + C^t u_k = 0$ (montrer qu'un tel x_k existe et est unique) et on pose $u_{k+1} = P_{\mathcal{C}^+}(u_k + \rho(Cx_k - d))$.

Dans la suite, on s'intéresse à la convergence de la suite $(x_k, u_k)_{k\in\mathbb{N}}$ donnée par cet algorithme.

8. Soit ρ t.q. $0 < \rho < 2\alpha/\|C\|^2$ avec $\|C\| = \sup\{|Cx|,\ x \in \mathbb{R}^n \text{ t.q. } |x| = 1\}$. Soit $(\overline{x}, \overline{u}) \in \mathbb{R}^n \times \mathcal{C}^+$ un point selle de L sur $\mathbb{R}^n \times \mathcal{C}^+$ (c'est-à-dire vérifiant (3.77)) et $(x_k, u_k)_{k\in\mathbb{N}}$ la suite donnée par l'algorithme de la question précédente. Montrer que

$$|u_{k+1} - \overline{u}|^2 \leq |u_k - \overline{u}|^2 - \rho(2\alpha - \rho\|C\|^2)|x_k - \overline{x}|^2,\ \forall k \in \mathbb{R}^n.$$

   En déduire que $x_k \to \overline{x}$ quand $k \to \infty$.

   Montrer que la suite $(u_k)_{k\in\mathbb{N}}$ est bornée et que, si $\tilde{u}$ est une valeur d'adhérence de la suite $(u_k)_{k\in\mathbb{N}}$, on a $\nabla f(\overline{x}) + C^t\tilde{u} = 0$. En déduire que, si rang$(C) = p$, on a $u_k \to \overline{u}$ quand $k \to \infty$ et que $\overline{u}$ est *l'unique* élément de $\mathcal{C}^+$ t.q. $\nabla f(\overline{x}) + C^t\overline{u} = 0$.

Exercice 167 (Méthode de relaxation avec Newton pour un problème d'optimisation contrainte)**.**
On considère le problème :

$$\begin{cases} \overline{x} \in K, \\ f(\overline{x}) \leq f(x), \forall x \in K, \end{cases} \tag{3.79}$$

où $K \subset \mathbb{R}^n$.

(a) On prend ici $K = \prod_{i=1,n}[a_i, b_i]$, où $(a_i, b_i) \in \mathbb{R}^2$ est tel que $a_i \leq b_i$. On considère l'algorithme suivant :

$$\left\{\begin{array}{ll} \text{Initialisation :} & x^{(0)} \in E, \\ \text{Itération } n : & x^{(k)} \text{ connu, } (n \geq 0) \\ & \text{Calculer } x_1^{(k+1)} \in [a_1, b_1] \text{ tel que :} \\ & f(x_1^{(k+1)}, x_2^{(k)}, x_3^{(k)}, \ldots, x_n^{(k)}) \leq f(\xi, x_2^{(k)}, x_3^{(k)}, \ldots, x_n^{(k)}), \text{ pour tout } \xi \in [a_1, b_1], \\ & \text{Calculer } x_2^{(k+1)} \in [a_2, b_2] \text{ tel que :} \\ & f(x_1^{(k+1)}, x_2^{(k+1)}, x_3^{(k)}, \ldots, x_n^{(k)}) \leq f(x_1^{(k+1)}, \xi, x_3^{(k)}, \ldots, x_n^{(k)}), \\ & \qquad \text{pour tout } \xi \in [a_2, b_2], \\ & \qquad\qquad \ldots \\ & \text{Calculer } x_k^{(k+1)} \in [a_k, b_k], \text{ tel que :} \\ & f(x_1^{(k+1)}, \ldots, x_{k-1}^{(k+1)}, x_k^{(k+1)}, x_{(k+1)}^{(k)}, \ldots, x_n^{(k)}) \\ & \qquad \leq f(x_1^{(k+1)}, \ldots, x_{k-1}^{(k+1)}, \xi, x_{(k+1)}^{(k)}, \ldots, x_n^{(k)}), \text{ pour tout } \xi \in [a_k, b_k], \\ & \qquad\qquad \ldots \\ & \text{Calculer } x_n^{(k+1)} \in [a_n, b_n] \text{ tel que :} \\ & f(x_1^{(k+1)}, x_2^{(k+1)}, \ldots, x_{n-1}^{(k+1)}, x_n^{(k+1)}) \leq f(x_1^{(k+1)}, \ldots, x_{n-1}^{(k+1)}, \xi), \\ & \qquad \text{pour tout } \xi \in [a_n, b_n]. \end{array}\right. \tag{3.80}$$

   Montrer que la suite $x^{(k)}$ construite par l'algorithme (3.80) est bien définie et converge vers $\overline{x}$ lorsque n tend vers $+\infty$, où $\overline{x} \in K$ est tel que $f(\overline{x}) \leq f(x)$ pour tout $x \in K$.

(b) On prend maintenant $n = 2$, f la fonction de $\mathbb{R}^2$ dans $\mathbb{R}$ définie par $f(x) = x_1^2 + x_2^2$, et $K = \{(x_1, x_2)^t \in \mathbb{R}^2; x_1 + x_2 \geq 2\}$. Montrer qu'il existe un unique élément $\overline{x} = (\overline{x}_1, \overline{x}_2)^t$ de K tel que $f(\overline{x}) = \inf_{x \in \mathbb{R}^2} f(x)$. Déterminer $\overline{x}$.

On considère l'algorithme suivant pour la recherche de $\overline{x}$:

$$\left\{\begin{array}{ll} \text{Initialisation :} & x^{(0)} \in E, \\ \text{Itération } n : & x^{(k)} \text{ connu, } (n \geq 0) \\ & \text{Calculer } x_1^{(k+1)} \geq 2 - x_2^{(k)} \text{ tel que :} \\ & f(x_1^{(k+1)}, x_2^{(k)}) \leq f(\xi, x_2^{(k)}), \text{ pour tout } \xi \geq 2 - x_2^{(k)}, \\ & \text{Calculer } x_2^{(k+1)} \geq 2 - x_1^{(k)} \text{ tel que :} \\ & f(x_1^{(k+1)}, x_2^{(k+1)}) \leq f(x_1^{(k+1)}, \xi), \text{ pour tout } \xi \geq 2 - x_1^{(k)}. \end{array}\right. \tag{3.81}$$

Montrer (éventuellement graphiquement) que la suite construite par l'algorithme ci-dessus ne converge vers $\overline{x}$ que si l'une des composantes de $x^{(0)}$ vaut 1.

3.5.4 Corrigés

Exercice 166 page 251 (Convergence de l'algorithme d'Uzawa)

1. Cette question a déjà été corrigée. Soit $x, y \in \mathbb{R}^n$. En posant $\varphi(t) = f(ty + (1-t)x)$, on remarque que

$$f(y) - f(x) = \varphi(1) - \varphi(0) = \int_0^1 \varphi'(t)dt = \int_0^1 \nabla f(x + t(y-x)) \cdot (y-x)dt,$$

et donc

$$f(y) - f(x) - \nabla f(x) \cdot (y-x) = \int_0^1 (\nabla f(x+t(y-x)) - \nabla f(x)) \cdot t(y-x)\frac{1}{t}dt \geq \alpha \int_0^1 t|y-x|^2 dt = \frac{\alpha}{2}|y-x|^2.$$

2. Montrer que f est strictement convexe et que $f(x) \to \infty$ quand $|x| \to \infty$. En déduire qu'il existe une et une seule solution au problème (3.75).
Cette question a aussi déjà été corrigée. La question précédente donne $f(y) \geq f(x) + \nabla f(x) \cdot (y-x)$ pour tout $x, y \in \mathbb{R}^n$. Ce qui montre que f est strictement convexe. Elle donne aussi, pour tout $x \in \mathbb{R}^n$,

$$f(x) \geq f(0) + \nabla f(0) \cdot x + \frac{\alpha}{2}|x|^2 \to +\infty \text{ quand } |x| \to +\infty.$$

De ces deux propriétés de f on déduit l'existence et l'unicité de la solution au problème (3.75).

3. L'application $x \mapsto u \cdot (Cx - d)$ est affine et donc convexe. Comme f est strictement convexe, on en déduit que $x \mapsto L(x, u)$ est aussi strictement convexe.
La question précédente donne $L(x, u) \geq f(0) + \nabla f(0) \cdot x + u \cdot (Cx - d) + \frac{\alpha}{2}|x|^2$. On en déduit que $L(x, u) \to +\infty$ quand $|x| \to +\infty$.
De ces deux propriétés de $L(\cdot, u)$ on déduit l'existence et l'unicité de la solution au problème (3.76).
Comme $L(\cdot, u)$ est strictement convexe, x_u est aussi l'unique point qui annule $\nabla L(\cdot, u)$ (c'est-à-dire le gradient de l'application $x \mapsto L(x, u)$). Ceci donne bien que x_u est l'unique point de $\mathbb{R}^n$ tel que $\nabla f(x_u) + C^t u = 0$.

4. La question précédente nous dit que $x_{\bar{u}}$ est l'unique point de $\mathbb{R}^n$ tel que $\nabla f(x_{\bar{u}}) + C^t \bar{u} = 0$. Comme $\nabla f(\overline{x}) + C^t \overline{u} = 0$, on a donc $\bar{x} = x_{\bar{u}}$ et donc

$$L(\overline{x}, \overline{u}) \leq L(y, \overline{u}) \text{ pour tout } y \in \mathbb{R}^n.$$

Soit maintenant $v \in \mathcal{C}^+$. On a, comme $C\overline{x} \leq d$ (car $\overline{x} \in D$) et $\overline{u} \cdot (C\overline{x} - d) = 0$,

$$L(\overline{x}, v) = f(\overline{x}) + v \cdot (C\overline{x} - d) \leq f(\overline{x}) = f(\overline{x}) + \overline{u} \cdot (C\overline{x} - d) = L(\overline{x}, \overline{u}).$$

5. On a $L(x,v) \leq L(x,u)$ pour tout $v \in \mathcal{C}^+$ et donc

$$(v-u)\cdot(Cx-d) \leq 0 \text{ pour tout } v \in \mathcal{C}^+.$$

On note $u = (u_1,\dots,u_p)^t$. Soit $i \in \{1,\dots,p\}$. en prenant $v = (v_1,\dots,v_p)^t$ avec $v_j = u_j$ si $j \neq i$ et $v_i = u_i + 1$ (on a bien $v \in \mathcal{C}^+$), la formule précédente nous montre que la i-ieme composante de $(Cx-d)$ est négative. On a donc $x \in D$.
Soit maintenant $i \in \{1,\dots,p\}$ tel que $u_i > 0$. En prenant $v = (v_1,\dots,v_p)^t$ avec $v_j = u_j$ si $j \neq i$ et $v_i = 0$, la formule précédente nous montre que la i-ieme composante de $(Cx-d)$ est positive. Elle donc nécessairement nulle. Ceci nous donne bien que $u\cdot(Cx-d) = 0$.
On utilise maintenant le fait que $L(x,u) \leq L(y,u)$ pour tout $y \in \mathbb{R}^n$. Ceci donne, bien sûr, que $x = x_u$. Cela donne aussi que

$$f(x) + u\cdot(Cx-d) \leq f(y) + u\cdot(Cy-d).$$

Comme on sait que $u\cdot(Cx-d) = 0$ et comme $u\cdot(Cy-d) \leq 0$ si $y \in D$, on en déduit que $f(x) \leq f(y)$ pour tout $y \in D$, et donc $x = \overline{x}$.
Enfin, $L(x,u) = L(x_u,u) = M(u)$ et $L(x,v) \geq L(x_v,v) = M(v)$. Comme $L(x,v) \leq L(x,u)$ pour tout $v \in \mathcal{C}^+$, on a donc $M(v) \leq M(u)$ pour tout $v \in \mathcal{C}^+$.

On passe maintenant à la seconde partie de cette question. On a vu à la question 3 que $\nabla f(x_u) + C^t u = 0$. Comme $\overline{x} = x_u$, on a donc $\nabla f(\overline{x}) + C^t u = 0$.
Puis pour montrer que $u = P_{\mathcal{C}^+}(u + \rho(C\overline{x} - d))$, on utilise le rappel. Pour tout $z \in \mathcal{C}^+$ on a, en utilisant $u\cdot(C\overline{x}-d) = 0$ et $\overline{x} \in D$,

$$(u+\rho(C\overline{x}-d)-u)\cdot(u-z) = \rho(C\overline{x}-d)\cdot(u-z) = -\rho(C\overline{x}-d)\cdot z \geq 0.$$

Ceci donne bien que $u = P_{\mathcal{C}^+}(u + \rho(C\overline{x}-d))$.

6. La question 2 donne l'existence de $\overline{x}$ solution de (3.75). Puis la question 4 donne l'existence de $\overline{u}$ tel que $(\overline{x},\overline{u})$ est solution de (3.77). Enfin, la question 5 donne alors que $\overline{u}$ est solution de (3.78).
7. Les itérations de l'algorithme du gradient à pas fixe avec projection pour trouver la solution de (3.78) s'écrivent

$$u_{k+1} = P_{\mathcal{C}^+}(u_k + \rho\nabla M(u_k)).$$

Comme $M(u) = L(x_u,u)$ et x_u annule le gradient de l'application $x \mapsto L(x,u)$, la dérivation de fonctions composées (que l'on peut appliquer car l'application $u \mapsto x_u$ est supposée dérivable) nous donne $\nabla M(u) = Cx_u - d$. Comme $x_{u_k} = x_k$, on en déduit que

$$u_{k+1} = P_{\mathcal{C}^+}(u_k + \rho(Cx_k - d)).$$

8. Comme $(\overline{x},\overline{u})$ est un point selle de L sur $\mathbb{R}^N \times \mathcal{C}^+$, la question 5 nous donne

$$\overline{u} = P_{\mathcal{C}^+}(\overline{u} + \rho(C\overline{x}-d)).$$

L'opérateur $P_{\mathcal{C}^+}$ étant contractant, on obtient, avec la question précédente, pour tout k,

$$\begin{aligned}|u_{k+1}-\overline{u}|^2 &\leq |u_k + \rho(Cx_k-d) - (\overline{u}+\rho(C\overline{x}-d))|^2\\ &= |u_k-\overline{u}|^2 + 2\rho(u_k-\overline{u})\cdot C(x_k-\overline{x}) + \rho^2|C(x_k-\overline{x})|^2.\end{aligned}$$

Comme $C^t u_k = -\nabla f(x_k)$ et $C^t\overline{u} = -\nabla f(\overline{x})$, on obtient (avec l'hypothèse sur ∇f)

$$\begin{aligned}|u_{k+1}-\overline{u}|^2 &\leq |u_k-\overline{u}|^2 - 2\rho(\nabla f(x_k) - \nabla f(\overline{x}))\cdot(x_k-\overline{x}) + \rho^2|C(x_k-\overline{x})|^2\\ &\leq |u_k-\overline{u}|^2 - 2\rho\alpha|x_k-\overline{x}|^2 + \rho^2|C(x_k-\overline{x})|^2 \leq |u_k-\overline{u}|^2 - \rho(2\alpha-\rho\|C\|^2)|x_k-\overline{x}|^2.\end{aligned}$$

Comme $2\alpha - \rho\|C\|^2 > 0$, ceci montre que la suite $(u_k - \overline{u})_{k\in\mathbb{N}}$ est décroissante (positive) et donc convergente. Il suffit alors de remarquer que

$$|x_k - \overline{x}|^2 \leq \frac{1}{\rho(2\alpha - \rho\|C\|^2)}(|u_k - \overline{u}|^2 - |u_{k+1} - \overline{u}|^2)$$

pour en déduire que $x_k \to \overline{x}$ quand $k \to +\infty$.
La suite $(u_k - \overline{u})_{k\in\mathbb{N}}$ est convergente. La suite $(u_k)_{k\in\mathbb{N}}$ est donc bornée. Si $\tilde{u}$ est une valeur d'adhérence de la suite $(u_k)_{k\in\mathbb{N}}$, en passant à la limite sur l'équation $\nabla f(x_k) + C^t u_k = 0$ on obtient $\nabla f(\overline{x}) + C^t\tilde{u} = 0$. Si rang$(C) = p$, on a aussi rang$(C^t) = p$. L'application $u \mapsto C^t u$ est de $\mathbb{R}^p$ dans $\mathbb{R}^n$, on a donc $\dim(\ker C^t) = p - \text{rang}(C^t) = 0$. Ceci prouve qu'il existe un unique $\tilde{u}$ tel que $C^t\tilde{u} = -\nabla f(\overline{x})$. La suite $(u_k)_{k\in\mathbb{N}}$ n'a alors qu'une seule valeur d'adhérence et elle est donc convergente vers $\overline{u}$ et $\overline{u}$ est l'unique élément de $\mathcal{C}^+$ t.q. $\nabla f(\overline{x}) + C^t\overline{u} = 0$.

Chapitre 4

Equations différentielles

4.1 Introduction

On s'intéresse ici à la résolution numérique d'équations différentielles avec conditions initiales (ou problème de Cauchy) :

$$\left\{ \begin{array}{l} x'(t) = f(x(t), t) \ \ t > 0, \\ x(0) = \bar{x}_0. \end{array} \right. \tag{4.1}$$

où f est une fonction de $\mathbb{R}^n \times \mathbb{R}$ à valeurs dans $\mathbb{R}^n$, avec $n \geq 1$. L'inconnue est la fonction x de $\mathbb{R}$ dans $\mathbb{R}^n$. Souvent, t représente le temps, et on cherche donc x fonction de $\mathbb{R}_+$ à valeurs dans $\mathbb{R}^n$. On a donc affaire à un système différentiel d'ordre 1. De nombreux exemples de problèmes s'écrivent sous cette forme. Citons entre autres les lois qui régissent la cinétique d'un ensemble de réactions chimiques, ou encore les équations régissant la dynamique des populations. Notons qu'un système différentiel faisant intervenir des différentielles d'ordre supérieur peut toujours s'écrire sous la forme (4.1). Prenons par exemple l'équation du second ordre décrivant le comportement de l'amortisseur d'une voiture :

$$\left\{ \begin{array}{l} my'' + cy' + ky = 0, \\ y(0) = \bar{x}_0, \\ y'(0) = 0. \end{array} \right. \tag{4.2}$$

où m est la masse de la voiture, c le coefficient d'amortissement et k la force de rappel. L'inconnue y est le déplacement de l'amortisseur par rapport à sa position d'équilibre. Pour se ramener à un système d'ordre 1, on pose $x_1 = y$, $x_2 = y'$, et le système amortisseur s'écrit alors, avec comme inconnue $x = (x_1, x_2)^t$:

$$\left\{ \begin{array}{l} x'(t) = f(x(t), t), \\ x(0) = (\bar{x}_0, 0)^t, \end{array} \right. \quad \text{avec } f(x,t) = \left(\begin{array}{c} x_2, \\ -\dfrac{1}{m}(cx_2 + kx_1) \end{array} \right). \tag{4.3}$$

On rappelle que par le théorème de Cauchy-Lipschitz, si $f \in C^1(\mathbb{R}^n \times \mathbb{R}, \mathbb{R}^n)$ alors il existe $T_M > 0$ et $x \in C^2([0, T_M[, \mathbb{R}^n)$ solution maximale de (4.1), c'est–à–dire que x est solution de (4.1) sur $[0, T_M[$, et que s'il existe $\alpha > 0$ et $y \in C^2([0, \alpha[, \mathbb{R}^n)$ solution de (4.1) sur $[0, \alpha[$ alors $\alpha \leq T_M$ et $y = x$ sur $[0, \alpha[$. De plus, par le théorème d'explosion en temps fini, si $T_M < +\infty$ alors $|x(t)| \to +\infty$ quand $t \to T_M$.

Remarque 4.1 (Hypothèse sur f)**.** *On rappelle d'abord qu'une fonction φ de $\mathbb{R}$ dans $\mathbb{R}$ est dite lipschitzienne si*

$$\forall A > 0, \exists M_A \in \mathbb{R}_+ \textit{ tel que } , \forall (x,y) \in \mathbb{R}, |\varphi(x) - \varphi(y)| \leq M_A |x - y|. \tag{4.4}$$

Par exemple, toute fonction linéaire est lipschitzienne, et la fonction valeur absolue l'est aussi. Mais la fonction $x \mapsto x^2$ ne l'est pas. Il est donc utile d'introduire la notion plus faible suivante : on dit qu'une fonction φ de $\mathbb{R}$ dans $\mathbb{R}$ est dite lipschitzienne sur les bornés si

$$\forall A > 0, \exists M_A \in \mathbb{R}_+ \textit{ tel que } \forall (x,y) \in B_A^2 |\varphi(x) - \varphi(y)| \leq M_A |x - y|. \tag{4.5}$$

Par exemple la fonction $x \mapsto x^2$ est lipschitzienne sur les bornés, mais la fonction $x \mapsto \sqrt{x}$ ne l'est pas (sa dérivée explose en 0).

On peut se servir de cette notion pour affaiblir l'hypothèse sur f pour avoir existence et unicité d'une solution maximale de (4.1)*; on remplace l'hypothèse $f \in C^1(\mathbb{R}^n \times \mathbb{R}, \mathbb{R}^n)$ par $f \in C(\mathbb{R}^n \times \mathbb{R}, \mathbb{R}^n)$ "lipschitzienne sur les bornés", c'est–à–dire qui vérifie :*

$$\begin{array}{l} \forall A > 0, \exists M_A \in \mathbb{R}_+ \text{ tel que } \forall t \in [0, T[, \forall (x, y) \in B_A \times B_A, \\ |f(x,t) - f(y,t)| \leq M_A |x - y|. \end{array} \tag{4.6}$$

où $|.|$ désigne une norme sur $\mathbb{R}^n$ et B_A la boule de centre 0 et de rayon A. Il est clair que si $f \in C^1(\mathbb{R}^n \times \mathbb{R}, \mathbb{R}^n)$ alors f vérifie (4.6)*, alors qu'elle n'est évidemment pas forcément globalement lipschitzienne (prendre $f(x) = x^2$ pour s'en convaincre). De même la propriété* (4.6) *est encore vérifiée si f est "C^1 par morceaux", propriété toutefois délicate à démontrer dans le cas général.*

Exemple 4.2. *On suppose $n = 1$; soit la fonction f définie par $f(z,t) = z^2$. On considère le problème de Cauchy :*

$$\left\{ \begin{array}{l} \dfrac{dx}{dt}(t) = x^2(t) \\ x(0) = 1 \end{array} \right.$$

La fonction f est de classe C^1, donc lipschitzienne sur les bornés (mais pas globalement lipschitzienne). On peut donc appliquer le théorème de Cauchy-Lipschitz qui nous donne existence et unicité d'une solution maximale. On cherche alors à calculer une solution locale. Un calcul simple donne $x(t) = \frac{1}{1-t}$, et cette fonction tend vers $+\infty$ lorsque t tend vers 1^-. On en déduit que le temps maximal de la solution est $T_M = 1$, et on a donc comme solution maximale $x(t) = \frac{1}{1-t} \quad t \in [0, 1[$.

Exemple 4.3. *Supposons que $f \in C^1(\mathbb{R}^n \times \mathbb{R}, \mathbb{R}^n)$, et soit x la solution maximale de* (4.1) *sur $[0, T_M[$. On suppose que pour tout $0 < T < +\infty$, il existe $a_T > 0$ et $b_T > 0$ tels que*

$$|f(z,t)| \leq a_T |z| + b_T \quad \forall z \in \mathbb{R}^n, \ \forall t \in [0, T]$$

On a donc : $x'(t) \leq a_T |x(t)| + b_T$ pour tout t, en intégrant entre 0 et t, on obtient :

$$x(t) \leq a_T \int_0^t |x(s)| ds + + b_T t + \bar{x}_0,$$

et donc :

$$|x(t)| \leq a_T \int_0^t |x(s)| ds + |b_T| T + |\bar{x}_0|, \ \forall t \in [0, T[.$$

On peut alors appliquer le lemme de Gronwall[1] *à la fonction $t \mapsto |x(t)|$. On obtient que : $|x(t)| \leq (|b_T| T + |\bar{x}_0|) e^{a_T t}$ pour tout $t \in [0, T[$. On en déduit que x reste bornée sur tout intervalle $[0, T]$, $T \in \mathbb{R}$. Le temps d'existence T_M est donc égal à $+\infty$.*

Dans de nombreux cas, il n'est pas possible d'obtenir une expression analytique de la solution de (4.1). L'objet de ce chapitre est de présenter des méthodes pour obtenir des solutions (numériques) approchées de la solution de (4.1). Plus précisément, on adopte les notations et hypothèses suivantes :

1. On rappelle que le lemme de Gronwall permet de dire que si $\varphi \in C([0, T], \mathbb{R}_+)$ est telle que $\varphi(t) \leq \alpha \int_0^t \varphi(s) ds + \beta$, avec $\alpha \geq 0$, $\beta > 0$ alors $\varphi(t) \leq \beta e^{\alpha t}$ pour $t \in [0, T]$.

Notations et hypothèses :

$$\left\{\begin{array}{l} \text{Soit } f \text{ vérifiant l'hypothèse (4.6))} \\ \text{et soit } x \text{ solution maximale de (4.1) (définie sur } [0,T_M[), \\ \text{on se donne } T \in]0,T_M[, \text{ on cherche à calculer } x \text{ sur } [0,T], \\ \text{où } x \in C^1([0,T],\mathbb{R}^n) \text{ est solution de (4.1).} \\ \text{On se donne une discrétisation de } [0,T], \text{ i.e. } n \in \mathbb{N} \text{ et} \\ (t_0,t_1,\ldots,t_k) \in \mathbb{R}^{n+1} \text{ tels que } 0 < t_0 < t_1 < \ldots < t_k = T. \\ \text{On pose } h_k = t_{k+1} - t_k,\ \forall k = 0,\ldots,n-1, \\ \text{et } h = \max\{h_0,\ldots,h_{k-1}\}. \text{ Pour } k = 1,\ldots n, \text{ on cherche } x_k \\ \text{valeur approchée de } x(t_k) = \bar{x}_k, \\ \text{et on appelle } e_k = \bar{x}_k - x_k \text{ l'erreur de discrétisation.} \end{array}\right. \tag{4.7}$$

On cherche alors une méthode qui permette le calcul de x_k, pour $k = 1,\ldots,n$, et telle que la solution approchée ainsi calculée converge, en un sens à définir, vers la solution exacte. On cherchera de plus à évaluer l'erreur de discrétisation e_k, et plus précisément, à obtenir des estimations d'erreur de la forme $|e_k| \leq Ch^\alpha$, où C ne dépend que de la solution exacte (et pas de h) ; α donne alors l'ordre de la convergence.
On étudiera ici les méthodes de discrétisation des équations différentielles dits "schéma à un pas" qui s'écrivent sous la forme suivante :

Définition 4.4 (Schéma à un pas)**.** *Avec les hypothèses et notations* (4.7)*, on appelle schéma à un pas pour la résolution numérique de* (4.1)*, un algorithme de construction des valeurs* $(x_k)_{k=1,n}$ *qui s'écrit sous la forme suivante :*

$$\left\{\begin{array}{l} x_0 \text{ donné (approximation de } \bar{x}_0) \\ \\ \dfrac{x_{k+1}-x_k}{h_k} = \phi(x_k,t_k,h_k), \quad k = 0,\ldots n-1, \end{array}\right. \tag{4.8}$$

où ϕ *est une fonction de* $\mathbb{R}^n \times \mathbb{R}_+ \times \mathbb{R}_+$ *à valeurs dans* $\mathbb{R}$.

Dans la définition du schéma (4.8), il est clair que le terme $\frac{x_{k+1}-x_k}{h_k}$ est obtenu en cherchant une approximation de $x'(t_k)$ et que $\phi(x_k,t_k,h_k)$ est obtenu en cherchant une approximation de $f(x_k,t_k)$. Le schéma numérique est défini par cette fonction ϕ.

Exemples :

1. Schéma d'Euler explicite Le schéma d'Euler explicite est défini par (4.8) avec la fonction ϕ très simple suivante :
$$\phi(x_k,t_k,h_k) = f(x_k,t_k). \tag{4.9}$$

2. Schéma Euler implicite

$$\left\{\begin{array}{l} x_0 \text{ donné} \\ \\ \dfrac{x_{k+1}-x_k}{h_k} = f(x_{k+1},t_{k+1}). \quad k = 0,\ldots n-1, \end{array}\right. \tag{4.10}$$

On remarque que dans le schéma d'Euler implicite, le calcul de x_{k+1} n'est pas explicite, il est donné de manière implicite par (4.8) (d'où le nom du schéma). La première question à se poser pour ce type de schéma est l'existence de x_{k+1}. On montrera au théorème 4.15 que si l'hypothèse suivante est vérifiée :

$$D_1 f(y,t)z \cdot z \leq 0 \quad \forall y \in \mathbb{R}^n, \ \forall z \in \mathbb{R}^n, \quad \forall t \geq 0, \tag{4.11}$$

alors x_{k+1} calculé par (4.10) est bien défini en fonction de x_k, t_k, et h_k. On peut donc bien écrire le schéma (4.10) sous la forme (4.8) avec

$$\frac{x_{k+1}-x_k}{h_k} = \phi(x_k, t_k, h_k),$$

bien que la fonction ϕ ne soit définie ici qu'implicitement et non explicitement. Sous l'hypothèse (4.11), ce schéma entre donc bien dans le cadre des schémas (4.8) étudiés ici ; néanmoins, une propriété supplémentaire dite de "stabilité inconditionnelle", est vérifiée par ce schéma. Cette propriété peut s'avérer très importante en pratique et justifie une étude séparée (voir section 4.6).

4.2 Consistance, stabilité et convergence

Définition 4.5 (Consistance)**.** *On se place sous les hypothèses et notations* (4.7) *et on étudie le schéma* (4.8)*.*

1. *Pour $k = 0, \ldots, n$, on définit l'erreur de consistance du schéma* (4.8) *en t_k par :*

$$R_k = \frac{\overline{x}_{k+1}-\overline{x}_k}{h_k} - \phi(\overline{x}_k, t_k, h_k). \tag{4.12}$$

2. *Le schéma est consistant si*

$$\max\{|R_k|, k = 0 \ldots n-1\} \to 0 \quad \textit{lorsque } h \to 0. \tag{4.13}$$

3. *Soit $p \in \mathbb{N}^*$, le schéma est consistant d'ordre p s'il existe $C \in \mathbb{R}_+$ ne dépendant que de f,T, $\bar{x}_0$ (et pas de h) tel que $|R_k| \leq Ch^p$, $\forall k = 1, \ldots, n-1$.*

Donnons maintenant une condition nécessaire sur ϕ pour que le schéma (4.8) soit consistant.

Proposition 4.6 (Caractérisation de la consistance)**.** *Sous les hypothèses et notations* (4.7)*, si $\phi \in C(\mathbb{R}^n \times \mathbb{R}_+ \times \mathbb{R}_+, \mathbb{R}^n)$ et si $\phi(z,t,0) = f(z,t)$ pour tout $z \in \mathbb{R}^n$ et pour tout $t \in [0,T]$, alors le schéma* (4.8) *est consistant.*

DÉMONSTRATION – Comme $x \in C^1([0,T], \mathbb{R}^n)$ est la solution exacte de (4.1), on peut écrire que

$$x(t_{k+1}) - x(t_k) = \int_{t_k}^{t_{k+1}} x'(s)ds = \int_{t_k}^{t_{k+1}} f(x(s), s)ds.$$

On en déduit que

$$R_k = \frac{x(t_{k+1})-x(t_k)}{h_k} - \phi(\overline{x}_k, t_k, h_k) = \frac{1}{h_k}\int_{t_k}^{t_{k+1}} (f(x(s), s) - \phi(\overline{x}_k, t_k, h_k))ds.$$

Soit $\varepsilon > 0$, comme f est continue et $\phi(\overline{x}_k, t_k, 0) = f(\overline{x}_k, t_k)$, il existe η_1 tel que si $h_k \leq \eta_1$ alors : $|\phi(\overline{x}_k, t_k, h_k) - f(\overline{x}_k, t_k)| \leq \varepsilon$. On a donc par inégalité triangulaire,

$$|R_k| \leq \varepsilon + \frac{1}{h_k}\int_{t_k}^{t_{k+1}} |f(x(s), s) - f(\overline{x}_k, t_k)|ds.$$

La fonction $s \mapsto f(x(s), s)$ est continue et donc uniformément continue sur $[t_k, t_{k+1}]$. Il existe donc η_2 tel que si $h \leq \eta_2$, alors

$$\frac{1}{h_k}\int_{t_k}^{t_{k+1}} |f(x(s), s) - f(\overline{x}_k, t_k)|ds \leq \varepsilon.$$

On a ainsi montré que si $h \leq \min(\eta_1, \eta_2)$, alors $|R_k| \leq 2\varepsilon$, ce qui termine la preuve de la proposition. ∎

Notons que pour obtenir une consistance d'ordre $p > 1$, il est nécessaire de supposer que la solution x de (4.1) est dans $C^p(\mathbb{R}_+, \mathbb{R}^n)$.

Définition 4.7 (Stabilité). *Sous les hypothèses* (4.7)*, on dit que le schéma* (4.8) *est stable s'il existe* $h^* > 0$ *et* $R \in \mathbb{R}_+$ *tels que* $x_k \in B_R$ *pour tout* $k = 0, \dots, n$ *et pour tout* $h \in [0, h^*[$*, où* B_R *désigne la boule de centre 0 et de rayon R. On dit que le schéma est inconditionnellement stable si de plus,* $h^* = +\infty$.

Définition 4.8 (Convergence). *On se place sous les hypothèses et notations* (4.7)*.*

1. *Le schéma* (4.8) *est convergent si, lorsqu'on suppose* $|e_0| = 0$*, on a*
$$\max_{k=0,\dots,n} |e_k| \to 0 \text{ lorsque } h \to 0.$$
2. *Soit* $p \in \mathbb{N}^*$*, le schéma est convergent d'ordre* p *s'il existe* $C \in \mathbb{R}_+$ *ne dépendant que de* f*,*T*,* $\bar{x}_0$ *(et pas de* h*) tel que si on suppose* $|e_0| = 0$*, alors*
$$\max_{k=0,\dots,n} |e_k| \leq Ch^p.$$

Nous donnons à présent une notion de stabiité souvent utilisée dans les ouvrages classiques, mais qui ne semble pas être la plus efficace en termes d'analyse d'erreur (voir remarque 4.14.

Définition 4.9 (Stabilité par rapport aux erreurs). *Sous les hypothèses et notations* (4.7)*, on dit que le schéma* (4.8) *est stable par rapport aux erreurs s'il existe* $h^* \in \mathbb{R}_+^*$ *et* $K \in \mathbb{R}_+$ *dépendant de* $\bar{x}_0$*,* f *et* ϕ *(mais pas de* h*) tels que si* $h \leq h^*$ *et si*
$$\begin{aligned} x_{k+1} &= x_k + h_k \phi(t_k, x_k, h_k), \\ y_{k+1} &= y_k + h_k \phi(t_k, y_k, h_k) + \varepsilon_k, \end{aligned} \quad \text{pour } k = 0, \dots, n-1, \tag{4.14}$$
où $(\varepsilon_k)_{k \in \mathbb{N}} \subset \mathbb{R}_+$ *est donnée, alors*
$$|x_k - y_k| \leq K(|x_0 - y_0| + \sum_{i=0}^{k-1} |\varepsilon_i|), \text{ pour tout } k = 0, \dots, n-1.$$

On peut alors énoncer le théorème de convergence suivant, dont la démonstration, très simple, fait partie de l'exercice 173 page 268.

Théorème 4.10 (Convergence). *Sous les hypothèses et notations* (4.7)*, on suppose que le schéma* (4.8) *est stable par rapport aux erreurs au sens de la définition 4.9 et qu'il est consistant d'ordre* p *au sens de la définition 4.12. Alors il existe* $K \in \mathbb{R}_+$ *ne dépendant que de* $\bar{x}_0$*,* f *et* ϕ *(mais pas de* h*) tel que* $|e_k| \leq Kh^p + |e_0|$*, pour tout* $k = 0, \dots, n$.

Comme on l'a dit dans la remarque 4.14, ce théorème est d'une portée moins générale que le théorème 4.12 car il n'est pas toujours facile de montrer la stabilité par rapport aux erreurs, en dehors de la condition suffisante donnée dans la proposition qui suit, et qui est rarement vérifiée en pratique.

Proposition 4.11 (Condition suffisante de stabilité)**.** *Sous les hypothèses et notations* (4.7)*, une condition suffisante pour que le schéma* (4.8) *soit stable par rapport aux erreurs est que*

$$\begin{array}{l}\exists h^* > 0, \exists M > 0; \forall (x,y) \in \mathbb{R}^n \times \mathbb{R}^n, \forall h < h^*, \ \forall t \in [0,T], \\ |\phi(x,t,h) - \phi(y,t,h)| \leq M|x-y|.\end{array} \tag{4.15}$$

La démonstration de cette proposition est laissée en exercice (exercice 173 page 268).

4.3 Théorème général de convergence

Théorème 4.12. *On se place sous les hypothèses et notations* (4.7)*.*

1. *On suppose que le schéma* (4.8) *est consistant d'ordre p (i.e. il existe $p \in \mathbb{N}^*$ et $C \in \mathbb{R}_+$ ne dépendant que de T, f, $\bar{x}_0$ tel que $|R_k| \leq Ch^p$.)*
2. *On suppose qu'il existe $h^* > 0$ tel que pour tout $A \in \mathbb{R}_+^*$, il existe $M_A > 0$ tel que*

$$\begin{array}{l}\forall (y,z) \in B_A \times B_A, \ \forall t \in [0,T], \ \ \forall h \in [0,h^*], \\ |\phi(y,t,h) - \phi(z,t,h)| \leq M_A|y-z|,\end{array} \tag{4.16}$$

où B_A désigne la boule de rayon A. (Noter que cette hypothèse sur ϕ est semblable à l'hypothèse (4.6) *"Lipschitz sur les bornés" faite sur f dans la remarque 4.1 page 256).*

Alors il existe $h^{**} > 0$ ($h^{**} \leq h^*$), $\varepsilon > 0$, et $K > 0$ (ne dépendant que de f,$\bar{x}_0$,T,h^*,M_A) tels que si

$$0 < h \leq h^{**} \text{ et } |e_0| \leq \varepsilon,$$

alors

1. *le schéma est "stable", au sens où $x_k \in B_{2A}$ pour tout $k = 0, \dots n$, avec $A = \max\{|x(t)|, t \in [0,T]\} < +\infty$.*
2. *le schéma converge, et plus précisément, on a l'estimation d'erreur suivante : $|e_k| \leq K(h^p + |e_0|)$, pour tout $k = 0, \dots, n$. (En particulier si $e_0 = 0$ on a $|e_k| \leq Kh^p$ donc $e_k$tend vers 0 au moins comme h^p.)*

DÉMONSTRATION – Soit $x \in C^1([0,T], \mathbb{R}^n)$ solution de (4.1), et soit $A = \max\{|x(t)|, t \in [0,T]\} < +\infty$ (car x est continue et $[0,T]$ est compact). On a donc $\bar{x}_k \in B_A = \{y \in \mathbb{R}^n, |y| \leq A\}$.
On va "parachuter" ici un choix de ε et h^{**} qui permettront de montrer le théorème par récurrence sur k, on montrera dans la suite de la démonstration pourquoi ce choix convient. On choisit :

1. $h^{**} > 0$ tel que $Ce^{T(M_{2A}+1)}(h^{**})^p \leq \dfrac{A}{2}$, où M_{2A} est la constante de Lipschitz de ϕ sur B_{2A} dans l'hypothèse (4.16),
2. $\varepsilon > 0$ tel que $e^{TM_{2A}}\varepsilon \leq \dfrac{A}{2}$.

On va maintenant montrer par récurrence sur k que si $h \leq h^{**}$ et $|e_0| \leq \varepsilon$, alors :

$$\left\{\begin{array}{l} |e_k| \leq \alpha_k h^p + \beta_k |e_0|, \\ x_k \in B_{2A}, \end{array}\right. , \tag{4.17}$$

$$\text{avec } \alpha_k = Ce^{t_k M_{2A}}\,(1+h_0)\dots(1+h_{k-1}) \text{ et } \beta_k = e^{t_k M_{2A}}. \tag{4.18}$$

Si on suppose (4.17) vraie, on peut terminer la démonstration du théorème : en effet pour $x \geq 0$, on a $1 + x \leq e^x$, et donc :

$$(1+h_0)(1+h_1)\dots(1+h_{k-1}) \leq e^{h_0+h_1+\dots h_{k-1}} = e^{t_k} \leq e^T.$$

On en déduit que

$$\alpha_k \leq Ce^{TM_{2A}}e^T = Ce^{T(M_{2A}+1)}, \text{ et que } \beta_k \leq e^{TM_{2A}}.$$

On déduit alors de (4.17) et (4.18) que

$$\begin{aligned} |e_k| &\leq Ce^{T(M_{2A}+1)}h^p + e^{TM_{2A}}|e_0| \\ &\leq K(h^p + |e_0|) \text{ avec } K = \max(Ce^{T(M_{2A}+1)}, e^{T(M_2A)}), \end{aligned}$$

et que $x_k \in B_{2A}$. Il ne reste donc plus qu'à démontrer (4.17) par récurrence sur k.

— Pour $k = 0$, les formules (4.18) donnent $\alpha_0 = C$ et $\beta_0 = 1$. Or on a bien $|e_0| \leq \alpha_0 h^p + |e_0|$ car $C \geq 0$. De plus, par définition de e_0, on a $x_0 = \bar{x}_0 - e_0$, et donc : $|x_0| \leq |\bar{x}_0| + |e_0| \leq A + \varepsilon \leq A + \frac{A}{2} \leq 2A$ car, par hypothèse $\varepsilon e^{TM_{2A}} \leq \frac{A}{2}$ et donc $\varepsilon \leq \frac{A}{2}$. On en déduit que $x_0 \in B_{2A}$.

— Supposons maintenant que les relations (4.17) et (4.18) sont vraies jusqu'au rang k et démontrons qu'elles le sont encore au rang $k + 1$.

Par définition du schéma (4.8) et de l'erreur de consistance (4.12), on a :

$$\begin{aligned} x_{k+1} &= x_k + h_k\phi(x_k, t_k, h_k) \\ \bar{x}_{k+1} &= \bar{x}_k + h_k\phi(\bar{x}_k, t_k, h_k) + h_k R_k. \end{aligned}$$

On a donc $e_{k+1} = e_k + h_k(\phi(\bar{x}_k, t_k, h_k) - \phi(x_k, t_k, h_k)) + h_k R_k$, ce qui entraîne que

$$|e_{k+1}| \leq |e_k| + h_k|\phi(\bar{x}_k, t_k, h_k) - \phi(x_k, t_k, h_k)| + h_k|R_k|. \tag{4.19}$$

Comme $x_k \in B_{2A}$ et $\bar{x}_k \in B_A$, en utilisant la propriété (4.16) de ϕ, on a

$$|\phi(\bar{x}_k, t_k, h_k) - \phi(x_k, t_k, h_k)| \leq M_{2A}|\bar{x}_k - x_k|.$$

De plus, comme le schéma (4.8) est supposé consistant d'ordre p, on a $|R_k| \leq Ch^p$. On peut donc déduire de (4.19) que

$$|e_{k+1}| \leq |e_k|(1 + M_{2A}h_k) + h_k Ch^p,$$

et, en utilisant l'hypothèse de récurrence (4.17) :

$$|e_{k+1}| \leq (1 + h_k M_{2A})(\alpha_k h^p + \beta_k|e_0|) + h_k Ch^p.$$

Comme $1 + u \leq e^u$ pour tout $u \geq 0$, ceci entraîne

$$|e_{k+1}| \leq \bar{\alpha}_{k+1}h^p + \beta_{k+1}|e_0|,$$

où $\bar{\alpha}_{k+1} = \alpha_k e^{h_k M_{2A}} + Ch_k$ et $\beta_{k+1} = \beta_k e^{h_k M_{2A}} = e^{t_{k+1}M_{2A}}$. Or

$$\alpha_k = Ce^{t_k M_{2A}}(1 + h_0) + \cdots (1 + h_{k-1}) \geq C,$$

et donc

$$\bar{\alpha}_{k+1} \leq \alpha_k(e^{h_k M_{2A}} + h_k) \leq \alpha_k e^{h_k M_{2A}}(1 + h_k),$$

ce qui entraîne

$$Ce^{t_k M_{2A}} e^{h_k M_{2A}}(1 + h_0) \cdots (1 + h_{k-1})(1 + h_k) = \alpha_{k+1} \text{ car } t_k + h_k = t_{k+1}.$$

Donc

$$|e_{k+1}| \leq \alpha_{k+1}h^p + \beta_k|e_0|.$$

Il reste à montrer que $x_{k+1} \in B_{2A}$. On a

$$|x_{k+1}| \leq |\bar{x}_{k+1}| + |e_{k+1}| \leq A + |e_{k+1}| \text{ car } \bar{x}_k \in B_A.$$

Or on vient de montrer que $|e_{k+1}| \leq \alpha_{k+1}h^p + \beta_{k+1}|e_0|$, et

$$\alpha_{k+1} \leq Ce^{T(M_{2A+1})} \text{ et } \quad \beta_{k+1} \leq e^{TM_{2A}}.$$

Donc

$$|e_{k+1}| \leq Ce^{T(M_{2A}+1)}h^{**p} + e^{TM_{2A}}\varepsilon \leq \frac{A}{2} + \frac{A}{2}$$

car on a choisi h^{**} et ε pour !... On a donc finalement $|x_{k+1}| \leq A + A$, c'est–à–dire $x_{k+1} \in B_{2A}$.

On a donc bien montré (4.17) pour tout $k = 0, \ldots n$. Ce qui donne la conclusion du théorème. ∎

Remarque 4.13. *Dans le théorème précédent, on a montré que $x_k \in B_{2A}$ pour tout $k = 1, \ldots n$. Ceci est un résultat de* **stabilité** *(c'est–à–dire une estimation sur la solution approchée ne dépendant que des données T, $\bar{x}_0$, f et ϕ (ne dépend pas du pas de discrétisation h))* **conditionnelle**, *car on a supposé pour le démontrer que $h \leq h^{**}$, où h^{**} ne dépend que de T, $\bar{x}_0$, f et ϕ.*

Remarque 4.14 (Sur la démonstration du théorème de convergence)**.**
Dans la plupart des ouvrages d'analyse numérique, la convergence des schémas de discrétisation des équations différentielles est obtenue à partir de la notion de consistance et de la notion de stabilité par rapport aux erreurs (vue au paragraphe précédent, voir définition 4.9, et souvent appelée stabilité tout court). Il est en effet assez facile de voir (cf exercice 173 page 268) que si le schéma (4.8) *est consistant d'ordre* p *et stable par rapport aux erreurs comme défini dans la définition 4.9, alors il est convergent, et plus précisément,* $|e_k| \leq K(h^p + |e_0|)$*, pour tout* $k = 0, \ldots, n$.

Il y a deux avantages à utiliser plutôt le théorème précédent. D'une part, ce théorème est d'une portée très générale et s'applique facilement à de nombreux schémas, comme on le verra sur des exemples (voir section 4.4).
D'autre part la preuve de convergence par la notion de stabilité par rapport aux erreurs présente un défaut majeur : la seule condition suffisante qu'on connaisse en général pour montrer qu'un schéma est stable par rapport aux erreurs est que la fonction $\phi(.,t,h)$ *soit globalement lipschitzienne pour tout* $t \in [0,T]$ *et pour tout* $h \in [0,h^*]$ *(voir proposition 4.11). Ceci revient à dire, dans le cas du schéma d'Euler explicite par exemple, que* f *est globalement lipschitizienne. Cette hypothèse est très forte et rarement vérifiée en pratique. Bien sûr, comme la solution* x *de* (4.1) *est bornée sur* $[0,T]$*,* x *vit dans un compact et on peut toujours modifier* f *sur le complémentaire de ce compact pour la rendre globalement lipschitzienne. Cependant, cette manipulation nécessite la connaissance des bornes de la solution exacte, ce qui est souvent loin d'être facile à obtenir dans les applications pratiques.*

4.4 Exemples

On se place sous les hypothèses (4.7) et on étudie le schéma (4.8). On donne quatre exemples de schémas de la forme (4.8) :

Exemple 1 Euler explicite On rappelle que le schéma s'écrit (voir (4.9)) :

$$\frac{x_{k+1} - x_k}{h_k} = f(x_k, t_k),$$

On a donc $\phi(x_k, t_k, h_k) = f(x_k, t_k)$.
On peut montrer (voir exercice 172 page 268) que :

- si $f \in C^1(\mathbb{R}^n \times \mathbb{R}_+, \mathbb{R}^n)$, le schéma est consistant d'ordre 1,
- le théorème 4.12 s'applique $|e_k| \leq K(h + |e_0|)$ pour $h < h^{**}$. (La convergence est assez lente, et le schéma n'est stable que conditionnellement.)

Exemple 2 Euler amélioré Le schéma s'écrit :

$$\frac{x_{k+1} - x_k}{h_k} = f\left(x_k + \frac{h_k}{2} f(x_k, t_k), t_k + \frac{h_k}{2}\right) = \phi(x_k, t_k, h_k) \tag{4.20}$$

- si $x \in C^2(\mathbb{R}_+, \mathbb{R}^n)$,le schéma est consistant d'ordre 2,
- le théorème 4.12 s'applique et $|e_k| \leq K(h^2 + |e_0|)$ pour $h \leq h^{**}$.

La convergence est plus rapide.

Exemple 3 Heun

$$\frac{x_{k+1} - x_k}{h_k} = \frac{1}{2} f(x_k, t_k) + \frac{1}{2}[f(x_k + h_k f(x_k, t_k), t_{k+1})]. \tag{4.21}$$

- si $x \in C^2(\mathbb{R}_+, \mathbb{R}^n)$, le schéma est consistant d'ordre 2,
- Le théorème 4.12 s'applique et $|e_k| \leq K(h^2 + |e_0|)$, pour $h \leq h^{**}$.

Exemple 4 RK4 (Runge et Kutta, 1902) Les schémas de type Runge Kutta peuvent être obtenus en écrivant l'équation différentielle sous la forme $\bar{x}_{k+1} - \bar{x}_k = \int_{t_k}^{t_{k+1}} f(x(t), t)dt$, et en construisant un schéma numérique à

partir des formules d'intégration numérique pour le calcul approché des intégrales. Le schéma RK4 s'obtient à partir de la formule d'intégration numérique de Simpson :

A x_k connu,

$$\begin{array}{ll} x_{k,0} & = x_k \\ x_{k,1} & = x_k + \frac{h_k}{2} f(x_{k,0}, t_k) \\ x_{k,2} & = x_k + \frac{h_k}{2} f(x_{k,1}, t_k + \frac{h_k}{2}) \\ x_{k,3} & = x_k + h_k f(x_{k,2}, t_k + \frac{h_k}{2}) \\ \frac{x_{k+1} - x_k}{h_k} & = \frac{1}{6} f(x_{k,0}, t_k) + \frac{1}{3} f(x_{k,1}, t_k + \frac{h_k}{2}) \\ & + \frac{1}{3} f(x_{k,2}, t_k + \frac{h_k}{2}) + \frac{1}{6} f(x_{k,3}, t_{k+1}) \\ & = \phi(x_k, t_k, h_k) \end{array}$$

On peut montrer (avec pas mal de calculs...) que si $x \in C^4([0,T])$ alors le schéma est consistant d'ordre 4. Le théorème 4.12 s'applique et $|e_k| \leq K(h^4 + |e_0|)$, pour $h \leq h^{**}$.

4.5 Explicite ou implicite ?

On lit souvent que "les schémas implicites sont plus stables". Il est vrai que lorsque la condition (4.11) donnée plus haut est vérifiée, le schéma d'Euler implicite (4.10) est inconditionnellement stable, comme nous le verrons dans la section suivante. Il est donc naturel de le préférer au schéma explicite pour lequel on n'a qu'un résultat de stabilité conditionnelle. Cependant, dans le cas général, le choix n'est pas si évident, comme nous allons le voir sur des exemples, en étudiant le comportement respectif des schémas d'Euler explicite et implicite.

4.5.1 L'implicite gagne...

Prenons d'abord $f(x,t) = -x$, $n = 1$ et $x_0 = 1$. L'équation différentielle est donc :

$$\left\{ \begin{array}{l} \frac{dx}{dt} = -x(t), \\ x(0) = 1, \end{array} \right.$$

dont la solution est clairement donnée par $x(t) = e^{-t}$. On suppose que le pas est constant, c'est–à–dire $h_k = h \; \forall k$. Le schéma d'Euler explicite s'écrit dans ce cas :

$$\begin{array}{l} x_{k+1} = x_k - h x_k = (1-h) x_k \text{ et donc} \\ x_k = (1-h)^k, \quad \forall k = 0, \ldots, n, \text{ avec } nh = T. \end{array} \tag{4.22}$$

(On a donc n points de discrétisation.) La valeur x_k est censée être une approximation de $x(t_k) = e^{-t_k}$, et de fait, on remarque que pour $n = \frac{T}{h}$, on a

$$x_k = (1-h)^{T/h} \to e^{-T} \text{ quand } h \to 0.$$

Lorsqu'on cherche par exemple à obtenir le comportement de la solution d'une équation différentielle "dans les grands temps", on peut être amené à utiliser des pas de discrétisation relativement grands. Ceci peut être aussi le cas dans des problèmes de couplage avec d'autres équations, les "échelles de temps" des équations pouvant être très différentes pour les différentes équations. Que se passe-t-il dans ce cas ? Dans le cas de notre exemple, si on prend $h = 2$, on obtient alors $x_k = (-1)^k$, ce qui n'est clairement pas une bonne approximation de la solution. Un des problèmes majeurs est la perte de la positivité de la solution. Dans un problème d'origine physique où x serait une concentration ou une densité, il est indispensable que le schéma respecte cette positivité. On peut noter que

ceci n'est pas en contradiction avec le théorème 4.12 qui donne un résultat de convergence (*i.e.* de comportement lorsque h tend vers 0). Dans l'exemple présent, le schéma d'Euler explicite (4.22) ne donne pas une solution approchée raisonnable pour h grand.
Si on essaye maintenant de calculer une solution approchée à l'aide du schéma d'Euler implicite (4.10), on obtient

$$x_{k+1} = x_k - hx_{k+1}, \text{ c.à.d. } x_{k+1} = \frac{1}{1+h}x_k \text{ et donc}$$
$$x_k = \frac{1}{(1+h)^k}, \quad \forall k = 0, \dots n, \text{ avec } nh = T.$$

Dans ce cas, la solution approchée reste "proche" de la solution exacte, et positive, même pour des pas de discrétisation grands. On pourrait en conclure un peu hâtivement que le schéma implicite est "meilleur" que le schéma explicite. On va voir dans l'exemple qui suit qu'une telle conclusion est peu rapide.

4.5.2 L'implicite perd...

On considère maintenant le problème de Cauchy (4.1) avec $f(y,t) = +y$, $\bar{x}_0 = 1$. La solution est maintenant $x(t) = e^t$. Si on prend un pas de discrétisation constant égal à h, le schéma d'Euler explicite s'écrit :

$$x_{k+1} = x_k + hx_k = (1+h)x_k, \text{ c.à.d. } x_k = (1+h)^k.$$

On a donc

$$x_k = (1+h)^n \to e^T \text{ c.à.d. lorsque } n \to +\infty.$$

Contrairement à l'exemple précédent, la solution approchée donnée par le schéma d'Euler explicite reste "raisonnable" même pour les grands pas de temps.
Si on essaye maintenant de calculer une solution approchée à l'aide du schéma d'Euler implicite (4.10), on obtient

$$x_{k+1} = x_k + hx_{k+1}, \text{ c.à.d. } x_{k+1} = \frac{1}{1-h}x_k.$$

On remarque d'une part que le schéma implicite n'est pas défini pour $h = 1$, et que d'autre part si h est proche de 1 (par valeurs supérieures ou inférieures), la solution approchée "explose". De plus pour les valeurs de h supérieures à 1, on perd la positivité de la solution (pour $h = 2$ par exemple la solution approchée oscille entre les valeurs +1 et -1).
Dans le cadre de cet exemple, le choix explicite semble donc plus approprié.

4.5.3 Match nul

En conclusion de ces deux exemples, il semble que le "meilleur" schéma n'existe pas dans l'absolu. Le schéma de discrétisation doit être choisi en fonction du problème ; ceci nécessite une bonne compréhension du comportement des schémas en fonction des problèmes donnés, donc une certaine expérience...

4.6 Etude du schéma d'Euler implicite

On peut écrire le schéma d'Euler implicite sous la forme d'un schéma (4.8), si pour tout $k = 0 \dots n-1$, x_k étant donné, il existe x_{k+1} qui satisfait :

$$\frac{x_{k+1} - x_k}{h_k} = f(x_{k+1}, t_{k+1}), \quad k = 0, \dots, \ n-1.$$

On va montrer dans le théorème suivant que ceci est le cas si la condition (4.11) qu'on rappelle ici est vérifiée :

$$D_1 f(y,t)z \cdot z \leq 0, \quad \forall y, z \in \mathbb{R}^n, \quad \forall t \in [0,T].$$

On montrera aussi que sous cette hypothèse, on obtient un résultat de stabilité inconditionnelle pour le schéma d'Euler implicite.

Théorème 4.15. *On se place sous les hypothèses* (4.7) *et* (4.11)*. Alors*

1. $(x_k)_{k=0\dots n}$ *est bien définie par* (4.10)*,*
2. $|e_k| \leq |e_0| + h\displaystyle\int_0^{t_k} |x''(s)|ds, \quad \forall k = 0, \dots, n.$

DÉMONSTRATION – 1. Soit φ la fonction définie de $[0,1]$ à valeurs dans $\mathbb{R}^n$ par $\varphi(t) = f((1-t)y + tz)$; en écrivant que $\varphi(1) - \varphi(0) = \int_0^1 \varphi'(s)ds$, et en utilisant l'hypothèse (4.11), on déduit que :

$$(f(y,t) - f(z,t), (y-z)) \leq 0, \ \forall y, z \in \mathbb{R}^n, \ \forall t \in [0,T]. \tag{4.23}$$

On veut alors montrer que si x_k, h_k, t_k sont donnés, il existe un et un seul y tel que $\dfrac{y - x_k}{h_k} = f(y, t_k + h_k)$. A x_k et t_k fixés, soit F la fonction de $\mathbb{R}_+ \times \mathbb{R}^n$ à valeurs dans $\mathbb{R}^n$ définie par $F(h,y) = y - x_k - hf(y, t_k + h)$. On considère alors l'équation

$$F(h,y) = 0. \tag{4.24}$$

Pour $h = 0$, cette équation admet évidemment une unique solution $y = x_k$. Soit $I = \{\bar{h} \in \mathbb{R}_+^*$ t.q. (4.24) admette une solution pour tout $h < \bar{h}\}$. On va montrer par l'absurde que $\sup I = +\infty$, ce qui démontre l'existence et l'unicité de y solution de (4.24).

Supposons que $\sup I = H < +\infty$. Montrons d'abord que H est atteint. Soit $(h_k)_{k\in\mathbb{N}} \subset I$ telle que $h_k \to H$ lorsque $n \to +\infty$, alors la suite $(y_k)_{k\in\mathbb{N}}$ définie par $y_k = x_k + h_k f(y_k, t_k + h_k)$ est bornée : en effet,

$$y_k = x_k + h_k(f(y_k, t_k + h_k) - f(0, t_k + h)) + h_k f(0, t_k + h),$$

en prenant le produit scalaire des deux membres de cette égalité avec y_k et en utilisant (4.23) et l'inégalité de Cauchy-Schwarz, on obtient que :

$$|y_k| \leq |x_k| + H|f(0, t_k + h)|.$$

Il existe donc une sous-suite $(y_{n_k})_{k\in\mathbb{N}}$ qui converge vers un certain Y lorsque $n \to +\infty$. Par continuité de f, on a $Y = x_k + Hf(Y, t_k + H)$, et donc $H = \max I$.

Montrons maintenant que H ne peut pas être égal à $\sup I$. On applique pour cela le théorème des fonctions implicites à F définie en (4.24). On a bien $F(H,Y) = 0$, et $D_2F(H,Y) = \mathrm{Id} - HD_1f(Y, t_k + H)$ est inversible grâce à l'hypothèse (4.11). Donc il existe un voisinage de (H,Y) sur lequel (4.24) admet une solution, ce qui contredit le fait que $H = \sup I$.

2. La démonstration de 2 se fait alors par récurrence sur k. Pour $k = 0$ la relation est immédiate. L'hypothèse de récurrence s'écrit

$$|e_k| \leq |e_0| + h\int_0^{t_k} |x''(s)|ds.$$

Par définition du schéma (4.10) et de l'erreur de consistance, on a :

$$\begin{aligned} x_{k+1} &= x_k + h_k f(x_{k+1}, t_{k+1}), \\ \bar{x}_{k+1} &= \bar{x}_k + h_k f(\bar{x}_{k+1}, t_{k+1}) + h_k R_k. \end{aligned}$$

avec (par intégration par parties)

$$|R_k| \leq \int_{t_k}^{t_{k+1}} |x''(s)|ds.$$

On a donc :

$$e_{k+1} = \bar{x}_{k+1} - x_{k+1} = \bar{x}_k - x_k + h_k(f(\bar{x}_{k+1}, t_{k+1}) - f(x_{k+1}, t_{k+1})) + h_k R_k,$$

et donc

$$e_{k+1} \cdot e_{k+1} = e_k \cdot e_{k+1} + h_k R_k \cdot e_{k+1} + h_k(f(\bar{x}_{k+1}, t_{k+1}) - f(x_{k+1}, t_{k+1})) \cdot e_{k+1}.$$

Grâce à l'hypothèse (4.11) ceci entraîne (par (4.23)) que

$$|e_{k+1}| \leq |e_k| + h|R_k|,$$

et donc

$$|e_{k+1}| \leq |e_0| + h\int_0^{t_k} |x''(s)|ds + \int_{t_k}^{t_{k+1}} |x''(s)|ds = |e_0| + h\int_0^{t_{k+1}} |x''(s)|ds.$$

Ce qui démontre le point 2.

■

Remarque 4.16 (Stabilité inconditionnelle du schéma Euler implicite)**.** *Le schéma d'Euler implicite* (4.10) *est inconditionnellement stable, au sens où la suite* $(x_k)_{k=0,\ldots,n}$ *est majorée indépendamment de* h*. En effet :*

$$\begin{aligned} &|e_k| \leq |e_0| + T\int_0^T |x''(s)|ds = \beta, \\ &|x_k| \leq |\bar{x}_k| + \beta \leq \max\{|x(s)|,\ \ s \in [0,T]\} + \beta = \gamma. \end{aligned}$$

4.7 Exercices

Exercice 168 (Condition de Lipschitz et unicité)**.** *Corrigé en page 274*

Pour $a \geq 0$, on définit la fonction $\varphi_a : \mathbb{R}_+ \to \mathbb{R}_+$ par : $\varphi_a(x) = x^a$. Pour quelles valeurs de a la fonction φ_a est-elle lipschitzienne sur les bornés ?

On considère le problème de Cauchy suivant :

$$\begin{aligned} &y'(t) = \varphi_a(y(t)),\ t \in [0, +\infty[\\ &y(0) = 0. \end{aligned} \tag{4.25}$$

Montrer que si φ_a est lipschitzienne sur les bornés alors le problème de Cauchy (4.25) admet une solution unique, et que si φ_a n'est pas lipschitzienne sur les bornés alors le problème de Cauchy (4.25) admet au moins deux solutions.

Exercice 169 (Fonctions lipschitziennes sur les bornés)**.**

Les fonctions suivantes sont elles lipschitziennes sur les bornés ?

1. $$\begin{aligned} \varphi_1 :\ & \mathbb{R} \to \mathbb{R} \\ & x \mapsto \min(x^2, \sqrt{(x^2+1)}) \end{aligned}$$
2. $$\begin{aligned} \varphi_2 :\ & \mathbb{R}^2 \to \mathbb{R}^2 \\ & (x,y) \mapsto (x^2 - xy, |y + 2xy|) \end{aligned}$$
3. $$\begin{aligned} \varphi_3 :\ & \mathbb{R}_+^2 \to \mathbb{R}_+^2 \\ & (x,y) \mapsto (\sqrt{x+y}, x^2 + y^2) \end{aligned}$$

Exercice 170 (Loi de Malthus)**.** *Corrigé en page 275*

On considère une espèce dont la population (i.e. le nombre d'individus) a doublé en 100 ans et triplé en 200 ans. Montrer que cette population ne peut pas satisfaire la loi de Malthus (on rappelle que la loi de Malthus s'écrit $p'(t) = ap(t)$ avec $a > 0$ indépendant de t).

Exercice 171 (Histoire de sardines)**.** *Corrigé en page 275*

Une famille de sardines tranquillement installées dans les eaux du Frioul a une population qui croît selon la loi de Malthus, $p'(t) = 4p(t)$ où t est exprimé en jours. A l'instant $t = 0$, un groupe de bonites voraces vient s'installer dans ces eaux claires, et se met à attaquer les pauvres sardines. Le taux de perte chez ces dernières s'élève à $10^{-4}p^2(t)$ par jour, où $p(t)$ est la population des sardines au temps t. De plus, au bout d'un mois de ce traitement, suite au dégazement intempestif d'un super tanker au large du phare du Planier, les sardines décident d'émigrer vers des eaux plus claires au rythme de 10 pour cent de la population par jour (on supposera que les bonites sont insensibles au gas oil, et donc que le nombre de bonites reste constant...).

1. Modifier la loi de Malthus pour prendre en compte les deux phénomènes.

2. En supposant qu'à $t = 0$ le nombre de sardines est de 1 million, calculer le nombre de sardines pour $t > 0$. Quel est le comportement de $p(t)$ à l'infini ?

Exercice 172 (Consistance et ordre des schémas)**.** *Corrigé en page 276*

On reprend les hypothèses et notations (4.7).

1. On suppose que $f \in C^1(\mathbb{R}^n \times \mathbb{R}_+, \mathbb{R}^n)$. Montrer que le schéma d'Euler explicite s'écrit (4.9)) est consistant et convergent d'ordre 1.
2. On suppose que $f \in C^1(\mathbb{R}^n \times \mathbb{R}_+, \mathbb{R}^n)$. Montrer que les schémas d'Euler amélioré (4.4), et de Heun (4.4) sont consistants et convergents d'ordre 2.
3. On suppose que $f \in C^4(\mathbb{R}^n \times \mathbb{R}_+, \mathbb{R}^n)$.Montrer que le schéma RK4 est consistant et convergent d'ordre 4 (pour les braves...)
4. On suppose que $f \in C^1(\mathbb{R}^n \times \mathbb{R}_+, \mathbb{R}^n)$.Montrer que le schéma d'Euler implicite est consistant d'ordre 1.

Exercice 173 (Stabilité par rapport aux erreurs et convergence)**.** *Corrigé donné en page 277*

On se place sous les hypothèses et notations (4.7) page 258, et on considère le schéma (4.8) page 258 pour la résolution numérique de l'équation différentielle (4.1) page 256.

1. Montrer que si le schéma (4.8) est stable par rapport aux erreurs au sens de la définition 4.9 page 260, et qu'il est consistant d'ordre p au sens de la définition 4.5 page 259, alors il existe $K \in \mathbb{R}_+$ ne dépendant que de $\bar{x}_0$, f et ϕ (mais pas de h) tel que $|e_k| \leq Kh^p + |e_0|$, pour tout $k = 0 \dots n$. En déduire que si $e_0 = 0$ le schéma converge.
2. Montrer que si ϕ est globalement lipschitzienne, c.à.d. si

$$\begin{array}{l}\exists h^* > 0, \exists M > 0;\ \forall (x,y) \in \mathbb{R}^n \times \mathbb{R}^n, \forall h < h^*,\ \forall t \in [0,T],\\ |\phi(x,t,h) - \phi(y,t,h)| \leq M|x-y|,\end{array}$$

alors le schéma est stable par rapport aux erreurs.

Exercice 174 (Schéma d'ordre 2)**.**

Soit $f \in C^2(\mathbb{R}^n \times \mathbb{R}, \mathbb{R}^n)$, $n \geq 1$, $\bar{x}_0 \in \mathbb{R}^n$, et soit x solution maximale de (E) (définie sur $[0, T_M[$) :

$$\left\{ \begin{array}{l} \dfrac{dx}{dt}(t) = f(x(t), t),\ \ t > 0, \\ x(0) = \bar{x}_0. \end{array} \right. \qquad (E)$$

On se donne $T \in]0, T_M[$, et une discrétisation de $[0, T]$, définie par $n \in \mathbb{N}$ et $(t_0, t_1, \dots, t_k) \in \mathbb{R}^{n+1}$ tels que $0 = t_0 < t_1 < \dots < t_k = T$. On pose $h_k = t_{k+1} - t_k$, $\forall k = 0, \dots, n-1$.
On considère le schéma de discrétisation

$$\left\{ \begin{array}{l} x_0 \text{ donné (approximation de } \bar{x}_0), \\ \frac{x_{k+1}-x_k}{h_k} = \frac{1}{2}[f(x_k, t_k) + f(x_k + h_k f(x_k, t_k), t_{k+1})], \quad k = 0, \dots n-1, \end{array} \right.$$

pour la résolution numérique de l'équation différentielle (E). Montrer que ce schéma est convergent d'ordre 2.

Exercice 175 (Algorithme du gradient à pas fixe et schéma d'Euler)**.**

Soit $f \in C^2(\mathbb{R}^n, \mathbb{R})$ strictement convexe et t.q. $f(x) \to \infty$ quand $|x| \to \infty$. Soit $x_0 \in \mathbb{R}^n$. On considère les 2 problèmes :

$$\begin{array}{l} \overline{x} \in \mathbb{R}^n, \\ f(\overline{x}) \leq f(x),\ \forall x \in \mathbb{R}^n, \end{array} \tag{4.26}$$

$$\begin{array}{l} \dfrac{dx}{dt}(t) = -\nabla f(x(t)),\ t \in \mathbb{R}^+, \\ x(0) = x_0. \end{array} \tag{4.27}$$

1. Montrer que l'algorithme du gradient à pas fixe (de pas noté ρ) pour trouver la solution de (4.26) (avec point de départ x_0) est le schéma d'Euler explicite pour la résolution approchée de (4.27) (avec pas de temps ρ).
2. Montrer qu'il existe un unique $\overline{x}$ solution de (4.26).
3. Montrer que (4.27) admet une et une seule solution sur $\mathbb{R}_+$ et que cette solution converge vers $\overline{x}$ (solution de (4.26)) quand $t \to \infty$.
4. Expliciter le cas $f(x) = (1/2)Ax \cdot x - b \cdot x$ avec A symétrique définie positive et $b \in \mathbb{R}^n$.

Exercice 176 (Méthode de Taylor)**.** *Corrigé en page 278*

Soit $f \in C^\infty(\mathbb{R} \times \mathbb{R}, \mathbb{R})$, et $\bar{x}_0 \in \mathbb{R}$, on considère le problème de Cauchy (4.1), dont on cherche à calculer la solution sur $[0, T]$, où $T > 0$ est donné. On se donne un pas de discrétisation $h = \frac{T}{n}$, avec $n \geq 1$.

Dans toute la suite, on note $x^{(k)}$ la dérivée d'ordre k de x, $\partial_i^k f$ la dérivée partielle d'ordre k de f par rapport à la i-ème variable, $\partial_i^k \partial_j^\ell f$ la dérivée partielle de f d'ordre k par rapport à la i-ème variable et d'ordre ℓ par rapport à la j-ème variable (on omettra les symboles k et ℓ lorsque $k = 1$ ou $\ell = 1$).
On définit $f^{(m)} \in C^\infty(\mathbb{R} \times \mathbb{R}, \mathbb{R})$ par

$$\begin{array}{l} f^{(0)} = f, \\ f^{(m+1)} = \left(\partial_1 f^{(m)}\right) f + \partial_2 f^{(m)},\ \text{pour } m \geq 0. \end{array} \tag{4.28}$$

1. Montrer que pour tout $m \in \mathbb{N}$, la solution x du problème de Cauchy (4.1) satisfait :

$$x^{(m+1)}(t) = f^{(m)}(x(t), t).$$

2. Calculer $f^{(1)}$ et $f^{(2)}$ en fonction des dérivées partielles $\partial_1 f$, $\partial_2 f$,$\partial_1 \partial_2 f$, $\partial_1^2 f$, $\partial_2^2 f$, et de f.

On définit pour $p \geq 1$ la fonction ψ_p de $\mathbb{R} \times \mathbb{R}$ à valeurs dans $\mathbb{R}$ par

$$\psi_p(y, t, h) = \sum_{j=0}^{p-1} \frac{h^j}{(j+1)!} f^{(j)}(y, t).$$

Pour $k = 1, \ldots, n$, on note $t_k = kh$. On définit alors la suite $(x_k)_{k=0,n+1} \subset \mathbb{R}$ par

$$\left\{ \begin{array}{l} x_0 = \bar{x}_0, \\ x_{k+1} = x_k + h\psi_p(x_k, t_k, h),\ \text{pour } k = 1, \ldots, n. \end{array} \right. \tag{4.29}$$

3. Montrer que dans le cas $p = 1$, le système (4.29) définit un schéma de discrétisation vu en cours, dont on précisera le nom exact.

4. On suppose, dans cette question uniquement, que $f(y, t) = y$ pour tout $(y, t) \in \mathbb{R} \times \mathbb{R}$, et que $\bar{x}_0 = 1$.

4.a/ Calculer $\psi_p(y, t, h)$ en fonction de y et h.

4.b/ Montrer que $x_k = \left(\sum_{j=0}^{p} \frac{h^j}{j!} \right)^k$, pour $k = 1, \ldots, n$.

4.c/ Montrer que $|x_k - x(t_k)| \leq \frac{h^p}{(p+1)!} t_k e^{t_k}$.

5. On revient au cas général $f \in C^\infty(\mathbb{R} \times \mathbb{R}, \mathbb{R})$. Montrer que le schéma (4.29) est consistant d'ordre p. Montrer qu'il existe $\bar{h} > 0$, et $C > 0$ ne dépendant que de $\bar{x}_0$, T et f, tels que si $0 < h < \bar{h}$, alors $|x_k - x(t_k)| \leq Ch^p$, pour tout $k = 0, \ldots, n+1$.

Exercice 177 (Schéma d'Euler implicite)**.**

Soit $f \in C^1(\mathbb{R}, \mathbb{R})$ telle que $f(y) < 0$ pour tout $y \in]0,1[$ et $f(0) = f(1) = 0$. Soit $y_0 \in]0,1[$. On considère le problème suivant :

$$y'(t) = f(y(t)),\ t \in \mathbb{R}_+, \tag{4.30}$$
$$y(0) = y_0. \tag{4.31}$$

Question 1.
1.1 Soit $T \in \overline{\mathbb{R}}_+$; on suppose que $y \in C^1([0,T[, \mathbb{R})$ est solution de (4.30)-(4.31). Montrer que $0 < y(t) < 1$ pour tout $t \in [0,T[$ (On pourra raisonner par l'absurde et utiliser le théorème d'unicité).

1.2 Montrer qu'il existe une unique fonction $y \in C^1([0,+\infty[, \mathbb{R})$ solution de (4.30)-(4.31) et que y est une fonction strictement positive et strictement décroissante.

Dans les questions suivantes on désigne par y cette unique solution définie sur $[0,+\infty[$.

Question 2.
2.1 Montrer que y admet une limite $\ell \in \mathbb{R}$ lorsque $t \to +\infty$.

2.2 Montrer que $\ell = 0$. (On pourra remarquer que, pour tout $t \geq 0$, on a $y(t+1) = y(t) + \int_t^{t+1} f(y(s))ds$).

Question 3. Soit $y_0 \in]0,1[$, on cherche à approcher la solution exacte de (4.30)-(4.31) par le schéma d'Euler implicite de pas $h \in \mathbb{R}_+^*$, qui s'écrit :

$$y_{n+1} = y_k + hf(y_{n+1}), n \in \mathbb{N}. \tag{4.32}$$

3.1 Soit $a \in]0,1[$. Montrer qu'il existe $b \in]0,1[$ t.q.

$$\frac{b-a}{h} = f(b).$$

En déduire que pour $y_0 \in]0,1[$ fixé, il existe $(y_k)_{n \in \mathbb{N}}$ solution du schéma d'Euler implicite (4.32) telle que $y_k \in]0,1[$ pour tout $n \in \mathbb{N}$.

3.2 Soit $(y_k)_{n \in \mathbb{N}}$ une suite construite à la question 3.1. Montrer que cette suite est décroissante et qu'elle tend vers 0 lorsque n tend vers l'infini.

Question 4. On suppose dans cette question que

$$f'(0) = -\alpha < 0$$

Soit $\beta \in]0,\alpha[$.

4.1 Montrer que pour t suffisamment grand,

$$\frac{f(y(t))}{y(t)} < -\beta.$$

4.2 En déduire qu'il existe $C \in \mathbb{R}_+$ t.q.

$$y(t) \leq Ce^{-\beta t}, \forall t \geq 0.$$

4.3 Montrer qu'il existe $C \in \mathbb{R}_+^*$ t.q. la solution du schéma d'Euler implicite construite à la question 3 vérifie :

$$y_k \leq C\Big(\frac{1}{1+h\beta}\Big)^n, \ \forall n \in \mathbb{N}.$$

Exercice 178 (Méthodes semi-implicite et explicite)**.** *Corrigé en page 280*

On s'intéresse dans cet exercice au système différentiel :

$$\left\{ \begin{array}{ll} x_1'(t) = -x_1(t) - x_1(t)x_2(t), & \\ x_2'(t) = -\dfrac{x_2(t)}{x_1(t)}, & t > 0, \end{array} \right. \tag{4.33}$$

avec les conditions initiales

$$x_1(0) = a, \ x_2(0) = b, \tag{4.34}$$

où a et b appartiennent à l'intervalle $]0, 1[$.

1. On pose $x = (x_1, x_2)^t$. Montrer que le système (4.33)-(4.34) s'écrit

$$\left\{ \begin{array}{l} x'(t) = f(x(t)), \ t > 0, \\ x(0) = (a, b)^t, \end{array} \right. \tag{4.35}$$

   avec $f \in C^1((\mathbb{R}_+^*)^2, \mathbb{R}^2)$.

2. Les questions suivantes sont facultatives : elles permettent de montrer que le système (4.35) admet une solution maximale $x \in C^1([0, +\infty[, (\mathbb{R}_+^*)^2)$. Le lecteur pressé par le temps pourra admettre ce résultat et passer à la question 3.

   (a) Montrer qu'il existe $\alpha > 0$ et $x \in C^1([0, \alpha[, (\mathbb{R}_+^*)^2)$ solution de (4.35) (on pourra utiliser, ainsi que dans la question suivante, le fait que f est lipschitzienne sur tout pavé $[\varepsilon, A]^2$ avec $0 < \varepsilon \leq A < +\infty$).

   (b) Soit $\beta > 0$, montrer qu'il existe au plus une solution de (4.35) appartenant à $C^1([0, \beta[, (\mathbb{R}_+^*)^2)$.

   (c) Montrer que le système (4.35) admet une solution maximale $x \in C^1([0, +\infty[, (\mathbb{R}_+^*)^2)$. (Cette question est difficile : il faut raisonner par l'absurde, supposer que $T < +\infty$, montrer que dans ce cas x n'est pas solution maximale…)

   (d) Montrer que la solution maximale x vérifie $x \in C^\infty([0, +\infty[, (\mathbb{R}_+^*)^2)$.

3. On considère le schéma suivant de discrétisation du système (4.33)-(4.34) : soit k le pas de discrétisation, choisi tel que $0 < k < \frac{1}{2}$.

$$\left\{ \begin{array}{l} \dfrac{x_1^{(k+1)} - x_1^{(k)}}{k} = -x_1^{(k)} - x_1^{(k)} x_2^{(k+1)}, \\ \dfrac{x_2^{(k+1)} - x_2^{(k)}}{k} = -\dfrac{x_2^{(k+1)}}{x_1^{(k)}}, \\ x_1^{(0)} = a, \ x_2^{(0)} = b. \end{array} \right. \tag{4.36}$$

   (a) Montrer par récurrence sur n que les suites $(x_1^{(k)})_{n\in\mathbb{N}}$ et $(x_2^{(k)})_{n\in\mathbb{N}}$ données par (4.36) sont bien définies, décroissantes et strictement positives.

   (b) Montrer que le schéma numérique (4.36) s'écrit sous la forme

$$\frac{x^{(k+1)} - x^{(k)}}{k} = \phi(x^{(k)}, k), \tag{4.37}$$

   avec $x^{(k)} = (x_1^{(k)}, x_2^{(k)})^t$, $\phi \in C^\infty((\mathbb{R}_+^*)^2 \times \mathbb{R}_+, \mathbb{R}^2)$ et $\phi(x, 0) = f(x)$.

(c) (Consistance)
Soit $T > 0$. Pour $n \in \mathbb{N}$, on note $t_k = nk$. Montrer qu'il existe $C(T) \in \mathbb{R}_+$ tel que

$$\frac{x(t_{n+1}) - x(t_n)}{k} = \phi(x(t_n), k) + R_k^{(k)}, \text{ pour tout } n \text{ tel que } nk \leq T, \tag{4.38}$$

avec $|R_k^{(k)}| \leq C(T)k$.

(d) (Stabilité)
Soit $T > 0$.
(i) Montrer que $x_1^{(k)} \geq (1 - k - kb)^{\frac{T}{k}}$ pour tout entier n tel que $nk \leq T$.
(ii) Montrer que

$$(1 - k - kb)^{\frac{T}{k}} \to e^{-(1+b)T} \text{ lorsque } k \to 0,$$

et en déduire que $\inf_{0<k<\frac{1}{2}}(1 - k - kb)^{\frac{T}{k}} > 0$.
(iii) En déduire qu'il existe $a(T) > 0$ et $b(T) > 0$ tels que

$$\begin{cases} a(T) \leq x_1^{(k)} \leq a, \\ b(T) \leq x_2^{(k)} \leq b, \end{cases} \text{ pour tout } n \text{ tel que } nk \leq T. \tag{4.39}$$

(e) (Convergence)
Soit $T > 0$. Montrer qu'il existe $D(T) \in \mathbb{R}_+$ tel que

$$|x^{(k)} - x(t_k)| \leq D(T)k, \text{ pour tout } n \text{ tel que } nk \leq T. \tag{4.40}$$

En déduire la convergence du schéma (4.36).

(f) On remplace maintenant le schéma (4.36) par le schéma d'Euler explicite pour le système (4.35). Ecrire ce schéma. Montrer que pour tout pas de discrétisation $k > 0$, il existe des valeurs de n telles que $x_1^{(k)} \leq 0$ ou $x_2^{(k)} \leq 0$. (On pourra montrer que si $x_1^{(k)} > 0$ et $x_2^{(k)} > 0$ pour tout $n \in \mathbb{N}$, alors $x_1^{(k)}$ tend vers 0 lorsque n tend vers $+\infty$, et donc qu'il existe n tel que $x_2^{(k)} \leq 0$, ce qui contredit l'hypothèse). Commenter.

Exercice 179.

Soit $f \in C^2(\mathbb{R}^n \times \mathbb{R}_+, \mathbb{R}^n)$, $T > 0$, et $y^{(0)} \in \mathbb{R}^n$. On désigne par $(.,.)$ le produit scalaire euclidien sur $\mathbb{R}^n$ et $\|.\|$ la norme associée. On suppose que :

$$\forall (y, z) \in (\mathbb{R}^n)^2, \ (f(y, t) - f(z, t), y - z) \leq 0. \tag{4.41}$$

On considère le système différentiel :

$$y'(t) = f(y(t), t) \, \forall t \in [0, T[, \tag{4.42}$$

$$y(0) = y^{(0)}. \tag{4.43}$$

1. Montrer que pour tout $y \in \mathbb{R}^n$ et $t \in [0, T[$, on a :

$$(f(y, t), y) \leq \frac{1}{2}(\|f(0, t)\|^2 + \|y\|^2). \tag{4.44}$$

En déduire qu'il existe une unique solution $y \in C^1([0, T[, \mathbb{R}^n)$ vérifiant (4.42)-(4.43).

On se propose de calculer une solution approchée de y sur $[0,T]$. Pour cela, on considère une discrétisation de l'intervalle $[0,T]$ de pas constant, noté h, avec $h=\frac{T}{n}$, où $n\in\mathbb{N}^*$. Pour $k=0,\dots,n$, on note $t_k=kh$, et on se propose d'étudier l'algorithme suivant, où $0\leq\theta\leq 1$.

$$y_0\in\mathbb{R}^n \text{ est donné} \tag{4.45}$$

$$y_{k,1}=y_k+\theta h f(y_{k,1},t_k+\theta h),\ \text{ pour } k=0,\dots,n-1, \tag{4.46}$$

$$y_{k+1}=y_k+hf(y_{k,1},t_k+\theta h)\ \text{ pour } k=0,\dots,n-1, \tag{4.47}$$

2. Montrer qu'il existe une unique solution $(y_k)_{k=0,\dots,n}\subset\mathbb{R}^n$ de (4.45)-(4.46)-(4.47).

Pour $k=0,\dots,n-1$, on pose $y(t_k)=\overline{y}_k$, où y est la solution exacte de (4.42)-(4.43), $t_{k,1}=t_k+\theta h$, on définit $\tilde{y}_{k,1}$ par :

$$\tilde{y}_{k,1}=\overline{y}_k+\theta h f(\tilde{y}_{k,1},t_{k,1}), \tag{4.48}$$

et on définit l'erreur de consistance R_k du schéma (4.45)-(4.46)-(4.47) au point t_k par :

$$R_k=\frac{\overline{y}_{k+1}-\overline{y}_k}{h}-f(\tilde{y}_{k,1},t_{k,1}) \tag{4.49}$$

3. Pour $k=0,\dots,n$, on pose $\overline{y}_{k,1}=y(t_{k,1})$, et, pour $k=0,\dots,n-1$ on pose :

$$\tilde{R}_k=\frac{1}{h}(\overline{y}_{k,1}-\overline{y}_k)-\theta f(\overline{y}_{k,1},t_{k,1}). \tag{4.50}$$

Montrer que pour tout $k=0,\dots,n-1$:

$$\tilde{y}_{k,1}-\overline{y}_{k,1}=\theta h\big(f(\tilde{y}_{k,1},t_{k,1})-f(\overline{y}_{k,1},t_{k,1})\big)+h\tilde{R}_k, \tag{4.51}$$

En déduire qu'il existe C_1 ne dépendant que de y et de T t.q. : $\|\tilde{y}_{k,1}-\overline{y}_{k,1}\|\leq C_1h^2$.

4. Montrer qu'il existe C_2 ne dépendant que de f,y et T t.q.

$$\|\overline{y}_{k+1}-\overline{y}_k-hf(\overline{y}_k,t_{k,1})-hR_k\|\leq C_2h^3,\ \forall k=0,\dots,n-1. \tag{4.52}$$

5. Déduire des questions précédentes qu'il existe C_3 ne dépendant que de y, f et T t.q. :

$$\|R_k\|\leq C_3((\theta-\frac{1}{2})h+h^2) \tag{4.53}$$

et en déduire l'ordre du schéma (4.45)-(4.46)-(4.47).

6. Montrer que pour tout $k=1,\dots,n$, on a :

$$\big(\overline{y}_k-y_k,f(y_{k,1},t_{k,1})-f(\tilde{y}_{k,1},t_{k,1})\big)\leq-\theta h\|f(y_{k,1},t_{k,1})-f(\tilde{y}_{k,1},t_{k,1})\|^2. \tag{4.54}$$

7. Montrer que pour tout $k=0,\dots,n$, on a :

$$\|e_{k+1}-hR_k\|^2=\|e_k\|^2+2h(f(y_{k,1},t_{k,1})-f(\tilde{y}_{k,1},t_{k,1}),e_k)+h^2\|f(y_{k,1},t_{k,1})-f(\tilde{y}_{k,1},t_{k,1})\|^2. \tag{4.55}$$

8. Montrer que si $\theta\geq\frac{1}{2}$, on a :

$$\|e_k\|\leq\|e_0\|+C_3(h^2+(\theta-\frac{1}{2})h),\ \forall k=1,\dots,n. \tag{4.56}$$

9. Soient $(\varepsilon_k)_{k\in\mathbb{N}} \subset \mathbb{R}^n$ donnée et $(z_k)_{k\in\mathbb{N}} \subset \mathbb{R}^n$ définie par :

$$z_0 \in \mathbb{R}^n \text{ donné} \tag{4.57}$$

$$z_{k,1} = z_k + \theta h f(z_{k,1}, t_{k,1}), \text{ pour } k = 0, \ldots, n-1, \tag{4.58}$$

$$z_{k+1} = z_k + +\varepsilon_k + h f(z_{k,1}, t_{k,1}) \text{ pour } k = 0, \ldots, n-1, \tag{4.59}$$

En s'inspirant des questions 6 et 7, montrer que si $\theta \geq \frac{1}{2}$, on a :

$$\|y_{k+1} - z_{k+1} + \varepsilon_k\|^2 \leq \|y_k - z_k\|^2, \tag{4.60}$$

et en déduire que

$$\|y_k - z_k\| \leq \|y_0 - z_0\| + \sum_{i=0}^{k-1} \|\varepsilon_i\|. \tag{4.61}$$

Exercice 180 (Le pendule)**.**

On considère l'équation différentielle suivante, qui décrit le mouvement d'un pendule.

$$\begin{array}{l} x''(t) + \sin x(t) = 0, \qquad t > 0, \\ x(0) = \xi, \\ x'(0) = 0, \end{array} \tag{4.62}$$

où $\xi \in [0, 2\pi[$ est la position initiale du pendule.

1. Ecrire le problème sous la forme d'un système différentiel d'ordre 1.

2. Ecrire la méthode d'Euler implicite pour la résolution du système différentiel d'ordre 1 obtenu à la question 1, donnant les approximations x_{n+1} et y_{n+1} de x et y au temps t_{n+1} en fonction des approximations x_k et y_k de x et y au temps t_n. En déduire qu'à chaque pas de temps, on doit résoudre un système non linvaire de la forme

$$\begin{array}{l} x - ay = \alpha, \\ a \sin x + y = \beta. \end{array} \tag{4.63}$$

On exprimera a, α et β en fonction du pas de temps δt et des approximations x_k et y_k de x et y au temps $t_k = n\delta t$.

3. Mettre le système (4.63) sous la forme $F(X) = 0$ où F est une fonction de $\mathbb{R}^2$ dans $\mathbb{R}^2$, et écrire la méthode de Newton pour la résolution $F(X) = 0$. Déterminer les valeurs de a pour lesquelles la méthode de Newton permet de construire une suite qui est toujours bien définie (quelque soit le choix initial).

4. Soit $(\overline{x}, \overline{y})$ une solution du problème (4.63). En supposant que a est tel que la méthode de Newton est bien définie, montrer qu'il existe $\varepsilon > 0$ tel que si (x_0, y_0) est dans la boule B_ε de centre $(\overline{x}, \overline{y})$ et de rayon ε, alors la suite $(x_k, y_k)_{n\in\mathbb{N}}$ construite par la méthode de Newton converge vers $(\overline{x}, \overline{y})$ lorsque n tends vers $+\infty$.

5. a Montrer que le système (4.63) est équivalent au système

$$\begin{array}{l} x - ay = \alpha \\ f(x) = 0 \end{array}$$

où f est une fonction de $\mathbb{R}$ dans $\mathbb{R}$ dont on donnera l'expression.

5.b Ecrire la méthode de Newton pour la résolution de l'équation $f(x) = 0$, et donner les valeurs de a pour lesquelles la méthode de Newton permet de construire une suite qui est toujours bien définie (quelque soit le choix initial). Comparer les itérés obtenus avec cette méthode avec les itérés obtenus par la méthode de la question 3.

6. En s'inspirant des question précédentes, étudier la méthode de Newton pour le schéma d'Euler implicite pour le problème suivant qui modélise le pendule avec amortissement :

$$\begin{array}{l} x''(t) + \mu x'(t) + \sin x(t) = 0, \qquad t > 0, \\ x(0) = \xi, \\ x'(0) = 0, \end{array} \tag{4.64}$$

où $\xi \in [0, 2\pi[$ est la position initiale du pendule et $\mu \in \mathbb{R}_+$. le coefficient d'amortissement.

4.8 Corrigés

Exercice 168 page 267 (Condition de Lipschitz et unicité)

Pour $a \geq 1$, la fonction $\varphi_a : \mathbb{R}_+ \to \mathbb{R}_+$ définie par : $\varphi_a(x) = x^a$ est continûment différentiable, et sa dérivée est $\varphi'_a(x) = ax^{a-1}$. Elle est donc lipschitzienne sur les bornés. Si $a = 0$, la fonction φ_a est constante et égale à 1, et donc encore lipschitzienne sur les bornés.
Soit maintenant $a \in]0, 1[$, supposons que soit lipschitzienne sur les bornés. Alors, pour tout $A > 0$, il existe $M_A > 0$ tel que $|\varphi_a(x)| \leq \mathcal{M}_A|x|$. Ceci entraîne que la fonction $x \mapsto |\frac{\varphi_a(x)}{x}|$ est bornée sur $B(0, A)$. Mais $|\frac{\varphi_a(x)}{x}| = |x^{a-1}| \to +\infty$ lorsque $x \to 0$. Ceci montre que la fonction φ_a n'est pas lipachitzienne sur les bornés si $a \in]0, 1[$.
Par le théorème de Cauchy-Lipschitz, si φ_a est lipschitzienne sur les bornés, alors le problème (4.25) admet une unique solution qui est la solution constante et égaleà zéro.
Si φ_a n'est pas lipschitzienne sur les bornés, *i.e.* si $a \in]0, 1[$, la fonction nulle est encore solution du problème (4.25), mais on peut obtenir une autre solution définie par (calcul élémentaire de séparation de variable) :

$$y_a(t) = [(1 - a)t]^{\frac{1}{1-a}} .$$

(Notons que cette fonction n'est définie que pour $a \in]0, 1[$.)

Exercice 170 page 267 (Loi de Malthus)

Soit p_0 le nombre d'individus au temps $t = 0$. On a donc $p(100) = 2p_0$, et $p(200) = 3p_0$. Or la loi de Malthus s'écrit $p'(t) = ap(t)$ avec $a > 0$, et donc $p(t) = p_0 e^{at}$. On a donc

$$p(100) = p_0 e^{100\ a} = 2p_0$$
$$p(200) = p_0 e^{200\ a} = 3p_0,$$

mais on a aussi $p(200) = p(100)e^{100\ a}$, et donc $p(200) = 3p_0 e^{100\ a}$. On obtient donc que $p(200) = 3p_0 e^{100\ a} = p_0 e^{200\ a} = 3p_0$, ce qui est vrai si

$$\begin{cases} e^{100\ a} = \frac{3}{2} \\ e^{200\ a} = 3. \end{cases}$$

Ceci est impossible car $\ln \frac{3}{2} \neq \frac{1}{2} \ln 3$. Donc la population ne peut pas satisfaire la loi de Malthus.

Exercice 171 page 267 (Histoire de sardines)

1. Pendant le premier mois, le taux d'accroissement de la population est celui de la loi de Malthus $(4p(t))$ auquel il faut retrancher les pertes dues auxbonites, soit $10^{-4}p^2(t)$. Donc pour $0 \leq t \leq T = 30$, on a :

$$\begin{cases} p'_1(t) = 4p_1(t) - 4\ 10^{-4}p_1^2(t) \\ p_1(0) = p_0. \end{cases}$$

A partir de $T = 30$, le taux diminue en raison de l'émigration, soit $10^{-1}p(t)$. On a donc :

$$\begin{cases} p'_2(t) = 4p_2(t) - 10^{-4}p_2^2(t) - 10^{-1}p_2(t), \quad t > 30 \\ p_2(30) = p_1(30). \end{cases}$$

c'est–à–dire :

$$\begin{cases} p'_2(t) = 3.9p_2(t) - 10^{-4}p_2^2(t), \quad t > 30 \\ p_2(30) = p_1(30). \end{cases}$$

2. Les équations à résoudre sont de la forme :

$$\begin{cases} x'(t) = ax(t) + bx^2(t) \\ x(0) = ax_0. \end{cases}$$

qui sont du type "Bernoulli". En supposant que x ne s'annule pas (ce qu'on vérifiera a posteriori) on divise par x^2, on pose $z = \frac{1}{x}$, et on obtient

$$\begin{cases} -z'(t) = az(t) + b \\ z(0) = \dfrac{1}{x_0} \end{cases}$$

Notons qu'on suppose $x_0 \neq 0$. Si $x_0 = 0$, la solution unique est $x(t) \iff 0 \quad \forall t \in \mathbb{R}_+$, par le théorème de Cauchy-Lipschitz. On cherche une solution sous la forme : $z(t) = C(t)e^{-at}$. On a donc $z'(t) = C'(t)e^{-at} - aC(t)e^{-at} = C'(t)e^{-at} - az(t)$. Pour que z soit solution, il faut donc que :

$$-C'(t)e^{-at} = b, \text{ soit encore } C'(t) = -be^{at}.$$

On en déduit que $C(t) = -\frac{b}{a}e^{at} + K$ où $K \in \mathbb{R}$. La fonction z est donc de la forme $z(t) = \left(-\frac{b}{a}e^{+at} + K\right)$. On détermine K à l'aide de la condition initiale $z(0) = \frac{1}{x_0}$, ce qui entraine $-\frac{b}{a} + K = \frac{1}{x_0}$, soit $K = \frac{b}{a} + \frac{1}{x_0}$. On en déduit que la solution de (4.8) s'écrit

$$z(t) = \left(\frac{1}{x_0} + \frac{b}{a}\right)e^{-at} - \frac{b}{a},$$

après avoir vérifié que cette fonction ne s'annule pas (ce qu'on avait supposé pour pouvoir la calculer). On a donc :

$$x(t) = \frac{1}{\left(\frac{1}{x_0} + \frac{b}{a}\right)e^{-at} - \frac{b}{a}}.$$

En reportant les données de notre problème dans cette expression, on a donc :

$$\begin{cases} x(t) = \frac{1}{\left(10^{-6} - \frac{10^{-4}}{4}\right)e^{-4t} + \frac{10^{-4}}{4}} & 0 \leq t \leq 30 \\ x_{30} = x(30), & \\ x(t) = \frac{1}{\left(\frac{1}{x_{30}} - \frac{10^{-4}}{3.9}\right)e^{-3.9t} + \frac{10^{-4}}{3.9}} & t \geq 30. \end{cases}$$

On en conclut que $\lim_{t\to+\infty} x(t) = 3.9\ 10^4$.

Exercice 172 page 268 (Consistance et ordre des schémas)

1. Un développement de Taylor à l'ordre 1 donne que

$$\left|\frac{x(t_{k+1}) - x(t_k)}{h_k} - f(x(t_k), t_k)\right| \leq \sup_{[0,T]} |f'| h_k.$$

L'erreur de consistance est donc d'ordre 1, et le schéma est convergent par le théorème 4.12 page 261.
2. Pour les schémas d'Euler amélioré et de Heun, le théorème 4.12 page 261 s'applique encore, à condition de montrer qu'ils sont consistants. Calculons l'erreur de consistance pour ces deux schémas :

1. Le schéma d'Euler amélioré s'écrit :

$$\frac{x_{k+1} - x_k}{h_k} = f\left(x_k + \frac{h_k}{2}f(x_k, t_k), t_k + \frac{h_k}{2}\right)$$

Soit $\bar{x}_k = x(t_k)$ la solution exacte de l'équation différentielle $x'(t) = f(x(t), t)$ (avec condition initiale $x(0) = x_0$) en t_k. En remarquant que $f(\bar{x}_k, t_k) = x'(t_k)$, on a :

$$\bar{x}_k + \frac{h_k}{2}f(\bar{x}_k, t_k) = \bar{x}_k + \frac{h_k}{2}x'(t_k) = x\left(t_k + \frac{h_k}{2}\right) - \frac{1}{8}h_k^2 x''(\xi_k),$$

avec $\xi_k \in \left[t_k, t_k + \frac{h_k}{2}\right]$. En posant $X = f(\bar{x}_k + \frac{h_k}{2} f(\bar{x}_k, t_k), t_k + \frac{h_k}{2})$, on remarque que $X = f(x(t_k + \frac{h_k}{2}) - \frac{1}{8}h^2 x''(\xi_k), t_k + \frac{h_k}{2})$ et donc qu'il existe $\zeta_k \in \mathbb{R}$ tel que :

$$X = f(x(t_k + \frac{h_k}{2}), t_k + \frac{h_k}{2}) - \frac{1}{8}h_k^2 x''(\xi_k)\partial_1 f(\zeta_k, t_k + \frac{h_k}{2}).$$

On en déduit que

$$X = x'(t_k + \frac{h_k}{2}) - \frac{1}{8}h_k^2 x''(\xi_k)\partial_1 f(\zeta_k, t_k + \frac{h_k}{2})$$

De plus, par développement de Taylor d'ordre 2, on a :

$$\left|\frac{\bar{x}_{k+1} - \bar{x}_k}{h_k} - x'(t_k + \frac{h_k}{2})\right| \leq Ch^2 \tag{4.65}$$

où C ne dépend que de f. On en déduit que l'erreur de consistance R_k vérifie :

$$\begin{array}{rcl} R_k & = & \left|\frac{\bar{x}_{k+1} - \bar{x}_k}{h_k} - f(\bar{x}_k + \frac{h_k}{2} f(\bar{x}_k, t_k), t_k + \frac{h_k}{2})\right| \\ & \leq & \frac{1}{8}h^2 x"(\xi_k)\partial_1 f(\zeta_k, t_k + \frac{h_k}{2}) + \tilde{C}h^2 \end{array}$$

où C ne dépend que de f. On en déduit que le schéma est bien d'ordre 2.

2. Le schéma de Heun s'écrit :

$$\frac{x_{k+1} - x_k}{h_k} = \frac{1}{2}f(x_k, t_k) + \frac{1}{2}[f(x_k + h_k f(x_k, t_k), t_{k+1})].$$

Ecrivons d'abord qu'il existe donc $\theta_k \in [t_k t_{k+1}]$ tel que

$$\bar{x}_k + h_k f(\bar{x}_k, t_k) = x(t_{k+1}) + \frac{h_k^2}{2}x''(\theta_k).$$

De plus, il existe $\zeta_k \in \mathbb{R}$ tel que :

$$f(\bar{x}_k + h_k f(\bar{x}_k, t_k), t_{k+1}) = f(\bar{x}_{k+1}, t_{k+1}) + \partial_1 f(\zeta_k, t_{k+1})\frac{h_k^2}{2}x"(\theta_k).$$

Or $\frac{1}{2}(f(\bar{x}_k, t_k) + f(\bar{x}_{k+1}\ t_{k+1})) = \frac{1}{2}(x'(t_{k+1}) + x'(t_k))$ et par développement de Taylor, il existe $C \in \mathbb{R}$ ne dépendant que de x tel que

$$|\frac{1}{2}(x'(t_{k+1}) + x'(t_k)) - x'(t_k + \frac{h_k}{2})| \leq Ch^2.$$

En utilisant à nouveau (4.65), on en déduit que l'erreur de consistance est d'ordre 2 (à condition que x soit trois fois dérivable . . .).

Exercice 173 page 268 (Stabilité par rapport aux erreurs et convergence)

1. Par définition du schéma (4.8) et de l'erreur de consistance (4.12), on a :

$$\begin{array}{rl} x_{k+1} & = x_k + h_k\phi(x_k, t_k, h_k) \\ \bar{x}_{k+1} & = \bar{x}_k + h_k\phi(\bar{x}_k, t_k, h_k) + h_k R_k. \end{array}$$

Comme le schéma (4.8) est supposé stable par rapport aux données, on a en prenant $y_k = \bar{x}_k$ et $\varepsilon_k = h_k R_k$ dans (4.14) page 260 :

$$e_{k+1} \leq K(|x_0 - \bar{x}_0| + \sum_{i=0}^{k-1} |h_i R_i|) \text{ pour tout } k = 0, \ldots, n-1.$$

Comme le schéma est consistant d'ordre p, on a $R_i \leq Ch^p$ et donc par l'inégalité précédente

$$e_{k+1} \leq K|e_0| + \tilde{C}h^p,$$

où $\tilde{C} \in R_+$ ne dépend que de f,T, $\bar{x}_0$ (et pas de h). On en déduit que le schéma est convergent d'ordre p.
2. Soient $(x_k)_{k=0,\dots,n-1}$ et $(y_k)_{k=0,\dots,n-1}$ vérifiant (4.14), c'est–à–dire :

$$\begin{array}{l} x_{k+1} = x_k + h_k\phi(x_k, t_k, h_k), \\ y_{k+1} = y_k + h_k\phi(y_k, t_k, h_k) + \varepsilon_k, \end{array} \quad \text{pour } k = 0, \dots, n-1,$$

alors grâce à l'hypothèse sur le caractère lipschitzien de ϕ, on a :

$$|x_{k+1} - y_{k+1}| \leq (1 + h_k M)|x_k - y_k| + |\varepsilon_k| \leq e^{h_k M}|x_k - y_k| + |\varepsilon_k|.$$

On en déduit par récurrence sur k que

$$|x_k - y_k| \leq e^{t_k M}|e_0| + \sum_{i=0}^{k-1} e^{(t_k - t_{i+1})M}|\varepsilon_i| \leq K(|e_0| + \sum_{i=0}^{k} |\varepsilon_i|),$$

avec $K = e^{TM}$. On a donc ainsi montré que le schéma (4.8) est stable par rapport aux erreurs.

Exercice 175 page 268 (Algorithme du gradient à pas fixe et schéma d'Euler)

1. L'algorithme du gradient à pas fixe s'écrit :
pour x_0 donné, et à x_k connu, $n \geq 0$, $w_k = -\nabla f(x_k)$ et $x_{n+1} = x_k + \rho \boldsymbol{w}^{(k)}$.
L'algorithme d'Euler explicite pour la résolution de l'équation $x'(t) = -\nabla f(x(t))$ avec $x(0) = x_0$, et pour le pas de discrétisation ρ s'écrit :

$$\begin{aligned} x_0 &= x(0), \\ \frac{x_{n+1} - x_k}{\rho} &= -\nabla f(x(t_k)) \end{aligned}$$

Il est donc clair que les deux algorithmes sont équivalents.

2. Ceci est une conséquence directe du théorème d'existence et unicité 3.12 page 196.

3. Comme f est de classe C^1, par le théorème de Cauchy-Lipschitz il existe une unique solution maximale. On sait de plus que si le temps maximal d'existence T_M est fini, alors $x(t) \to +\infty$ lorsque $t \to +\infty$. Montrons que $x(t)$ est borné, ce qui entraîne donc existence d'une solution globale.
On pose $\varphi(t) = f(x(t))$. On a donc $\varphi'(t) = -|\nabla f(x(t))|^2 < 0$. La fonction φ est donc décroissante et bornée inférieurement (car f est bornée inférieurement, puisque f est continue et tend vers $+\infty$ en $\pm\infty$). Il existe donc $\ell \in \mathbb{R}$ tel que φ tend en décroissant vers ℓ quand t tend vers l'infini. Ceci prouve en particulier que l'ensemble $\{x(t); t \in \mathbb{R}\}$ est borné. On obtient donc ainsi existence et unicité d'une solution globale.
Montrons maintenant que que $\ell = \min_{\mathbb{R}} f$, et que $x(t) \to \bar{x}$ où $\bar{x} = \text{Argmin} f$. On raisonne par l'absurde ; sSi $\ell > \min_{\mathbb{R}} f$, on pose $\varepsilon = \ell - \min_{\mathbb{R}} f > 0$. Il existe $\eta > 0$ tel que $|x - \bar{x}| \leq \eta \Rightarrow \|f(x) - f(\bar{x})\| = \|f(x) - \min_{\mathbb{R}} f\| < \varepsilon$.
Comme $f(x(t)) \geq \ell = \varepsilon + \min_{\mathbb{R}} f$, on a donc $|x(t) - \bar{x}| > \eta$ pour tout $t \in \mathbb{R}$. Donc $|\nabla f(x(t))| \geq \delta = \min\{|\nabla f(y)|, |y - \bar{x}| \geq \eta, |y| \leq M\}$, où M est une borne de $\{x(t), t \in \mathbb{R}\}$.
Comme $\delta > 0$, on obtient une contradiction avec le fait que φ tend en décroissant vers ℓ quand t tend vers l'infini, puisque : $\int_0^t \varphi'(s)ds \leq -\int_0^t \delta^2 ds \leq -t\delta^2$, avec $\int_0^t \varphi'(s)ds \to \ell - \varphi(0)$ et $-t\delta^2 \to +\infty$ lorque $t \to +\infty$.

4. Dans le cas $f(x) = \frac{1}{2}Ax \cdot x - b \cdot x$, où A est s.d.p., il est facile de voir que la fonctionnelle f vérifie les hypothèses de l'exercice. On a $\nabla f(x) = Ax - b$, et donc $x(t)$ tend vers $A^{-1}b$. L'algorithme s'écrit dans ce cas :

$$x_{n+1} = x_k - \rho(Ax - b).$$

Exercice 176 page 269 (Méthode de Taylor)

1. Soit x solution du problème de Cauchy (4.1). Montrons par récurrence que

$$x^{(m+1)}(t) = f^{(m)}(x(t), t).$$

Pour $m = 0$, on a $x^{(1)}(t) = f(x(t), t) = f^{(0)}(x(t), t)$. Supposons que

$$x^{(p+1)}(t) = f^{(p)}(x(t), t) \text{ pour } p = 0, \dots, m,$$

et calculons $x^{(m+2)}(t)$. On a

$$\begin{array}{rcl} x^{(m+2)}(t) & = & \partial_1 f^{(m)}(x(t), t)x'(t) + \partial_2 f^{(m)}(x(t), t) \\ & = & \partial_1 f^{(m)}(x(t), t)f(x(t), t) + \partial_2 f^{(m)}(x(t), t) \\ & = & f^{(m+1)}(x(t), t). \end{array}$$

2. On a $f^{(1)} = \partial_2 f + (\partial_1 f)f$, et $f^{(2)} = (\partial_1 f^{(1)})f + (\partial_2 f^{(1)})$, soit encore

$$f^{(2)} = \left(\partial_1\partial_2 f + (\partial_1^2)f + (\partial_1 f)^2\right) + \partial_2^2 + (\partial_1\partial_2 f)f + (\partial_1 f)(\partial_2 f).$$

3.Dans le cas $p = 1$, on a $\psi_p(y, t, h) = f(y, t)$ et donc le schéma (4.29) s'écrit :

$$\left\{ \begin{array}{l} x_0 = \bar{x}_0, \\ x_{k+1} = x_k + hf(x_k, t_k), \text{ pour } k = 1, \dots, n. \end{array} \right.$$

On reconnaît le schéma d'Euler explicite.

4.a/ Puisque $f(y, t) = y$, on a $f^{(k)} = f$ pour tout k, et donc

$$\psi_p(y, t, h) = \sum_{j=0}^{p-1} \frac{h^j}{(j+1)!} f(y, t).$$

4.b/ Par définition,

$$x_1 = \bar{x}_0 + hf(\bar{x}_0, 0) = \bar{x}_0 + h\sum_{j=0}^{p-1} \frac{h^j}{(j+1)!}\bar{x}_0 = 1 + h\sum_{j=0}^{p-1} \frac{h^j}{(j+1)!} = \sum_{j=0}^{p} \frac{h^j}{(j+1)!}.$$

Supposons que

$$x_k = \left(\sum_{j=0}^{p} \frac{h^j}{j!}\right)^k \text{ pour } k = 1, \dots, \ell,$$

et montrons que cette relation est encore vérifiée au rang $\ell + 1$. On a bien :

$$x_{\ell+1} = x_\ell + h\sum_{j=0}^{p-1} \frac{h^j}{j!}x_\ell = \sum_{j=0}^{p} \frac{h^j}{j!}x_\ell,$$

ce qui termine la récurrence.

4.c/ Comme x est la solution de (4.1) pour $f(y, t) = y$ et $\bar{x}_0 = 1$, on a évidemment $x(t) = e^t$, et donc $x(t_k) = e^{hk}$. Le résultat de la question 4.b/ permet de déduire que

$$\begin{array}{rcl} x_k & = & \left(\sum_{j=0}^{p} \frac{h^j}{j!}\right)^k \\ & = & \left(e^h - R(h)\right)^k, \end{array}$$

avec $0 < R(h) < e^h \frac{h^{p+1}}{(p+1)!}$. On a donc

$$\begin{aligned} x_k &= e^k h \left(1 - \tfrac{R(h)}{e^h}\right)^k \\ &= e^k h (1-a)^k, \end{aligned}$$

avec $a = \frac{R(h)}{e^h} \in]0,1[$. On en déduit que

$$0 \leq \bar{x}_k - x_k \leq e^k h \left(1 - (1-a)^k\right).$$

Comme $k \geq 1$ et $a \in]0,1[$, on en déduit (par récurrence sur k) que $(1-a)^k \geq 1 - ka$. On a donc

$$0 \leq \bar{x}_k - x_k \leq kae^{kh} \leq ke^{t_k} \frac{h^{p+1}}{(p+1)!} \leq t_k e^{t_k} \frac{h^p}{(p+1)!}.$$

5. Un développement de Taylor montre que

$$\begin{aligned} \bar{x}_{k+1} &= \sum_{j=0}^{p} \frac{h^j}{j!} x^{(j)}(t_k) + C_{k,h} h^{p+1} \\ &= \bar{x}_k + \sum_{j=1}^{p} \frac{h^{j-1}}{j!} f^{(j-1)}(\bar{x}_k, t_k) + C_{k,h} h^{p+1}, \end{aligned}$$

avec $C_{k,h} \leq C \in \mathbb{R}_+$. On a donc

$$\begin{aligned} \frac{\bar{x}_{k+1} - \bar{x}_k}{h} &= \sum_{j=1}^{p} \frac{h^{j-1}}{j!} f^{(j-1)}(\bar{x}_k, t_k) + C_{k,h} h^p \\ &= \sum_{j=0}^{p-1} \frac{h^j}{(j+1)!} f^{(j)}(\bar{x}_k, t_k) + C_{k,h} h^p \\ &= \psi_p(\bar{x}_k, t_k, h) + C_{k,h} h^p. \end{aligned}$$

Le schéma est donc consistant d'ordre p. Il suffit alors d'appliquer le théorème 4.12 page 261 (car ψ_p est de classe C^∞ donc lipschitzienne sur les bornés) pour obtenir l'existence de $\bar{h} > 0$ et $C > 0$ ne dépendant que de $\bar{x}_0$, T et f, tels que si $0 < h < \bar{h}$, alors $|x_k - x(t_k)| \leq Ch^p$, pour tout $k = 0, \ldots, n+1$.

Exercice 178 page 270 (Méthodes semi-implicite et explicite)

1.

$$\left\{ \begin{array}{l} \dfrac{x_1^{(k+1)} - x_1^{(k)}}{k} = -x_1^{(k)} - x_1^{(k)} x_2^{(k+1)}, \\ \dfrac{x_2^{(k+1)} - x_2^{(k)}}{k} = -\dfrac{x_2^{(k+1)}}{x_1^{(k)}}, \\ x_1^{(0)} = a,\ x_2^{(0)} = b. \end{array} \right. \tag{4.66}$$

On a $x_1^{(0)} = a > 0$ et $x_2^{(0)} = b > 0$. De plus, on a

$$x_2^{(1)} = \frac{1}{1 + \frac{k}{a}} b,$$

donc $x_2^{(1)}$ est bien défini, et $0 < x_2^{(1)} < x_2^{(0)} = b$. Or $x_1^{(1)} = a - k(a + ab)$ et comme a et b appartiennent à $]0,1[$, on a $a + ab \in]0,2[$, et comme $0 < k < 1/2$, on en déduit que $0 < x_1^{(1)} < x_1^{(0)} = a$.

Supposons que les suites soient bien définies, décroissantes et strictement positives jusqu'au rang n, et vérifions-le au rang $n+1$. On a

$$x_2^{(k+1)} = \frac{1}{1+\frac{k}{x_1^{(k)}}} x_2^{(k)}, \tag{4.67}$$

et donc en utilisant l'hypothèse de récurrence, on obtient que $x_2^{(k+1)} < x_2^{(k)}$ et $0 < x_2^{(k+1)} < b$.
De plus

$$x_1^{(k+1)} = x_1^{(k)} - kx_1^{(k)} - kx_1^{(k)}x_2^{(k+1)} = x_1^{(k)}(1-k-kx_2^{(k+1)}), \tag{4.68}$$

et donc grâce au fait que $0 < x_2^{(k+1)} < b$ (et donc $1-k-kx_2^{(k+1)} > 1-k-kb$, et à l'hypothèse de récurrence, on déduit que $x_1^{(k+1)} < x_1^{(k)}$ et $0 < x_1^{(k+1)} < a$.

2. Après calcul, on obtient que le schéma numérique (4.36) s'écrit sous la forme

$$\frac{x^{(k+1)} - x^{(k)}}{k} = \phi(x^{(k)}, k), \tag{4.69}$$

avec $x^{(k)} = (x_1^{(k)}, x_2^{(k)})^t$, et où $\phi \in C^\infty((\mathbb{R}_+^*)^2 \times \mathbb{R}_+, \mathbb{R}^2)$ est définie par

$$\phi(x,k) = \begin{pmatrix} -x_1(1+\dfrac{x_1x_2}{x_1+k}) \\ -\dfrac{x_2}{x_1+k} \end{pmatrix}, \tag{4.70}$$

et on vérifie bien que $\phi \in C^\infty((\mathbb{R}_+^*)^2 \times \mathbb{R}_+, \mathbb{R}^2)$ (en fait ϕ est de classe C^∞ sur $\mathbb{R}_+^2 \times \mathbb{R}_+ \setminus \{0\} \times \mathbb{R}_+ \times \{0\}$.) et que $\phi(x,0) = f(x)$. Ceci montre que pour $(x_1^{(k)}, x_2^{(k)}) \in (\mathbb{R}_+^*)^2$ et $k > 0$, le couple $(x_1^{(k+1)}, x_2^{(k+1)})$ est bien défini par (4.36) de manière unique.

3. Comme $x \in C^\infty([0,+\infty[, (\mathbb{R}_+^*)^2)$, on a

$$|\frac{x(t_{n+1}) - x(t_n)}{k} - x'(t_n)| \leq k \max_{[0,T]} |x''|,$$

et

$$|\phi(x(t_n),k) - \phi(x(t_n),0)| \leq k \max_{[0,T]} |D_2\phi(x(t),t)|.$$

Or la solution exacte x sur $[0,T]$ vit dans un borné $[\alpha,\beta]^2$ de R_+^*, et ses dérivées atteignent ses bornes sur le compact $[0,T]$, donc il existe $C(T) \in \mathbb{R}_+$ tel que $\max_{[0,T]} |x''| \leq C(T)$ et $\max_{[0,T]} |D_2\phi(x(t),t)| \leq C(T)$. Comme de plus $\phi(x(t_n),0) = f(x(t_n)$, on en déduit par inégalité triangulaire que $|R_k^{(k)}| \leq C(T)k$.

4. (Stabilité)

   (i) Soit $T > 0$. De (4.68) et du fait que $0 < x_2^{(k)} < b$ on déduit que

   $$x_1^{(k+1)} \geq x_1^{(k)}(1-k-kb),$$

   et donc par récurrence sur n que

   $$x_1^{(k)} \geq x_1^{(0)}(1-k-kb)^n,$$

   Donc pour tout entier n tel que $nk \leq T$, on a $n \leq \frac{T}{k}$, et comme $1-k-kb > 0$ (car $k < 1/2$), on a $x_1^{(k)} \geq (1-k-kb)^{\frac{T}{k}}$.

(ii) On a $(1-k-kb)^{\frac{T}{k}} = exp(\frac{T}{k}\ln(1-k-kb))$, et $\ln(1-k-kb)$ est équivalent à $k-kb$ dans un voisinage de $k=0$. On en déduit que $(1-k-kb)^{\frac{T}{k}} \to e^{-(1+b)T}$ lorsque $k \to 0$.

La fonction φ définie par $\varphi(k) = (1-k-kb)^{\frac{T}{k}}$ est continue, strictement positive sur $[0, 1/2]$, et sa limite lorsque k tend vers 0 est minorée par un nombre strictement positif. Donc la fonction est elle-même minorée par un nombre strictement positif. On en déduit que $\inf_{0<k<\frac{1}{2}}(1-k-kb)^{\frac{T}{k}} > 0$.

(iii) D'après les résultats des questions 3 (a) et 3 (d) (ii), on a $a(T) \leq x_1^{(k)} \leq a$, pour tout n tel que $nk \leq T$, avec $a(T) = \inf_{0<k<\frac{1}{2}}(1-k-kb)^{\frac{T}{k}}$.

En utilisant ce résultat (et la question 3 (a)), on déduit alors de (4.67) que

$$x_2^{(k+1)} \geq \frac{1}{1+\frac{k}{a(T)}} x_2^{(k)},$$

et donc que

$$x_2^{(k)} \geq \Big(\frac{1}{1+\frac{k}{a(T)}}\Big)^{\frac{T}{k}} x_2^{(0)},$$

Une étude similaire à celle de la question précédente montre que la fonction

$$k \mapsto \Big(\frac{1}{1+\frac{k}{a(T)}}\Big)^{\frac{T}{k}}$$

est continue et strictement positive sur $[0, 1/2]$ et sa limite lorsque k tend vers 0 est strictement positive. On en déduit que $b(T) \leq x_2^{(k)} \leq b$, pour tout n tel que $nk \leq T$, avec

$$b(T) = b \inf_{k\in[0,1/2]} \Big(\frac{1}{1+\frac{k}{a(T)}}\Big)^{\frac{T}{k}} > 0.$$

5. (Convergence) Soit $T > 0$. On ne peut pas appliquer directement le théorème du cours car ϕ n'est pas lipschitzienne sur les bornés, mais il suffit de remarquer que :
   — la solution exacte sur $[0, T]$ vit dans un borné $[\alpha, \beta]^2$ de R_+^*.
   — le schéma est inconditionnellement stable : $x^{(k)} \in [a(T), a] \times [b(T), b]$.

   Or la fonction ϕ est de classe C^1 sur $[A, B]^2 \times R_+^*$, où $A = \min(\alpha, a(T), b(T))$ et $B = \max(\beta, a, b)$. Donc elle est lipschitzienne par rapport à la première variable sur le pavé $[A, B]^2$. La démonstration par récurrence faite en cours dans le cas ϕ globalement lipschitzienne s'adapte donc très facilement. (elle est même plus facile car $e_0 = 0$ et le pas de discrétisation est constant...)

6. On remplace maintenant le schéma (4.36) par le schéma d'Euler explicite. Celui s'écrit :

$$\begin{cases} \dfrac{x_1^{(k+1)} - x_1^{(k)}}{k} = -x_1^{(k)} - x_1^{(k)} x_2^{(k)}, \\ \dfrac{x_2^{(k+1)} - x_2^{(k)}}{k} = -\dfrac{x_2^{(k)}}{x_1^{(k)}}, \\ x_1^{(0)} = a,\ x_2^{(0)} = b. \end{cases} \tag{4.71}$$

Supposons $x_1^{(k)} > 0$ et $x_2^{(k)} > 0$ pour tout n. La première équation de (4.36) donne alors que

$$\frac{x_1^{(k+1)} - x_1^{(k)}}{k} = -x_1^{(k)},$$

et donc $x_1^{(k+1)} < (1-k)x_1^{(k)}$. On en déduit par récurrence que $x_1^{(k)} < (1-k)^n a \to 0$ lorsque $n \to 0$ (on supposera que $k < 1$ pour que le schéma soit bien défini. Donc pour un pas de temps k donné, il existe n tel que $x_1^{(k)} \leq k$. Or pour cette valeur de n,

$$x_2^{(k+1)} = x_2^{(k)}(1 - \frac{k}{x_1^{(k)}}) \leq 0,$$

ce qui contredit l'hypothèse $x_2^{(k)} > 0$ pour tout n.
Ceci montre que le schéma d'Euler explicite n'est franchement pas bon dans ce cas. (Etudier si le coeur vous en dit le schéma totalement implicite...)

Chapitre 5

Quelques problèmes supplémentaires

5.1 Méthode de Jacobi et optimisation

Corrigé détaillé à la suite de l'énoncé, 284
Rappel Soit $f \in C^1(\mathbb{R}^n, \mathbb{R})$; on appelle **méthode de descente à pas fixe** $\alpha \in \mathbb{R}_+^*$ pour la minimisation de f, une suite définie par

$$\begin{aligned} &\boldsymbol{x}^{(0)} \in \mathbb{R}^n \text{ donné,} \\ &\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha \boldsymbol{w}^{(k)}, \text{ pour } k \geq 0, \end{aligned}$$

où $\boldsymbol{w}^{(k)}$ est une **direction de descente** en $\boldsymbol{x}^{(k)}$.
Dans toute la suite, on considère la fonction f de $\mathbb{R}^n$ dans $\mathbb{R}$ définie par

$$f(\boldsymbol{x}) = \frac{1}{2} A\boldsymbol{x} \cdot \boldsymbol{x} - \boldsymbol{b} \cdot \boldsymbol{x}, \tag{5.1}$$

où A une matrice carrée d'ordre n, symétrique définie positive, et $\boldsymbol{b} \in \mathbb{R}^n$. On pose $\bar{x} = A^{-1}b$.

1. Montrer que la méthode de Jacobi pour la résolution du système $A\boldsymbol{x} = \boldsymbol{b}$ peut s'écrire comme une méthode de descente à pas fixe pour la minimisation de la fonction f définie par (5.1). Donner l'expression du pas α et de la direction de descente $\boldsymbol{w}^{(k)}$ à chaque itération k et vérifier que c'est bien une direction de descente stricte si $\boldsymbol{x}^{(k)} \neq A^{-1}\boldsymbol{b}$.
2. On cherche maintenant à améliorer la méthode de Jacobi en prenant non plus un pas fixe dans l'algorithme de descente ci-dessus, mais un pas optimal qui est défini à l'itération k par

   $$f(\boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{w}^{(k)}) = \min_{\alpha > 0} f(\boldsymbol{x}^{(k)} + \alpha \boldsymbol{w}^{(k)}), \tag{5.2}$$

   où $\boldsymbol{w}^{(k)}$ est défini à la question précédente. On définit alors une méthode de descente à pas optimal par :

   $$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{w}^{(k)}.$$

   On appelle cette nouvelle méthode "méthode de Jacobi à pas optimal".
   (a) Justifier l'existence et l'unicité du pas optimal défini par (5.2), et donner son expression à chaque itération.
   (b) Montrer que $|f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^{(k+1)})| = \dfrac{|\boldsymbol{r}^{(k)} \cdot \boldsymbol{w}^{(k)}|^2}{2A\boldsymbol{w}^{(k)} \cdot \boldsymbol{w}^{(k)}}$ si $w^{(k)} \neq 0$.
   (c) Montrer que $\boldsymbol{r}^{(k)} \to \boldsymbol{0}$ lorsque $k \to +\infty$, et en déduire que la suite donnée par la méthode de Jacobi à pas optimal converge vers la solution $\bar{\boldsymbol{x}}$ du système linéaire $A\boldsymbol{x} = \boldsymbol{b}$.

(d) On suppose que la diagonale extraite D de la matrice A (qui est symétrique définie positive) est de la forme $D = \alpha \mathrm{Id}$ avec $\alpha \in \mathbb{R}$.

i. Ecrire l'algorithme de descente à pas optimal dans ce cas.

ii. Comparer les algorithmes de descente obtenus par Jacobi et Jacobi à pas optimal avec les algorithmes de gradient que vous connaissez.

Corrigé

1. La méthode de Jacobi peut s'écrire

$$\begin{aligned} \boldsymbol{x}^{(k+1)} &= (\mathrm{Id} - D^{-1}A)\boldsymbol{x}^{(k)} + D^{-1}\boldsymbol{b} \\ &= \boldsymbol{x}^{(k)} + D^{-1}(\boldsymbol{b} - A\boldsymbol{x}^{(k)}) \\ &= \boldsymbol{x}^{(k)} + \boldsymbol{w}^{(k)} \end{aligned}$$

avec $\boldsymbol{w}^{(k)} = D^{-1}(\boldsymbol{b} - A\boldsymbol{x}^{(k)}) = D^{-1}\boldsymbol{r}^{(k)}$. On a $\boldsymbol{w}^{(k)} \cdot \nabla f(\boldsymbol{x}^{(k)}) = -D^{-1}\boldsymbol{r}^{(k)} \cdot \boldsymbol{r}^{(k)}$, et comme A est s.d.p., D^{-1} l'est également, et donc $\boldsymbol{w}^{(k)} \cdot \nabla f(\boldsymbol{x}^{(k)}) < 0$ si $\boldsymbol{x}^{(k)} \neq A^{-1}\boldsymbol{b}$. Ceci montre que $\boldsymbol{w}^{(k)}$ est une direction de descente stricte en $x^{(k)}$.

2. (a) Le pas optimal α_k est celui qui minimise la fonction φ définie de $\mathbb{R}$ dans $\mathbb{R}$ par $f(\boldsymbol{x}^{(k)} + \alpha \boldsymbol{w}^{(k)})$, qui est de classe C^1, strictement convexe et croissante à l'infini, ce qui donne l'existence et l'unicité ; de plus α_k vérifie :

$$\nabla f(\boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{w}^{(k)}) \cdot \boldsymbol{w}^{(k)} = 0, \text{ c.à.d. } (A\boldsymbol{x}^{(k)} + \alpha_k A\boldsymbol{w}^{(k)} + \boldsymbol{b}) \cdot \boldsymbol{w}^{(k)} = 0.$$

On en déduit que (si $w^{(k)} \neq 0$)

$$\alpha_k = \frac{(\boldsymbol{b} - A\boldsymbol{x}^{(k)}) \cdot \boldsymbol{w}^{(k)}}{A\boldsymbol{w}^{(k)} \cdot \boldsymbol{w}^{(k)}} = \frac{\boldsymbol{r}^{(k)} \cdot \boldsymbol{w}^{(k)}}{A\boldsymbol{w}^{(k)} \cdot \boldsymbol{w}^{(k)}} = \frac{\boldsymbol{r}^{(k)} \cdot D^{-1}\boldsymbol{r}^{(k)}}{AD^{-1}\boldsymbol{r}^{(k)} \cdot D^{-1}\boldsymbol{r}^{(k)}}.$$

(Si $w^{(k)} = 0$, on a alors $r^{(k)} = 0$ et $x^{(k)} = \bar{x}$, l"algorithme s'arrête.)

(b) On a :

$$f(\boldsymbol{x}^{(k+1)}) = f(\boldsymbol{x}^{(k)}) - \gamma\alpha_k + \delta\alpha_k^2,$$

avec $\gamma = \boldsymbol{r}^{(k)} \cdot \boldsymbol{w}^{(k)}$ et $\delta = \frac{1}{2}A\boldsymbol{w}^{(k)} \cdot \boldsymbol{w}^{(k)}$. Comme α_k minimise ce polynôme de degré 2 en α, on a

$$\begin{aligned} f(\boldsymbol{x}^{(k+1)}) - f(\boldsymbol{x}^{(k)}) &= -\frac{\gamma^2}{4\delta} \\ &= -\frac{|\boldsymbol{r}^{(k)} \cdot \boldsymbol{w}^{(k)}|^2}{2A\boldsymbol{w}^{(k)} \cdot \boldsymbol{w}^{(k)}}, \end{aligned}$$

d'où le résultat.

(c) On suppose que $w^{(k)} \neq 0$ pour tout k. La suite $(f(x^{(k)}))_{k\in\mathbb{N}}$ est décroissante et bornée inférieurement (car la fonction f est bornée inférieurement). Elle est donc convergente. Ce qui prouve que $\lim_{k\to+\infty} f(\boldsymbol{x}^{(k+1)}) - f(\boldsymbol{x}^{(k)}) = 0$.
On sait que $\boldsymbol{w}^{(k)} = D^{-1}\boldsymbol{r}^{(k)}$. On a donc, par la question précédente,

$$\frac{|\boldsymbol{r}^{(k)} \cdot \boldsymbol{w}^{(k)}|^2}{A\boldsymbol{w}^{(k)} \cdot \boldsymbol{w}^{(k)}} = \frac{|\boldsymbol{r}^{(k)} \cdot D^{-1}\boldsymbol{r}^{(k)}|^2}{AD^{-1}\boldsymbol{r}^{(k)} \cdot D^{-1}\boldsymbol{r}^{(k)}} = 2|f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^{(k+1)})|.$$

Or

$$0 < AD^{-1}\boldsymbol{r}^{(k)} \cdot D^{-1}\boldsymbol{r}^{(k)} \leq \zeta|\boldsymbol{r}^{(k)}|^2$$

avec $\zeta = \|A\|_2\|D^{-1}\|_2^2$ et

$$\boldsymbol{r}^{(k)} \cdot D^{-1}\boldsymbol{r}^{(k)} \geq \theta|\boldsymbol{r}^{(k)}|^2,$$

où $\theta = \min_{i\in\{1,\dots,n\}} 1/a_{i,i}$. (Les $a_{i,i}$ étant les termes diagonaux de A.) On en déduit que

$$\frac{\theta^2}{\zeta}|\boldsymbol{r}^{(k)}|^2 \leq |f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^{(k+1)})| \to 0 \text{ lorsque } k \to +\infty,$$

et donc $\boldsymbol{r}^{(k)} \to \mathbf{0}$ lorsque $k \to +\infty$.
Comme $\boldsymbol{x}^{(k)} - \bar{\boldsymbol{x}} = -A^{-1}\boldsymbol{r}^{(k)}$, on en déduit la convergence de la suite $\boldsymbol{x}^{(k)}$ vers la solution du système.

(d) i. Si $D = \alpha\mathrm{Id}$, on a

$$\alpha_k = \frac{(\boldsymbol{b} - A\boldsymbol{x}^{(k)})\cdot\boldsymbol{w}^{(k)}}{A\boldsymbol{w}^{(k)}\cdot\boldsymbol{w}^{(k)}} = \frac{\boldsymbol{r}^{(k)}\cdot\boldsymbol{w}^{(k)}}{A\boldsymbol{w}^{(k)}\cdot\boldsymbol{w}^{(k)}} = \frac{1}{\alpha}\frac{\boldsymbol{r}^{(k)}\cdot\boldsymbol{r}^{(k)}}{A\boldsymbol{r}^{(k)}\cdot\boldsymbol{r}^{(k)}}.$$

ii. Jacobi simple= algorithme de gradient avec $\rho = \frac{1}{\alpha}$
Jacobi à pas optimal= algorithme de gradient à pas optimal.

5.2 Assemblage des matrices éléments finis

5.2.1 Notations et rappels

Pour $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, on note $\boldsymbol{x}\cdot\boldsymbol{y} = \boldsymbol{x}^t\boldsymbol{y} = \boldsymbol{y}^t\boldsymbol{x}$ leur produit scalaire.
Soit $m \in \mathbb{N}^\star$ quelconque. On rappelle qu'une matrice symétrique $B \in \mathcal{M}_m(\mathbb{R})$ est :

— **définie** *si* $(B\boldsymbol{u}\cdot\boldsymbol{u} = 0 \Longrightarrow \boldsymbol{u} = 0)$,

— **positive** *si* $\forall \boldsymbol{u} \in \mathbb{R}^m,\ \boldsymbol{u}^t B\boldsymbol{u} \geq 0$.

Enfin, on rappelle qu'une matrice $B \in \mathcal{M}_m(\mathbb{R})$ est une **ICP-matrice** *si B est inversible et tous les coefficients de B^{-1} sont positifs ou nuls (voir exercice 14 du polycopié).*

5.2.2 Méthode des éléments finis et matrices d'assemblage

La méthode des éléments finis de Lagrange est une méthode de discrétisation des équations aux dérivées partielles qui permet d'approcher des valeurs de la fonction inconnue en des points d'un maillage. Ces points sont appelés "noeuds" et les valeurs en ces noeuds sont appelés "degrés de liberté". L'utilisation de la méthode des éléments finis pour la résolution d'une équation aux dérivées partielles conduit à résoudre un système linéaire. Nous ne détaillerons pas ici le principe de la méthode ni son implémentation, mais nous allons nous intéresser aux propriétés du système linéaire résultant. Dans ce but, nous définissons les grandeurs suivantes :

(i) Les inconnues sont représentés par un vecteur $\boldsymbol{u} \in \mathbb{R}^n$, $n \in \mathbb{N}^\star$ (valeurs approchées de la fonction inconnue aux noeuds, ou degrés de liberté .

(ii) Il existe un "maillage" de M éléments, avec $M \in \mathbb{N}^\star$, et, pour tout $k = 1, \dots, M$, il existe

— *une matrice $A_k \in \mathcal{M}_{n_k}(\mathbb{R})$ avec $n_k \in \mathbb{N}^\star$, appelée"matrice élémentaire d'assemblage", telle que A_k est non nulle, symétrique positive mais pas nécessairement définie,*

— *une application injective $\varphi_k : \{1, \dots, n_k\} \to \{1, \dots, n\}$ qui définit "les degrés de liberté de la maille k". Les ensembles définis, pour tout $i = 1, \dots, n$, par*

$$E_i = \Big\{k \in \{1, \dots, M\} :\ i \in \varphi_k(\{1, \dots, n_k\})\Big\},$$

sont alors tous supposés non vides, c.à.d. que pour tout noeud i du maillage, il existe au moins une maille $\mathfrak{m}_k$ dont il est un noeud, voir Figure 5.1.

(iii) un vecteur $\boldsymbol{b}_k \in \mathbb{R}^{n_k}$, appelés "second membre élémentaire d'assemblage".

(iv) Le système linéaire à résoudre s'écrit $A\boldsymbol{u} = \boldsymbol{b}$*, avec* $A \in \mathcal{M}_n(\mathbb{R})$ *et* $\boldsymbol{b} = \begin{bmatrix} b_1 & \dots & b_n \end{bmatrix}^t \in \mathbb{R}^n$ *tels que*

$$A = \sum_{k=1}^{M} H_k^t A_k H_k \text{ et } \boldsymbol{b} = \sum_{k=1}^{M} H_k^t b_k, \tag{5.3}$$

avec $H_k \in \mathcal{M}_{n_k,n}(\mathbb{R})$ *(matrice associé à l'application* φ_k*) définie par*

$$\forall i = 1, \dots, n_k,\ \forall j = 1, \dots, n,\ (H_k)_{i,j} = 1 \text{ si } j = \varphi_k(i) \text{ sinon } (H_k)_{i,j} = 0. \tag{5.4}$$

5.2.3 Un exemple 1D simple

Soit l'intervalle $\Omega =]0,1[$*, sur lequel on cherche à résoudre le problème*

$$-u'' = 2 \text{ dans } \Omega, \tag{5.5a}$$
$$u(0) = u(1) = 0. \tag{5.5b}$$

1. Donner la solution exacte du problème (5.5).

On décompose Ω *en trois mailles,* $\mathfrak{m}_1 =]0, \frac{1}{3}[$, $\mathfrak{m}_2 =]\frac{1}{3}, \frac{2}{3}[$, $\mathfrak{m}_3 =]\frac{2}{3}, 1[$*. On a donc :* $M = 3$*. On a quatre "noeuds" du maillage :* $x_1 = 0, x_1 = 1, x_3 = 2$ *et* $x_4 = 3$*, dont deux (*x_0 *et* x_3*, repérés par des* $\circ$ *sur la figure 5.1), pour lesquelles la valeur de la solution est connue en raison des conditions aux limites* (5.5b).

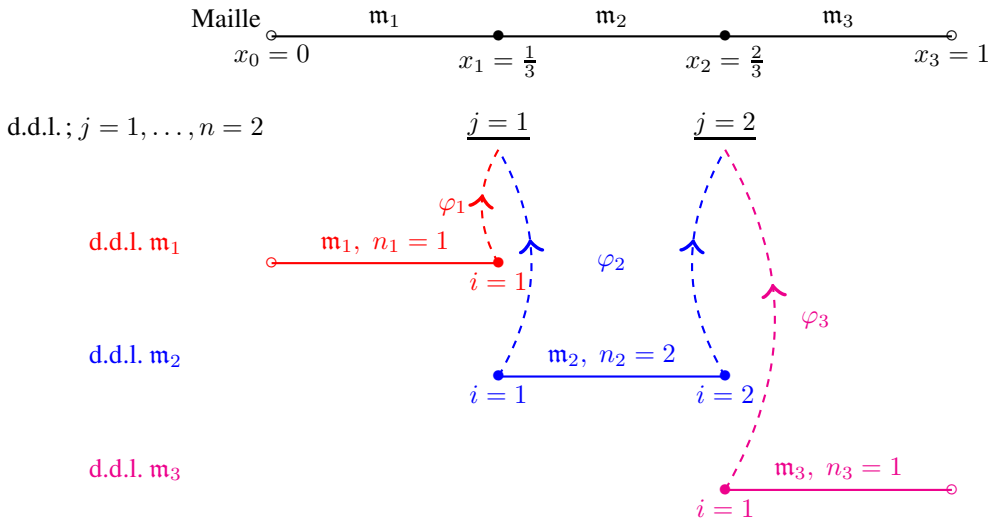


FIGURE 5.1: Le maillage et les degrés de liberté (d.d.l.) du maillage et de chaque maille dans le cas $M = 3, n = 2$.

On utilise des éléments finis de Lagrange linéaires par morceaux (on ne détaillera pas ici la procédure, qui relève du cours de M1). Les degrés de liberté sont les valeurs aux noeuds intérieurs à Ω*, qui sont indiqués par un* $\bullet$ *sur la figure 5.1 (car on a vu que les valeurs aux noeuds du bord sont fixées par* (5.5b)*). On a donc* $n = 2$*. Toujours en raison des conditions limites, les mailles du bord* $\mathfrak{m}_1$ *et* $\mathfrak{m}_3$ *ont un seul degré de liberté et la maille intérieure* $\mathfrak{m}_2$ *en a 2. On a donc* $n_1 = n_3 = 1$ *et* $n_2 = 2$*.*
Avec des éléments finis linéaires par morceaux, on obtient les matrices élémentaires, second membres élémentaires et degrés de liberté suivants pour chaque maille $\mathfrak{m}_k$*,* $k = 1, 2, 3$*, où on a posé* $h = \frac{1}{3}$ *(pas du maillage) :*

1. $A_1 = \dfrac{1}{h}\begin{bmatrix} 1 \end{bmatrix}$, $\varphi_1(1) = 1$, $b_1 = \begin{bmatrix} h \end{bmatrix}$;
2. $A_2 = \dfrac{1}{h}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$, $\begin{cases} \varphi_2(1) = 1 \\ \varphi_2(2) = 2 \end{cases}$, $b_2 = h\begin{bmatrix} 1 \\ 1 \end{bmatrix}$;

3. $A_3 = \dfrac{1}{h}\begin{bmatrix}1\end{bmatrix}$, $\varphi_3(1) = 2$, $b_3 = \begin{bmatrix}h\end{bmatrix}$.
2. Montrer que les matrices A_k, $k = 1, \ldots, 3$ sont symétriques positives et que les matrices A_1 et A_3 sont de plus définies positives.
3. Expliciter des ensembles E_i pour $i = 1, 2$.
4. *Construction de la matrice A.*
   (a) Ecrire les matrices H_k pour $k = 1, \ldots, 3$.
   (b) Donner la matrice A et le second membre b.
   (c) Comparer la matrice A avec la matrice K_2 vue en cours pour la discrétisation de (5.5a) par différences finies et comparer le second membre $\boldsymbol{b}$ avec celui obtenu par différences finies. En déduire que le schéma éléments finis donné ci-dessus donne le même système linéaire que celui obtenu par différences finies.
   (d) Calculer les valeurs propres de la matrice $A = K_2$.

Corrigé de la partie 5.2.3

1. $u(x) = -x(x-1)$.
2. Les matrices $A_1 = A_3 = \begin{bmatrix}\dfrac{1}{h}\end{bmatrix}$ sont évidemment s.d.p.; la matrice A_2 est évidemment symétrique, et pour tout $\boldsymbol{u} = (u_1, u_2)^t \in \mathbb{R}^2$, on a $\boldsymbol{u}^t A_2 \boldsymbol{u} = \dfrac{1}{h}(u_1 - u_2)^2 \geq 0$, ce qui prouve qu'elle est positive.
3. Si $i = 1$: on a uniquement $1 = \varphi_1(1)$ et $1 = \varphi_2(1)$ donc $E_1 = \{1, 2\}$. Si $i = 2$: on a uniquement $2 = \varphi_2(2)$ et $2 = \varphi_3(1)$ donc $E_2 = \{2, 3\}$.
4. *Construction de la matrice A.*
   (a) $H_1 = \begin{bmatrix}1 & 0\end{bmatrix}$, $H_2 = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$, et enfin $H_3 = \begin{bmatrix}0 & 1\end{bmatrix}$.
   (b)

$$\begin{aligned}
A &= H_1^t A_1 H_1 + H_2^t A_2 H_2 + H_3^t A_3 H_3 \\
&= \begin{bmatrix}1\\0\end{bmatrix}\begin{bmatrix}\frac{1}{h}\end{bmatrix}\begin{bmatrix}1 & 0\end{bmatrix} + \frac{1}{h}\begin{bmatrix}1 & -1\\ -1 & 1\end{bmatrix} + \begin{bmatrix}0\\1\end{bmatrix}\begin{bmatrix}\frac{1}{h}\end{bmatrix}\begin{bmatrix}0 & 1\end{bmatrix} \\
&= \frac{1}{h}\begin{bmatrix}2 & -1\\ -1 & 2\end{bmatrix}. \\
\boldsymbol{b} &= H_1^t \boldsymbol{b}_1 + H_2^t \boldsymbol{b}_2 + H_3^t \boldsymbol{b}_3 \\
&= \begin{bmatrix}1\\0\end{bmatrix}\begin{bmatrix}h\end{bmatrix} + h\begin{bmatrix}1\\1\end{bmatrix} + \begin{bmatrix}0\\1\end{bmatrix}\begin{bmatrix}h\end{bmatrix}\begin{bmatrix}0 & 1\end{bmatrix} \\
&= h\begin{bmatrix}2\\2\end{bmatrix}
\end{aligned}$$

   (c) On a : $A = \dfrac{1}{h}K_2$, où K_n est la matrice vue en cours pour la matrice du schéma des différences finies appliqué au problème (5.5), avec $h = \dfrac{1}{n}$, n étant le nombre de points de discétisation. On a vu que $\boldsymbol{b} = h\begin{bmatrix}2\\2\end{bmatrix}$, or le second membre dans la méthode des différences finies est égal à $\widetilde{\boldsymbol{b}} = \begin{bmatrix}2\\2\end{bmatrix} = \frac{1}{h}\boldsymbol{b}$. Le schéma éléments finis s'écrit donc $A\boldsymbol{u} = \boldsymbol{b}$, c.à.d. $\dfrac{1}{h}K_2\boldsymbol{u} = \dfrac{1}{h}\tilde{\boldsymbol{b}}$, ou encore $K_2\boldsymbol{u} = \tilde{\boldsymbol{b}}$. C'est donc le même système linéaire que celui des différences finies.
   (d) $\lambda_1 = \dfrac{1}{h}, \lambda_2 = \dfrac{3}{h}$.

5.2.4 Propriétés de la matrice A

On revient ici au cadre général décrit en section 5.2.2.

1. *Propriétés spectrales de A.*
   (a) Soit $m \in \mathbb{N}^\star$, soit $B \in \mathcal{M}_m(\mathbb{R})$ une matrice symétrique et soient $\underline{\mu}$ la plus petite valeur propre de B et $\overline{\mu}$ la plus grande valeur propre de B : montrer que

$$\underline{\mu} = \min_{\boldsymbol{x}\in\mathbb{R}^m, \boldsymbol{x}^t\boldsymbol{x}=1} \boldsymbol{x}^t B\boldsymbol{x} \text{ et } \overline{\mu} = \max_{\boldsymbol{x}\in\mathbb{R}^m, \boldsymbol{x}^t\boldsymbol{x}=1} \boldsymbol{x}^t B\boldsymbol{x}.$$

   On note $\underline{\lambda}$ (resp. $\underline{\lambda}_k$) la plus petite valeur propre de A (resp. A_k) et $\overline{\lambda}$ (resp. $\overline{\lambda}_k$) la plus grande valeur propre de A (resp. A_k).
   (b) Soit $\boldsymbol{u} \in \mathbb{R}^n$ tel que $\boldsymbol{u}^t\boldsymbol{u} = 1$, montrer que

$$\underline{\lambda}_k (H_k\boldsymbol{u})^t(H_k\boldsymbol{u}) \le (H_k\boldsymbol{u})^t A_k (H_k\boldsymbol{u}) \le \overline{\lambda}_k (H_k\boldsymbol{u})^t(H_k\boldsymbol{u}),$$

   (c) Montrer que la matrice $H_k^t H_k \in \mathcal{M}_n(\mathbb{R})$ peut s'écrire

$$H_k^t H_k = \mathrm{diag}\Big((\chi_{E_i}(k))_{i=1,\dots,n}\Big),$$

   avec $\chi_{E_i}(k) = 1$ si $k \in E_i$ et 0 sinon, et où l'on note $\mathrm{diag}((d_i)_{i=1,\dots,n}) \in \mathcal{M}_n(\mathbb{R})$ la matrice diagonale d'ordre n dont les coefficients diagonaux sont donnés par $(d_i)_{i=1,\dots,n}$.
   En déduire que pour toute famille de réels $(\alpha_k)_{k=1,\dots,M}$, on a

$$\sum_{k=1}^{M} \alpha_k H_k^t H_k = \mathrm{diag}\Big((\sum_{k\in E_i} \alpha_k)_{i=1,\dots,n}\Big), \tag{5.6}$$

   (d) Montrer que

$$\min_{i=1,\dots,n} \sum_{k\in E_i} \underline{\lambda}_k \le \underline{\lambda} \le \overline{\lambda} \le \max_{i=1,\dots,n} \sum_{k\in E_i} \overline{\lambda}_k.$$

   (e) En déduire que, si A est inversible, A est symétrique définie positive, et que, si chaque matrice A_k est symétrique définie positive, A est également symétrique définie positive.
   (f) On suppose que A est inversible. Pour $\varepsilon > 0$ donné on pose

$$P = (\frac{1}{2} + \varepsilon)\mathrm{diag}\Big((\sum_{k\in E_i} \overline{\lambda}_k)_{i=1,\dots,n}\Big).$$

   Prouver, au moyen d'un critère du cours, que la méthode itérative définie par $P\boldsymbol{u}^{(k+1)} = (P-A)\boldsymbol{u}^{(k)} + \boldsymbol{b}$ est convergente.
2. *On suppose dans cette question que chaque matrice A_k est telle que tous ses termes hors diagonale sont négatifs ou nuls, et que*

$$\forall p = 1,\dots,n_k,\ (A_k)_{pp} \ge -\sum_{\substack{q=1,\dots,n_k \\ q\neq p}} (A_k)_{pq}.$$

   (a) Prouver que chaque matrice A_k est (symétrique) positive.
   (b) Prouver que la matrice A est symétrique, que tous ses termes hors diagonale sont négatifs ou nuls, et que

$$\forall i = 1,\dots,n,\ A_{ii} \ge -\sum_{j=1,\dots,n,\ j\neq i} A_{ij}.$$

   [Pour montrer cette dernière égalité, on pourra commencer par remarquer que pour tout $k = 1,\dots,M$ et pour tout $q = 1,\dots,n_k$, on a $\sum_{j=1}^{n} (H_k)_{qj} = 1$.]
3. On suppose que A est inversible. Prouver que la matrice A est une ICP-matrice (cette propriété permet en particulier de déduire la monotonie de la méthode d'approximation, voir le cours).

Corrigé de la partie 5.2.4

1. *Propriétés spectrales de A.*

(a) Soit $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_m$ une base orthonormée de $\mathbb{R}^m$ telle que chaque $\boldsymbol{f}_i$ est un vecteur propre de B associé à la valeur propre μ_i, et $\mu_1 \leq \ldots \leq \mu_n$. Pour $\boldsymbol{x} \in \mathbb{R}^m$ qui s'écrit $\boldsymbol{u} = \sum_{i=1}^m a_i f_i$ avec $\boldsymbol{u}^t\boldsymbol{u} = \sum_{i=1}^m a_i^2 = 1$, on a

$$\boldsymbol{u}^t B \boldsymbol{u} = \sum_{i=1}^m \mu_i a_i^2.$$

On a donc $\boldsymbol{u}^t B \boldsymbol{u} \in [\mu_1, \mu_n]$, et les deux extrémités de cet intervalle sont respectivement atteintes pour $\boldsymbol{u} = \boldsymbol{f}_1$ et $\boldsymbol{u} = \boldsymbol{f}_m$.

(b) Comme les matrices A_k sont symétriques, la question 1(a) donne que pour tout $\boldsymbol{v} \in \mathbb{R}^{n_k}$ tel que $\boldsymbol{v}^t\boldsymbol{v} = 1$,

$$\underline{\lambda}_k \leq \boldsymbol{v}^t A_k \boldsymbol{v} \leq \overline{\lambda}_k,$$

En prenant $\boldsymbol{v} = \dfrac{H_k\boldsymbol{u}}{((H_k\boldsymbol{u})^t H_k\boldsymbol{u})^{\frac{1}{2}}}$, qui est bien tel que $\boldsymbol{v}^t\boldsymbol{v} = 1$, on obtient

$$\underline{\lambda}_k (H_k\boldsymbol{u})^t(H_k\boldsymbol{u}) \leq (H_k\boldsymbol{u})^t A_k (H_k\boldsymbol{u}) \leq \overline{\lambda}_k (H_k\boldsymbol{u})^t(H_k\boldsymbol{u}).$$

(c) On a, pour $k = 1, \ldots, M$ et $i, j = 1, \ldots, n$

$$(H_k^t H_k)_{ij} = \sum_{\ell=1}^{n_k} (H_k)_{\ell,i}(H_k)_{\ell,j}.$$

On ne peut avoir $(H_k)_{\ell,i}(H_k)_{\ell,j}$ non nul que si $i = \varphi_k(\ell)$ et $j = \varphi_k(\ell)$, donc si $i = j \in \varphi_k(\{1, \ldots, n_k\})$. Dans ce cas, comme φ_k est injective, il n'existe qu'un seul $\ell = 1, \ldots, n_k$ tel que $i = \varphi_k(\ell)$. Donc

$$H_k^t H_k = \mathrm{diag}\Big((\chi_{E_i}(k))_{i=1,\ldots,n}\Big),$$

avec $\chi_{E_i}(k) = 1$ si $k \in E_i$ et 0 sinon.

La matrice $\displaystyle\sum_{k=1}^M \alpha_k H_k^t H_k$ est donc aussi une matrice diagonale d'ordre n dont le i-ème terme diagonal est égal à $\displaystyle\sum_{k=1}^M \alpha_k \chi_{E_i}(k)$. Par définition de χ_{E_i}, on a donc bien (5.6).

(d) Par définition de A, on a

$$\boldsymbol{u}^t A \boldsymbol{u} = \sum_{k=1}^M \boldsymbol{u}^t H_k^t A_k H_k \boldsymbol{u} = \sum_{k=1}^M (H_k\boldsymbol{u})^t A_k (H_k\boldsymbol{u}).$$

donc

$$\boldsymbol{u}^t\Big(\sum_{k=1}^M \underline{\lambda}_k H_k^t H_k\Big)\boldsymbol{u} \leq \boldsymbol{u}^t A \boldsymbol{u} \leq \boldsymbol{u}^t\Big(\sum_{k=1}^M \overline{\lambda}_k H_k^t H_k\Big)\boldsymbol{u}.$$

Grâce à la question 1(b), on a

$$\boldsymbol{u}^t\Big(\sum_{k=1}^M \underline{\lambda}_k H_k^t H_k\Big)\boldsymbol{u} = \sum_{i=1}^n u_i^2 \sum_{k\in E_i} \underline{\lambda}_k \geq \Big(\min_{i=1,\ldots,n} \sum_{k\in E_i} \underline{\lambda}_k\Big) \sum_{i=1}^n u_i^2,$$

et

$$\boldsymbol{u}^t\Big(\sum_{k=1}^M \overline{\lambda}_k H_k^t H_k\Big)\boldsymbol{u} = \sum_{i=1}^n u_i^2 \sum_{k\in E_i} \overline{\lambda}_k \leq \Big(\max_{i=1,\ldots,n} \sum_{k\in E_i} \overline{\lambda}_k\Big) \sum_{i=1}^n u_i^2,$$

ce qui permet de conclure compte tenu que $\sum_{i=1}^n u_i^2 = \boldsymbol{u}^t\boldsymbol{u} = 1$.

(e) Comme pour tout $k = 1, \dots, M$, on a supposé que $\underline{\lambda}_k \geq 0$, on en déduit que $\underline{\lambda} \geq 0$ donc que A est symétrique positive. Si A est inversible, $\underline{\lambda} \neq 0$, donc A est symétrique définie positive. Si chaque matrice A_k est symétrique définie positive, alors $\underline{\lambda}_k > 0$, donc $\underline{\lambda} > 0$, d'où la conclusion.

(f) Comme les matrices A_k sont supposées non nulles, on a donc $\overline{\lambda}_k > 0$, donc tous les termes diagonaux de la matrice P sont non nuls, et la matrice P est donc inversible. Comme A est symétrique définie positive d'après la question précédente, on peut essayer d'utiliser la condition suffisante donné par le lemme 1.56 du cours. Examinons donc la matrice $P^t + N$ (avec les notations du cours). On a

$$P^t + N = 2P - A = \sum_{k=1}^{M} H_k^t \big((1+2\varepsilon)\overline{\lambda}_k \mathrm{I}_{n_k} - A_k\big) H_k.$$

On peut alors appliquer les conclusions de la question c) en remplaçant A_k par les matrices $(1 + 2\varepsilon)\overline{\lambda}_k \mathrm{I}_{n_k} - A_k$. Ces matrices sont symétriques positives, de plus petite valeur propre $2\varepsilon\overline{\lambda}_k > 0$, et sont donc symétriques définies positives. La question d) permet alors de conclure que $P^t + N$ est symétrique définie positive, donc le critère du cours (lemme 1.56) permet de conclure que $\rho(P^{-1}(P - A)) < 1$, donc la méthode itérative converge.

2. (a) Soit $u \in \mathbb{R}^{n_k}$. On a, en additionnant et en retranchant le terme diagonal,

$$\boldsymbol{u}^t A_k \boldsymbol{u} = \sum_{p=1}^{n_k}\sum_{q=1}^{n_k} (A_k)_{pq} u_p u_q = \sum_{p=1}^{n_k} \Big((A_k)_{pp} + \sum_{\substack{q=1,\dots,n_k \\ q\neq p}} (A_k)_{pq}\Big) u_p^2 + \sum_{p=1}^{n_k} \sum_{\substack{q=1,\dots,n_k \\ q\neq p}} (A_k)_{pq}(u_p u_q - u_p^2).$$

Le dernier terme du membre de droite de l'équation ci-dessus, qu'on note X, s'écrit, compte tenu de la symétrie de A_k,

$$\begin{aligned}
X &= \sum_{p=1}^{n_k} \sum_{\substack{q=1,\dots,n_k \\ q\neq p}} (A_k)_{pq}(u_p u_q - u_p^2) \\
&= \sum_{p=1}^{n_k}\sum_{q>p} (A_k)_{pq}(u_p u_q - u_p^2) + \sum_{p=1}^{n_k}\sum_{q<p} (A_k)_{pq}(u_p u_q - u_p^2) \\
&= \sum_{p=1}^{n_k}\sum_{q>p} (A_k)_{pq}(u_p u_q - u_p^2) + \sum_{p=1}^{n_k}\sum_{q>p} (A_k)_{pq}(u_p u_q - u_q^2) (\text{ car } (A_k)_{qp} = (A_k)_{pq}) \\
&= \sum_{p=1}^{n_k}\sum_{q>p} (A_k)_{pq}\Big(- (u_q - u_p)^2\Big) \\
&\geq 0
\end{aligned}$$

car $(A_k)_{pq} \leq 0$ par hypothèse. On a donc

$$\boldsymbol{u}^t A_k \boldsymbol{u} = \sum_{p=1}^{n_k} \Big((A_k)_{pp} + \sum_{\substack{q=1,\dots,n_k \\ q\neq p}} (A_k)_{pq}\Big) u_p^2 \geq 0 + \sum_{p=1}^{n_k}\sum_{q>p} (-(A_k)_{pq})(u_p - u_q)^2 \geq 0,$$

ce qui conclut la preuve de la positivité de A_k.

(b) Pour tout $k = 1, \dots M$, la matrice $H_k^t A_k H_k$ est symétrique car elle vérifie l'identité $(H_k^t A_k H_k)^t = H_k^t A_k H_k$. La matrice A est donc elle aussi symétrique.

Pour $i, j = 1, \ldots, n$ on a

$$A_{ij} = \sum_{k=1}^{M} \sum_{p=1}^{n_k} \sum_{q=1}^{n_k} (H_k)_{pi} (A_k)_{pq} (H_k)_{qj}$$

Supposons $i \neq j$. Si $(H_k)_{pi}(H_k)_{qj} \neq 0$, on a $i = \varphi_k(p)$ et $j = \varphi_k(q)$, et comme $i \neq j$ et que φ est injective, on a $p \neq q$, et donc $(A_k)_{pq} \leq 0$, ce qui prouve que $A_{ij} \leq 0$.
Maintenant pour $i = 1, \ldots, n$, on a

$$\sum_{j=1}^{n} A_{ij} = \sum_{j=1}^{n} \sum_{k=1}^{M} \sum_{p=1}^{n_k} \sum_{q=1}^{n_k} (H_k)_{pi} (A_k)_{pq} (H_k)_{qj} = \sum_{k=1}^{M} \sum_{p=1}^{n_k} \sum_{q=1}^{n_k} (H_k)_{pi} (A_k)_{pq} \sum_{j=1}^{n} (H_k)_{qj}$$

Or $\displaystyle\sum_{j=1}^{n} (H_k)_{qj} = 1$, et donc

$$\sum_{j=1}^{n} A_{ij} = \sum_{k=1}^{M} \sum_{p=1}^{n_k} (H_k)_{pi} \sum_{q=1}^{n_k} (A_k)_{pq}.$$

Or par hypothèse sur A_k, $\displaystyle\sum_{q=1}^{n_k} (A_k)_{pq} \geq 0$ et comme $(H_k)_{pi} \geq 0$, on a $\displaystyle\sum_{j=1^n} A_{ij} \geq 0$, soit encore $A_{ii} \geq - \displaystyle\sum_{j=1,\ldots,n,\ j\neq i} A_{ij}$.

3. Le résultat de la question précédente permet d'appliquer la question 5 de l'exercice 14 du polycopié, ce qui prouve que A est une ICP-matrice.

5.2.5 Deux cas particuliers

1. **Un exemple en dimension** 1**.** *Soient $n \geq 1$ et $M = n + 1$. Pour $k = 1, \ldots, M$, on définit la matrice carrée $A_k \in \mathcal{M}_{n_k}(\mathbb{R})$ (matrice d'assemblage) et le second membre d'assemblage $\boldsymbol{b}_k \in \mathbb{R}^{n_k}$ par*

   (i) Si $k = 1$: $A_k = \dfrac{1}{h}\begin{bmatrix}1\end{bmatrix}$, $\varphi_k(1) = 1$, $b_k = \begin{bmatrix}h\end{bmatrix}$;

   (ii) Si $k = 2, \ldots, n$: $A_k = \dfrac{1}{h}\begin{bmatrix}1 & -1\\ -1 & 1\end{bmatrix}$, $\begin{cases}\varphi_k(1) = k-1\\ \varphi_k(2) = k\end{cases}$, $b_k = h\begin{bmatrix}1\\1\end{bmatrix}$;

   (iii) Si $k = n + 1$: $A_k = \dfrac{1}{h}\begin{bmatrix}1\end{bmatrix}$, $\varphi_k(1) = n$, $b_k = \begin{bmatrix}h\end{bmatrix}$.

   (a) Reprendre la figure 5.1 pour donner une représentation graphique du maillage et des degrés de liberté pour un n quelconque.

   (b) Expliciter les ensembles E_i, $i = 1, \ldots, n$.

   (c) Prouver que tous les vecteurs propres de A sont de la forme transposé de $(\sin(k\frac{p\pi}{n+1}))_{k=1,\ldots,n}$, pour tout $p = 1, \ldots, n$), et donner les valeurs propres correspondantes. Comparer les bornes inférieure et supérieure des valeurs propres de A au majorant et minorant trouvés dans la partie 5.2.4.

   (d) Expliciter la matrice P construite dans la partie 5.2.4. Les méthodes de Jacobi, de Gauss-Seidel ou la méthode SOR sont-elles convergentes pour un système linéaire dont la matrice est A ?

2. **Un exemple en dimension** 2**.** *Soient $m \geq 2$, $M = (m + 1)^2$, $n = m^2$. On maille le carré unité $\Omega =]0, 1[\times]0, 1[$ en n petits carrés identiques de côté $h = \dfrac{1}{m+1}$, pour chercher des valeurs approchées de la solution (dont on on admettra l'existence et l'unicité) de l'équation aux dérivées partielles suivante :*

$$-\partial_{xx}u - \partial_{yy}u = 2 \text{ dans } \Omega,$$
$$u = 0 \text{ sur } \partial\Omega.$$

où $\partial_{xx}u$ (resp. $\partial_{yy}u$) désigne la dérivée partielle seconde par rapport à x (resp. y) de u.

(a) Dessiner ce maillage pour $m = 2$ et donner les valeurs correspondantes de M, n et h. Repérer les noeuds correspondant à des degrés de liberté par un $\bullet$ et les autres par un $\circ$ comme dans la figure 5.1.

Pour $i = 1, \dots, m+1$ et $j = 1, \dots, m+1$, on pose $k = i + (m+1)(j-1)$, (noter que $k \in \{1, \dots, M\}$), et on définit la matrice d'assemblage) $A_k \in \mathcal{M}_{n_k}(\mathbb{R})$ et le second membre d'assemblage $\boldsymbol{b}_k \in \mathbb{R}^{n_k}$ par

(a) Si $\begin{cases} i = 1 \\ j = 1 \end{cases}$, $n_k = 1$ *et* $A_k = \dfrac{1}{h}\begin{bmatrix}2\end{bmatrix}$, $\varphi_k(1) = 1$, $b_k = \begin{bmatrix}h\end{bmatrix}$;

(b) Si $\begin{cases} i = 2, \dots, m \\ j = 1 \end{cases}$, $n_k = 2$ *et* $A_k = \dfrac{1}{h}\begin{bmatrix}2 & -1\\ -1 & 2\end{bmatrix}$, $\begin{cases} \varphi_k(1) = i-1 \\ \varphi_k(2) = i \end{cases}$, $b_k = h\begin{bmatrix}1\\1\end{bmatrix}$;

(c) Si $\begin{cases} i = m+1 \\ j = 1 \end{cases}$, $n_k = 1$ *et* $A_k = \dfrac{1}{h}\begin{bmatrix}2\end{bmatrix}$, $\varphi_k(1) = m$, $b_k = h\begin{bmatrix}1\end{bmatrix}$;

(d) Si $\begin{cases} i = 1 \\ j = 2, \dots, m \end{cases}$, $n_k = 2$ *et* $A_k = \dfrac{1}{h}\begin{bmatrix}2 & -1\\ -1 & 2\end{bmatrix}$, $\begin{cases} \varphi_k(1) = 1 + (j-2)m \\ \varphi_k(2) = 1 + (j-1)m \end{cases}$, $b_k = h\begin{bmatrix}1\\1\end{bmatrix}$;

(e) Si $\begin{cases} i = 2, \dots, m \\ j = 2, \dots, m \end{cases}$, $n_k = 4$ *et* $A_k = \dfrac{1}{h}\begin{bmatrix}2 & -1 & -1 & 0\\ -1 & 2 & 0 & -1\\ -1 & 0 & 2 & -1\\ 0 & -1 & -1 & 2\end{bmatrix}$,

$$\begin{cases} \varphi_k(1) = i - 1 + (j-2)m \\ \varphi_k(2) = i + (j-2)m \\ \varphi_k(3) = i - 1 + (j-1)m \\ \varphi_k(4) = i + (j-1)m \end{cases}, \ b_k = h\begin{bmatrix}1\\1\\1\\1\end{bmatrix};$$

(f) Si $\begin{cases} i = m+1 \\ j = 2, \dots, m \end{cases}$, $n_k = 2$ *et* $A_k = \dfrac{1}{h}\begin{bmatrix}2 & -1\\ -1 & 2\end{bmatrix}$, $\begin{cases} \varphi_k(1) = m + (j-2)m \\ \varphi_k(2) = m + (j-1)m \end{cases}$, $b_k = h\begin{bmatrix}1\\1\end{bmatrix}$;

(g) Si $\begin{cases} i = 1 \\ j = m+1 \end{cases}$, $n_k = 1$ *et* $A_k = \dfrac{1}{h}\begin{bmatrix}2\end{bmatrix}$, $\varphi_k(1) = 1 + m(m-1)$, $b_k = \begin{bmatrix}h\end{bmatrix}$;

(h) Si $\begin{cases} i = 2, \dots, m \\ j = m+1 \end{cases}$, $n_k = 2$ *et* $A_k = \dfrac{1}{h}\begin{bmatrix}2 & -1\\ -1 & 2\end{bmatrix}$, $\begin{cases} \varphi_k(1) = i - 1 + m(m-1) \\ \varphi_k(2) = i + m(m-1) \end{cases}$, $b_k = h\begin{bmatrix}1\\1\end{bmatrix}$;

(i) Si $\begin{cases} i = m+1 \\ j = m+1 \end{cases}$, $n_k = 1$ *et* $A_k = \dfrac{1}{h}\begin{bmatrix}2\end{bmatrix}$, $\varphi_k(1) = m^2$, $b_k = h\begin{bmatrix}1\end{bmatrix}$.

(b) Prouver que la matrice A_k est symétrique positive dans le cas génfal et symétrique d ?inie positive si i ou $j \in \{1, m+1\}$.

(c) Pour tout $I = 1, \dots, n$, en vérifiant $I = i + (j-1)m$, $i = 1, \dots, m$ et $j = 1, \dots, m$, expliciter les ensembles E_I à l'aide de i et j.

(d) Donner la matrice A et le second membre $\boldsymbol{b}$ dans le cas particulier $m = 2$.

(e) Prouver que tous les vecteurs propres de A sont de la forme transposé de

$$\Big(\sin(i\frac{p\pi}{m+1})\sin(j\frac{q\pi}{m+1})\Big)_{(i+(j-1)m)=1,\dots,m^2},$$

pour tout $p, q = 1, \dots, m$), et donner les valeurs propres correspondantes. Comparer les bornes inférieur e et supérieur e des valeurs propres de A au majorant et minorant trouvé dans la partie 5.2.4, question 1(d).

(f) Expliciter la matrice P construite dans la partie 5.2.4, question 1(f).

(g) Les méthodes de Jacobi, de Gauss-Seidel ou la méthode SOR sont-elles convergentes pour un système linéaire dont la matrice est A ?

Corrigé de la partie 5.2.5

1. **Un exemple en dimension** 1**.**
   (a)
   (b) On a $E_i = \{i, i+1\}$ pour tout $i = 1, \dots, n$.
   (c) On a, en passant par exemple par les nombres complexes, pour tout $k = 1, \dots, n$
   $$-\sin((k-1)\frac{p\pi}{n+1}) + 2\sin(k\frac{p\pi}{n+1}) - \sin((k+1)\frac{p\pi}{n+1}) = \sin(k\frac{p\pi}{n+1})(2 - 2\cos(\frac{p\pi}{n+1})).$$
   Comme, pour $k = 1$, $\sin((k-1)\frac{p\pi}{n+1}) = 0$, et que, pour $k = n$, $\sin((k+1)\frac{p\pi}{n+1}) = 0$, on obtient que le vecteur transposé de $(\sin(k\frac{p\pi}{n+1}))_{k=1,\dots,n}$ est vecteur propre de A associé à la valeur propre $2 - 2\cos(\frac{p\pi}{n+1}) = 4\sin^2(\frac{p\pi}{2(n+1)})$. Toutes ces valeurs propres étant distinctes et au nombre de n pour $p = 1, \dots, n$, on a donc trouvé tous les vecteurs propres de la matrice A, et les vecteurs forment une base orthogonale de $\mathbb{R}^n$.
   Ces valeurs propres sont comprises entre $4\sin^2(\frac{\pi}{2(n+1)})$ et $4(1 - \sin^2(\frac{\pi}{2(n+1)}))$. Or on a
   $$\min_{i=1,\dots,n} \sum_{k\in E_i} \underline{\lambda}_k = 0 \text{ et } \max_{i=1,\dots,n} \sum_{k\in E_i} \overline{\lambda}_k = 2 + 2 = 4.$$
   On obtient que ces minorants et majorants sont proches d'un facteur en C/n^2 de la plus petite et plus grande valeur propre de A.
   (d) On a
   $$P = (\frac{1}{2} + \varepsilon)\text{diag}\Big(3, 4, \dots, 4, 3\Big) = \text{diag}\Big(\frac{3}{2} + 2\varepsilon, 2 + 2\varepsilon, \dots, 2 + 2\varepsilon, \frac{3}{2} + 2\varepsilon\Big).$$
   Puisque la matrice A est inversible et donc symétrique définie positive, la méthode de Gauss-Seidel et la méthode SOR (pour $\omega \in]0, 2[$) en vertu du théorème 1.57. Comme la matrice est tridiagonale, l'exercice 70 montre que $\rho(B_J)^2 = \rho(B_{GS}) < 1$, donc la méthode de Jacobi converge également. Noter que, pour la méthode de Jacobi, la matrice P est égale à $2\mathrm{I_n}$.

2. **Un exemple en dimension** 2**.**
   (a) On divise le carré $[0, 1]^2$ en $(m+1)^2$ carrés égaux (les éléments), et on numérote les noeuds intérieurs de 1 à m^2 par ligne.
   (b) Le spectre des matrices $A_k = \begin{bmatrix} 2 \end{bmatrix}$ est égal à $\{2\}$, celui des matrices $A_k = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$ est égal à $\{1, 3\}$, et celui des matrices $A_k = \begin{bmatrix} 2 & -1 & -1 & 0 \\ -1 & 2 & 0 & -1 \\ -1 & 0 & 2 & -1 \\ 0 & -1 & -1 & 2 \end{bmatrix}$ est $\{0, 2, 4\}$ (2 est valeur propre double dans ce cas).
   (c) On a $E_I = \{i + (j-1)(m+1), i + 1 + (j-1)(m+1), i + j(m+1), i + 1 + j(m+1)\}$.
   (d) $A = \begin{bmatrix} 8 & -2 & -2 & 0 \\ -2 & 8 & 0 & -2 \\ -2 & 0 & 8 & -2 \\ 0 & -2 & -2 & 8 \end{bmatrix}$ et $\boldsymbol{b} = \begin{bmatrix} 4 \\ 4 \\ 4 \\ 4 \end{bmatrix}$.
   (e) Le calcul effectué dans le cas particulier 1D permet de montrer que ces vecteurs sont associés à la valeur propre $8 - 4\cos(\frac{p\pi}{m+1}) - 4\cos(\frac{q\pi}{m+1})$. Montrons qu'ils forment une famille libre. Supposons que les coefficients α_{pq} soient tels que
   $$\forall i, j = 1, \dots, m, \ \sum_{p=1}^{m}\sum_{q=1}^{m} \alpha_{pq} \sin(i\frac{p\pi}{m+1}) \sin(j\frac{q\pi}{m+1})$$
   $$= \sum_{p=1}^{m} \sin(i\frac{p\pi}{m+1}) \sum_{q=1}^{m} \alpha_{pq} \sin(j\frac{q\pi}{m+1}) = 0.$$

On a alors, comme les vecteurs $(\sin(i\frac{p\pi}{m+1}))_i$ forment une famille libre pour $p = 1, \ldots, m$,

$$\forall j, p = 1, \ldots, m, \sum_{q=1}^{m} \alpha_{pq} \sin(j\frac{q\pi}{m+1}) = 0,$$

et comme les $(\sin(j\frac{q\pi}{m+1}))_j$ forment une famille libre pour $q = 1, \ldots, m$,

$$\forall p, q = 1, \ldots, m, \alpha_{pq} = 0.$$

Nous avons donc une base de vecteurs propres. Les valeurs propres sont comprises entre C/n^2 et $16 - C/n^2$, et les bornes données par la partie 5.2.4 sont 0 et 16.

(f) La matrice P est la matrice diagonale telle que, pour $k = i + m(j-1)$, avec $i = 1, \ldots, m$ et $j = 1, \ldots, m$:

i. Si $\left\{ \begin{array}{l} i = 1 \\ j = 1 \end{array} \right.$: $P_{kk} = (\frac{1}{2} + \varepsilon)(2 + 3 + 3 + 4)$;

ii. Si $\left\{ \begin{array}{l} i = 2, \ldots, m-1 \\ j = 1 \end{array} \right.$: $P_{kk} = (\frac{1}{2} + \varepsilon)(3 + 3 + 4 + 4)$;

iii. Si $\left\{ \begin{array}{l} i = m \\ j = 1 \end{array} \right.$: $P_{kk} = (\frac{1}{2} + \varepsilon)(2 + 3 + 3 + 4)$;

iv. Si $\left\{ \begin{array}{l} i = 1 \\ j = 2, \ldots, m-1 \end{array} \right.$: $P_{kk} = (\frac{1}{2} + \varepsilon)(3 + 3 + 4 + 4)$;

v. Si $\left\{ \begin{array}{l} i = 2, \ldots, m-1 \\ j = 2, \ldots, m-1 \end{array} \right.$: $P_{kk} = (\frac{1}{2} + \varepsilon)(4 + 4 + 4 + 4)$;

vi. Si $\left\{ \begin{array}{l} i = m \\ j = 2, \ldots, m-1 \end{array} \right.$: $P_{kk} = (\frac{1}{2} + \varepsilon)(3 + 3 + 4 + 4)$;

vii. Si $\left\{ \begin{array}{l} i = 1 \\ j = m \end{array} \right.$: $P_{kk} = (\frac{1}{2} + \varepsilon)(2 + 3 + 3 + 4)$;

viii. Si $\left\{ \begin{array}{l} i = 2, \ldots, m-1 \\ j = m \end{array} \right.$: $P_{kk} = (\frac{1}{2} + \varepsilon)(3 + 3 + 4 + 4)$;

ix. Si $\left\{ \begin{array}{l} i = 1 \\ j = m \end{array} \right.$: $P_{kk} = (\frac{1}{2} + \varepsilon)(2 + 3 + 3 + 4)$.

(g) Puisque la matrice A est inversible donc symétrique définie positive, le cours montre que la méthode de Gauss-Seidel et la méthode SOR (pour $\omega \in]0, 2[$) convergent. En revanche, contrairement au cas 1D, la matrice n'est plus tridiagonale,et il faut calculer $\rho(B_J)$. Les vecteurs propres de la matrice B_J sont les mêmes que ceux de A, et les valeurs propres associées de la matrice B_J sont $\frac{1}{4}\Big(2\cos(\frac{p\pi}{m+1}) + 2\cos(\frac{q\pi}{m+1})\Big)$, donc toutes strictement comprises entre 0 et 1. La méthode de Jacobi est donc convergente.

5.2.6 Un peu d'optimisation pour terminer...

Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible et soit $\boldsymbol{b} \in \mathbb{R}^n$. On considère les deux fonctions suivantes de $\mathbb{R}^n$ dans $\mathbb{R}$:

$$J : \boldsymbol{u} \in \mathbb{R}^n \mapsto J(\boldsymbol{u}) = \frac{1}{2}\boldsymbol{u}^t A\boldsymbol{u} - \boldsymbol{b}^t\boldsymbol{u} \;\; \textit{et} \;\; \widetilde{J} : \boldsymbol{u} \in \mathbb{R}^n \mapsto \widetilde{J}(\boldsymbol{u}) = (A\boldsymbol{u} - \boldsymbol{b})^t(A\boldsymbol{u} - \boldsymbol{b})$$

1. On suppose dans cette question que A est symétrique définie positive. Montrer que la fonction J (resp. $\widetilde{J}$) admet un unique minimum et que ce minimum est l'unique solution de $\nabla J(\boldsymbol{u}) = 0$ (resp. $\nabla\widetilde{J}(\boldsymbol{u}) = 0$) ; on calculera ces gradients pour expliciter les équations correspondantes.
2. Donner un exemple d'une ICP-matrice 2×2 non symétrique.

3. Donner un exemple d'une ICP-matrice 2×2 symétrique qui ne soit pas symétrique définie positive.
4. On suppose dans cette question que A est une ICP-matrice qui est symétrique mais non définie positive ;
   (a) Montrer que
   $$\inf_{\boldsymbol{u}\in\mathbb{R}^n} J(\boldsymbol{u}) = -\infty$$
   (b) Montrer que $\widetilde{J}$ admet un unique minimum que l'on caractérisera.

Corrigé de la partie 5.2.6

1. La fonction J est quadratique, et comme A est s.d.p. la fonction $\widetilde{J}$ peut s'écrire
   $$\widetilde{J}(\boldsymbol{u}) = A^2\boldsymbol{u}\cdot\boldsymbol{u} - 2A\boldsymbol{b}\cdot\boldsymbol{u} - bfb\cdot\boldsymbol{b}$$
   Les deux fonctions sont donc quadratiques, et les matrices A et A^2 étant s.d.p., le théorème 3.16 sur la minimisation d'une fonctionnelle quadratique s'applique. Les deux fonctions admettent donc chacune un unique minimum qui annule leur gradient. Dans le cas de la fonction J, ce minimum est donc la solution de
   $$\nabla J(\boldsymbol{u}) = A\boldsymbol{u} - \boldsymbol{b} = 0,$$
   et dans le cas de la fonction J, ce minimum est la solution de
   $$\nabla \widetilde{J}(\boldsymbol{u}) = 2A^2\boldsymbol{u} - 2A\boldsymbol{b} = 0,$$
   et comme A est inversible, ceci revient à résoudre
   $$A\boldsymbol{u} - \boldsymbol{b} = 0,$$
2. $A = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}$ et $A^{-1} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$.
3. Soit $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, A^{-1} = \begin{bmatrix} 0 & 1 \\ 1 & 0. \end{bmatrix}$ Les valeurs propres de A (qui est une matrice de permutation) sont 1 et -1, et la matrice A est donc une ICP matrice symétrique qui n'est pas s.d.p.
4. (a) Si A est une IP matrice symétrique mais non s.d.p., ses valeurs propres sont toutes réelles non nulles, et il y en a au moins une strictement négative. Soit λ une valeur propre strictement négative de A, et $\boldsymbol{f}$ un vecteur propre associé, et soit $t > 0$. Alors
   $$J(\boldsymbol{f}) = \frac{1}{2}\lambda t^2 \boldsymbol{f}^t\boldsymbol{f} - t\boldsymbol{b}^t\boldsymbol{f} \to -\infty \text{ lorsque } t \to +\infty,$$
   ce qui prouve le résultat.
   (b) Si la matrice A est une ICP-matrice symétrique, la matrice A^2 est s.d.p. et donc le théorème 3.16 sur la minimisation d'une fonctionnelle quadratique s'applique encore.