Université Aix Marseille 1

Licence de mathématiques

Cours d'Analyse numérique

Raphaèle Herbin

17 avril 2010

Table des matières

1 Systèmes linéaires **4**
- 1.1 Objectifs . . . 4
- 1.2 Quelques rappels d'algèbre linéaire . . . 4
  - 1.2.1 Norme induite . . . 4
  - 1.2.2 Rayon spectral . . . 6
  - 1.2.3 Matrices diagonalisables . . . 7
- 1.3 Les méthodes directes . . . 9
  - 1.3.1 Définition . . . 9
  - 1.3.2 Méthode de Gauss et méthode LU . . . 9
  - 1.3.3 Version programmable des algorithmes de Gauss et LU pour un système carré . . . 12
  - 1.3.4 Méthode de Choleski . . . 14
  - 1.3.5 Quelques propriétés . . . 19
- 1.4 Conditionnement . . . 21
  - 1.4.1 Le problème des erreurs d'arrondis . . . 21
  - 1.4.2 Conditionnement et majoration de l'erreur d'arrondi . . . 21
  - 1.4.3 Discrétisation d'équations différentielles, conditionnement "efficace" . . . 23
- 1.5 Méthodes itératives . . . 25
  - 1.5.1 Origine des systèmes à résoudre . . . 25
  - 1.5.2 Définition et propriétés . . . 28
  - 1.5.3 Méthodes de Jacobi, Gauss-Seidel et SOR/SSOR . . . 30
  - 1.5.4 Recherche de valeurs propres et vecteurs propres . . . 36
- 1.6 Exercices . . . 36
- 1.7 Suggestions . . . 53
- 1.8 Corrigés . . . 57

2 Systèmes non linéaires **82**
- 2.1 Les méthodes de point fixe . . . 82
  - 2.1.1 Point fixe de contraction . . . 82
  - 2.1.2 Point fixe de monotonie . . . 86
  - 2.1.3 Vitesse de convergence . . . 87
- 2.2 Méthode de Newton . . . 89
  - 2.2.1 Construction et convergence de la méthode . . . 89
  - 2.2.2 Variantes de la méthode de Newton . . . 93
- 2.3 Exercices . . . 96
- 2.4 Suggestions . . . 103
- 2.5 Corrigés . . . 105

3 Optimisation **118**
3.1 Définitions et rappels . . . 118
3.1.1 Définition des problèmes d'optimisation . . . 118
3.1.2 Rappels et notations de calcul différentiel . . . 118
3.2 Optimisation sans contrainte . . . 119
3.2.1 Définition et condition d'optimalité . . . 119
3.2.2 Résultats d'existence et d'unicité . . . 120
3.3 Algorithmes d'optimisation sans contrainte . . . 124
3.3.1 Méthodes de descente . . . 124
3.3.2 Algorithmes du gradient conjugué . . . 127
3.3.3 Méthodes de Newton et Quasi–Newton . . . 134
3.3.4 Résumé sur les méthodes d'optimisation . . . 137
3.4 Optimisation sous contraintes . . . 138
3.4.1 Définitions . . . 138
3.4.2 Existence – Unicité – Conditions d'optimalité simple . . . 138
3.4.3 Conditions d'optimalité dans le cas de contraintes égalité . . . 139
3.4.4 Contraintes inégalités . . . 142
3.5 Algorithmes d'optimisation sous contraintes . . . 143
3.5.1 Méthodes de gradient avec projection . . . 143
3.5.2 Méthodes de dualité . . . 145
3.6 Exercices . . . 149
3.7 Suggestions . . . 161
3.8 Corrigés . . . 164

4 Equations différentielles **178**
4.1 Introduction . . . 178
4.2 Consistance, stabilité et convergence . . . 181
4.3 Théorème général de convergence . . . 182
4.4 Exemples . . . 185
4.5 Explicite ou implicite ? . . . 186
4.5.1 L'implicite gagne... . . . 186
4.5.2 L'implicite perd... . . . 187
4.5.3 Match nul . . . 187
4.6 Etude du schéma d'Euler implicite . . . 188
4.7 Exercices . . . 189
4.8 Corrigés . . . 197

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 !) :
– 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.
– J. Hubbard, B. West, Equations différentielles et systèmes dynamiques, Cassini.
– 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.
Ce cours a été rédigé pour la licence de mathématiques par télé-enseignement de l'université d'Aix-Marseille 1. Chaque chapitre est suivi d'un certian nombre d'exercices. On donne ensuite des suggestions pour effectuer les exercices, puis des corrigés détaillés. Il est fortement conseillé d'essayer de faire les exercices d'abord sans ces indications, et de ne regarder les corrigés détaillés qu'une fois l'exercice achevé (même si certaines questions n'ont pas pu être effectuées), ceci pour se préparer aux conditions d'examen.

Chapitre 1

Systèmes linéaires

1.1 Objectifs

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

$$\left\{ \begin{array}{l} x \in \mathbb{R}^N \\ Ax = b \end{array} \right. \tag{1.1.1}$$

Comme A est inversible, il existe un unique vecteur $x \in \mathbb{R}^N$ solution de (1.1.1). Nous allons étudier dans les deux chapitres suivants des méthodes de calcul de ce vecteur 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. Entre 1980 et 2000, la taille de la mémoire des ordinateurs a augmenté de façon drastique. La taille des systèmes qu'on peut résoudre sur ordinateur a donc également augmenté, selon l'ordre de grandeur suivant :

1980 :	matrice "pleine" (tous les termes sont non nuls)	$N = 10^2$
	matrice "creuse"	$N = 10^6$
2000 :	matrice "pleine"	$N = 10^6$
	matrice "creuse"	$N = 10^8$

Le développement des méthodes de résolution de systèmes linéaires est liée à l'évolution des machines informatiques. Un grand nombre de recherches sont d'ailleurs en cours pour 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 Quelques rappels d'algèbre linéaire

1.2.1 Norme induite

Définition 1.1 (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.*

$$\|AB\| \leq \|A\|\|B\|, \ \forall A, B \in \mathcal{M}_N(\mathbb{R}) \tag{1.2.2}$$

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 :*

$$\|A\| = \sup\{\|Ax\|; \quad x \in \mathbb{R}^N, \|x\| = 1\}, \ \forall A \in \mathcal{M}_N(\mathbb{R}) \tag{1.2.3}$$

Proposition 1.2 *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. $\|Ax\| \leq \|A\| \ \|x\|, \quad \forall x \in \mathbb{R}^N$,
2. $\|A\| = \max\left\{\|Ax\| \ ; \ \|x\| = 1, \ x \in \mathbb{R}^N\right\}$,
3. $\|A\| = \max\left\{\dfrac{\|Ax\|}{\|x\|} \ ; \ x \in \mathbb{R}^N \setminus \{0\}\right\}$.
4. *$\|\cdot\|$ est une norme matricielle.*

Démonstration :

1. Soit $x \in \mathbb{R}^N \setminus \{0\}$, posons $y = \dfrac{x}{\|x\|}$, alors $\|y\| = 1$ donc $\|Ay\| \leq \|A\|$. On en déduit $\dfrac{\|Ax\|}{\|x\|} \leq \|A\|$ et donc $\|Ax\| \leq \|A\| \ \|x\|$. Si maintenant $x = 0$, alors $Ax = 0$, et donc $\|x\| = 0$ et $\|Ax\| = 0$; l'inégalité $\|Ax\| \leq \|A\| \ \|x\|$ est encore vérifiée.
2. L'application φ définie de $\mathbb{R}^N$ dans $\mathbb{R}$ par : $\varphi(x) = \|Ax\|$ est continue sur la sphère unité $S_1 = \{x \in \mathbb{R}^N \mid \|x\| = 1\}$ qui est un compact de $\mathbb{R}^N$. Donc φ est bornée et atteint ses bornes : il existe $x_0 \in \mathbb{R}^N$ tel que $\|A\| = \|Ax_0\|$
3. La dernière égalité résulte du fait que $\dfrac{\|Ax\|}{\|x\|} = \|A\dfrac{x}{\|x\|}\|$ et $\dfrac{x}{\|x\|} \in S_1$ pour tout $x \neq 0$.

■

Proposition 1.3 *Soit $A = (a_{i,j})_{i,j\in\{1,\ldots,N\}} \in \mathcal{M}_N(\mathbb{R})$.*

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

$$\|A\|_\infty = \max_{i\in\{1,\ldots,N\}} \sum_{j=1}^{N} |a_{i,j}|. \tag{1.2.4}$$

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*

$$\|A\|_1 = \max_{j\in\{1,\ldots,N\}} \sum_{i=1}^{N} |a_{i,j}| \tag{1.2.5}$$

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

$$\|A\|_2 = (\rho(A^tA))^{\frac{1}{2}}. \tag{1.2.6}$$

La démonstration de cette proposition fait l'objet de l'exercice 3 page 36

1.2.2 Rayon spectral

Définition 1.4 (Valeurs propres et rayon spectral) *Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice inversible. On appelle valeur propre de A tout $\lambda \in \mathbb{C}$ tel qu'il existe $x \in \mathbb{C}^N$, $x \neq 0$ tel que $Ax = \lambda x$. L'élément x est appelé vecteur propre de A associé à λ. On appelle rayon spectral de A la quantité $\rho(A) = \max\{|\lambda|;\ \lambda \in \mathbb{C},\ \lambda$ valeur propre de $A\}$.*

Proposition 1.5 *Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice carrée quelconque et soit $\|\cdot\|$ une norme matricielle (induite ou non). Alors*

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

La preuve de cette proposition fait l'objet de l'exercice 5 page 37. Elle nécessite un résultat d'approximation du rayon spectral par une norme induite bien choisie, que voici :

Proposition 1.6 (Rayon spectral et norme induite) *Soient $A \in \mathcal{M}_N(\mathbb{R})$ et $\varepsilon > 0$. 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 : Soit $A \in \mathcal{M}_N(\mathbb{R})$, alors par le lemme 1.7 donné ci-après, il existe une base $(f_1, \ldots, f_N)$ de $\mathbb{C}^N$ et une famille de complexes $(\lambda_{i,j})_{i,j=1,\ldots,N, j\leq i}$ telles que $Af_i = \sum_{j\leq i} \lambda_{i,j} f_j$. Soit $\eta \in]0,1[$, pour $i = 1, \ldots, N$, on définit $e_i = \eta^{i-1} f_i$. La famille $(e_i)_{i=1,\ldots,N}$ forme une base de $\mathbb{C}^N$. On définit alors une norme sur $\mathbb{R}^N$ par $\|x\| = (\sum_{i=1}^N \alpha_i \overline{\alpha_i})^{1/2}$, où les α_i sont les composantes de x dans la base $(e_i)_{i=1,\ldots,N}$. Notons que cette norme dépend de A et de η. Soit $\varepsilon > 0$. Montrons que si η bien choisi, on a $\|A\| \leq \rho(A) + \varepsilon$. Soit $x = \sum_{i=1,\ldots,N} \alpha_i e_i$. Comme

$$Ae_i = \sum_{1\leq j\leq i} \eta^{i-j} \lambda_{i,j} e_j,$$

on a donc :

$$Ax = \sum_{i=1}^N \sum_{1\leq j\leq i} \eta^{i-j} \lambda_{i,j} \alpha_i e_j = \sum_{j=1}^N \Big(\sum_{i=j}^N \eta^{i-j} \lambda_{i,j} \alpha_i\Big) e_j$$

On en déduit que

$$\|Ax\|^2 = \sum_{j=1}^N \Big(\sum_{i=j}^N \eta^{i-j} \lambda_{i,j} \alpha_i\Big) \Big(\sum_{i=j}^N \eta^{i-j} \overline{\lambda}_{i,j} \overline{\alpha}_i\Big),$$

soit encore

$$\|Ax\|^2 = \sum_{j=1}^N \lambda_{j,j} \overline{\lambda_{j,j}} \alpha_j \overline{\alpha_j} + \sum_{j=1}^N \sum_{\substack{k,\ell\geq j \\ (k,\ell)\neq(j,j)}} \eta^{k+\ell-2j} \lambda_{k,j} \overline{\lambda_{\ell,j}} \alpha_k \overline{\alpha_\ell}.$$

Comme $\eta \in]0,1[$, on en conclut que :

$$\|Ax\|^2 \leq \rho(A)\|x\|^2 + N^3 \rho(A)^2 \|x\|^2 \eta.$$

D'où le résultat, en prenant η tel que $\eta N^3 \rho(A)^2 < \varepsilon$.

■

Lemme 1.7 (Triangularisation d'une matrice) *Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice carrée quelconque, alors il existe une base $(f_1, \ldots, f_N)$ de $\mathbb{C}$ et une famille de complexes $(\lambda_{i,j})_{i=1,\ldots,N, j=1,\ldots,N, j<i}$ telles que $Af_i = \lambda_{i,i} f_i + \sum_{j<i} \lambda_{i,j} f_j$. De plus $\lambda_{i,i}$ est valeur propre de A pour tout $i \in \{1, \ldots, N\}$.*

On admettra ce lemme.
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 1.8 (Matrices de la forme $Id + A$)

1. *Soit une norme matricielle induite, Id la matrice identité de $\mathcal{M}_N(\mathbb{R})$ et $A \in \mathcal{M}_N(\mathbb{R})$ telle que $\|A\| < 1$. Alors la matrice $Id + A$ est inversible et*

$$\|(Id + A)^{-1}\| \leq \frac{1}{1 - \|A\|}.$$

2. *Si une matrice de la forme $Id + A \in \mathcal{M}_N(\mathbb{R})$ est singulière, alors $\|A\| \geq 1$ pour toute norme matricielle $\|\cdot\|$.*

Démonstration :

1. La démonstration du point 1 fait l'objet de l'exercice 9 page 37.
2. Si la matrice $Id + A \in \mathcal{M}_N(\mathbb{R})$ est singulière, alors $\lambda = -1$ est valeur propre, et donc en utilisant la proposition 1.6, on obtient que $\|A\| \geq \rho(A) \geq 1$.

■

1.2.3 Matrices diagonalisables

Définition 1.9 (Matrice diagonalisable) *Soit A une matrice réelle carrée d'ordre n. On dit que A est diagonalisable dans $\mathbb{R}$ si il existe une base $(\phi_1, \ldots, \phi_n)$ et des réels $\lambda_1, \ldots, \lambda_n$ (pas forcément distincts) tels que $A\phi_i = \lambda_i\phi_i$ pour $i = 1, \ldots, n$. Les réels $\lambda_1, \ldots, \lambda_n$ sont les valeurs propres de A, et les vecteurs $\phi_1, \ldots, \phi_n$ sont les vecteurs propres associés.*

Vous connaissez sûrement aussi la diagonalisation dans $\mathbb{C}$: une matrice réelle carrée d'ordre n est diagonalisable dans $\mathbb{C}$ si il existe une base $(\phi_1, \ldots, \phi_n)$ de $\mathbb{C}^n$ et des nombres complexes $\lambda_1, \ldots, \lambda_n$ (pas forcément distincts) tels que $A\phi_i = \lambda_i\phi_i$ pour $i = 1, \ldots, n$. Ici et dans toute la suite, comme on résout des systèmes linéaire réels, on préfère travailler avec la diagonlaisation 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 1.10 *Soit A une matrice réelle carrée d'ordre n, diagonalisable dans $\mathbb{R}$. Alors*

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

où P est la matrice dont les vecteurs colonnes sont égaux aux vecteurs $\phi_1, \ldots, \phi_n$.

Démonstration Soit P la matrice définie (de manière unique) par $Pe_i = \phi_i$, où $(e_i)_{i=1,\ldots,n}$ est la base canonique de $\mathbb{R}^n$, c'est-à-dire que $(e_i)_j = \delta_{i,j}$. La matrice P est appelée matrice de passage de la base $(e_i)_{i=1,\ldots,n}$ à la base $(\phi_i)_{i=1,\ldots,n}$; la i-ème colonne de P est constituée des composantes de ϕ_i dans la base canonique $(e_1, \ldots, e_N)$. La matrice P est évidemment inversible, et on peut écrire :

$$A\phi_i = APe_i = \lambda_i\phi_i,$$

de sorte que :

$$P^{-1}APe_i = \lambda_i P^{-1}\phi_i = \lambda_i e_i.$$

On a donc bien $P^{-1}AP = \mathrm{d}iag(\lambda_1, \ldots, \lambda_n)$ (ou encore $A = P\mathrm{d}iag(\lambda_1, \ldots, \lambda_n)P^{-1}$).

La diagonalisation des matrices réelles symétriques est un outil qu'on utilisera souvent dans la suite, en particulier dans les exercices.

Lemme 1.11 *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*

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

qui vérifie :

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

Ce produit scalaire induit une norme sur E, $\|x\| = \sqrt{(x \mid x)_E}$.
Soit T une application linéaire de E dans E. On suppose que T est symétrique, c.à.d. que $(T(x) \mid y)_E = (x \mid T(y))_E$, $\forall (x, y) \in E^2$. Alors il existe une base orthonormée $(f_1 \dots f_n)$ de E (c.à.d. telle que $(f_i, f_j)_E = \delta_{i,j}$) et $(\lambda_1 \dots \lambda_n) \in \mathbb{R}^n$ tels que $T(f_i) = \lambda_i f_i$ pour tout $i \in \{1 \dots n\}$.

Conséquence immédiate : Dans le cas où $E = \mathbb{R}^N$, le produit scalaire canonique de $x = (x_1, \dots, x_N)^t$ et $y = (y_1, \dots, y_N)^t$ est défini par $(x \mid y)_E = x \cdot y = \sum_{i=1}^N x_i y_i$. Si $A \in \mathcal{M}_N(\mathbb{R})$ est une matrice symétrique, alors l'application T définie de E dans E par : $T(x) = Ax$ est linéaire, et : $(Tx|y) = Ax \cdot y = x \cdot A^t y = x \cdot Ay = (x \mid Ty)$. Donc T est linéaire symétrique. Par le lemme précédent, il existe $(f_1 \dots f_N)$ et $(\lambda_1 \dots \lambda_N) \in \mathbb{R}$ tels que $Tf_i = Af_i = \lambda_i f_i \, \forall \, i \in \{1, \dots, N\}$ et $f_i \cdot f_j = \delta_{i,j}, \forall \, (i, j) \in \{1, \dots, N\}^2$.

Interprétation algébrique : Il existe une matrice de passage P de $(e_1, \dots, e_N)$ base canonique dans $(f_1, \dots, f_N)$ dont la première colonne de P est constituée des coordonnées de f_i dans $(e_1 \dots e_N)$. On a : $Pe_i = f_i$. On a alors $P^{-1}APe_i = P^{-1}Af_i = P^{-1}(\lambda_i f_i) = \lambda_i e_i = diag(\lambda_1, \dots, \lambda_N)e_i$, où $diag(\lambda_1, \dots, \lambda_N)$ désigne la matrice diagonale de coefficients diagonaux $\lambda_1, \dots, \lambda_N$. On a donc :

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

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

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

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

Démonstration du lemme 1.11 Cette démonstration se fait par récurrence sur la dimension de E.
1ère étape.
On suppose $dimE = 1$. Soit $e \in E$, $e \neq 0$, alors $E = \mathbb{R}e = f_1$ avec $f_1 = \dfrac{e}{\|e\|}$. Soit $T : E \to E$ linéaire symétrique, on a : $Tf_1 \in \mathbb{R}f_1$ donc il existe $\lambda_1 \in \mathbb{R}$ tel que $Tf_1 = \lambda_1 f_1$.

2ème étape.
On suppose le lemme vrai si $dimE < n$. On montre alors le lemme si $dimE = n$. Soit E un espace vectoriel normé sur $\mathbb{R}$ tel que $dimE = n$ et $T : E \to E$ linéaire symétrique. Soit φ l'application définie par :

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

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

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

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

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

Comme $t > 0$, ceci donne :

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

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

Soit $F = \{x \in E; (x \mid e) = 0\}$, on a donc $F \neq E$, et $E = F \bigoplus \mathbb{R}e$: on peut décomposer $x \in E$ comme $(x = x - (x \mid e)e + (x \mid e)e)$. L'application $S = T|_F$ est linéaire symétrique et on a $dimF = n - 1$. et $S(F) \subset F$. On peut donc utiliser l'hypothèse de récurrence : $\exists(\lambda_1 \dots \lambda_{n-1}) \in \mathbb{R}^n$ et $\exists(f_1 \dots f_{n-1}) \in E^n$ tels que $\forall\, i \in \{1 \dots n-1\}$, $Sf_i = Tf_i = \lambda_i f_i$, et $\forall i, j \in \{1 \dots n-1\}, (f_i \mid f_j) = \delta_{i,j}$. Et donc $(\lambda_1 \dots \lambda_N)$ et $(f_1 \dots f_N)$ conviennent.

1.3 Les méthodes directes

1.3.1 Définition

Définition 1.12 (Méthode directe) *On appelle méthode directe de résolution de (1.1.1) une méthode qui donne exactement x (A et b étant connus) solution de (1.1.1) après un nombre fini d'opérations élémentaires $(+, -, \times, /)$.*

Parmi les méthodes de résolution du système (1.1.1) on citera :

- la méthode de Gauss (avec pivot)
- la méthode LU, qui est une réécriture de la méthode Gauss.

Nous rappelons la méthode de Gauss et 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 et méthode LU

Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice inversible, et $b \in \mathbb{R}^N$. On cherche à calculer $x \in \mathbb{R}^N$ tel que $Ax = b$. Le principe de la méthode de Gauss est de se ramener, par des opérations simples (combinaisons linéaires), à un système triangulaire équivalent, qui sera donc facile à inverser. On pose $A^{(1)} = A$ et $b^{(1)} = b$. Pour $i = 1, \dots, N-1$, on cherche à calculer $A^{(i+1)}$ et $b^{(i+1)}$ tels que les systèmes $A^{(i)}x = b^{(i)}$ et $A^{(i+1)}x = b^{(i+1)}$ soient équivalents, où $A^{(i)}$ est une matrice de la forme suivante :
Une fois la matrice $A^{(N)}$ (triangulaire supérieure) et le vecteur $b^{(N)}$ calculés, il sera facile de résoudre le système $A^{(N)}x = b^{(N)}$. Le calcul de $A^{(N)}$ est l'étape de "factorisation", le calcul de $b^{(N)}$ l'étape de "descente", et le calcul de x l'étape de "remontée". Donnons les détails de ces trois étapes.

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

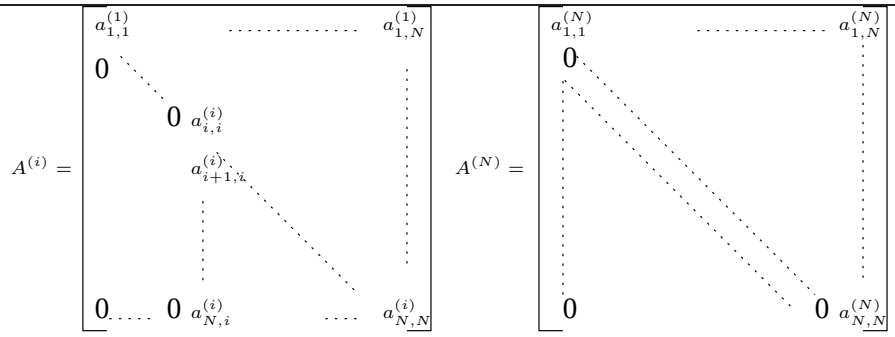

FIG. 1.1 – Allure des matrices de Gauss à l'étape i et à l'étape N

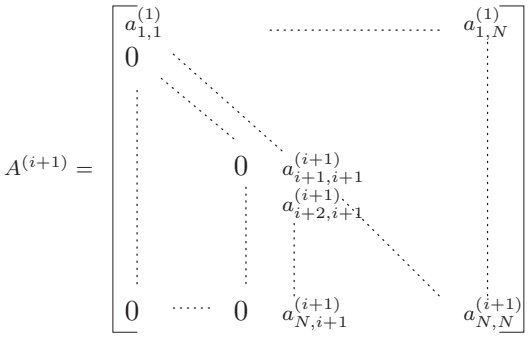

FIG. 1.2 – Allure de la matrice de Gauss à l'étape $i+1$

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

$$a^{(i+1)}_{k,j} = a^{(i)}_{k,j} - \frac{a^{(i)}_{k,i}}{a^{(i)}_{i,i}} a^{(i)}_{i,j}, \text{ pour } k = j, \ldots, N, \tag{1.3.7}$$

$$b^{(i+1)}_{k} = b^{(i)}_{k} - \frac{a^{(i)}_{k,i}}{a^{(i)}_{i,i}} b^{(i)}_{i}. \tag{1.3.8}$$

La matrice $A^{(i+1)}$ est de la forme donnée sur la figure 1.3.2. Remarquons que le système $A^{(i+1)}x = b^{(i+1)}$ est bien équivalent au système $A^{(i)}x = b^{(i)}$.
Si la condition $a^{(i)}_{i,i} \neq 0$ est vérifiée pour $i = 1$ à N, on obtient par le procédé de calcul ci-dessus un système linéaire $A^{(N)}x = b^{(N)}$ équivalent au système $Ax = b$, avec une matrice $A^{(N)}$ triangulaire supérieure facile à inverser. On verra un peu plus loin les techniques de pivot qui permettent de régler le cas où la condition $a^{(i)}_{i,i} \neq 0$ n'est pas vérifiée.

Etape de remontée Il reste à résoudre le système $A^{(N)}x = b^{(N)}$. Ceci est une étape facile. Comme $A^{(N)}$ est une matrice inversible, on a $a^{(i)}_{i,i} \neq 0$ pour tout $i1, \ldots, N$, et comme $A^{(N)}$ est une matrice triangulaire supérieure, on peut donc calculer les composantes de x en "remontant", c'est–à–dire de la composante x_N à la composante x_1 :

$$x_N = \frac{b^{(N)}}{a^{(N)}_{N,N}},$$

$$x_i = \frac{1}{a^{(N)}_{i,i}}(b^{(i)} - \sum_{j=i+1,N} a^{(N)}_{i,j} x_j), i = N-1, \ldots, 1.$$

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 pour effectuer les étapes de factorisation, descente et remontée est $\frac{2}{3}N^3 + O(N^2)$.
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 $Ax = b$ doit être résolu pour plusieurs second membres b, il est évident qu'on a intérêt à ne faire l'étape de factorisation (*i.e.* le calcul de $A^{(N)}$), qu'une seule fois, alors que les étapes de descente et remontée (*i.e.* le calcul de $b^{(N)}$ et x) seront faits pour chaque vecteur b. L'étape de factorisation peut se faire en décomposant la matrice A sous la forme LU.
On admettra le théorème suivant (voir par exemple le livre de Ciarlet cité en début de cours.

Théorème 1.13 (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*

$$PA = LU.$$

Cette décomposition peut se calculer très facilement à partir de la méthode de Gauss. Pour simplifier l'écriture, on supposera ici que lors de la méthode de Gauss, la condition $a^{(i)}_{i,i} \neq 0$ est vérifiée pour tout $i = 1, \ldots, N$. Dans ce cas, la matrice de permutation P du théorème 1.13 est la matrice identité. La matrice L a comme coefficients $\ell_{i,j} = \frac{a^{(i)}_{i,j}}{a^{(i)}_{i,i}}$ pour $i > j$, $\ell_{i,i} = 1$ pour tout $i = 1, \ldots, N$, et $\ell_{i,j} = 0$ pour $j > i$, et la matrice U est égale à la matrice $A^{(N)}$. On peut vérifier que $A = LU$ grâce au fait que le système $A^{(N)}x = b^{(N)}$ est équivalent au système $Ax = b$. En effet, comme $A^{(N)}x = b^{(N)}$ et $b^{(N)} = L^{-1}b$, on en déduit que $LUx = b$, et comme A et LU sont inversibles, on en déduit que $A^{-1}b = (LU)^{-1}b$ pour tout $b \in \mathbb{R}^N$. Ceci démontre que $A = LU$.

Techniques de pivot Dans la présentation de la méthode de Gauss et de la décomposition LU, on a supposé que la condition $a^{(i)}_{i,i} \neq 0$ était vérifiée à chaque étape. Or il peut s'avérer que ce ne soit pas le cas, ou que, même si la condition est vérifiée, le "pivot" $a^{(i)}_{i,i}$ soit très petit, ce qui peut entraîner des erreurs d'arrondi importantes dans les calculs. On peut résoudre ce problème en utilisant les techniques de "pivot partiel" ou "pivot total", qui reviennent à choisir une matrice de permutation P qui n'est pas forcément égale à la matrice identité dans le théorème 1.13.
Plaçons-nous à l'itération i de la méthode de Gauss. Comme la matrice $A^{(i)}$ est forcément non singulière, on a :

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

On a donc en particulier

$$det \begin{pmatrix} a^{(i)}_{i,i} & \ldots & a^{(i)}_{i,N} \\ \vdots & \ddots & \vdots \\ a^{(i)}_{N,i} & \ldots & a^{(i)}_{N,N} \end{pmatrix} \neq 0.$$

Pivot partiel On déduit qu'il existe $i_0 \in \{i, \ldots, N\}$ tel que $a^{(i)}_{i_0,i} \neq 0$. On choisit alors $i_0 \in \{i, \ldots, N\}$ tel que $|a^{(i)}_{i_0,i}| = \max\{|a^{(i)}_{k,i}|, k = i, \ldots, N\}$. On échange alors les lignes i et i_0 (dans la matrice A et le second membre b) et on continue la procédure de Gauss décrite plus haut.

Pivot total On choisit maintenant i_0 et $j_0 \in \{i, \ldots, N\}$ tels que $|a^{(i)}_{i_0,j_0}| = \max\{|a^{(i)}_{k,j}|, k = i, \ldots, N, j = i, \ldots, N\}$, et on échange alors les lignes i et i_0 (dans la matrice A et le second membre b), les colonnes j et j_0 de A et les inconnues x_j et x_{j_0}.

L'intérêt de ces stratégies de pivot est qu'on aboutit toujours à la résolution du système (dès que A est inversible). 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 $A^{(N)}$.

1.3.3 Version programmable des algorithmes de Gauss et LU pour un système carré

On donne dans ce paragraphe la version "programmable" de l'algorithme de Gauss et de la maéthode LU pour une matrice carrée d'ordre n.
On souhaite résoudre $Ax = b$, avec $A \in M_n(\mathbb{R})$ inversible et un ou plusieurs second membres $b \in \mathbb{R}^n$.

Méthode de Gauss sans pivot

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

Méthode LU sans pivot

La méthode LU se déduit de la méthode de Gauss en remarquant simplement que, ayant conservé la matrice L, on peut effectuer les calculs sur b après les calculs sur A. Ce qui donne :

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

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

Méthode LU avec pivot partiel

La méthode LU avec pivot partiel consiste simplement à remarquer que l'ordre dans lequel les équations sont prises n'a aucune importance pour l'algorithme. Au départ de l'algorithme, on se donne la bijection t de $\{1,\dots,n\}$ dans $\{1,\dots,n\}$ définie par $t(i) = i$, et qui va être modfiée au cours de l'algorithme pour tenir compte du choix du pivot. On écrit l'algorithme précédent avec les nouveaux indices de ligne $t(i)$, ce qui donne :

1. (Factorisation) Pour i allant de 1 à n, on effectue les calculs suivants :
   (a) Choix du pivot (et de $t(i)$) : on cherche $k \in \{i,\dots,n\}$ t.q. $|a_{t(k),i}| = \max\{|a_{t(l),i}|, l \in \{i,\dots,n\}\}$. On change alors t en prenant $t(i) = t(k)$ et $t(k) = t(i)$.
   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 $t(j)$ autres que les lignes $t(1),\dots,t(n)$ (et le second membre), en utilisant la ligne $t(i)$. Donc pour $t(j) \in \{t(i+1),\dots,t(n)\}$ (noter qu'on a uniquement besoin de connaï¿½tre l'ensemble , et pas l'ordre) :
   $l_{t(j),i} = \frac{a_{t(j),i}}{a_{t(i),i}}$
   pour k de $i+1$ à n, $a_{t(j),k} = a_{t(j),k} - l_{t(j),i}a_{t(i),k}$ (noter que $a_{t(j),i} = 0$)
2. (Descente) On calcule y
   Pour i allant de 1 à n,
   $y_i = b_{t(i)} - \sum_{j=1}^{i-1} l_{t(j),k}y_k$
3. (Remontée) On calcule x
   Pour i allant de n à 1,
   $x_i = \frac{1}{u_{t(i),i}}(y_i - \sum_{j=i+1}^{n} u_{t(i),j}x_j)$

NB : On a change l'ordre dans lequel les équations sont considérées (le tableau t donne cet ordre). Donc, on a aussi changé l'ordre dans lequel interviennent les composantes du second membre. Par contre, on n'a pas touché à l'ordre dans lequel interviennent les composantes de x et y.

Il reste maintenant à signaler le "miracle" de cet algorithme... Il est inutile de connaitre complètement t pour faire 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). Pour minimiser des erreurs d'arrondis, on a intérêt à choisir $t(i)$ pour que $|a_{t(i),i}|$ soit le plus grand possible. Ceci suggère de faire le choix suivant de $t(i)$ à l'étape i de l'item $1(a)$ de l'algorithme (et c'est à cette étape que $t(i)$ doit être défini) :

on cherche $k \in \{i,\dots,n\}$ t.q. $|a_{t(k),i}| = \max\{|a_{t(l),i}|, l \in \{i,\dots,n\}\}$. On change alors t en prenant $t(i) = t(k)$ et $t(k) = t(i)$.

Remarque : 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 b, que l'on modifie au cours de l'algorithme). L'introduction de L, U, x et y peut toutefois aider à comprendre la méthode. Le principe retenu est que, dans les algorithmes (Gauss ou LU), on modifie la matrice A et le second membre b (en remplaçant le système à résoudre par un système équivalent) mais on ne modifie jamais L, U, y et x (qui sont définis au cours de l'algorithme).

Remarque L'algorithme se ramene donc à résoudre $LUx = b$, en faisant $L\boldsymbol{y} = b$ et $Ux = \boldsymbol{y}$. Quand on fait $L\boldsymbol{y} = b$ les equations sont dans l'ordre $t(1), \ldots, t(n)$ (les composantes de b sont donc aussi prises dans cet ordre), mais le vecteur $\boldsymbol{y}$ est bien $(y_1, \ldots, y_n)^t$ (t signifie ici "transposé" pour obtenir un vecteur colonne). Puis, on fait $Ux = \boldsymbol{y}$, les equations sont toujours dans l'ordre $t(1), \ldots, t(n)$ mais les vecteurs x et $\boldsymbol{y}$ sont $(x_1, \ldots, x_n)^t$ et $(y_1, \ldots, y_n)^t$.

1.3.4 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

La méthode de Choleski consiste à trouver une décomposition de A de la forme $A = LL^t$, où L est triangulaire inférieure de coefficients diagonaux strictement positifs. On résout alors le système $Ax = b$ en résolvant d'abord $Ly = b$ puis le système $L^t x = y$. 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 $Ly = b$ s'écrit :

$$Ly = \begin{bmatrix} \ell_{1,1} & 0 & \\ \vdots & \ddots & \vdots \\ \ell_{N,1} & \ldots & \ell_{N,N} \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_N \end{bmatrix}.$$

   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, $\ldots$, y_N.

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

$$L^t x = \begin{bmatrix} \ell_{1,1} & \ell_{2,1} & \ldots & \ell_{N,1} \\ 0 & \ddots & & \\ \vdots & & & \vdots \\ 0 & \ldots & & \ell_{N,N} \end{bmatrix} \begin{bmatrix} x_1 \\ \\ \vdots \\ \\ x_N \end{bmatrix} = \begin{bmatrix} y_1 \\ \\ \vdots \\ \\ y_N \end{bmatrix}.$$

On a donc :

$$\ell_{N,N}\, x_N = y_N \text{ donc } x_N = \frac{y_N}{\ell_{N,N}}$$

$$\ell_{N-1,N-1} x_{N-1} + \ell_{N,N-1} x_N = y_{N-1} \text{ donc } x_{N-1} = \tfrac{y_{N-1} - \ell_{N,N-1} x_N}{\ell_{N-1,N-1}}$$

$$\vdots$$

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

On calcule ainsi $x_N, x_{N-1}, \ldots, x_1$.

Existence et unicité de la décomposition

On donne ici le résultat d'unicité de la décomposition LL^t d'une matrice symétrique définie positive ainsi qu'un procédé constructif de la matrice L.

Théorème 1.14 (Décomposition de Choleski) *Soit $A \in \mathcal{M}_N(\mathbb{R})$ avec $N > 1$. On suppose que A est 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 : On sait déjà par le théorème 1.13 page 11, qi'il existe une matrice de permutation et L triangulaire inférieure et U triangulaire supérieure $PA = LU$. Ici on va montrer que dans le cas où la matrice est symétrique, la décomposition est toujours possible sans permutation. Nous donnons ici une démonstration directe de l'existence et de l'unicité de la décomposition LL^t qui a l'avantage d'être constructive.
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 :

$$A = \left[\begin{array}{c|c} B & a \\ \hline a^t & \alpha \end{array}\right] \tag{1.3.9}$$

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 :

$$L = \left[\begin{array}{c|c} M & 0 \\ \hline b^t & \lambda \end{array}\right] \tag{1.3.10}$$

avec $b \in \mathbb{R}^N$, $\lambda \in \mathbb{R}_+^*$ tels que $LL^t = A$. Pour déterminer b et λ, calculons LL^t où L est de la forme (1.3.10) et identifions avec A :

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

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

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

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

$$a^t B^{-1} a + \lambda^2 = \alpha \tag{1.3.11}$$

Pour que (1.3.11) soit vérifiée, il faut que

$$\alpha - a^t B^{-1} a > 0 \tag{1.3.12}$$

Montrons que la condition (1.3.12) 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 :

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

On a donc $Az \cdot z = \alpha - a^t B^{-1} a > 0$ ce qui montre que (1.3.12) est vérifiée. On peut ainsi choisir $\lambda = \sqrt{\alpha - a^t B^{-1} a}$ (> 0) de telle sorte que (1.3.11) 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".

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 :

$$a_{i,j} = \sum_{k=1}^{N} \ell_{i,k}\ell_{j,k}, \quad \forall (i,j) \in \{1 \dots N\}^2. \tag{1.3.13}$$

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

$$\begin{aligned} &a_{1,1} = \ell_{1,1}\ell_{1,1} \text{ donc } \ell_{1,1} = \sqrt{a_{1,1}} \ \ (a_{1,1} > 0 \text{ car } \ell_{1,1} \text{ existe }), \\ &a_{2,1} = \ell_{2,1}\ell_{1,1} \text{ donc } \ell_{2,1} = \frac{a_{2,1}}{\ell_{1,1}}, \\ &a_{i,1} = \ell_{i,1}\ell_{1,1} \text{ donc } \ell_{i,1} = \frac{a_{i,1}}{\ell_{1,1}} \ \forall i \in \{2, \dots, N\}. \end{aligned}$$

2. On suppose avoir calculé les p premières colonnes de L. On calcule la colonne $(q+1)$ en prenant $j = q+1$ dans (1.3.13)

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

$$\ell_{q+1,q+1} = (a_{q+1,q+1} - \sum_{k=1}^{q} \ell_{q+1,k}\ell_{q+1,k})^{1/2} > 0. \tag{1.3.14}$$

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

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

   et donc

$$\ell_{i,q+1} = \left(a_{i,q+1} - \sum_{k=1}^{q} \ell_{i,k}\ell_{q+1,k} \right) \frac{1}{\ell_{q+1,q+1}}. \tag{1.3.15}$$

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

■

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 $p = 0, \dots, N-1$, 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 (1.3.14) nécessite p multiplications, p soustractions et une extraction de racine, soit $2p+1$ opérations ; le calcul de $\ell_{i,p+1}$ par la formule (1.3.15) 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 (1.1.1) 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 (1.1.1) 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 1.15 (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" ($\times, +, -$).*

1.3.5 Quelques propriétés

Comparaison Gauss/Choleski

Soit $A \in \mathcal{M}_N(\mathbb{R})$ inversible, la résolution de (1.1.1) 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 (1.1.1) consiste à calculer les composantes de x par les formules :

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

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

Conservation du profil de A

Dans de nombreuses applications, par exemple lors de la résolution de systèmes linéaires issus de la discrétisation[1] de problèmes réels, la matrice $A \in \mathcal{M}_N(\mathbb{R})$ est "creuse", au sens où un grand nombre de ses coefficients sont nuls. Il est intéressant dans ce cas pour des raisons d'économie de mémoire de connaître le "profil" de la matrice, donné dans le cas où la matrice est symétrique, par les indices $j_i = \min\{j \in \{1, \ldots, N\}$ tels que $a_{i,j} \neq 0\}$. Le profil de la matrice est donc déterminé par les diagonales contenant des coefficients non nuls qui sont les plus éloignées de la diagonale principale. Dans le cas d'une matrice creuse, il est avantageux de faire un stockage "profil" de A, en stockant, pour chaque ligne i la valeur de j_i et des coefficients $a_{i,k}$, pour $k = i - j_i, \ldots, i$, ce qui peut permettre un large gain de place mémoire, comme on peut s'en rendre compte sur la figure 1.3.5.
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 1.14) 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 (1.1.1) 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

[1] On appelle discrétisation le fait de se ramer d'un problème où l'inconnue est une fonction en un problème ayant un nombre fini d'inconnues.

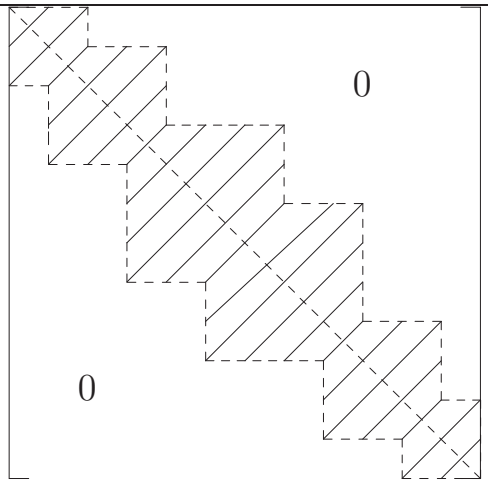

FIG. 1.3 – Exemple de profil d'une matrice symétrique

$(x, y) \in (\mathbb{R}^N)^2$. En prenant $B = A^t A$, on en déduit que $A^t A$ est symétrique. De plus, comme A est inversible, $A^t Ax \cdot x = Ax \cdot Ax = |Ax|^2 > 0$ si $x \neq 0$. La matrice $A^t A$ est donc bien symétrique définie positive.
La méthode de Choleski dans le cas d'une matrice non symétrique consiste donc à calculer $A^t A$ et $A^t b$, puis à résoudre le système linéaire$A^t A \cdot x = A^t b$ 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 13).

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

$$Ax = b. \tag{1.3.16}$$

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 (1.3.16) "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 (1.3.16). Comme on l'a dit, un tel x n'existe pas forcément, et on cherche alors un vecteur x qui vérifie (1.3.16) "au mieux", au sens où $f(x)$ soit le plus proche de 0. On cherche donc $x \in \mathbb{R}^N$ satisfaisant (1.3.16) en minimisant f, c.à.d. en cherchant $x \in \mathbb{R}^N$ solution du problème d'optimisation :

$$f(x) \leq f(y) \quad \forall y \in \mathbb{R}^N \tag{1.3.17}$$

On peut réécrire f sous la forme : $f(x) = A^t Ax \cdot x - 2b \cdot Ax + b \cdot b$. On montrera au chapitre III que s'il existe une solution au problème (1.3.17), elle est donnée par la résolution du système linéaire suivant :

$$AA^t x = A^t b \ \in \mathbb{R}^N,$$

qu'on appelle équations normales du problème de minimisation. La résolution approchée du problème (1.3.16) 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 (1.3.5).

1.4 Conditionnement

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.

1.4.1 Le problème des erreurs d'arrondis

Soient $A \in \mathcal{M}_N(\mathbb{R})$ inversible et $b \in \mathbb{R}^N$; supposons que les données A et 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 23), 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 catastrophique 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 x solution de (1.1.1) à partir des erreurs commises sur b et A. Notons $\delta_b \in \mathbb{R}^N$ l'erreur commise sur b et $\delta_A \in \mathcal{M}_N(\mathbb{R})$ l'erreur commise sur A. On cherche alors à évaluer δ_x où $x + \delta_x$ est solution (si elle existe) du système :

$$\begin{cases} x + \delta_x \in \mathbb{R}^N \\ (A + \delta_A)(x + \delta_x) = b + \delta_b. \end{cases} \tag{1.4.18}$$

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 δ_x en fonction de δ_A et δ_b.

1.4.2 Conditionnement et majoration de l'erreur d'arrondi

Définition 1.16 (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 1.17 (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 $A \in \mathcal{M}_N(\mathbb{R})$ 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 $\mathrm{cond}(AB) \leq \mathrm{cond}(A)\mathrm{cond}(B)$.*

Proposition 1.18 (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. 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* $\mathrm{cond}_2(A) = \sqrt{\sigma_N/\sigma_1}$.
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.*
3. *Si A et B sont deux matrices symétriques définies positives, alors* $\mathrm{cond}_2(A+B) \leq \max(\mathrm{cond}_2(A), \mathrm{cond}_2(B))$.
4. *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 $Q^t = Q^{-1}$).*
5. *Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice inversible. On suppose que $A = QR$ où Q est une matrice orthogonale. Alors* $\mathrm{cond}_2(A) = \mathrm{cond}_2(R)$.
6. *Soient A, $B \in \mathcal{M}_N(\mathbb{R})$ deux matrices symétriques définies positives. Montrer que* $\mathrm{cond}_2(A+B) \leq \max\{\mathrm{cond}_2(A), \mathrm{cond}_2(B)\}$.

La démonstration de deux propositions précédentes fait l'objet de l'exercice 17 page 40.

Théorème 1.19 *Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice inversible, et $b \in \mathbb{R}^N$, $b \neq 0$. 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 $\delta_b \in \mathbb{R}^N$. On suppose que $\|\delta_A\| < \dfrac{1}{\|A^{-1}\|}$. Alors la matrice $(A+\delta_A)$ est inversible et si x est solution de* (1.1.1) *et $x+\delta_x$ est solution de* (1.4.18)*, alors*

$$\frac{\|\delta_x\|}{\|x\|} \leq \frac{\mathrm{cond}(A)}{1 - \|A^{-1}\|\ \|\delta_A\|}\left(\frac{\|\delta_b\|}{\|b\|} + \frac{\|\delta_A\|}{\|A\|}\right). \tag{1.4.19}$$

Démonstration :
On peut écrire $A + \delta_A = A(Id + B)$ 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 1.8 page 7 et l'exercice 9 page 37) $(Id + B)$ est inversible et $(Id+B)^{-1} = \sum_{n=0}^{\infty}(-1)^n B^n$. On a aussi $\|(Id+B)^{-1}\| \leq \sum_{n=0}^{\infty}\|B\|^n = \dfrac{1}{1-\|B\|} \leq \dfrac{1}{1-\|A^{-1}\|\ \|\delta_A\|}$. On en déduit que $A+\delta_A$ est inversible, car $A+\delta_A = A(Id+B)$ et comme A est inversible, $(A+\delta_A)^{-1} = (Id+B)^{-1}A^{-1}$.

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

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

On peut donc écrire la majoration suivante :

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

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

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

ce qui termine la démonstration. ■

Remarquons que l'estimation (1.4.19) est optimale. En effet, en supposant que $\delta_A = 0$. L'estimation (1.4.19) devient alors :

$$\frac{\|\delta_x\|}{\|x\|} \leq cond(A)\frac{\|\delta_b\|}{\|b\|}. \tag{1.4.20}$$

Peut-on avoir égalité dans (1.4.20)? Pour avoir égalité dans (1.4.20) il faut choisir convenablement b et δ_b. Soit $x \in \mathbb{R}^N$ tel que $\|x\| = 1$ et $\|Ax\| = \|A\|$. Notons qu'un tel x existe parce que $\|A\| = \sup\{\|Ax\|; \|x\| = 1\} = \max\{\|Ax\|; \|x\| = 1\}$ (voir proposition 1.2 page 5) Posons $b = Ax$; on a donc $\|b\| = \|A\|$. De même, grâce à la proposition 1.2, il existe $y \in \mathbb{R}^N$ tel que $\|y\| = 1$, et $\|A^{-1}y\| = \|A^{-1}\|$. On choisit alors δ_b tel que $\delta_b = \varepsilon y$ où $\varepsilon > 0$ est donné. Comme $A(x + \delta_x) = b + \delta_b$, on a $\delta_x = A^{-1}\delta_b$ et donc : $\|\delta_x\| = \|A^{-1}\delta_b\| = \varepsilon\|A^{-1}y\| = \varepsilon\|A^{-1}\| = \|\delta_b\| \ \|A^{-1}\|$. On en déduit que

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

Par ce choix de b et δ_b on a bien égalité dans (1.4.20). L'estimation (1.4.20) est donc optimale.

1.4.3 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 d'une équation différentielle introduite ci-dessous (voir (1.4.21)). On peut alors montrer (voir exercice 25 page 44 du chapitre 1) 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 (1.4.20) donne :

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

Une erreur de 1% sur 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 (1.4.20) n'est pas significative pour l'étude de la propagation des erreurs lors de larésolution des systèmes linéraires provenant de la discrétisation d'une équation différentielle ou d'une équation aux dérivées partielles[2]. Pour illustrer notre propos, nous allons étudier un système linéaire très simple provenant d'un problème de mécanique et sa discrétisation par différences finies.

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

$$\left\{ \begin{array}{l} -u''(x) = f(x) \\ u(0) = u(1) = 0. \end{array} \right. \tag{1.4.21}$$

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 1.4.3), qui sont les points où l'on va écrire l'équation $-u'' = f$ en vue de se ramer à un système discret, c.à.d. à un système avec un nombre fini d'inconnues. Remarquons que $x_0 = 0$ et $x_{N+1} = 1$. Soit $u(x_i)$ la valeur exacte de u en x_i. On écrit la première équation de (1.4.21) en chaque point x_i, pour $i = 1 \ldots N$.

$$-u''(x_i) = f(x_i) = b_i \ \forall i \in \{1 \ldots N\}.$$

[2] 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 $\mathbb{R}$

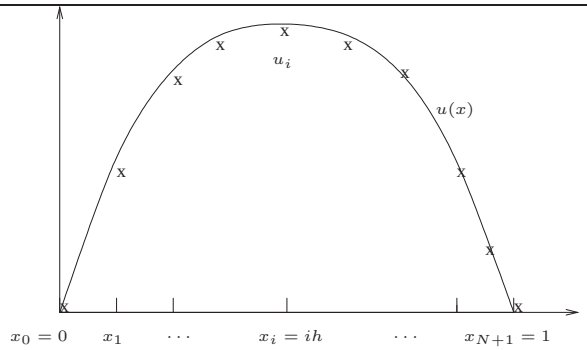

FIG. 1.4 – Solution exacte et approchée de $-u'' = f$

On peut facilement montrer, par un développement de Taylor, que si $u \in C^4([0,1], \mathbb{R})$ (ce qui est vrai si $f \in C^2$) alors[3]

$$-\frac{u(x_{i+1}) + u(x_{i-1}) - 2u(x_i)}{h^2} = -u''(x_i) + R_i \text{ avec } |R_i| \leq \frac{h^2}{12}\|u^{(4)}\|_\infty. \tag{1.4.22}$$

La valeur R_i s'appelle erreur de consistance au point x_i.
On introduit alors les inconnues $(u_i)_{i=1,\ldots,N}$ qu'on espère être des valeurs approchées de u aux points x_i et qui sont les composantes de la solution (si elle existe) du système suivant

$$\begin{cases} -\dfrac{u_{i+1} + u_{i-1} - 2u_i}{h^2} = b_i, \quad \forall i \in \{1 \leq N\}, \\ u_0 = u_{N+1} = 0. \end{cases} \tag{1.4.23}$$

On cherche donc $u = \begin{bmatrix} u_1 \\ \vdots \\ u_N \end{bmatrix} \in \mathbb{R}^N$ solution de (1.4.23). Ce système peut s'écrire sous forme matricielle :$Au = b$ avec$b = (b_1, \ldots, b_N)^t$ et A la matrice carrée d'ordre N de coefficients $(a_{i,j})_{i,j=1,N}$ définis par :

$$\begin{cases} a_{i,i} &= \dfrac{2}{h^2}, \forall\, i = 1, \ldots, N, \\ a_{i,j} &= -\dfrac{1}{h^2}, \forall\, i = 1, \ldots, N, \;\; j = i \pm 1, \\ a_{i,j} &= 0, \forall\, i = 1, \ldots, N, \; |i - j| > 1. \end{cases} \tag{1.4.24}$$

On remarque immédiatement que A est tridiagonale. On peut montrer que A est symétrique définie positive (voir exercice 25 page 44). On peut aussi montrer 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 (1.4.22)) $A(u - \overline{u}) = R$, et donc $\|u - \overline{u}\|_\infty \leq$

[3] En effet, par développement de Taylor, on a :
$u(x_{i+1}) = u(x_i) + hu'(x_i) + \frac{h^2}{2}u''(x_i) + \frac{h^3}{6}u'''(x_i) + \frac{h^4}{24}u^{(4)}(\xi_i)$ où $\xi_i \in (x_i, x_{i+1})$ et
$u(x_{i-1}) = u(x_i) - hu'(x_i) + \frac{h^2}{2}u''(x_i) - \frac{h^3}{6}u'''(x_i) + \frac{h^4}{24}u^{(4)}(\eta_i)$ où $\eta_i \in (x_i, x_{i+1})$.
En faisant la somme de ces deux égalités, on en déduit que : $u(x_{i+1}) + u(x_{i-1}) = 2u(x_i) + \frac{h^2}{u}''(x_i) + \frac{h^4}{24}u^{(4)}(\xi_i) + \frac{h^4}{24}u^{(4)}(\eta_i)$, et donc R_i défini par (1.4.22) vérifie :
$|R_i| = |-\frac{u(x_{i+1})+u(x_{i-1})-2u(x_i)}{h^2} + u''(x_i)| \leq |\frac{h^4}{24}u^{(4)}(\xi_i) + \frac{h^4}{24}u^{(4)}(\eta_i)| \leq \frac{h^2}{12}\|u^{(4)}\|_\infty$.

$\|A^{-1}\|_\infty \|R\|_\infty$. Or on peut montrer (voir exercice 27 page 45, partie I) que $\|A^{-1}\|_\infty \leq 1/8$, et on obtient donc avec (1.4.22) que $\|u - \overline{u}\|_\infty \leq h^2/96 \|u^{(4)}\|_\infty$.

Cette inégalité donne la précision de la méthode. 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. Or on a déjà dit qu'on peut montrer (voir exercice 25 page 44) 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 (1.4.20) 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 (1.4.20) 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 27 page 45. 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

$$\frac{\|\delta_u\|}{\|u\|} \leq C \frac{\|\delta_b\|}{\|b\|} \text{ avec } b = \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_N) \end{bmatrix}. \tag{1.4.25}$$

L'estimation (1.4.25) est évidemment bien meilleure que l'estimation (1.4.20) puisqu'elle montre que l'erreur relative commise sur u est du même ordre que celle commise sur b. En particulier, elle n'augmente pas avec le nombre de points de discrétisation. En conclusion, l'estimation (1.4.20) est peut-être optimale dans le cas d'une matrice quelconque, (on a montré ci-dessus qu'il peut y avoir égalité dans (1.4.20)) mais elle n'est pas significative pour l'étude des systèmes linéaires issus de la discrétisation des équations aux dérivées partielles.

1.5 Méthodes itératives

1.5.1 Origine des systèmes à résoudre

Les méthodes directes que nous avons étudiées dans le paragraphe précédent 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 19). Prenons par exemple le cas d'une discrétisation du Laplacien sur un carré par différences finies. On cherche à résoudre le problème :

$$\begin{array}{l} -\Delta u = f \text{ sur } \Omega =]0,1[\times]0,1[, \\ u = 0 \text{ sur } \partial\Omega, \end{array} \tag{1.5.26}$$

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 1.5.1 pour $M = 6$. On peut alors approcher les

dérivées secondes par des quotients différentiels comme dans le cas unidimensionnel (voir page 23), pour obtenir un système linéaire : $AU = b$ où $A \in \mathcal{M}_N(\mathbb{R})$ et $b \in \mathbb{R}^N$ 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 $b_k = f(x_i, y_j)$.

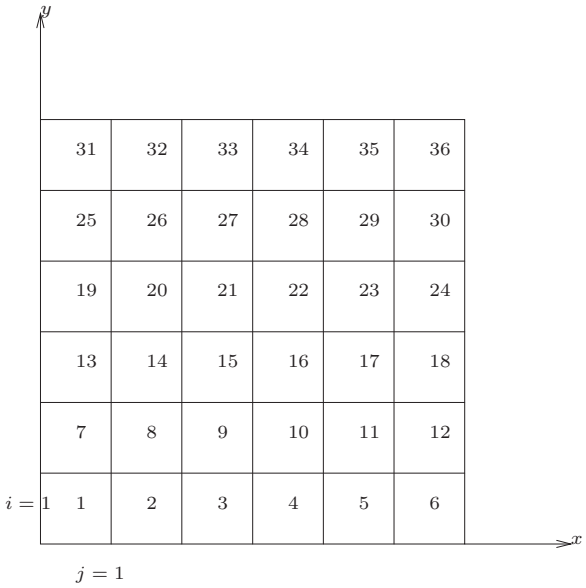

FIG. 1.5 – Ordre lexicographique des inconnues, exemple dans le cas $M = 6$

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

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

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

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

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

$$A = \begin{pmatrix} D & -Id & 0 & \dots & \dots & 0 \\ -Id & D & -Id & 0 & \dots & 0 \\ 0 & -Id & D & -Id & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & & \ddots & -Id & D & -Id \\ 0 & \dots & & 0 & -Id & D \end{pmatrix}, \tag{1.5.27}$$

où Id désigne la matrice identité d'ordre M.
On peut remarquer que la matrice A a une "largeur de bande" de M. Si on utilise une méthode directe genre Choleski, on aura donc besoin d'une place mémoire de $N \times M = M^3$. (Notons que pour une matrice pleine on a besoin de M^4.)
Lorsqu'on a affaire à de très gros systèmes issus par exemple de l'ingénierie (calcul des structures, mécanique des fluides, ...), où N peut être de l'ordre de plusieurs milliers, on cherche à utiliser des méthodes nécessitant le moins de mémoire possible. On a intérêt dans ce cas à utiliser des méthodes itératives. Ces méthodes ne font appel qu'à des produits matrice vecteur, et ne nécessitent donc pas le stockage du profil de la matrice mais uniquement des termes non nuls. Dans l'exemple précédent, on a 5 diagonales non nulles, donc la place mémoire nécessaire pour un produit matrice vecteur est $5N$ = $5M^2$. Ainsi pour les gros systèmes, il est souvent avantageux d'utiliser des méthodes itératives qui ne donnent pas toujours la solution exacte du système en un nombre fini d'itérations, mais qui donnent une solution approchée à coût moindre qu'une méthode directe, car elles ne font appel qu'à des produits matrice vecteur.

Remarque 1.20 (Sur la méthode du gradient conjugué)
Il existe une méthode itérative "miraculeuse" de résolution des systèmes linéaires lorsque la matrice A est symétrique définie positive : c'est la méthode du gradient conjugué. Elle est miraculeuse en ce sens qu'elle donne la solution exacte du système $Ax = b$ en un nombre fini d'opérations (en ce sens c'est une méthode directe) : moins de N itérations où N est l'ordre de la matrice A, bien qu'elle ne nécessite que des produits matrice vecteur ou des produits scalaires. La méthode du gradient conjugué est en fait une méthode d'optimisation pour la recherche du minimum dans $\mathbb{R}^N$ de la fonction de $\mathbb{R}^N$ dans $\mathbb{R}$ définie par : $f(x) = \frac{1}{2}Ax \cdot x - b \cdot x$. Or on peut montrer que lorsque A est symétrique définie positive, la recherche de x minimisant f dans $\mathbb{R}^N$ est équivalent à la résolution du système $Ax = b$. (Voir paragraphe 3.2.2 page 124.) En fait, la méthode du gradient conjugué n'est pas si miraculeuse que cela en pratique : en effet, le nombre N est en général très grand et on ne peut en 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. On a alors recours aux techniques de "préconditionnement". Nous reviendrons sur ce point au chapitre 3.
La méthode itérative du gradient à pas fixe, qui est elle aussi obtenue comme méthode de minimisation de la fonction f ci-dessus, fait l'objet des exercices 28 page 46 et 63 page 149.

1.5.2 Définition et propriétés

Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice inversible et $b \in \mathbb{R}^N$, on cherche toujours ici à résoudre le système linéaire (1.1.1) c'est–à–dire à trouver $x \in \mathbb{R}^N$ tel que $Ax = b$.

Définition 1.21 *On appelle méthode itérative de résolution du système linéaire* (1.1.1) *une méthode qui construit une suite $(x^{(k)})_{k\in\mathbb{N}}$ (où "l'itéré" $x^{(k)}$ est calculé à partir des itérés $x^{(0)} \dots x^{(k-1)}$) censée converger vers x solution de* (1.1.1)*.*

Définition 1.22 *On dit qu'une méthode itérative est convergente si pour tout choix initial $x^{(0)} \in \mathbb{R}^N$, on a :*

$$x^{(k)} \longrightarrow x \text{ quand } n \to +\infty$$

Puisqu'il s'agit de résoudre un sytème linéaire, il est naturel d'essayer de construire la suite des itérés sous la forme $x^{(k+1)} = Bx^{(k)} + c$, où $B \in \mathcal{M}_N(\mathbb{R})$ et $c \in \mathbb{R}^N$ seront choisis de manière à ce que la méthode itérative ainsi définie soit convergente. On appellera ce type de méthode *Méthode I*, et on verra par la suite un choix plus restrictif qu'on appellera *Méthode II*.

Définition 1.23 (Méthode I) *On appelle méthode itérative de type I pour la résolution du système linéaire* (1.1.1) *une méthode itérative où la suite des itérés $(x^{(k)})_{k\in\mathbb{N}}$ est donnée par :*

$$\begin{cases} \textit{Initialisation} & x^{(0)} \in \mathbb{R}^N \\ \textit{Itération } n & x^{(k+1)} = Bx^{(k)} + c. \end{cases}$$

où $B \in \mathcal{M}_N(\mathbb{R})$ et $c \in \mathbb{R}^N$.

Remarque 1.24 (Condition nécessaire de convergence) *Une condition nécessaire pour que la méthode I converge est que $c = (Id - B)A^{-1}b$. En effet, supposons que la méthode converge. En passant à la limite lorsque n tend vers l'infini sur l'itération n de l'algorithme, on obtient $x = Bx + c$ et comme $x = A^{-1}b$, ceci entraîne $c = (Id - B)A^{-1}b$.*

Remarque 1.25 (Intérêt pratique) *La "méthode I" est assez peu intéressante en pratique, car il faut calculer $A^{-1}b$, sauf si $(Id - B)A^{-1} = \alpha Id$, avec $\alpha \in \mathbb{R}$. On obtient dans ce cas :*

$$\begin{aligned} & B = -\alpha A + Id \\ \textit{et} \quad & c = \alpha b \end{aligned}$$

c'est–à–dire

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

Le terme $b - Ax^n$ est appelé résidu et la méthode s'appelle dans ce cas la méthode d'extrapolation de Richardson.

Théorème 1.26 (Convergence de la méthode de type I) *Soit $A \in \mathcal{M}_N(\mathbb{R})$ A inversible, $b \in \mathbb{R}^N$. On considère la méthode I avec $B \in \mathcal{M}_N(\mathbb{R})$ et*

$$c = (Id - B)A^{-1}b. \tag{1.5.28}$$

Alors la méthode I converge si et seulement si le rayon spectral $\rho(B)$ de la matrice B vérifie $\rho(B) < 1$.

Démonstration
Soit $B \in \mathcal{M}_N(\mathbb{R})$.
Soit x la solution du système linéaire (1.1.1) ; grâce à (1.5.28), $x = Bx + c$, et comme $x^{(k+1)} = Bx^{k)} + c$, on a donc $x^{(k+1)} - x = B(x^{k)} - x)$ et par récurrence sur k,

$$x^{(k)} - x = B^k(x^{(0)} - x), \quad \forall k \in \mathbb{N}. \tag{1.5.29}$$

On rappelle (voir exercice 5 page 37) que $\rho(B) < 1$ si et seulement si $B^n \to 0$ et que donc, si $\rho(B) \geq 1$ alors $B^n \not\to 0$. On rappelle aussi que $B^n \to 0$ si et seulement si $B^n y \to 0, \quad \forall y \in \mathbb{R}^n$.

(⇒) On démontre l'implication par contraposée. Supposons que $\rho(B) \geq 1$ et montrons que la méthode I ne converge pas. Si $\rho(B) \geq 1$ il existe $y \in \mathbb{R}^N$ tel que $B^k y \underset{k\to+\infty}{\not\longrightarrow} 0$.

En choisissant $x^{(0)} = x + y = A^{-1}b + y$, l'égalité (1.5.29) devient : $x^{(k)} - x = B^k y \underset{k\to+\infty}{\not\longrightarrow} 0$ par hypothèse et donc la méthode n'est pas convergente.

(⇐) Supposons maintenant que $\rho(B) < 1$ alors l'égalité (1.5.29) donne

$$x^{(k)} - x = B^k(x^{(0)} - x) \underset{k\to+\infty}{\to} 0$$

car $\rho(B) < 1$. Donc $x^{(k)} \underset{k\to+\infty}{\longrightarrow} x = A^{-1}b$. La méthode est bien convergente.

■

Définition 1.27 (Méthode II) *Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice inversible, $b \in \mathbb{R}^N$. Soient $\tilde{M}$ et $\tilde{N} \in \mathcal{M}_N(\mathbb{R})$ des matrices telles que $A = \tilde{M} - \tilde{N}$ et $\tilde{M}$ est inversible (et facile à inverser).*
On appelle méthode de type II pour la résolution du système linéaire (1.1.1) *une méthode itérative où la suite des itérés $(x^{(k)})_{k\in\mathbb{N}}$ est donnée par :*

$$\begin{cases} \textit{Initialisation} & x^{(0)} \in \mathbb{R}^N \\ \textit{Itération } n & \tilde{M}x^{(k+1)} = \tilde{N}x^{(k)} + b. \end{cases} \tag{1.5.30}$$

Remarque 1.28 *Si $\tilde{M}x^{(k+1)} = \tilde{N}x^{(k)} + b$ pour tout $k \in \mathbb{N}$ et $x^{(k)} \longrightarrow y$ quand $n \longrightarrow +\infty$ alors $\tilde{M}y = \tilde{N}y + b$, c.à.d. $(\tilde{M} - \tilde{N})y = b$ et donc $Ay = b$. En conclusion, si la méthode de type II converge, alors elle converge bien vers la solution du système linéaire.*

Théorème 1.29 (Convergence de la méthode II) *Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice inversible, $b \in \mathbb{R}^N$. Soient $\tilde{M}$ et $\tilde{N} \in \mathcal{M}_N(\mathbb{R})$ des matrices telles que $A = \tilde{M} - \tilde{N}$ et $\tilde{M}$ est inversible. Alors :*

1. *La méthode définie par* (1.5.30) *converge si et seulement si $\rho(\tilde{M}^{-1}\tilde{N}) < 1$.*
2. *La méthode itérative définie par* (1.5.30) *converge si et seulement si il existe une norme induite notée $\|\cdot\|$ telle que $\|\tilde{M}^{-1}\tilde{N}\| < 1$.*

Démonstration

1. Pour démontrer le point 1, il suffit de réécrire la méthode II avec le formalisme de la méthode I. En effet,
$$\tilde{M}x^{(k+1)} = \tilde{N}x^{(k)} + b \Longleftrightarrow x^{(k+1)} = \tilde{M}^{-1}\tilde{N}x^{(k)} + \tilde{M}^{-1}b = Bx^{(k)} + c,$$
avec $B = \tilde{M}^{-1}\tilde{N}$ et $c = \tilde{M}^{-1}b$.
2. Si il existe une norme induite notée $\|\cdot\|$ telle que $\|\tilde{M}^{-1}\tilde{N}\| < 1$, alors en vertu de la proposition 1.5, $\rho(\tilde{M}^{-1}\tilde{N}) < 1$ et donc la méthode converge ce qui précède.
Réciproquement, si la méthode converge alors $\rho(\tilde{M}^{-1}\tilde{N}) < 1$, et donc il existe $\eta > 0$ tel que $\rho(\tilde{M}^{-1}\tilde{N}) = 1 - \eta$. varepsilon Prenons maintenant $\varepsilon = \frac{\eta}{2}$ et appliquons la proposition 1.6 : il existe une norme induite $\|\cdot\|$ telle que $\|\tilde{M}^{-1}\tilde{N}\| \leq \rho(\tilde{M}^{-1}\tilde{N}) + \varepsilon < 1$, ce qui démontre le résultat.

■

Pour trouver des méthodes itératives de résolution du système (1.1.1), on cherche donc une décomposition de la matrice A de la forme : $A = \tilde{M} - \tilde{N}$, où $\tilde{M}$ est inversible, telle que le système $\tilde{M}y = d$ soit un système facile à résoudre (par exemple $\tilde{M}$ soit triangulaire).

Théorème 1.30 (Condition suffisante de convergence, méthode II) *Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice symétrique définie positive, et soient $\tilde{M}$ et $\tilde{N} \in \mathcal{M}_N(\mathbb{R})$ telles que $A = \tilde{M} - \tilde{N}$ et $\tilde{M}$ est inversible. Si la matrice $\tilde{M}^t + \tilde{N}$ est symétrique définie positive alors $\rho(\tilde{M}^{-1}\tilde{N}) < 1$, et donc la méthode II converge.*

Démonstration On rappelle (voir exercice (5) page 37) 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

$$\|\tilde{M}^{-1}\tilde{N}\|_* = \max\{\|\tilde{M}^{-1}\tilde{N}x\|_*,\ x \in \mathbb{R}^N,\ \|x\|_* = 1\} < 1,$$

(où on désigne encore par $\|.\|_*$ la norme induite sur $\mathcal{M}_N(\mathbb{R})$) ou encore :

$$\|\tilde{M}^{-1}\tilde{N}x\|_* < \|x\|_*,\ \forall x \in \mathbb{R}^N,\ x \neq 0. \tag{1.5.31}$$

On définit la norme $\|\cdot\|_*$ par $\|x\|_* = \sqrt{Ax \cdot x}$, pour tout $x \in \mathbb{R}^N$. Comme A est symétrique définie positive, $\|\cdot\|_*$ est bien une norme sur $\mathbb{R}^N$, induite par le produit scalaire $(x|y)_A = Ax \cdot y$. On va montrer que la propriété (1.5.31) est vérifiée par cette norme. Soit $x \in \mathbb{R}^N$, $x \neq 0$, on a : $\|\tilde{M}^{-1}\tilde{N}x\|_*^2 = A\tilde{M}^{-1}\tilde{N}x \cdot \tilde{M}^{-1}\tilde{N}x$. Or $\tilde{N} = \tilde{M} - A$, et donc : $\|\tilde{M}^{-1}\tilde{N}x\|_*^2 = A(Id - \tilde{M}^{-1}A)x \cdot (Id - \tilde{M}^{-1}A)x$. Soit $y = \tilde{M}^{-1}Ax$; remarquons que $y \neq 0$ car $x \neq 0$ et $\tilde{M}^{-1}A$ est inversible. Exprimons $\|\tilde{M}^{-1}\tilde{N}x\|_*^2$ à l'aide de y.

$$\|\tilde{M}^{-1}\tilde{N}x\|_*^2 = A(x-y) \cdot (x-y) = Ax \cdot x - 2Ax \cdot y + Ay \cdot y = \|x\|_*^2 - 2Ax \cdot y + Ay \cdot y.$$

Pour que $\|\tilde{M}^{-1}\tilde{N}x\|_*^2 < \|x\|_*^2$ (et par suite $\rho(\tilde{M}^{-1}\tilde{N}) < 1$), il suffit donc de montrer que $-2Ax \cdot y + Ay \cdot y < 0$. Or, comme $\tilde{M}y = Ax$, on a : $-2Ax \cdot y + Ay \cdot y = -2\tilde{M}y \cdot y + Ay \cdot y$. En écrivant : $\tilde{M}y \cdot y = y \cdot \tilde{M}^t y = \tilde{M}^t y \cdot y$, on obtient donc que : $-2Ax \cdot y + Ay \cdot y = (-\tilde{M} - \tilde{M}^t + A)y \cdot y$, et comme $A = \tilde{M} - \tilde{N}$ on obtient $-2Ax \cdot y + Ay \cdot y = -(\tilde{M}^t + \tilde{N})y \cdot y$. Comme $\tilde{M}^t + \tilde{N}$ est symétrique définie positive par hypothèse et que $y \neq 0$, on en déduit que $-2Ax \cdot y + Ay \cdot y < 0$, ce qui termine la démonstration. ■

Estimation de la vitesse de convergence On montre dans l'exercice 41 page 51 que si la suite $(x^{(k)})_{k\in\mathbb{N}} \subset \mathbb{R}$ est donnée par une "méthode I" (voir définition 1.23 page 28) convergente, *i.e.* $x^{(k+1)} = Bx^{(k)} + C$ (avec $\rho(B) < 1$), et si on suppose que x est la solution du système (1.1.1), et que $x^{(k)} \to x$ quand $k \to \infty$, alors $\dfrac{\|x^{(k+1)} - x\|}{\|x^{(k)} - x\|} \longrightarrow \rho(B)$ quand $k \to +\infty$ (sauf cas particuliers) indépendamment de la norme sur $\mathbb{R}^N$. Le rayon spectral $\rho(B)$ de la matrice B est donc une bonne estimation de la vitesse de convergence. Pour estimer cette vitesse de convergence lorsqu'on ne connaît pas x, on peut utiliser le fait (voir encore l'exercice 41 page 51) qu'on a aussi

$$\frac{\|x^{(k+1)} - x^{(k)}\|}{\|x^{(k)} - x^{(k-1)}\|} \longrightarrow \rho(B) \quad \text{ lorsque } n \to +\infty,$$

ce qui permet d'évaluer la vitesse de convergence de la méthode par le calcul des itérés courants.

1.5.3 Méthodes de Jacobi, Gauss-Seidel et SOR/SSOR

Décomposition par blocs de A :

Dans de nombreux cas pratiques, les matrices des systèmes linéaires à résoudre ont une structure "par blocs", et on se sert de cette structure lors de la résolution par une méthode itérative.

Définition 1.31 *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*

$$A = \begin{bmatrix} A_{1,1} & A_{1,2} & \dots & & \dots & A_{1,S} \\ A_{2,1} & \ddots & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & A_{S-1,S} \\ A_{S,1} & \dots & & \dots & A_{S,S-1} & A_{S,S} \end{bmatrix} \tag{1.5.32}$$

Remarque 1.32

1. *Si $S = N$ et $n_i = 1\ \forall i \in \{1, \dots, S\}$, chaque bloc est constitué d'un seul coefficient.*
2. *Si A est symétrique définie positive, la condition $A_{i,i}$ inversible dans la définition 1.31 est inutile car $A_{i,i}$ est nécessairement symétrique définie positive donc inversible. Prenons par exemple $i = 1$; soit $y \in \mathbb{R}^{n_1}$, $y \neq 0$ et $x = (y, 0 \dots, 0)^t \in \mathbb{R}^N$. Alors $A_{1,1}y \cdot y = Ax \cdot x > 0$ donc $A_{1,1}$ est symétrique définie positive.*
3. *Si A est une matrice triangulaire par blocs, c.à.d. de la forme* (1.5.32) *avec $A_{i,j} = 0$ si $j > i$, alors*

$$\det(A) = \prod_{i=1}^{S} \det(A_{i,i}).$$

Par contre si A est décomposée en 2×2 blocs carrés (i.e. tels que $n_i = m_j$, $\forall (i,j) \in \{1,2\}$), on a en général : $\det(A) \neq \det(A_{1,1})\det(A_{2,2}) - \det(A_{1,2})\det(A_{2,1})$.

Méthode de Jacobi

On peut remarquer que le choix le plus simple pour le système $\tilde{M}x = d$ soit facile à résoudre (on rappelle que c'est un objectif de la mise sous forme méthode de type II) est de prendre pour $\tilde{M}$ une matrice diagonale. La méthode de Jacobi[4] consiste à prendre pour $\tilde{M}$ 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.

[4]Carl Gustav Jakob Jacobi, (Potsdam, 1804 - Berlin, 1851), mathématicien allemand. Issu d'une famille juive, il étudie à l'Université de Berlin, où il obtient son doctorat en 1825, à peine âgé de 21 ans. Sa thèse est une discussion analytique de la théorie des fractions. En 1829, il devient professeur de mathématique à l'Université de Kï¿½nigsberg, 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. Dans une lettre du 2 juillet 1830 adressée à Legendre, Jacobi écrit : ï¿½ 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. ï¿½ L'expression est restée, et renvoie à un débat toujours d'actualité.

On a alors $\tilde{N} = E + F$, où E et F sont constitués des blocs triangulaires inférieurs et supérieurs de la matrice A :

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

et

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

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

$$\begin{cases} x^{(0)} \in \mathbb{R}^N \\ Dx^{(k+1)} = (E+F)x^{(k)} + b. \end{cases} \tag{1.5.33}$$

Lorsqu'on écrit la méthode de Jacobi comme une méthode I, on a $B = D^{-1}(E+F)$; on notera J cette matrice. En introduisant la décomposition par blocs de x, solution recherchée de (1.1.1), c.à.d. : $x = [x_1, \dots, x_S]^t$, où $x_i \in \mathbb{R}^{n_i}$, on peut aussi écrire la méthode de Jacobi sous la forme :

$$\begin{cases} x_0 \in \mathbb{R}^N \\ A_{i,i}x_i^{(k+1)} = -\sum_{j<i} A_{i,j}x_j^{(k)} - \sum_{j>i} A_{i,j}x_j^{(k)} + b_i \;\; i = 1, \dots, S. \end{cases} \tag{1.5.34}$$

Si $S = N$ et $n_i = 1 \; \forall i \in \{1, \dots, S\}$, chaque bloc est constitué d'un seul coefficient, et on obtient la méthode de Jacobi par points (aussi appelée méthode de Jacobi), qui s'écrit donc :

$$\begin{cases} x_0 \in \mathbb{R}^N \\ a_{i,i}x_i^{(k+1)} = -\sum_{j<i} a_{i,j}x_j^{(k)} - \sum_{j>i} a_{i,j}x_j^{(k)} + b_i \;\; i = 1, \dots, N. \end{cases} \tag{1.5.35}$$

Méthode de Gauss-Seidel

L'idée de la méthode de Gauss-Seidel[5] est d'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 (1.5.34) ; 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 (1.5.34) $x_j^{(k)}$ par $x_j^{(k+1)}$ si $j < i$. On obtient donc l'algorithme suivant :

[5] Philipp Ludwig von Seidel (Zweibrï¿½cken, 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.

$$\left\{ \begin{array}{l} x^{(0)} \in \mathbb{R}^N \\ A_{i,i}x_i^{(k+1)} = -\sum_{j<i} A_{i,j}x_j^{(k+1)} - \sum_{i<j} A_{i,j}x_j^{(k)} + b_i, \quad i = 1, \ldots, S. \end{array} \right. \tag{1.5.36}$$

Notons que l'algorithme de Gauss–Seidel par points (cas ou $S = N$ et $n_i = 1$) s'écrit donc :

$$\left\{ \begin{array}{l} x^{(0)} \in \mathbb{R}^N \\ a_{i,i}x_i^{(k+1)} = -\sum_{j<i} a_{i,j}x_j^{(k+1)} - \sum_{i<j} a_{i,j}x_j^{(k)} + b_i, \quad i = 1, \ldots, N. \end{array} \right. \tag{1.5.37}$$

La méthode de Gauss–Seidel s'écrit donc sous forme de méthode II avec $\tilde{M} = D - E$ et $\tilde{N} = F$:

$$\left\{ \begin{array}{l} x_0 \in \mathbb{R}^N \\ (D - E)x^{(k+1)} = Fx^{(k)} + b. \end{array} \right. \tag{1.5.38}$$

Lorsqu'on écrit la méthode de Gauss–Seidel comme une méthode I, on a $B = (D - E)^{-1}F$; on notera $\mathcal{L}_1$ cette matrice, dite matrice de Gauss-Seidel.

Méthodes SOR et SSOR

L'idée de la méthode de sur-relaxation (SOR = Successive Over Relaxation) est d'utiliser la méthode de Gauss-Seidel pour calculer un itéré intermédiaire $\tilde{x}^{(k+1)}$ qu'on "relaxe" ensuite pour améliorer la vitesse de convergence de la méthode. On se donne $0 < \omega < 2$, et on modifie l'algorithme de Gauss–Seidel de la manière suivante :

$$\left\{ \begin{array}{l} x_0 \in \mathbb{R}^N \\ A_{i,i}\tilde{x}_i^{(k+1)} = -\displaystyle\sum_{j<i} A_{i,j}x_j^{(k+1)} - \sum_{i<j} A_{i,j}x_j^{(k)} + b_i \\ x_i^{(k+1)} = \omega\tilde{x}_i^{(k+1)} + (1 - \omega)x_i^{(k)}, \quad i = 1, \ldots, S. \end{array} \right. \tag{1.5.39}$$

(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 (1.5.39)) :

$$\left\{ \begin{array}{l} x^{(0)} \in \mathbb{R}^N \\ A_{i,i}x_i^{(k+1)} = \omega\left[-\displaystyle\sum_{j<i} A_{i,j}x_j^{(k+1)} - \sum_{j>i} A_{i,j}x_j^{(k)} + b_i\right] \\ \qquad\qquad +(1 - \omega)A_{i,i}x_i^{(k)}. \end{array} \right. \tag{1.5.40}$$

On obtient donc

$$(D - \omega E)x^{(k+1)} = \omega Fx^{(k)} + \omega b + (1 - \omega)Dx^{(k)}.$$

L'algorithme SOR s'écrit donc comme une méthode II avec

$$\tilde{M} = \frac{D}{\omega} - E \text{ et } \tilde{N} = F + \left(\frac{1 - \omega}{\omega}\right) D.$$

Il est facile de vérifier que $A = \tilde{M} - \tilde{N}$.

L'algorithme SOR s'écrit aussi comme une méthode I avec

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

On notera $\mathcal{L}_\omega$ cette matrice.

Remarque 1.33 (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 35 page 48). 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

$$B = \underbrace{\left(\frac{D}{\omega} - F\right)^{-1}\left(E + \frac{1-\omega}{\omega}D\right)}_{\text{calcul dans l'ordre } \; S\ldots 1}\underbrace{\left(\frac{D}{\omega} - E\right)^{-1}\left(F + \frac{1-\omega}{\omega}D\right)}_{\text{calcul dans l'ordre } \; 1\ldots S}.$$

Etude théorique de convergence

On aimerait pouvoir répondre aux questions suivantes :

1. Les méthodes sont–elles convergentes ?
2. Peut-on estimer leur vitesse de convergence ?
3. Peut-on estimer le coefficient de relaxation ω optimal dans la méthode SOR, c.à.d. celui qui donnera la plus grande vitesse de convergence ?

On va maintenant donner des réponses, partielles dans certains cas, faute de mieux, à ces questions.

Convergence On rappelle qu'une méthode itérative de type I, *i.e.* écrite sous la forme $x^{(n+1)} = Bx^{(n)} + C$ converge si et seulement si $\rho(B) < 1$ (voir le théorème 1.26 page 28).

Théorème 1.34 (Sur la convergence de la méthode SOR)
Soit $A \in \mathcal{M}_N(\mathbb{R})$ qui admet une décomposition par blocs définie dans la définition 1.5.32 page 31 ; soient D la matrice constituée par les blocs diagonaux, $-E$ (resp. $-F$) la matrice constituée par les blocs triangulaires inférieurs (resp. supérieurs) ; on a donc : $A = D - E - F$. Soit $\mathcal{L}_\omega$ la matrice d'itération de la méthode SOR (et de la méthode de Gauss–Seidel pour $\omega = 1$) définie par :

$$\mathcal{L}_\omega = \left(\frac{D}{\omega} - E\right)^{-1}\left(F + \frac{1-\omega}{\omega}D\right), \quad \omega \neq 0.$$

Alors :

1. *Si $\rho(\mathcal{L}_\omega) < 1$ alors $0 < \omega < 2$.*
2. *Si on suppose de plus que A symétrique définie positive, alors :*

$$\rho(\mathcal{L}_\omega) < 1 \text{ si et seulement si } 0 < \omega < 2.$$

En particulier, si A est une matrice symétrique définie positive, la méthode de Gauss–Seidel converge.

Démonstration du théorème 1.34 :

1. Calculons $\det(\mathcal{L}_\omega)$. Par définition,

$$\mathcal{L}_\omega = \tilde{M}^{-1}\tilde{N}, \text{ avec } \tilde{M} = \frac{1}{\omega}D - E \text{ et } \tilde{N} = F + \frac{1-\omega}{\omega}D.$$

Donc $\det(\mathcal{L}_\omega) = (\det(\tilde{M}))^{-1}\det(\tilde{N})$. Comme $\tilde{M}$ et $\tilde{N}$ sont des matrices triangulaires par blocs, leurs déterminants sont les produits des déterminants des blocs diagonaux (voir la remarque 1.32 page 31). On a donc :

$$\det(\mathcal{L}_\omega) = \frac{(\frac{1-\omega}{\omega})^N \det(D)}{(\frac{1}{\omega})^N \det(D)} = (1-\omega)^N.$$

Or le déterminant d'une matrice est aussi le produit des valeurs propres de cette matrice (comptées avec leur multiplicités algébriques), dont les valeurs absolues sont toutes, par définition, inférieures au rayon spectral. On a donc : $|\det(\mathcal{L}_\omega)| = |(1-\omega)^N| \leq (\rho(\mathcal{L}_\omega))^N$, d'où le résultat.

2. Supposons maintenant que A est une matrice symétrique définie positive, et que $0 < \omega < 2$. Montrons que $\rho(\mathcal{L}_\omega) < 1$. Par le théorème 1.30 page 29, il suffit pour cela de montrer que $\tilde{M}^t + \tilde{N}$ est une matrice symétrique définie positive. Or,

$$\tilde{M}^t = \left(\frac{D}{\omega} - E\right)^t = \frac{D}{\omega} - F,$$

$$\tilde{M}^t + \tilde{N} = \frac{D}{\omega} - F + F + \frac{1-\omega}{\omega}D = \frac{2-\omega}{\omega}D.$$

La matrice $\tilde{M}^t + \tilde{N}$ est donc bien symétrique définie positive.

■

Remarque 1.35 (Comparaison Gauss–Seidel/Jacobi) *On a vu (théorème 1.34) 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 30 page 46.*

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 31 page 47, et un résultat de comparaison des méthodes pour les matrices tridiagonales par blocs, voir le théorème 1.36 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 1.34) tel que $\rho(\mathcal{L}_{\omega_0}) < \rho(\mathcal{L}_\omega)\ \forall \omega \in]0,2[$.
D'après le paragraphe précédent ce ω_0 donnera la meilleure convergence possible pour SOR. On sait le faire dans le cas assez restrictif des matrices tridiagonales par blocs. 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 1.36 (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 1.5.32 page 31 ; on suppose que la matrice A est tridiagonale par blocs, c.à.d. $A_{i,j} = 0$ si $|i-j| > 1$; soient $\mathcal{L}_1$ et J les matrices d'itération respectives des méthodes de Gauss-Seidel et Jacobi, alors :*

1. *$\rho(\mathcal{L}_1) = (\rho(J))^2$: la méthode de Gauss–Seidel converge (ou diverge) donc plus vite que celle de Jacobi.*
2. *On suppose de plus que toutes les valeurs propres de la matrice d'itération J de la méthode de Jacobi sont réelles. alors le paramètre de relaxation optimal, c.à.d. le paramètre ω_0 tel que $\rho(\mathcal{L}_{\omega_0}) = \min\{\rho(\mathcal{L}_\omega), \omega \in]0,2[\}$, s'exprime en fonction du rayon spectral $\rho(J)$ de la matrice J par la formule :*

$$\omega_0 = \frac{2}{1+\sqrt{1-\rho(J)^2}} > 1,$$

et on a : $\rho(\mathcal{L}_{\omega_0}) = \omega_0 - 1$.

1.5.4 Recherche de valeurs propres et vecteurs propres

Les techniques de recherche des éléments propres, c.à.d. des valeurs et vecteurs propres (voir Définition 1.4 page 6) 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, voir à ce propos l'exemple célèbre du "Tacoma Bridge", décrit dans les livres de M. Braun (en anglais) et M. Schatzman (en français) conseillés en début de cours.
On donne dans les exercices qui suivent deux méthodes assez classiques de recherche de valeurs propres d'une matrice qui sont la méthode de la puissance (exercice 41 page 51) et celui de la puissance inverse (exercice 42 page 52). Citons également une méthode très employée, la méthode QR, qui est présente dans de nombreuses bibliothèques de programmes. On pourra se référer aux ouvrages de Ph. Ciarlet et de M. Schatzman, de D. Serre et de P. Lascaux et R. Theodor (voir introduction).

1.6 Exercices

Exercice 1 (Matrices symétriques définies positives) *Suggestions en page 53, corrigé en page 57.*

On rappelle que toute matrice $A \in \mathcal{M}_N(\mathbb{R})$ symétrique est diagonalisable dans $\mathbb{R}$ (cf. lemme 1.11 page 8). 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 $\{f_1, \ldots, f_N\}$, et il existe $\lambda_1, \ldots, \lambda_N \in \mathbb{R}$ t.q. $Af_i = \lambda_i f_i$, pour tout $i \in \{1, \ldots, N\}$, et $f_i \cdot f_j = \delta_{i,j}$ pour tout $i, j \in \{1, \ldots, N\}$ ($x \cdot y$ désigne le produit scalaire de x avec y dans $\mathbb{R}^N$).

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

Exercice 2 (Normes de l'Identité)

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

Exercice 3 (Normes induites particulières) *Suggestions en page 53, corrigé détaillé en page 57.*

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

1. On munit $\mathbb{R}^N$ de la norme $\|\cdot\|_\infty$ et $\mathcal{M}_N(\mathbb{R})$ de la norme induite correspondante, notée aussi $\|\cdot\|_\infty$. Montrer que $\|A\|_\infty = \max_{i \in \{1,\ldots,N\}} \sum_{j=1}^N |a_{i,j}|$.
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 $\|A\|_1 = \max_{j \in \{1,\ldots,N\}} \sum_{i=1}^N |a_{i,j}|$.
3. On munit $\mathbb{R}^N$ de la norme $\|\cdot\|_2$ et $\mathcal{M}_N(\mathbb{R})$ de la norme induite correspondante, notée aussi $\|\cdot\|_2$. Montrer que $\|A\|_2 = (\rho(A^t A))^{\frac{1}{2}}$.

Exercice 4 (Norme non induite)

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

1. Montrer que $\|\cdot\|_s$ est une norme matricielle mais n'est pas une norme induite (pour $N > 1$).

2. Montrer que $\|A\|_s^2 = tr(A^tA)$. En déduire que $\|A\|_2 \leq \|A\|_s \leq \sqrt{N}\|A\|_2$ et que $\|Ax\|_2 \leq \|A\|_s\|x\|_2$, pour tout $A \in \mathcal{M}_N(\mathbb{R})$ et tout $x \in \mathbb{R}^N$.
3. Chercher un exemple de norme non matricielle.

Exercice 5 (Rayon spectral) *Suggestions en page 53, corrigé détaillé en page 58.*

On munit $\mathcal{M}_N(\mathbb{R})$ d'une norme, notée $\|\cdot\|$. Soit $A \in \mathcal{M}_N(\mathbb{R})$.

1. Montrer que $\rho(A) < 1$ si et seulement si $A^k \to 0$ quand $k \to \infty$.

2. Montrer que : $\rho(A) < 1 \Rightarrow \limsup_{k\to\infty} \|A^k\|^{\frac{1}{k}} \leq 1$.

3. Montrer que :$\liminf_{k\to\infty} \|A^k\|^{\frac{1}{k}} < 1 \Rightarrow \rho(A) < 1$.

4. Montrer que $\rho(A) = \lim_{k\to\infty} \|A^k\|^{\frac{1}{k}}$.

5. On suppose ici que $\|\cdot\|$ est une norme matricielle, déduire de la question précédente que $\rho(A) \leq \|A\|$. On a ainsi démontré la proposition 1.5.

Exercice 6 (Valeurs propres nulles d'un produit de matrices)

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

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

Exercice 7 (Rayon spectral) *Corrigé en page 59.*

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 8 (Sur le rayon spectral)

On définit les matrices carrées d'ordre 2 suivantes :

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

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

Exercice 9 (Série de Neumann) *Suggestions en page 53, corrigé détaillé en page 59.*

Soient $A \in \mathcal{M}_N(\mathbb{R})$ et $\|\cdot\|$ une norme matricielle.

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 $\rho(A) < 1$.

Exercice 10 (Normes matricielles)

Soit $\|.\|$ une norme matricielle quelconque, 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 11 (Décomposition LDL^t et LL^t) *Corrigé en page 60*

1. Soit $A = \begin{pmatrix} 2 & 1 \\ 1 & 0 \end{pmatrix}$.
Calculer la décomposition LDL^t de A. Existe-t-il une décomposition LL^t de A ?

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

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

Exercice 12 (Décomposition LU et LL^t de matrices 3×3)

1. Soit $A = \begin{pmatrix} 2 & -1 & 4 & 0 \\ 4 & -1 & 5 & 1 \\ -2 & 2 & -2 & 3 \\ 0 & 3 & -9 & 4 \end{pmatrix}$. Donner la décomposition LU de A.

2. Soit $B = \begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix}$. Donner la décomposition LL^t de B.

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

Exercice 13 (Sur la méthode LL^t) *Corrigé détaillé en page 62.*

Soit A une matrice carrée d'ordre N, symétrique définie positive et pleine. On cherche à résoudre le système $A^2 x = b$.
On propose deux méthodes de résolution de ce système :

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

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

Exercice 14 (Décomposition LL^t d'une matrice tridiagonale symétrique)

Soit $A \in \mathcal{M}_N(\mathbb{R})$ symétrique définie positive et tridiagonale (i.e. $a_{i,j} = 0$ si $i - j > 1$).

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

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

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

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

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

Exercice 15 (Choleski pour matrice bande) *Suggestions en page 53, corrigé en page 62*

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 23. Même question pour la discrétisation de l'équation $-\Delta u = f$ présentée page 25.

Exercice 16 (Un système par blocs)

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

$$\begin{aligned} &Ax + b\lambda = f, \\ &c \cdot x + \alpha\lambda = \gamma. \end{aligned} \tag{1.6.41}$$

1. Ecrire le système (1.6.41) 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 (1.6.41) soit inversible. Dans toute la suite, on supposera que cette relation est vérifiée.
3. On propose la méthode suivante pour la résolution du système (1.6.41) :
   (a) Soient z solution de $Az = b$, et h solution de $Ah = f$.
   (b) $x = h - \dfrac{\gamma - c \cdot h}{\alpha - c \cdot z} z, \qquad \lambda = \dfrac{\gamma - c \cdot h}{\alpha - c \cdot z}.$

   Montrer que $x \in \mathbb{R}^N$ et $\lambda \in \mathbb{R}$ ainsi calculés sont bien solutions du système (1.6.41).
4. On suppose dans cette question que A est une matrice bande, dont la largeur de bande est p.
   (a) Calculer le coût de la méthode de résolution proposée ci-dessus en utilisant la méthode LU pour la résolution des systèmes linéaires.

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

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

Exercice 17 (Propriétés générales du conditionnement) *Corrigé détaillé en page 63.*

Partie I
On munit $\mathbb{R}^N$ d'une norme, notée $\|\cdot\|$, et $\mathcal{M}_N(\mathbb{R})$ de la norme induite, notée aussi $\|\cdot\|$. Pour une matrice inversible $A \in \mathcal{M}_N(\mathbb{R})$, on note $\mathrm{cond}(A) = \|A\|\|A^{-1}\|$.

1. Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice inversible. Montrer que $\mathrm{cond}(A) \geq 1$.
2. Montrer que $\mathrm{cond}(\alpha A) = \mathrm{cond}(A)$ pour tout $\alpha \in \mathbb{R}^\star$.
3. Soit $A, B \in \mathcal{M}_N(\mathbb{R})$ deux matrices inversibles. Montrer que $\mathrm{cond}(AB) \leq \mathrm{cond}(A)\mathrm{cond}(B)$.

Partie II
On suppose maintenant 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 $\mathrm{cond}_2(A)$ conditionnement d'une matrice A inversible.

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). Montrer que $\mathrm{cond}_2(A) = \sqrt{\sigma_N/\sigma_1}$.
2. On suppose maintenant que A est symétrique définie positive, montrer que $\mathrm{cond}_2(A) = \lambda_N/\lambda_1$ où λ_N [resp. λ_1] est la plus grande [resp. petite] valeur propre de A.
3. 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 $Q^t = Q^{-1}$).
4. Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice inversible. On suppose que $A = QR$ où Q est une matrice orthogonale. Montrer que $\mathrm{cond}_2(A) = \mathrm{cond}_2(R)$.
5. 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 18 (Minoration 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 $\mathrm{cond}\,(A) \geq \sup_{\substack{B \in \mathcal{M}_N(\mathbb{R}) \\ detB=0}} \frac{\|A\|}{\|A-B\|}$

Exercice 19 (Minoration du conditionnement) *Corrigé détaillé en page 65.*

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

1. Montrer que si $A + \delta A$ est singulière, alors
$$\mathrm{cond}(A) \geq \frac{\|A\|}{\|\delta A\|}. \tag{1.6.42}$$
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 (1.6.42) 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 (1.6.42).
[*On pourra chercher δA 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 (1.6.42) pour trouver un minorant, qui tend vers $+\infty$ lorsque α tend vers 0, de $\mathrm{cond}(A)$ pour la matrice

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

Exercice 20 (Conditionnement du carré)

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

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

Exercice 21 (Calcul de l'inverse d'une matrice et conditionnement) *Corrigé détaillé en page 66.*

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

1. On suppose qu'on a calculé B, approximation (en raison par exemple d'erreurs d'arrondi) de la matrice A^{-1}. On pose :

$$\left\{ \begin{array}{lllll} e_1 & = & \dfrac{\|B - A^{-1}\|}{\|A^{-1}\|}, & e_2 & = & \dfrac{\|B^{-1} - A\|}{\|A\|} \\ \\ e_3 & = & \|AB - Id\|, & e_4 & = & \|BA - Id\| \end{array} \right. \tag{1.6.43}$$

   (a) Expliquer en quoi les quantités e_1, e_2, e_3 et e_4 mesurent la qualité de l'approximation de A^{-1}.

   (b) On suppose ici que $B = A^{-1} + E$, où $\|E\| \leq \varepsilon \|A^{-1}\|$, et que

$$\varepsilon \mathrm{cond}(A) < 1.$$

   Montrer que dans ce cas,

$$e_1 \leq \varepsilon,\ e_2 \leq \frac{\varepsilon \mathrm{cond}(A)}{1 - \varepsilon \mathrm{cond}(A)},\ e_3 \leq \varepsilon \mathrm{cond}(A) \text{ et } e_4 \leq \varepsilon \mathrm{cond}(A).$$

   (c) On suppose maintenant que $AB - Id = E'$ avec $\|E'\| \leq \varepsilon < 1$. Montrer que dans ce cas :

$$e_1 \leq \varepsilon,\ e_2 \leq \frac{\varepsilon}{1 - \varepsilon},\ e_3 \leq \varepsilon \text{ et } e_4 \leq \varepsilon \mathrm{cond}(A).$$

2. On suppose maintenant que la matrice A n'est connue qu'à une certaine matrice d'erreurs près, qu'on note δ_A.

   (a) Montrer que la matrice $A + \delta_A$ est inversible si $\|\delta_A\| < \dfrac{1}{\|A^{-1}\|}$.

   (b) Montrer que si la matrice $A + \delta_A$ est inversible,

$$\frac{\|(A + \delta_A)^{-1} - A^{-1}\|}{\|(A + \delta_A)^{-1}\|} \leq \mathrm{cond}(A) \frac{\|\delta_A\|}{\|A\|}.$$

Exercice 22 (Discrétisation)

On considère la discrétisation à pas constant par le schéma aux différences finies symétrique à trois points (vu en cours) du problème (1.4.21) page 23, 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 (1.4.23).

1. Montrer que si $u \in C^4([0,1]$, alors la propriété (1.4.22) est vérifiée, c.à.d. :

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

2. Montrer que si f est constante, alors

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

.

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

Exercice 23 (IP-matrice) *Corrigé en page 67*

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

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

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

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

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

$$\left\{ \begin{array}{l} ad < bc, \\ a < 0,\ d < 0 \\ b \geq 0,\ c \geq 0 \end{array} \right. \text{ ou } \left\{ \begin{array}{l} ad > bc, \\ a > 0,\ d > 0, \\ b \leq 0,\ c \leq 0. \end{array} \right. \tag{1.6.44}$$

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

4. Montrer que si A est telle que

$$\begin{array}{l} a_{i,j} \leq 0, \text{ pour tout } i, j = 1, \ldots, N,\ i \neq j, \\ a_{i,i} > \displaystyle\sum_{\substack{j=1 \\ j\neq i}}^{N} |a_{i,j}|, \text{ pour tout } i = 1, \ldots, N, \end{array} \tag{1.6.45}$$

alors A est une IP-matrice.

5. En déduire que si A est telle que

$$\begin{array}{l} a_{i,j} \leq 0, \text{ pour tout } i, j = 1, \ldots, N,\ i \neq j, \\ a_{i,i} > \displaystyle\sum_{\substack{j=1 \\ j\neq k}}^{N} |a_{j,i}|, \text{ pour tout } i = 1, \ldots, N, \end{array} \tag{1.6.46}$$

alors A est une IP-matrice.

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

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

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

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

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

9. 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 24 (M-matrice)

Dans ce qui suit, toutes les inégalités écrites sur des vecteurs ou des matrices sont à entendre au sens composante par composante.
Soit $A = (a_{i,j})_{i,j=1,\ldots,n}$ une matrice carrée d'ordre n. On dit que A est une M-matrice si A est une IP-matrice (A est inversible et $A^{-1} \geq 0$, voir exercice 23) qui vérifie de plus

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

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

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

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

$$\left\{ \begin{array}{l} ad > bc, \\ a > 0,\ d > 0, \\ b \leq 0,\ c \leq 0. \end{array} \right. \tag{1.6.47}$$

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

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

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

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

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

$$\left. \begin{array}{l} a_{i,i} \geq 0, \\ a_{i,j} \leq 0, \text{ pour } j = 1, \ldots, n,\ j \neq i, \\ a_{i,i} + \displaystyle\sum_{\substack{j=1 \\ j \neq i}}^{n} a_{i,j} = 0 \end{array} \right\} \text{ pour } i = 1, \ldots, m, \tag{1.6.48}$$

$$\left. \begin{array}{l} a_{i,i} = 1, \\ a_{i,j} = 0, \text{ pour } j = 1, \ldots, n,\ j \neq i, \end{array} \right\} \text{ pour } i = m+1, \ldots, n. \tag{1.6.49}$$

$$\forall i \leq m,\ \exists (k_\ell)_{\ell=1,\ldots,L_i}; k_1 = i, k_{L_i} > m, \text{ et } a_{k_\ell,k_{\ell+1}} < 0, \forall\, \ell = 1,\ldots,L_i. \tag{1.6.50}$$

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

$$Au = b \tag{1.6.51}$$

5.1 Montrer que le système (1.6.51) admet une et une seule solution.

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

6. On considère le problème de Dirichlet suivant :

$$-u'' = 0 \text{ sur } [0,1] \tag{1.6.52a}$$
$$u(0) = -1 \tag{1.6.52b}$$
$$u(1) = 1. \tag{1.6.52c}$$

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

6.2 Soit $m > 1$ et soient A et b et la matrice et le second membre du système linéaire d'ordre $n = m+2$ obtenu par discrétisation par différences finies avec un pas uniforme $h = \frac{1}{m}$ du problème (1.6.52) (en écrivant les conditions aux limites dans le système). Montrer que la solution $U \in \mathbb{R}^n$ du système $AU = b$ vérifie $-1 \leq u_i \leq 1$.

Exercice 25 (Valeurs propres du Laplacien discret 1D) *Suggestions en page 54, corrigé détaillé en page 68.*

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 (1.4.24) page 24, issue d'une discrétisation par différences finies (vue en cours) du problème (1.4.21) page 23.
1. Montrer que A est symétrique définie positive.
2. 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* $\varphi(0) = \varphi(1) = 0$].
3. Calculer $\mathrm{cond}_2(A)$ et montrer que $h^2\mathrm{cond}_2(A) \to \frac{4}{\pi^2}$ lorsque $h \to 0$.

Exercice 26 (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 aux limites suivant :

$$\begin{array}{l} -u''(x) + u(x) = f(x),\ x \in]0,1[, \\ u(0) = u(1) = 0. \end{array} \tag{1.6.53}$$

Soit $N \in \mathbb{N}^\star$. On note $U = (u_j)_{j=1,\ldots,N}$ une "valeur approchée" de la solution u du problème (1.6.53) 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 $A = (N+1)^2 B + Id$, 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
   $$\begin{array}{l} -u''(x) = \lambda u(x),\ x \in]0,1[, \\ u(0) = u(1) = 0. \end{array} \tag{1.6.54}$$
   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 $U_k = \left(u_k(\frac{j}{N+1})\right)_{j=1,\dots,N}$ 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 27 (Conditionnement "efficace".) *Suggestions en page 54, corrigé en page 69.*

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 (1.4.24) page 24, issue d'une discrétisation par différences finies (vue en cours) du problème (1.4.21) page 23.

Pour $u \in \mathbb{R}^N$, on note $u_1, \dots, u_N$ les composantes de u. Pour $u \in \mathbb{R}^N$, on dit que $u \geq 0$ si $u_i \geq 0$ pour tout $i \in \{1, \dots, N\}$. 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 $u \in \mathbb{R}^N$, $\|u\| = \max\{|u_i|, i \in \{1, \dots, N\}\}$. 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 :
   $$\left\{ \begin{array}{l} \frac{1}{h^2}(u_i - u_{i-1}) + \frac{1}{h^2}(u_i - u_{i+1}) = b_i,\ \forall i \in \{1, \dots, N\}, \\ u_0 = u_{N+1} = 0. \end{array} \right. \tag{1.6.55}$$
   Montrer que $b \geq 0 \Rightarrow u \geq 0$. [On pourra considèrer $p \in \{0, \dots, N+1\}$ t.q. $u_p = \min\{u_j, j \in \{0, \dots, N+1\}$.]

   En déduire que A est inversible.
2. (Préliminaire...) On considère la fonction $\varphi \in C([0,1], \mathbb{R})$ définie par $\varphi(x) = (1/2)x(1-x)$ pour tout $x \in [0,1]$. On définit alors $\phi \in \mathbb{R}^N$ par $\phi_i = \phi(ih)$ pour tout $i \in \{1, \dots, N\}$. Montrer que $(A\phi)_i = 1$ pour tout $i \in \{1, \dots, N\}$.
3. (calcul de $\|A^{-1}\|$) Soient $b \in \mathbb{R}^N$ et $u \in \mathbb{R}^N$ t.q. $Au = b$. Montrer que $\|u\| \leq (1/8)\|b\|$ [Calculer $A(u \pm \|b\|\phi)$ avec ϕ 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 b, $\delta_b \in \mathbb{R}^N$. 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 b et δ_b donne l'égalité dans l'inégalité précédente.

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

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

1. Montrer que
$$h\sum_{i=1}^{N} b_i\varphi_i \to \int_0^1 f(x)\phi(x)dx \text{ quand } N \to \infty$$
et que
$$\sum_{i=1}^{N} b_i\varphi_i > 0 \text{ pour tout } N \in \mathbb{N}\ *\,.$$
En déduire qu'il existe $\alpha > 0$, ne dépendant que de f, t.q. $h\sum_{i=1}^{N} b_i\varphi_i \geq \alpha$ pour tout $N \in \mathbb{N}^\star$.
2. Soit $u \in \mathbb{R}^N$ t.q. $Au = b$. Montrer que $N\|u\| \geq \sum_{i=1}^{N} u_i = u \cdot A\varphi \geq \frac{\alpha}{h}$ (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 $N \to \infty$).

Exercice 28 (Méthode itérative du "gradient à pas fixe"[6]) *Suggestions en page 54*

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 : $x^{(n+1)} = x^{(n)} + \alpha(b - Ax^{(n)})$.

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. $\rho(Id - \alpha_0 A) = \min\{\rho(Id - \alpha A),\ \alpha \in \mathbb{R}\}$.

Exercice 29 (Une matrice cyclique) *Suggestions en page 54, corrigé en page 71*

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 $(x^{(n)})_{n\in\mathbb{N}}$ la suite de vecteurs donnés par l'algorithme. On note $x_i^{(n)}$ pour $i = 1, \ldots, 4$ les composantes de $x^{(n)}$. Donner l'expression de $x_i^{(n+1)}$, $i = 1, \ldots, 4$, en fonction de $x_i^{(n)}$ et $b_i^{(n)}$, $i = 1, \ldots, 4$. 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 30 (Non convergence de la méthode de Jacobi) *Suggestions en page 55.*

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 31 (Jacobi pour les matrices à diagonale dominante stricte) *Suggestions en page 55, corrigé en page 73*

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 32 (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 :

$$\begin{cases} -u''(x) + \alpha u(x) = f(x), \ x \in [0,1], \\ u(0) = 0, u(1) = 1, \end{cases} \tag{1.6.56}$$

où $\alpha \geq 0$.

1. Donner l'expression de A et 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 33 (Jacobi et diagonale dominance forte) *Corrigé en page 74*

On considère une matrice $A \in \mathcal{M}_N(\mathbb{R})$ inversible.

1. Montrer que si A est symétrique définie positive alors tous ses coefficients diagonaux sont strictement positifs. En déduire que la méthode de Jacobi est bien définie.
2. On suppose maintenant que la matrice diagonale extraite de A, notée D, est inversible. On suppose de plus que
$$\forall i = 1, \ldots, N, |a_{i,i}| \geq \sum_{j \neq i} |a_{i,j}| \text{ et } \exists i_0; |a_{i_0,i_0}| > \sum_{j \neq i_0} |a_{i,j}|.$$
(On dit que la matrice est à diagonale fortement dominante). Soit J la matrice d'itération de la méthode de Jacobi.
   (a) Montrer que $\rho(J) \leq 1$.
   (b) Montrer que si $Jx = \lambda x$ avec $|\lambda| = 1$, alors $x_i = \|x\|$, $\forall i = 1, \ldots, N$. En déduire que $x = 0$ et que la méthode de Jacobi converge.
   (c) Retrouver ainsi le résultat de la question 2 de l'exercice 32.

Exercice 34 (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 $(x,y) \to (x,y)_T$ 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 1.11 page 8, qu'il existe une base de E, notée $\{f_1, \dots, f_N\}$, et il existe $\{\lambda_1, \dots, \lambda_N\} \subset \mathbb{R}$ t.q. $T^{-1}Sf_i = \lambda_i f_i$ pour tout $i \in \{1, \dots, N\}$ et t.q. $(Tf_i/f_j) = \delta_{i,j}$ pour tout $i, j \in \{1, \dots, N\}$.

Exercice 35 (Méthode de Jacobi et relaxation) *Suggestions en page 55, corrigé en page 75*

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 $n \in \mathbb{N}$, $Dx^{(n+1)} = (E+F)x^{(n)} + b$.

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 $\{f_1, \dots, f_N\}$, et il existe $\{\mu_1, \dots, \mu_N\} \subset \mathbb{R}$ t.q. $Jf_i = \mu_i f_i$ pour tout $i \in \{1, \dots, N\}$ et t.q. $Df_i \cdot f_j = \delta_{i,j}$ pour tout $i, j \in \{1, \dots, N\}$.

   En ordonnant les valeurs propres de J, on a donc $\mu_1 \leq \dots \leq \mu_N$, 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 $x^{(n)} \to x$ 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 $n \in \mathbb{N}$, $D\tilde{x}^{(n+1)} = (E+F)x^{(n)} + b$, $x^{(n+1)} = \omega\tilde{x}^{(n+1)} + (1-\omega)x^{(n)}$.
5. Calculer les matrices M_ω (inversible) et N_ω telles que $M_\omega x^{(n+1)} = N_\omega x^{(n)} + b$ 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 $x^{(n)} \to x$ 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 36 (Jacobi et Gauss-Seidel pour une matrice tridiagonale)

Soit $A = (a_{i,j})_{i,j=1,\dots,N} \in \mathcal{M}_N(\mathbb{R})$ une matrice carrée d'ordre N tridiagonale, c'est-à-dire telle que $a_{i,j} = 0$ si $|i-j| > 1$, et telle que la matrice diagonale $D = diag(a_{i,i})_{i=1,\dots,N}$ soit inversible. On note $A = D - E - F$ où $-E$ (resp. $-F$) est la partie triangulaire inférieure (resp. supérieure) de A, et on note J et G les matrices d'itération des méthodes de Jacobi et Gauss-Seidel associées à la matrice A.

1.a. Pour $\mu \in \mathbb{C}, \lambda \neq 0$ et $x \in \mathbb{C}^N$, on note

$$x_\mu = (x_1, \mu x_2, \dots, \mu^{k-1} x_k, \mu^{N-1} x_N)^t.$$

Montrer que si λ est valeur propre de J associée au vecteur propre x, alors x_μ vérifie $(\mu E + \frac{1}{\mu} F)x_\mu = \lambda D x_\mu$. En déduire que si $\lambda \neq 0$ est valeur propre de J alors λ^2 est valeur propre de G.

1.b Montrer que si λ^2 est valeur propre non nulle de G, alors λ est valeur propre de J .

2. Montrer que $\rho(G) = \rho(J)^2$. En déduire que lorsqu'elle converge, la méthode de Gauss-Seidel pour la résolution du système $Ax = b$ converge plus rapidement que la méthode de Jacobi.

3. Soit $\mathcal{L}_\omega$ 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 $\mathcal{L}_\omega$, où $\nu_\omega = \mu_\omega^2$ et μ_ω vérifie $\mu_\omega^2 - \lambda\omega\mu_\omega + \omega - 1 = 0$.
En déduire que

$$\rho(\mathcal{L}_\omega) = \max_{\lambda \text{ valeur propre de } J} \{|\mu_\omega|; \mu_\omega^2 - \lambda\omega\mu_\omega + \omega - 1 = 0\}.$$

Exercice 37 (Méthode de Jacobi pour des matrices particulières) *Suggestions en page 55, corrigé en page 77*

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 :

$$a_{i,j} \leq 0, \forall i, j = 1, \dots, N, i \neq j, \tag{1.6.57}$$

$$a_{i,i} > 0, \forall i = 1, \dots, N. \tag{1.6.58}$$

$$\sum_{i=1}^{N} a_{i,j} = 0, \forall j = 1, \dots, N. \tag{1.6.59}$$

Soit $\lambda \in \mathbb{R}_+^*$.

1. Pour $x \in \mathbb{R}^N$, on définit
$$\|x\|_A = \sum_{i=1}^{N} a_{i,i}|x_i|.$$
Montrer que $\|\cdot\|_A$ est une norme sur $\mathbb{R}^N$.
2. Montrer que la matrice $\lambda Id + A$ est inversible.
3. On considère le système linéaire suivant :
$$(\lambda Id + A)u = b \tag{1.6.60}$$
Montrer que la méthode de Jacobi pour la recherche de la solution de ce système définit une suite $(u^{(k)})_{k\in N} \subset \mathbb{R}^N$.
4. Montrer que la suite $(u^{(k)})_{k\in\mathbb{N}}$ vérifie :
$$\|u^{(k+1)} - u^{(k)}\|_A \leq (\frac{1}{1+\alpha})^k \|u^{(1)} - u^{(0)}\|_A,$$
où $\alpha = \min_{i=1,\dots,N} a_{i,i}$.

5. Montrer que la suite $(u^{(k)})_{k\in\mathbb{N}}$ est de Cauchy, et en déduire qu'elle converge vers la solution du système (1.6.60).

Exercice 38 (Une méthode itérative particulière)

Soient $\alpha_1, \ldots, \alpha_n$ des réels strictement positifs, et A la matrice $n \times n$ de coefficients $a_{i,j}$ définis par :

$$\begin{cases} a_{i,i} = 2 + \alpha_i \\ a_{i,i+1} = a_{i,i-1} = -1 \\ a_{i,j} = 0 \text{ pour tous les autres cas.} \end{cases}$$

Pour $\beta > 0$ on considère la méthode itérative $Mx^{(k+1)} = Nx^{(k)} + b$ avec $A = M - N$ et $N = \text{diag}(\beta - \alpha_i)$ (c.à.d $\beta - \alpha_i$ pour les coefficients diagonaux, et 0 pour tous les autres).

1. Soit $\lambda \in \mathbb{C}$ une valeur propre de la matrice $M^{-1}N$; montrer qu'il existe un vecteur $x \in \mathbb{C}^n$ non nul tel que $Nx \cdot \overline{x} = \lambda Mx \cdot \overline{x}$ (où $\overline{x}$ désigne le conjugué de x). En déduire que toutes les valeurs propres de la matrice $M^{-1}N$ sont réelles.

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

3. Déduire de la question 1. que si $\beta > \dfrac{\overline{\alpha}}{2}$, où $\overline{\alpha} = \max_{i=1,n} \alpha_i$, alors $\rho(M^{-1}N) < 1$, et donc que la méthode itérative converge.

4. Trouver le paramètre β minimisant $\max_{i=1,n} \dfrac{|\beta - \alpha_i|}{\beta}$.
(On pourra d'abord montrer que pour tout $\beta > 0$, $|\beta - \alpha_i| \leq \max(\beta - \underline{\alpha}, \bar{\alpha} - \beta)$ pour tout $i = 1, \ldots, n$, 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 39 (Une matrice 3×3) *Suggestions en page 55, corrigé en page 79*

Soit $A \in M_3(\mathbb{R})$ définie par $A = Id - E - F$ avec

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

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

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

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

   Il s'agit donc de la "méthode I" du cours avec $B = \mathcal{L}_\omega = (\frac{1}{\omega} Id - E)^{-1}(F + \frac{1-\omega}{\omega} Id)$.
3. Calculer, en fonction de ω, les valeurs propres de $\mathcal{L}_\omega$ et son rayon spectral.
4. Pour quelles valeurs de ω la méthode est-elle convergente ? Déterminer $\omega_0 \in]0, 2[$ t.q. $\rho(\mathcal{L}_{\omega_0}) = \min\{\rho(\mathcal{L}_\omega), \omega \in]0, 2[, \omega \neq \sqrt{2}/2\}$.

Exercice 40 (Méthode des directions alternées)

Soit $N \in \mathbb{N}$ et $N \geq 1$, Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice carrée d'ordre N symétrique inversible et $b \in \mathbb{R}^N$.
On cherche à calculer $u \in \mathbb{R}^N$, solution du système linéaire suivant :

$$Au = b, \tag{1.6.61}$$

On suppose connues des matrices X et $Y \in \mathcal{M}_N(\mathbb{R})$, symétriques. Soit $\alpha \in \mathbb{R}_+^*$, choisi tel que $X + \alpha Id$ et $Y + \alpha Id$ soient définies positives (où Id désigne la matrice identité d'ordre N) et $X + Y + \alpha Id = A$.

Soit $u^{(0)} \in \mathbb{R}^N$, on propose, pour résoudre (1.6.61), la méthode itérative suivante :

$$\begin{cases} (X + \alpha Id)u^{(k+1/2)} = -Yu^{(k)} + b, \\ (Y + \alpha Id)u^{(k+1)} = -Xu^{(k+1/2)} + b. \end{cases} \tag{1.6.62}$$

1. Montrer que la méthode itérative (1.6.62) définit bien une suite $(u^{(k)})_{k\in\mathbb{N}}$ et que cette suite converge vers la solution u de (1.1.1) si et seulement si

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

   (On rappelle que pour toute matrice carrée d'ordre N, $\rho(M)$ désigne le rayon spectral de la matrice M.)
2. Montrer que si les matrices $(X + \frac{\alpha}{2}Id)$ et $(Y + \frac{\alpha}{2}Id)$ sont définies positives alors la méthode (1.6.62) converge. On pourra pour cela (mais ce n'est pas obligatoire) suivre la démarche suivante :
   (a) Montrer que

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

   (On pourra utiliser l'exercice 6 page 37).
   (b) Montrer que

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

   (c) Montrer que $\rho\big(X(X + \alpha Id)^{-1}\big) < 1$ si et seulement si la matrice $(X + \frac{\alpha}{2}Id)$ est définie positive.
   (d) Conclure.
3. Soit $f \in C([0, 1] \times [0, 1])$ et soit A la matrice carrée d'ordre $N = M \times M$ obtenue par discrétisation de l'équation $-\Delta u = f$ sur le carré $[0, 1] \times [0, 1]$ avec conditions aux limites de Dirichlet homogènes $u = 0$ sur $\partial\Omega$, par différences finies avec un pas uniforme $h = \frac{1}{M}$, et b le second membre associé.
   (a) Donner l'expression de A et b.
   (b) Proposer des choix de X, Y et α pour lesquelles la méthode itérative (1.6.62) converge dans ce cas et qui justifient l'appellation "méthode des directions alternées" qui lui est donnée.

Exercice 41 (Méthode de la puissance) *Suggestions en page 56, corrigé en page 80*

1. Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice symétrique. Soit $\lambda_N \in \mathbb{R}$ valeur propre de A t.q. $|\lambda_N| = \rho(A)$ et soit $x^{(0)} \in \mathbb{R}^N$. On suppose que $-\lambda_N$ n'est pas une valeur propre de A et que $x^{(0)}$ n'est pas orthogonal à $\mathrm{Ker}(A - \lambda_N Id)$. On définit la suite $(x^{(n)})_{n\in\mathbb{N}}$ par $x^{(n+1)} = Ax^{(n)}$ pour $n \in \mathbb{N}$. Montrer que
   (a) $\frac{x^{(n)}}{(\lambda_N)^n} \to x$, quand $n \to \infty$, avec $x \neq 0$ et $Ax = \lambda_N x$.
   (b) $\frac{\|x^{(n+1)}\|}{\|x^{(n)}\|} \to \rho(A)$ quand $n \to \infty$.

   Cette méthode de calcul s'appelle "méthode de la puissance".

2. Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice inversible et $b \in \mathbb{R}^N$. Pour calculer x t.q. $Ax = b$, on considère la méthode itérative appelée “méthode I" en cours, et on suppose B symétrique. Montrer que, sauf cas particuliers à préciser,

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

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

Exercice 42 (Méthode de la puissance inverse) *Suggestions en page 56.*

Soient $A \in \mathcal{M}_N(\mathbb{R})$ une matrice symétrique et $\lambda_1, \dots, \lambda_p$ ($p \leq N$) les valeurs propres de A. Soit $i \in \{1, \dots, p\}$, on cherche à calculer λ_i. Soit $x^{(0)} \in \mathbb{R}^N$. On suppose que $x^{(0)}$ n'est pas orthogonal à $\mathrm{Ker}(A - \lambda_i Id)$. On suppose également connaître $\mu \in \mathbb{R}$ t.q. $0 < |\mu - \lambda_i| < |\mu - \lambda_j|$ pour tout $j \neq i$. On définit la suite $(x^{(n)})_{n\in\mathbb{N}}$ par $(A - \mu Id)x^{(n+1)} = x^{(n)}$ pour $n \in \mathbb{N}$. Montrer que

1. $x^{(n)}(\lambda_i - \mu)^n \to x$, quand $n \to \infty$, avec $x \neq 0$ et $Ax = \lambda_i x$.
2. $\frac{\|x^{(n+1)}\|}{\|x^{(n)}\|} \to \frac{1}{|\mu-\lambda_i|}$ quand $n \to \infty$.

1.7 Suggestions

Exercice 1 page 36 (Matrices symétriques définies positives)

3. Utiliser la diagonalisation sur les opérateurs linéaires associés.

Exercice 3 page 36 (Normes induites particulières)

1. Pour montrer l'égalité, prendre x tel que $x_j = \mathrm{sign}(a_{i_0,j})$ 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 $\mathrm{sign}(s)$ 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 $\sum_{i=1,\dots,N} |a_{i,j_0}| = \max_{j=1,\dots,N} \sum_{i=1,\dots,N} |a_{i,j}|$.

3. Utiliser le fait que $A^t A$ est une matrice symétrique positive pour montrer l'inégalité, et pour l'égalité, prendre pour x le vecteur propre associé à la plus grande valeur propre de A.

Exercice 5 page 37 (Rayon spectral)

1. Pour le sens direct, utiliser la proposition 1.6 page 6 du cours.

2. On rappelle que $\limsup_{k\to+\infty} u_k = \lim_{k\to+\infty} \sup_{n\geq k} u_n$, et $\liminf_{k\to+\infty} u_k = \lim_{k\to+\infty} \inf_{n\geq k} u_n$. Utiliser la question 1.

3. Utiliser le fait que $\liminf_{k\to+\infty} u_k$ 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 $uu_{k_n} \to \liminf_{k\to+\infty} u_k$ lorsque $k \to +\infty$.
4. Raisonner avec $\frac{1}{\alpha}A$ où $\alpha \in \mathbb{R}_+$ est tel que $\rho(A) < \alpha$ et utiliser la question 2 pour déduire que

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

Raisonner ensuite avec $\frac{1}{\beta}A$ où $\beta \in \mathbb{R}_+$ est tel que $\liminf_{k\to+\infty} \|A^k\|^{\frac{1}{k}} < \beta$ et utiliser la question 3.

Exercice 9 page 37 (Série de Neumann)

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

2. Utiliser le résultat de la question 1 de l'exercice 5.

Exercice 15 page 39

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 :

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

Exercice 17 page 40 (Propriétés générales du conditionnement)

Partie II
1. On rappelle que si A a comme valeurs propres $\lambda_1, \ldots, \lambda_N$, alors A^{-1} a comme valeurs propres $\lambda_1^{-1}, \ldots, \lambda_N^{-1}$ et A^t a comme valeurs propres $\lambda_1, \ldots, \lambda_N$.
2. Utiliser le fait que AA^t est diagonalisable.

5. Soient $0 < \lambda_1 \leq \lambda_2 \ldots \leq \lambda_N$ et $0 < \mu_1 \leq \mu_2 \ldots \leq \mu_N$ les valeurs propres de A et B (qui sont s.d.p.). Montrer d'abord que :

$$cond_2(A+B) \leq \frac{\lambda_N + \mu_N}{\lambda_1 + \mu_1}.$$

Montrer ensuite que

$$\frac{a+b}{c+d} \leq \max(\frac{a}{c}, \frac{b}{d}),\ \forall (a,b,c,d) \in (\mathbb{R}_+^*)^4.$$

et conclure

Exercice 25 page 44 (Valeurs propres et vecteurs propres de A.)

Chercher les vecteurs propres $\Phi \in \mathbb{R}^N$ de A sous la forme $\Phi_j = \varphi(x_j)$, $j = 1, \ldots, N$ où φ est introduite dans les indications de l'énoncé. Montrer que les valeurs propres associées à ces vecteurs propres sont de la forme : $\lambda_k = \frac{2}{h^2}(1 - \cos k\pi h) = \frac{2}{h^2}(1 - \cos \frac{k\pi}{N+1})$.

Exercice 27 page 45 (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. 2. Utiliser le fait que $A\phi = (1 \ldots 1)^t$.

Exercice 28 page 46(Méthode itérative du "gradient à pas fixe".)

1. Calculer le rayon spectral $\rho(B)$ de la matrice d'itération $B = Id - \alpha A$. Calculer les valeurs de α pour lesquelles $\rho(B) < 1$ et en déduire que la méthode itérative du gradient à pas fixe converge si $0 < \alpha < \frac{2}{\rho(A)}$.
2. Remarquer que $\rho(Id - \alpha A) = \max(|1 - \alpha\lambda_1|, |1 - \alpha\lambda_N - 1|$, où $\lambda_1, \ldots, \lambda_N$ sont les valeurs propres de A ordonnées dans le sens croissant. En traçant les graphes des valeurs prises par $|1 - \alpha\lambda_1|$ et $|1 - \alpha\lambda_N - 1|$ en fonction de α, en déduire que le min est atteint pour $\alpha = \frac{2}{\lambda_1 + \lambda_N}$.

Exercice 29 page 46 (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 30 page 46 (Non convergence de la méthode de Jacobi)

Considèrer 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 31 page 47 (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 35 page 48 (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 34 page 47 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 $\rho(J) = \max(-\mu_1, \mu_N)$, 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 37 page 49 (Méthode de Jacobi et relaxation.)

2. Utiliser l'exercice 31 page 47

Exercice 39 page 50 (Convergence de SOR.)

1. Calculer le déterminant de A.
2. Calculer le déterminant de $\frac{I}{d}\omega - E$.
3. Remarquer que les valeurs propres de $\mathcal{L}_\omega$ annulent $\det(\frac{1-\omega}{\omega} Id + F - \lambda(\frac{I}{d}\omega - E))$. Après calcul de ce déterminant, on trouve $\lambda_1 = 1 - \omega$, $\lambda_2 = \frac{1-\omega}{1+\sqrt{2}\omega}$, $\lambda_3 = \frac{1-\omega}{1-\sqrt{2}\omega}$.
Montrer que si $\omega < \sqrt{2}$, $\rho(\mathcal{L}_\omega) = |\lambda_3|$ et que $\rho(\mathcal{L}_\omega) = |\lambda_1|$ si $\omega \geq \sqrt{2}$.
4. Utiliser l'expression des valeurs propres pour montrer que la méthode converge si $\omega > \frac{2}{1+\sqrt{2}}$ et que le paramètre de relaxation optimal est $\omega_0 = 1$.

Exercice 41 page 51 (Méthode de la puissance pour calculer le rayon spectral de A.)

1. Décomposer x_0 sur une base de vecteurs propres orthonormée de A, et utiliser le fait que $-\lambda_N$ n'est pas valeur propre.
2. a/ Raisonner avec $y^{(n)} = x^{(n)} - x$ où x est la solution de $Ax = b$ et appliquer la question 1.
b/ Raisonner avec $y^{(n)} = x^{(n+1)} - x^{(n)}$.

Exercice 42 page 52 (Méthode de la puissance inverse)

Appliquer l'exercice précédent à la matrice $B = (A - \mu Id)^{-1}$.

1.8 Corrigés

Exercice 1 page 36 (Matrices symétriques définies positives)

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

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

$$Ax \cdot x = \sum_{i=1}^N \lambda_i x_i^2. \tag{1.8.63}$$

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 (1.8.63), $Ax \cdot x \geq 0$ et la matrice A est positive.
Supposons maintenant que $\lambda_i \geq 0$, $\forall i = 1, \ldots, N$. Alors pour $\forall x \in \mathbb{R}^N$, toujours d'après (1.8.63), $(Ax \cdot x = 0) \Rightarrow (x = 0)$, et et la matrice A est donc bien définie.

Montrons maintenant la réciproque :
Si A est positive, alors $Af_i \cdot f_i \geq 0$, $\forall i = 1, \ldots, N$ et donc $\lambda_i \geq 0$, $\forall i = 1, \ldots, N$.
Si A est définie, alors $(A\alpha f_i \cdot \alpha f_i = 0) \Rightarrow (\alpha = 0)$, $\forall i = 1, \ldots, N$ et donc $\lambda_i > 0$, $\forall i = 1, \ldots, N$.
3. Comme A est s.d.p., toutes ses valeurs propres sont strictement positives, et on peut donc définir l'application linéaire S dans la base orthonormée $(f_i)_{i=1,N}$ par : $S(f_i) = \sqrt{\lambda_i} f_i$, $\forall i = 1, \ldots, N$. On a évidemment $S \circ S = T$, et donc si on désigne par B la matrice représentative de l'application S dans la base canonique, on a bien $B^2 = A$.

Exercice 3 page 36 (Normes induites particulières)

1. Par définition, $\|A\|_\infty = \sup_{x \in \mathbb{R}^N, \|x\|_\infty = 1} \|Ax\|_\infty$, et

$$\|Ax\|_\infty = \max_{i=1,\ldots,N} |\sum_{j=1,\ldots,N} a_{i,j} x_j| \leq \max_{i=1,\ldots,N} |\sum_{j=1,\ldots,N} |a_{i,j}||x_j|.$$

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

$$\|Ax\|_\infty \leq \max_{i=1,\ldots,N} |\sum_{j=1,\ldots,N} |a_{i,j}|.$$

Posons maintenant $\alpha = \max_{i=1,\ldots,N} |\sum_{j=1,\ldots,N} |a_{i,j}|$ et montrons qu'il existe $x \in \mathbb{R}^N$, $\|x\|_\infty = 1$, tel que $\|Ax\|_\infty = \alpha$. Pour $s \in \mathbb{R}$, on note $\mathrm{sign}(s)$ le signe de s, c'est–à–dire $\mathrm{sign}(s) = s/|s|$ si $s \neq 0$ et $sign(0) = 0$. Choisissons $x \in \mathbb{R}^N$ défini par $x_j = \mathrm{sign}(a_{i_0,j})$ où i_0 est tel que $\sum_{j=1,\ldots,N} |a_{i_0,j}| \geq \sum_{j=1,\ldots,N} |a_{i,j}|$, $\forall i = 1, \ldots, N$. On a bien $\|x\|_\infty = 1$, et

$$\|Ax\|_\infty = \max_{i=1,\ldots,N} |\sum_{j=1}^N a_{i,j} sgn(a_{i_0,j})|.$$

Or, par choix de x, on a $\sum_{j=1,\ldots,N} |a_{i_0,j}| = max_{i=1,\ldots,N} \sum_{j=1,\ldots,N} |a_{i,j}|$. On en déduit que pour ce choix de x, on a bien $\|Ax\| = \max_{i=1,\ldots,N} |\sum_{j=1,\ldots,N} |a_{i,j}|$.

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

$$\begin{aligned}\|Ax\|_1 &= \sum_{i=1,\dots,N} |\sum_{j=1,\dots,N} a_{i,j}x_j| \leq \sum_{j=1,\dots,N} |x_j \left(|\sum_{i=1,\dots,N} |a_{i,j}| \right) \\ &\leq \max_{j=1,\dots,N} |\sum_{i=1,\dots,N} |a_{i,j}| \sum_{j=1,\dots,N} |x_j|.\end{aligned}$$

Et comme $\sum_{j=1,\dots,N} |x_j| = 1$, 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 $x \in \mathbb{R}^N$, $\|x\|_1 = 1$, tel que $\|Ax\|_1 = \sum_{i=1,\dots,N} |a_{i,j}|$. Il suffit de considérer pour cela le vecteur $x \in \mathbb{R}^N$ 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 $\|Ax\|_1 = \max_{j=1,\dots,N} \sum_{i=1,\dots,N} |a_{i,j}|$.

3. Par définition de la norme 2, on a :

$$\|A\|_2^2 = \sup_{x\in\mathbb{R}^N,\|x\|_2=1} Ax \cdot Ax = \sup_{x\in\mathbb{R}^N,\|x\|_2=1} A^tAx \cdot x.$$

Comme A^tA est une matrice symétrique positive (car $A^tAx \cdot x = Ax \cdot Ax \geq 0$), il existe une base orthonormée $(f_i)_{i=1,\dots,N}$ 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 $Af_i = \mu_i f_i$ pour tout $i \in \{1,\dots,N\}$. Soit $x = \sum_{i=1,\dots,N} \alpha_i f_i \in \mathbb{R}^N$. On a donc :

$$A^tAx \cdot x = \big(\sum_{i=1,\dots,N} \mu_i\alpha_i f_i\big) \cdot \big(\sum_{i=1,\dots,N} \alpha_i f_i\big) = \sum_{i=1,\dots,N} \alpha_i^2\mu_i \leq \mu_N\|x\|_2^2.$$

On en déduit que $\|A\|_2^2 \leq \rho(A^tA)$.
Pour montrer qu'on a égalité, il suffit de considérer le vecteur $x = f_N$; on a en effet $\|f_N\|_2 = 1$, et $\|Af_N\|_2^2 = A^tAf_N \cdot f_N = \mu_N = \rho(A^tA)$.

Exercice 5 page 37 (Rayon spectral)

1. Si $\rho(A) < 1$, grâce au résultat d'approximation du rayon spectral de la proposition 1.6 page 6, il existe $\varepsilon > 0$ tel que $\rho(A) < 1 - 2\varepsilon$ et une norme induite $\|.\|_{A,\varepsilon}$ tels que $\|A\|_{A,\varepsilon} = \mu \leq \rho(A) + \varepsilon = 1 - \varepsilon < 1$. Comme $\|.\|_{A,\varepsilon}$ est une norme matricielle, on a $\|A^k\|_{A,\varepsilon} \leq \mu^k \to 0$ lorsque $k \to \infty$. Comme l'espace $\mathcal{M}_N(\mathbb{R})$ est de dimension finie, toutes les normes sont équivalentes, et on a donc $\|A^k\| \to 0$ lorsque $k \to \infty$.
Montrons maintenant la réciproque : supposons que $A^k \to 0$ lorsque $k \to \infty$, et montrons que $\rho(A) < 1$. Soient λ une valeur propre de A et x un vecteur propre associé. Alors $A^kx = \lambda^k x$, et si $A^k \to 0$, alors $A^kx \to 0$, et donc $\lambda^k x \to 0$, ce qui n'est possible que si $|\lambda| < 1$.

2. Si $\rho(A) < 1$, d'après la question précédente 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, $\limsup_{k\to+\infty} \|A^k\|^{\frac{1}{k}} \leq 1$.

3. Comme $\liminf_{k\to+\infty} \|A^k\|^{1/k} < 1$, il existe une sous-suite $(k_n)_{n\in\mathbb{N}} \subset \mathbb{N}$ telle que $\|A^{k_n}\|^{1/k_n} \to \ell < 1$ lorsque $n \to +\infty$, et donc il existe N tel que pour $n \geq N$, $\|A^{k_n}\|^{1/k_n} \leq \eta$, avec $\eta \in]0,1[$. On en déduit que pour $n \geq N$, $\|A^{k_n}\| \leq \eta^{k_n}$, et donc que $A^{k_n} \to 0$ lorsque $n \to +\infty$. Soient λ une valeur propre de A et x un vecteur propre associé, on a : $A^{k_n}x = \lambda^{k_n}x$; on en déduit que $|\lambda| < 1$, et donc que $\rho(A) < 1$.

4. Soit $\alpha \in \mathbb{R}_+$ tel que $\rho(A) < \alpha$. Alors $\rho(\frac{1}{\alpha}A) < 1$, et donc par la question 2,

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

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

$$\limsup_{k\to+\infty} \|A^k\|^{\frac{1}{k}} \leq \rho(A). \tag{1.8.64}$$

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 par la question 3, $\rho(\frac{1}{\beta}A) < 1$, 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

$$\rho(A) \leq \liminf_{k\to+\infty} \|A^k\|^{\frac{1}{k}}. \tag{1.8.65}$$

De (1.8.64) et (1.8.65), on déduit que

$$\limsup_{k\to+\infty} \|A^k\|^{\frac{1}{k}} = \liminf_{k\to+\infty} \|A^k\|^{\frac{1}{k}} = \lim_{k\to+\infty} \|A^k\|^{\frac{1}{k}} = \rho(A). \tag{1.8.66}$$

5. Si $\|.\|$ est une norme matricielle, alors $\|A^k\| \leq \|A\|^k$ et donc d'après la question précédente, $\rho(A) \leq \|A\|$.

Exercice 7 page 37 (Rayon spectral)

Il suffit de prendre comme norme la norme définie par : $\|x\| = \sum_{i=1}^N \alpha_i^2$ 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 9 page 37 (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. Suppposons que $\rho(A) < 1$. Il est facile de remarquer que

$$(\sum_{k=0}^{N} A^k)(Id - A) = Id - A^{N+1}. \tag{1.8.67}$$

Si $\rho(A) < 1$, d'après la question 1. de l'exercice 5 page 37, on a $A^k \to 0$ lorsque $k \to 0$. De plus, $Id - A$ est inversible. On peut donc passer à la limite dans (1.8.67) et on a donc $(Id - A)^{-1} = \sum_{k=0}^{+\infty} A^k$.
Remarquons de plus que la série de terme général A^k est absolument convergente pour une norme $\|\cdot\|_{A,\epsilon}$ donnée par la proposition 1.6 page 6, avec ε choisi tel que $\rho(A) + \varepsilon < 1$. Par contre, la série n'est pas absolument convergente pour n'importe quelle norme. On pourra s'en convaincre facilement grâce au contre–exemple (en dimension 1) suivant : la série $s_k = 1 + x + \cdots + x^k$ est absolument convergente pour la norme $|\cdot|$ sur $\mathbb{R}$ pour $|x| < 1$, ce qui n'est évidemment plus le cas si l'on remplace la norme par la norme (pourtant équivalente) $\|\cdot\| = 10|\cdot|$.
Réciproquement, si $\rho(A) \geq 1$, la série ne peut pas converger en raison du résultat de la question 1 de l'exercice 5 page 37.

Exercice 11 page 38 (Décompositions LL^t et LDL^t)

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

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

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

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

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

$$A = \left[\begin{array}{c|c} B & a \\ \hline a^t & \alpha \end{array}\right] \tag{1.8.68}$$

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 :

$$L = \left[\begin{array}{c|c} M & 0 \\ \hline b^t & 1 \end{array}\right], \; D = \left[\begin{array}{c|c} \tilde{D} & 0 \\ \hline 0 & \lambda \end{array}\right], \tag{1.8.69}$$

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 (1.8.69) 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 $\lambda = \alpha - b^t M^{-1}a$. 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 :

$$a_{i,j} = \sum_{k=1}^{N} \ell_{i,k} d_{k,k} \ell_{j,k}, \quad \forall\, (i,j) \in \{1, \ldots, N\}^2. \tag{1.8.70}$$

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

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

2. On suppose avoir calculé les n premières colonnes de L. On calcule la colonne $(n+1)$ en prenant $j = n+1$ dans (1.3.13)

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

$$d_{n+1,n+1} = a_{n+1,n+1} - \sum_{k=1}^{n} \ell_{n+1,k}^2 d_{k,k}. \tag{1.8.71}$$

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 :

$$\ell_{i,n+1} = \left(a_{i,n+1} - \sum_{k=1}^{n} \ell_{i,k} d_{k,k} \ell_{n+1,k} \right) \frac{1}{d_{n+1,n+1}}. \tag{1.8.72}$$

Exercice 13 page 38 (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 15 page 39 (Décomposition LL^t d'une matrice bande)

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

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

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

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

On en déduit que le nombre d'opérations élémentaires pour le calcul de la colonne $n+1$, avec $1\leq n<N-1$, est de 4.

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

2. Cas d'une matrice à p diagonales.

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

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

$$\ell_{n+1,n+1}=(a_{n+1,n+1}-\sum_{k=1}^{n}\ell_{n+1,k}\ell_{n+1,k})^{1/2}>0, \tag{1.8.73}$$

$$\text{et } \ell_{i,n+1} = \left(a_{i,n+1} - \sum_{k=1}^{n} \ell_{i,k}\ell_{n+1,k} \right) \frac{1}{\ell_{n+1,n+1}}, \tag{1.8.74}$$

est toujours inférieur à $2q+1$, car la somme $\sum_{k=1}^{n}$ fait intervenir au plus q termes non nuls.
De plus, pour chaque colonne $n+1$, il y a au plus $q+1$ coefficients $\ell_{i,n+1}$ non nuls, donc au plus $q+1$ coefficients à calculer. Donc le nombre d'opérations pour chaque colonne peut être majoré par $(2q+1)(q+1)$.
On peut donc majorer le nombre d'opérations z_q pour les q premières colonnes et les q dernières par $2q(2q+1)(q+1)$, qui est indépendant de N (on rappelle qu'on cherche une estimation en fonction de N, et donc le nombre z_q est $O(1)$ par rapport à N.)
Calculons maintenant le nombre d'opérations x_n nécessaires une colonne $n = q+1$ à $N-q-1$. Dans (1.8.73) et (1.8.74), les termes non nuls de la somme sont pour $k = i-q, \ldots, n$, et donc on a $(n-i+q+1)$ multiplications et additions, une division ou extraction de racine. On a donc

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

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

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

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

$$(q+1)^2(N-2q) \le Z_p(N) \le (q+1)^2(N-2q) + 2q(2q+1)(q+1), \tag{1.8.75}$$

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

3. Dans le cas de la discrétisation de l'équation $-u'' = f$ traitée dans le cours page 18, 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 $-\Delta u = f$ traitée dans le cours page 25-26, on a $q = \sqrt{N}$ et la méthode de Choleski nécessite de l'ordre de N^2 opérations élémentaires (dans les deux cas N est le nombre d'inconnues).

On peut noter que l'encadrement (1.8.75) est intéressant dès que q est d'ordre inférieur à N^α, $\alpha < 1$.

Exercice 17 page 40 (propriétés générales du conditionnement)

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

$$\|Id\| \le \|A\| \ \|A^{-1}\|$$

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

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

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

car $\|\cdot\|$ est une norme matricielle. Donc $\mathrm{cond}(AB) \leq \mathrm{cond}(A)\mathrm{cond}(B)$.

Partie II

1. Par définition, on a $\mathrm{cond}_2(A) = \|A\|_2\|A^{-1}\|_2$. Or on a vu à l'exercice 3 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. Or 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 $cond_2(A) = \dfrac{\lambda_N}{\lambda_1}$.

3. 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 $(f_1 \dots f_N)$ telle que $A^tAf_i = \sigma f_i,\ \forall i$ et $\sigma > 0$ (car A^tA est s.d.p.). On a donc $A^tA = \sigma Id\ A^t = \alpha^2A^{-1}$ 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 $A^tA = \alpha^2Id$, $\frac{\sigma_N}{\sigma_1} = 1$, et donc $cond_2(A) = 1$.

4. $A \in \mathcal{M}_N(\mathbb{R})$ est une matrice inversible. On suppose que $A = QR$ où Q est une matrice orthogonale. On a donc :

$$cond_2(A) = \|A\|_2\ \|A^{-1}\|_2 = \|QR\|_2\ \|R^{-1}Q^t\|_2.$$

On a aussi $cond_2(A) = \sqrt{\dfrac{\sigma_N}{\sigma_1}}$ 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 $cond_2(A) = cond_2(R)$.

5. Soient $0 < \lambda_1 \leq \lambda_2 \ldots \leq \lambda_N$ et $0 < \mu_1 \leq \mu_2 \ldots \leq \mu_N$ les valeurs propres de A et B (qui sont s.d.p.). 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

$$cond_2(A+B) \leq \frac{\lambda_N + \mu_N}{\lambda_1 + \mu_1}.$$

Remarquons en premier lieu que si A est s.d.p., alors

$$cond_2(A) = \frac{\sup_{\|x\|=1} Ax \cdot x}{\inf_{\|x\|_2=1} Ax \cdot x}$$

En effet, si A est s.d.p., alors $\sup_{\|x\|_2=1} Ax \cdot x = \lambda_N$; il suffit pour s'en rendre compte de décomposer x sur la base $(f_i)_{i=1\ldots N}$. Soit $x = \sum_{i=1}^{N} \alpha_i f_i$. Alors : $Ax \cdot x = \sum_{i=1}^{N} \alpha_i^2 \lambda_i \leq \lambda_N \Sigma \alpha_i^2 = \lambda_N$. Et $Af_N \cdot f_N = \lambda_N$.

De même, $Ax \cdot x \geq \lambda_1 \Sigma \alpha_i^2 = \lambda_1$ et $Ax \cdot x = \lambda_1$ si $x = f_1$. Donc $\inf_{\|x\|=1} Ax \cdot x = \lambda_1$.

On en déduit que si A est s.d.p., $cond_2(A) = \dfrac{\sup_{\|x\|=1} Ax \cdot x}{\inf_{\|x\|=1} Ax \cdot x}$

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

$$\text{Or } \sup_{\|x\|=1} (Ax \cdot x + Bx \cdot x) \leq \sup_{\|x\|=1} Ax \cdot x + \sup_{\|x\|=1} Bx \cdot x = \lambda_N + \mu_N$$

$$\text{et } \inf_{\|x\|=1} (Ax \cdot x + Bx \cdot x) \geq \inf_{\|x\|=1} Ax \cdot x + \inf_{\|x\|=1} Bx \cdot x = \lambda_1 + \mu_1$$

donc

$$cond_2(A+B) \leq \frac{\lambda_N + \mu_N}{\lambda_1 + \mu_1}.$$

b) On va montrer que

$$\frac{a+b}{c+d} \leq \max(\frac{a}{c}, \frac{b}{d}),\ \forall (a,b,c,d) \in (\mathbb{R}_+^*)^4.$$

Supposons que $\dfrac{a+b}{c+d} \geq \dfrac{a}{c}$ 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 19 page 40 (Minoration du conditionnement)

1. Comme A est inversible, $A + \delta A = A(Id + A^{-1}\delta A)$, et donc si $A + \delta A$ est singulière, alors $Id + A^{-1}\delta A$ est singulière. Or on a vu en cours que toute matrice de la forme $Id + B$ est inversible si et seulement si $\rho(B) < 1$. On en déduit que $\rho(A^{-1}\delta A) \geq 1$, et comme

$$\rho(A^{-1}\delta A) \leq \|A^{-1}\delta A\| \leq \|A^{-1}\|\|\delta A\|,$$

on obtient

$$\|A^{-1}\|\|\delta A\| \geq 1, \text{ soit encore } \mathrm{cond}(A) \geq \frac{\|A\|}{\|\delta A\|}.$$

2. Soit $y \in \mathbb{R}^N$ tel que $\|y\| = 1$ et $\|A^{-1}y\| = \|A^{-1}\|$. Soit $x = A^{-1}y$, et $\delta A = \frac{-y\ x^t}{x^t\ x}$,on a donc

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

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 (1.6.42).

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

$$\mathrm{cond}(A) \geq \frac{\|A\|}{\|\delta A\|}. \tag{1.8.76}$$

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 21 page 41 (Calcul de l'inverse d'une matrice et conditionnement)

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

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

Les quantités e_1, e_2, e_3 et e_4 sont les erreurs relatives commises sur ces quatre équations lorsqu'on remplace X par B ; en ce sens, elles mesurent la qualité de l'approximation de A^{-1}.

(b) On remarque d'abord que comme la norme est matricielle, on a $\|MP\| \leq \|M\|\|P\|$ pour toutes matrices M et P de $\mathcal{M}_N(\mathbb{R})$. On va se servir de cette propriété plusieurs fois par la suite.

(α) Comme $B = A^{-1} + E$, on a

$$e_1 = \frac{\|E\|}{\|A^{-1}\|} \leq \varepsilon\frac{\|A^{-1}\|}{\|A^{-1}\|} = \varepsilon.$$

(β) Par définition,

$$e_2 = \frac{\|B^{-1} - A\|}{\|A\|} = \frac{\|(A^{-1} + E)^{-1} - A\|}{\|A\|}.$$

Or

$$\begin{aligned} (A^{-1} + E)^{-1} - A &= (A^{-1}(Id + AE))^{-1} - A \\ &= (Id + AE)^{-1}A - A \\ &= (Id + AE)^{-1}(Id - (Id + AE))A \\ &= -(Id + AE)^{-1}AEA. \end{aligned}$$

On a donc

$$e_2 \leq \|(Id + AE)^{-1}\|\|A\|\|E\|.$$

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

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

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

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

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

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

(β) Par définition,

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

car $\varepsilon < 1$ (théorème 1.1).

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

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

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

Exercice 23 page 42 (IP-matrice)

1. Supposons d'abord que A est inversible et que $A^{-1} \geq 0$; soit $x \in \mathbb{R}^n$ tel que $b = Ax \geq 0$. On a donc $x = A^{-1}b$, et comme tous les coefficients de A^{-1} et de b sont positifs ou nuls, on a bien $x \geq 0$.
   Réciproquement, si A est une IP-matrice, alors $Ax = 0$ entraine $x = 0$ ce qui montre que A est inversible. Soit e_i le i-ème vecteur de la base canonique de $\mathbb{R}^n$, on a : $AA^{-1}e_i = e_i \geq 0$, et donc par la propriété de IP-matrice, $A^{-1}e_i \geq 0$, ce qui montre que tous les coefficients de A^{-1} sont positifs.

2. La matrice inverse de A est $A^{-1} = \frac{1}{\Delta}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$ avec $\Delta = ad - bc$. Les coefficients de A^{-1} sont donc positifs ou nuls si et seulement si

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

   Or on a forcément $ad \neq 0$: en effet sinon on aurait dans le premier cas) $bc < 0$, or $b \leq 0$ et $c \leq 0$, ce qui aboutit à une contradiction. De même dans le deuxième cas, on aurait $bc > 0$, or $b \geq 0$ et $c \geq 0$. Les conditions précédentes sont donc équivalentes aux conditions (1.6.44).

3. La matrice A^t est une IP-matrice si et seulement A^t est inversible et $(A^t)^{-1} \geq 0$. Or $(A^t)^{-1} = (A^{-1})^t$. D'où l'équivalence.

4. Supposons que A vérifie (1.6.45), et soit $x \in \mathbb{R}^N$ tel que $Ax \geq 0$. Soit $k \in 1, \ldots, N$ tel que $x_k = \min\{x_i, i = 1, \ldots, N\}$. Alors

$$(Ax)_k = a_{k,k}x_k + \sum_{\substack{j=1 \\ j\neq k}}^{N} a_{k,j}x_j \geq 0.$$

Par hypothèse, $a_{k,j} \leq 0$ pour $k \neq j$, et donc $a_{k,j} = -|a_{k,j}|$. On peut donc écrire :

$$a_{k,k}x_k - \sum_{\substack{j=1 \\ j\neq k}}^{N} |a_{k,j}|x_j \geq 0,$$

et donc :

$$(a_{k,k} - \sum_{\substack{j=1 \\ j\neq k}}^{N} |a_{k,j}|)x_k \geq \sum_{\substack{j=1 \\ j\neq k}}^{N} |a_{k,j}|(x_j - x_k).$$

Comme $x_k = \min\{x_i, i = 1, \ldots, N\}$, on en déduit que le second membre de cette inégalité est positif ou nul, et donc que $x_k \geq 0$. On a donc $x \geq 0$.

5. Si la matrice A vérifie (1.6.46), alors la matrice A^t vérifie (1.6.45). On en déduit par les questions précédentes que A^t et A sont des IP-matrices.
6. Soit $\mathbf{1}$ le vecteur de $\mathbb{R}^N$ dont toutes les composantes sont égales à 1. Si $Ax > 0$, comme l'espace $\mathbb{R}^N$ est de dimension finie, il existe $\epsilon > 0$ tel que $Ax \geq \epsilon\mathbf{1}$. Soit $\eta = \epsilon A^{-1}\mathbf{1} \geq 0$; on a alors $A(x - \eta) \geq 0$ et donc $x \geq \eta$, car A est une IP-matrice.

   Montrons maintenant que $\eta > 0$: 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 $\eta_i = \epsilon \sum_{i=1}^{N} (A^{-1})_{i,j} > 0$ pour tout $i = 1, \ldots, N$. On a donc bien $x \geq \eta > 0$.
7. Soit A la matrice nulle, on a alors $\{x \in \mathbb{R}^N \text{ t.q. } Ax > 0\} = \emptyset$, et donc $\{x \in \mathbb{R}^N \text{ t.q. } Ax > 0\} \subset \{x \in \mathbb{R}^N \text{ t.q. } x > 0\}$. Pourtant A n'est pas inversible, et n'est donc pas une IP-matrice.
8. Soit x tel que $Ax \geq 0$, alors il existe $\varepsilon \geq 0$ tel que $Ax + \varepsilon\mathbf{1} \geq 0$. Soit maintenant $b = A^{-1}\mathbf{1}$; on a $A(x + \varepsilon b) > 0$ et donc $x + \varepsilon b > 0$. En faisant tendre ε vers 0, on en déduit que $x \geq 0$.
9. Soit $T \in \mathcal{L}(E)$ défini par $f \in E \mapsto Tf$, avec $Tf(x) = f(\frac{1}{x})$ si $x \neq 0$ et $f(0) = \ell$, avec $\ell = \lim_{\pm\infty} f$. On vérifie facilement que $Tf \in E$. Si $Tf \geq 0$, alors $f(\frac{1}{x}) \geq 0$ pour tout $x \in \mathbb{R}$; donc $f(x) \geq 0$ pour tout $x \in \mathbb{R} \setminus \{0\}$; on en déduit que $f(0) \geq 0$ par continuité. On a donc bien $f \geq 0$.

   Soit maintenant g définie de $\mathbb{R}$ dans $\mathbb{R}$ par $g(x) = |\arctan x|$. On a $g(0) = 0$, donc $g \not> 0$. Or $Tg(0) = \frac{\pi}{2}$ et $Tg(x) = |\arctan \frac{1}{x}| > 0$ si $x > 0$, donc $Tg > 0$.

Exercice 25 page 44 (Valeurs propres et vecteurs propres de A.)

1. Pour montrer que A est définie positive (car A est évidemment symétrique), on va montrer que $Ax \cdot x > 0$ si $x \neq 0$.
On a

$$Ax \cdot x = \frac{1}{h^2}\left[x_1(2x_1 - x_2) + \sum_{i=2}^{N-1} x_i(-x_{i-1} + 2x_i - x_{i+1}) + 2x_N^2 - x_{N-1}x_N\right]$$

On a donc

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

De plus, $Ax \cdot x = 0 \Rightarrow x_1^2 = x_N = 0$ et $x_i = x_{i-1}$ pour $i = 2$ à N, donc $x = 0$.
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

$$\begin{cases} -\varphi''(x) = \lambda\varphi(x) & x \in]0,1[\\ \varphi(0) = \varphi(1) = 0 \end{cases} \tag{1.8.77}$$

(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 (1.8.77) sont donc de la forme $(k^2\pi^2, \sin k\pi x)$.
2. Pour $k = 1$ à N, posons $\Phi_i^{(k)} = \sin k\pi x_i$, où $x_i = ih$, pour $i = 1$ à N, et calculons $A\Phi^{(k)}$:

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

En utilisant le fait que $\sin(a+b) = \sin a\cos b + \cos a\sin b$ pour développer $\sin k\pi(1-i)h$ et $\sin k\pi(i+1)h$, on obtient (après calculs) :

$$(A\Phi^{(k)})_i = \lambda_k\Phi_i^{(k)}, \quad i = 1, \ldots, N,$$

où $\lambda_k = \dfrac{2}{h^2}(1 - \cos k\pi h) = \dfrac{2}{h^2}(1 - \cos\dfrac{k\pi}{N+1})$
On a donc trouvé N valeurs propres $\lambda_1 \ldots \lambda_N$ associées aux vecteurs propres $\Phi^{(1)} \ldots \Phi^{(N)}$ de $\mathbb{R}^N$ tels que $\Phi_i^{(k)} = \sin\dfrac{k\pi i}{N+1}$, $i = 1 \ldots N$.
Remarque : Lorsque $N \to +\infty$ (ou $h \to 0$), on a

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

Donc

$$\begin{array}{ll} \lambda_k^{(h)} & \to k^2\pi^2 = \lambda_k \\ & h \to 0. \end{array}$$

Calculons $cond_2(A)$. Comme A est s.d.p., on a $cond_2(A) = \dfrac{\lambda_N}{\lambda_1} = \dfrac{1 - \cos\frac{N\pi}{N+1}}{1 - \cos\frac{\pi}{N+1}}$.
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 $h^2 cond_2(A) \to \frac{4}{\pi^2}$ lorsque $h \to 0$.

Exercice 27 page 45 (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 $p \in \{0, \ldots, N+1\}$ tel que $u_p = \min(u_i, \ i = 0, \ldots, N+1)$.
Si $p = 0$ ou $N+1$, alors $u_i \geq 0 \ \forall i = 0, N+1$ et donc $u \geq 0$.
Si $p \in \{1, \ldots, N\}$, alors

$$\frac{1}{h^2}(u_p - u_{p-1}) + \frac{1}{h^2}(u_p - u_{p+1}) \geq 0$$

et comme $u_p - u_{p-1} < 0$ et $u_p - u_{p+1} \leq 0$, on aboutit 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 $\varphi(x) = 1/2x(1-x)$ et $\phi_i = \varphi(x_i)$, $i = 1, N$, où $x_i = ih$.
$(A\phi)_i$ est le développement de Taylor à l'ordre 2 de $\varphi''(x_i)$, et comme φ est un polynôme de degré 2, ce développement est exact. Donc $(A\phi)_i = \varphi''(x_i) = 1$.

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 $-\|b\|\phi_i \leq \|b\|\phi_i$.
On en déduit que $\|u\|_\infty \leq \|b\| \ \|\phi\|_\infty$; or $\|\phi\|_\infty = \dfrac{1}{8}$. D'où $\|u\|_\infty \leq \dfrac{1}{8}\|b\|$.
On peut alors écrire que pour tout $b \in \mathbb{R}^N$,

$$\|A^{-1}b\|_\infty \leq \frac{1}{8}\|b\|, \text{ donc } \frac{\|A^{-1}b\|_\infty}{\|b\|_\infty} \leq \frac{1}{8}, \text{ d'où } \|A^{-1}\| \leq \frac{1}{8}.$$

On montre que $\|A^{-1}\| = \dfrac{1}{8}$ en prenant le vecteur 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 $\|\varphi\|_\infty = \varphi(\frac{1}{2}) = \frac{1}{8}$.

4. Par définition, on a $\|A\| = \sup_{\|x\|_\infty = 1} \|Ax\|$, 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 $\varphi^{(h)}$ 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 $h \to 0$. On a donc

$$h \sum_{i=1}^{N} b_i \varphi_i = \int_0^1 f^{(h)}(x) \varphi^{(h)}(x) dx \to \int_0^1 f(x)\varphi(x)dx \text{ lorsque } h \to 0.$$

Comme $b_i > 0$ et $f_i > 0\ \forall i = 1, \ldots, N$, 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 : $\dfrac{\|\delta_u\|}{\|u\|} \leq \dfrac{1}{8}\dfrac{hN}{\alpha}\|\delta_b\|\dfrac{\|f\|_\infty}{\|b\|}$. Or $hN = 1$ et on a donc bien :

$$\frac{\|\delta_u\|}{\|u\|} \leq \frac{\|f\|_\infty}{8\alpha}\frac{\|\delta_b\|}{\|b\|}.$$

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.

Exercice 29 page 46 (Une matrice cyclique)

1. On remarque que A s'écrit sous la forme $A = \alpha Id + K$, où

$$K = \begin{pmatrix} 0 & -1 & 0 & -1 \\ -1 & 0 & -1 & 0 \\ 0 & -1 & 0 & -1 \\ -1 & 0 & -1 & 0 \end{pmatrix}$$

Soit λ une valeur propre de A et x un vecteur propre associé. On a $Ax = \alpha x + Kx$. Les valeurs propres de A s'écrivent donc sous la forme $\lambda = \alpha + \mu$ où μ est valeur propre de K. Or K est une matrice qui comporte 2 fois 2 lignes identiques. La matrice K admet donc 0 comme valeur propre double. De plus la somme des colonnes de K

est égale au vecteur constant de composantes -2, donc -2 est valeur propre de K (pour le vecteur propre constant de composantes égales à 1). Enfin, la dernière valeur propre de K est 2 car la somme des valeurs propres est égale à la trace. On en déduit que les valeurs propres de A sont $\alpha+2$, $\alpha-2$ et α.

2. La matrice A est sdp si et seulement si toutes ses valeurs propres sont strictement positives, i.e. si $\alpha > 2$. Elle est singulière dès qu'une de ses valeurs propres est nulle, c.à.d. pour $\alpha = -2$, 0 et 2.

3. La méthode de Jacobi s'écrit :

$$\begin{aligned} x_1^{(n+1)} &= \frac{1}{\alpha}(b_1 + x_2^{(n)} + x_4^{(n)}), \\ x_2^{(n+1)} &= \frac{1}{\alpha}(b_2 + x_1^{(n)} + x_3^{(n)}), \\ x_3^{(n+1)} &= \frac{1}{\alpha}(b_3 + x_2^{(n)} + x_4^{(n)}), \\ x_4^{(n+1)} &= \frac{1}{\alpha}(b_4 + x_1^{(n)} + x_3^{(n)}), \end{aligned}$$

La matrice d'itération est

$$B = \frac{1}{\alpha}\begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix}$$

dont les valeurs propres sont 0 (double) $\frac{2}{\alpha}$ et $\frac{-2}{\alpha}$. On en déduit que la méthode de Jacobi converge si $\alpha > 2$.

4. L'algorithme de GS pour le système $Ax = b$ s'écrit :

$$\begin{aligned} \alpha x_1^{(n+1)} &= b_1 + x_2^{(n)} + x_4^{(n)}, \\ -x_1^{(n+1)} + \alpha x_2^{(n+1)} &= b_2 + x_3^{(n)}, \\ -x_2^{(n+1)} + \alpha x_3^{(n+1)} &= b_3 + x_4^{(n)}, \\ -x_1^{(n+1)} - x_3^{(n+1)} + \alpha x_4^{(n+1)} &= b_4. \end{aligned}$$

On obtient donc, avec $\beta = \frac{1}{\alpha}$:

$$\begin{aligned} x_1^{(n+1)} &= \beta(b_1 + x_2^{(n)} + x_4^{(n)}), \\ x_2^{(n+1)} &= \beta(b_2 + x_1^{(n+1)} + x_3^{(n)}), \\ x_3^{(n+1)} &= \beta(b_3 + x_2^{(n+1)} + x_4^{(n)}), \\ x_4^{(n+1)} &= \beta(b_4 + x_1^{(n+1)} + x_3^{(n+1)}). \end{aligned}$$

soit encore :

$$\begin{aligned} x_1^{(n+1)} &= \beta\left(b_1 + x_2^{(n)} + x_4^{(n)}\right), \\ x_2^{(n+1)} &= \beta\left(b_2 + \beta\left(b_1 + x_2^{(n)} + x_4^{(n)}\right) + x_3^{(n)}\right), \\ x_3^{(n+1)} &= \beta\left(b_3 + \beta\left(b_2 + \beta(b_1 + x_2^{(n)} + x_4^{(n)}) + x_3^{(n)}\right) + x_4^{(n)}\right), \\ x_4^{(n+1)} &= \beta\left(b_4 + \beta(b_1 + x_2^{(n)} + x_4^{(n)}) + \beta\left(b_3 + \beta\left(b_2 + \beta(b_1 + x_2^{(n)} + x_4^{(n)}) + x_3^{(n)}\right)) + x_4^{(n)}\right)\right). \end{aligned}$$

On obtient donc :

$$\begin{aligned}
x_1^{(n+1)} &= \beta\left(b_1 + x_2^{(n)} + x_4^{(n)}\right),\\
x_2^{(n+1)} &= \beta\left(b_2 + \beta b_1 + \beta x_2^{(n)} + x_3^{(n)} + \beta x_4^{(n)}\right),\\
x_3^{(n+1)} &= \beta\left(b_3 + \beta b_2 + \beta^2 b_1 + \beta^2 x_2^{(n)} + \beta x_3^{(n)} + (1+\beta^2) x_4^{(n)}\right),\\
x_4^{(n+1)} &= \beta\left(b_4 + \beta(1+\beta^2) b_1 + \beta^2 b_2 + \beta b_3 + \beta(1+\beta^2) x_2^{(n)} + \beta^2 x_3^{(n)} + (2+\beta^2) x_4^{(n)}\right)
\end{aligned}$$

Pour $\alpha > 2$, la matrice est symétrique définie positive et la méthode de Gauss–Seidel converge.

Exercice 31 page 47 (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 : Avec le formalisme de la méthode II du cours, on a

$$M = D = \begin{bmatrix} a_{1,1} & & 0 \\ & \ddots & \\ 0 & & a_{N,N} \end{bmatrix}, \text{ et } N = M - A.$$

La matrice d'itération est

$$\begin{aligned}
J = M^{-1}N = D^{-1}N &= \begin{bmatrix} a_{1,1}^{-1} & & 0 \\ & \ddots & \\ 0 & & a_{N,N}^{-1} \end{bmatrix} \begin{bmatrix} 0 & & -a_{i,j} \\ & \ddots & \\ -a_{i,j} & & 0 \end{bmatrix}\\
&= \begin{bmatrix} 0 & -\frac{a_{1,2}}{a_{1,1}} & \cdots \\ & \ddots & \\ -\frac{a_{1,1}}{a_{N,N}} & \cdots & 0 \end{bmatrix}.
\end{aligned}$$

Cherchons le rayon spectral de J : soient $x \in \mathbb{R}^N$ et $\lambda \in \mathbb{R}$ tels que $Jx = \lambda x$, 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 $|\lambda| \leq \frac{\sum_{j;i\neq j} |a_{i,j}|}{|a_{ii}|} < 1$ pour toute valeur propre λ. On a donc $\rho(J) < 1$. Donc la méthode de Jacobi converge.

Exercice 33 page 47 (Jacobi pour les matrices à diagonale dominante forte)

1. Si A est symétrique définie positive alors $Ae_i \cdot e_i > 0$ pour tout vecteur e_i de la base canonique. Tous les coefficients diagonaux sont donc strictement positifs, et donc aussi inversibles. On en déduit que la matrice D est inversible et que la méthode de Jacobi est bien définie.
2. (a) Soit λ une valeur propre de J associée au vecteur propre x. On a donc $Jx = \lambda x$, c'est-à-dire $D^{-1}(E+F)x = \lambda x$, soit encore $(E+F)x = \lambda Dx$. On a donc

$$|\sum_{j\neq i} a_{i,j}x_j| = |\lambda||a_{i,i}||x_i|.$$

Soit i tel que $|x_i| = \max_{j=1,N}|x_j|$. Notons que $|x_i| \neq 0$ car x est vecteur propre, donc non nul. En divisant l'égalité précédente par $|x_i|$, on obtient :

$$|\lambda| \leq \sum_{j\neq i} \frac{|a_{i,j}|}{|a_{i,i}|} \leq 1,$$

par hypothèse. On en déduit que $\rho(J) \leq 1$.

(b) Soit $x \in \mathbb{R}^N$ tel que $Jx = \lambda x$ avec $|\lambda| = 1$. Alors

$$|\sum_{j\neq i} a_{i,j}x_j| = |a_{i,i}||x_i|, \text{ pour tout } i = 1,\ldots,N. \tag{1.8.78}$$

On a donc

$$\begin{aligned} \sum_{j\neq i}|a_{i,j}||x_i| &\leq |a_{i,i}||x_i| = |\sum_{j\neq i} a_{i,j}x_j| \\ &\leq \sum_{j\neq i}|a_{i,j}||x_j| \text{ pour tout } i = 1,\ldots,N. \end{aligned} \tag{1.8.79}$$

Si A est diagonale, alors en vertu de (1.8.78), $x_i = 0$ pour tout $i = 1,\ldots,N$. Supposons maintenant A non diagonale. On déduit alors de (1.8.79) que

$$\frac{|x_i|}{|x_j|} \leq 1 \text{ pour tout } i \neq j.$$

Donc $|x_i| = |x_j|$ pour tout i, j.

Comme de plus, par hypothèse,

$$|a_{i_0,i_0}| > \sum_{j\neq i_0}|a_{i_0,j}|,$$

on a donc, si $|x_{i_0}| \neq 0$,

$$\sum_{j\neq i_0}|a_{i_0,j}||x_{i_0}| < |a_{i_0,i_0}x_{i_0}| \leq \sum_{j\neq i_0}|a_{i_0,j}x_{i_0}|,$$

ce qui est impossible. On en déduit que $x = 0$.

On a ainsi prouvé que J n'admet pas de valeur propre de module égal à 1, et donc par la question précédente, $\rho(J) < 1$, ce qui prouve que la méthode converge.

(c) La matrice A de l'exercice 32 est à diagonale fortement dominante. Donc la méthode de Jacobi converge.

Exercice 35 page 48 (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 précédent pour l'application linéaire T de matrice D, qui est, par hypothèse, définie positive (et évidemment symétrique puisque diagonale) et $S = E + F$, symétrique car A est symétrique.
   Il existe donc $(f_1 \dots f_N)$ base de E et $(\mu_1 \dots \mu_N) \in \mathbb{R}^N$ tels que

   $$Jf_i = D^{-1}(E+F)f_i = \mu_i f_i, \quad \forall i = 1, \dots, N, \text{ et } (Df_i, f_j) = \delta_{ij}.$$

3. Par définition de J, tous les éléments diagonaux de J sont nuls et donc sa trace également. Or $TrJ = \sum_{i=1}^{N} \mu_i$.

   Si $\mu_i > 0 \ \forall i = 1, \dots, N$, alors $TrJ > 0$, donc $\exists i_0; \ \mu_i \leq 0$ et comme $\mu_1 \leq \mu_{i_0}$, on a $\mu_1 \leq 0$. Un raisonnement similaire montre que $\mu_N \geq 0$.

4. La méthode de Jacobi converge si et seulement si $\rho(J) < 1$ (théorème 1.26 page 28). Or, par la question précédente, $\rho(A) = max(-\mu_1, \mu_N)$. Supposons que $\mu_1 \leq -1$, alors $\mu_1 = -\alpha$, avec $\alpha \geq 1$. On a alors $D^{-1}(E+F)f_1 = -\alpha f_1$ ou encore $(E+F)f_1 = -\alpha D f_1$, ce qui s'écrit aussi $(D+E+F)f_1 = D(1-\alpha)f_1$ c'est–à–dire $(2D-A)f_1 = \beta D f_1$ avec $\beta \leq 0$. On en déduit que $((2D-A)f_1, f_1) = \beta \leq 0$, ce qui contredit le fait que $2D - A$ est définie positive. En conséquence, on a bien $\mu_1 \geq -1$.

   Supposons maintenant que $\mu_N = \alpha \geq 1$. On a alors $D^{-1}(E+F)f_1 = -\alpha f_N$, soit encore $(E+F)f_N = -\alpha D f_N$. On en déduit que $Af_N = (D-E-F)f_N = D(1-\alpha)f_N = D\beta f_N$ avec $\beta \leq 0$. On a alors$(Af_N, f_N) \leq 0$, ce qui contredit le fait que A est définie positive.

5. Par définition, on a :$D\tilde{x}^{(n+1)}s = (E+F)x^{(n)} + b$ et $x^{(n+1)} = \omega\tilde{x}^{(n+1)} + (1-\omega)x^{(n)}$. On a donc $x^{(n+1)} = \omega[D^{-1}(E+F)x^{(n)} + D^{-1}b] + (1-\omega)x^{(n)}$ c'est–à–dire $x^{(n+1)} = [Id - \omega(Id - D^{-1}(E+F))]x^{(n)} + \omega D^{-1}b$,, soit encore $\frac{1}{\omega}Dx^{(n+1)} = [\frac{1}{\omega}D - (D-(E+F))]x^{(n)} + b$. On en déduit que $M_\omega x^{(n+1)} = N_\omega x^{(n)} + b$ avec $M_\omega = \frac{1}{\omega}D$ et $N_\omega = \frac{1}{\omega}D - A$.

6. La matrice d'itération est donc maintenant $J_\omega = M_\omega^{-1}N_\omega$ qui est symétrique pour le produit scalaire $(\cdot,\cdot)_{M_\omega}$ donc en reprenant le raisonnement de la question 2, il existe une base $(\tilde{f}_1, \dots, \tilde{f}_N) \in (\mathbb{R}^N)^N$ et $(\tilde{\mu}_1, \dots \tilde{\mu}_N) \subset \mathbb{R}^N$ tels que

   $$J_\omega \tilde{f}_i = {M_\omega}^{-1}N_\omega \tilde{f}_i = \omega D^{-1}\left(\frac{1}{\omega}D - A\right)\tilde{f}_i = \tilde{\mu}_i \tilde{f}_i, \quad \forall i = 1, \dots, N,$$

   $$\text{et } \frac{1}{\omega}D\tilde{f}_i \cdot \tilde{f}_j = \delta_{ij}, \quad \forall i, = 1, \dots, N, \ \forall j, = 1, \dots, N.$$

   Supposons $\tilde{\mu}_1 \leq -1$, alors $\tilde{\mu}_1 = -\alpha$, avec $\alpha \geq 1$ et $\omega D^{-1}(\frac{1}{\omega}D - A)\tilde{f}_1 = -\alpha\tilde{f}_1$, ou encore $\frac{1}{\omega}D - A\tilde{f}_1 = -\alpha\frac{1}{\omega}D\tilde{f}_1$. On a donc $\frac{2}{\omega}D - A\tilde{f}_1 = (1-\alpha)\frac{1}{\omega}D\tilde{f}_1$, ce qui entraîne $(\frac{2}{\omega}D - A)\tilde{f}_1 \cdot \tilde{f}_1 \leq 0$. Ceci contredit l'hypothèse $\frac{2}{\omega}D - A$ 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

$$\left(\frac{2}{\omega}D - A\right) x \cdot x > 0, \quad \forall x \neq 0, \tag{1.8.80}$$

ce qui est équivalent à

$$\left(\frac{2}{\omega}D - A\right) f_i \cdot f_i > 0, \quad \forall i = 1, \ldots, N, \tag{1.8.81}$$

où les $(f_i)_{i=1,N}$ sont les vecteurs propres de $D^{-1}(E+F)$. En effet, la famille $(f_i)_{i=1,\ldots,N}$ est une base de $\mathbb{R}^N$, et

$$\begin{aligned} \left(\frac{2}{\omega}D - A\right) f_i &= \left(\frac{2}{\omega}D - D + (E+F)\right) f_i \\ &= \left(\frac{2}{\omega} - 1\right) Df_i + \mu_i Df_i \\ &= \left(\frac{2}{\omega} - 1 + \mu_i\right) Df_i. \end{aligned} \tag{1.8.82}$$

On a donc en particulier $\left(\frac{2}{\omega}D - A\right) f_i \cdot f_j = 0$ is $i \neq j$, ce qui prouve que (1.8.80) est équivalent à (1.8.81). De (1.8.81), on déduit, grâce au fait que $(Df_i, f_i) = 1$,

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

On veut donc que $\frac{2}{\omega} - 1 + \mu_1 > 0$ car $\mu_1 = \inf \mu_i$, c'est–à–dire : $-\frac{2}{\omega} < \mu_1 - 1$, ce qui est équivalent à : $\omega < \dfrac{2}{1-\mu_1}$.

8. La matrice d'itération J_ω s'écrit :

$$J_\omega = \left(\frac{1}{\omega}D\right)^{-1}\left(\frac{1}{\omega}D - A\right) = \omega I_\omega, \text{ avec } I_\omega = D^{-1}(\frac{1}{\omega}D - A).$$

Soit λ une valeur propre de I_ω associée à un vecteur propre u ; alors :

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

On en déduit que

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

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

Or f_i est vecteur prore de $D^{-1}(E+F)$ associée à la valeur propre μ_i (question 2). On a donc :

$$D^{-1}(E+F)f_i = \mu_i f_i = \left(1 - \frac{1}{\omega} + \lambda\right) f_i,$$

ce qui est vrai si $\mu_i = 1 - \frac{1}{\omega} + \lambda$, c'est–à–dire $\lambda = \mu_i - 1 - \frac{1}{\omega}$. Donc $\mu_i^{(\omega)} = \omega\left(\mu_i - 1 - \dfrac{1}{\omega}\right)$ est valeur propre de J_ω associée au vecteur propre f_i.

On cherche maintenant à minimiser le rayon spectral

$$\rho(J_\omega) = \sup_i |\omega(\mu_i - 1 - \frac{1}{\omega})|$$

On a

$$\omega(\mu_1 - 1 - \frac{1}{\omega}) \leq \omega(\mu_i - 1 - \frac{1}{\omega}) \leq \omega(\mu_N - 1 - \frac{1}{\omega}),$$

et

$$-\omega(\mu_N - 1 - \frac{1}{\omega}) \leq -\omega(\mu_1 - 1 - \frac{1}{\omega}) \leq -\omega(\mu_i - 1 - \frac{1}{\omega}),$$

donc

$$\rho(J_\omega) = \max\left(|\omega(\mu_N - 1 - \frac{1}{\omega})|, |-\omega(\mu_1 - 1 - \frac{1}{\omega}|)\right)$$

dont le minimum est atteint (voir Figure 1.8) pour

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

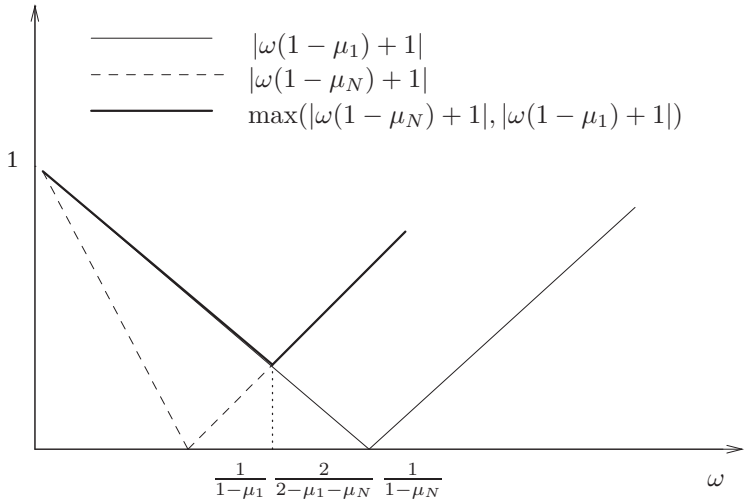

FIG. 1.6 – Détermination de la valeur de ω réalisant le minimum du rayon spectral.

Exercice 37 page 49 (Méthode de Jacobi pour des matrices particulières)

1. Soit $x \in \mathbb{R}^N$, supposons que

$$\|x\|_A = \sum_{i=1}^{N} a_{i,i}|x_i| = 0.$$

Comme $a_{i,i} > 0$, $\forall i = 1, \ldots, N$, on en déduit que $x_i = 0$, $\forall i = 1, \ldots, N$. D'autre part, il est immédiat de voir que $\|x+y\|_A \leq \|x\| + \|y\|_A$ pour tout $(x,y) \in \mathbb{R}^N \times \mathbb{R}^N$ et que $\|\lambda x\|_A = |\lambda|\|x\|_A$ pour tout $(x,\lambda) \in \mathbb{R}^N \times \mathbb{R}$. On en déduit que $\|\cdot\|_A$ est une norme sur $\mathbb{R}^N$.

2. Posons $\tilde{A} = \lambda Id + A$ et notons $\tilde{a}_{i,j}$ ses coefficients. Comme $\lambda \in \mathbb{R}_+^*$,et grâce aux hypothèses (1.6.57)–(1.6.59), ceux-ci vérifient :

$$\tilde{a}_{i,j} \leq 0, \forall i,j = 1,\dots,N,\ i \neq j, \tag{1.8.83}$$

$$\tilde{a}_{i,i} > \sum_{\substack{i=1\\ j\neq i}}^{N} a_{i,j},\ \forall i = 1,\dots,N. \tag{1.8.84}$$

La matrice $\tilde{A}$ est donc à diagonale dominante stricte, et par l'exercice 31 page 47, elle est donc inversible.

3. La méthode de Jacobi pour la résolution du système (1.6.60) s'écrit :

$$\tilde{D}u^{(k+1)} = (E+F)u^{(k)} + b, \tag{1.8.85}$$

avec $\tilde{D} = \lambda Id + D$, et $A = D - E - F$ est la décomposition habituelle de A en partie diagonale, triangulaire inférieure et triangulaire supérieure. Comme $a_{i,i} \geq 0$ et $\lambda \in \mathbb{R}_+^*$, on en déduit que $\tilde{D}$ est inversible, et que donc la suite $(u^{(k)})_{k\in\mathbb{N}}$ est bien définie dans $\mathbb{R}$.

4. Par définition de la méthode de Jacobi, on a :

$$u_i^{(k+1)} = \frac{1}{a_{i,i}+\lambda}\Big(\sum_{\substack{j=1,N\\ j\neq i}} a_{i,j}u_j^{(k)} + b_i\Big).$$

On en déduit que

$$u_i^{(k+1)} - u_i^{(k)} = \frac{1}{a_{i,i}+\lambda}\sum_{\substack{j=1,N\\ j\neq i}} a_{i,j}(u_j^{(k)} - u_j^{(k-1)}).$$

et donc

$$\|u^{(k+1)} - u^{(k)}\|_A \leq \sum_{i=1}^{N}\frac{a_{i,i}}{a_{i,i}+\lambda}\sum_{\substack{j=1,N\\ j\neq i}} a_{i,j}(u_j^{(k)} - u_j^{(k-1)}).$$

Or $\dfrac{a_{i,i}}{a_{i,i}+\lambda} \leq \dfrac{1}{1+\frac{\lambda}{a_{i,i}}} \leq \dfrac{1}{1+\alpha}$. 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 :

$$\begin{aligned}\|u^{(q)} - u^{(p)}\|_A &\leq \sum_{i=1}^{m}\|u^{(p+i)} - u^{(p-i-1)}\|_A \\ &\leq \|u^{(1)} - u^{(0)}\|_A\Big(\frac{1}{1+\alpha}\Big)^p\sum_{i=0}^{m}\Big(\frac{1}{1+\alpha}\Big)^i\end{aligned}$$

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 (1.8.85), on obtient que $\overline{u}$ est solution de (1.6.60).

Exercice 39 page 50 (Une méthode itérative particulière)

1. Det $(A) = -1$ et donc A est inversible.

2. Det $(\frac{1}{\omega}Id - E) = \frac{1}{\omega}\left(\frac{1}{\omega^2} - 2\right)$. Or $\omega \in]0, 2[$. Donc la matrice $\frac{1}{\omega}Id - E$ est inversible si $\omega \neq \frac{\sqrt{2}}{2}$.

3. Les valeurs propres de $\mathcal{L}_\omega$ sont les complexes λ tels qu'il existe $x \in C^3, x \neq 0$, t.q : $\mathcal{L}_\omega x = \lambda x$, c'est–à–dire :

$$\left(F + \frac{1-\omega}{\omega}Id\right)x = \lambda\left(\frac{1}{\omega}Id - E\right)x,$$

soit encore $M_{\lambda,\omega}x = 0$, avec $M_{\lambda,\omega} = \omega F + \lambda\omega E + (1-\omega-\lambda)Id$.

Or

$$\begin{aligned}Det\,(M_{\lambda,\omega}) &= (1-\omega-\lambda)((1-\omega-\lambda)^2 - 2\lambda^2\omega^2) \\ &= (1-\omega-\lambda)(1-\omega-(1+\sqrt{2}\omega)\lambda)(1-\omega-(1-\sqrt{2}\omega)\lambda)\end{aligned}$$

Les valeurs propres de $\mathcal{L}_\omega$ sont donc réelles, et égales à

$$\lambda_1 = 1-\omega, \lambda_2 = \frac{1-\omega}{1+\sqrt{2}\omega} \text{ et } \lambda_3 = \frac{1-\omega}{1-\sqrt{2}\omega}.$$

Par définition, le rayon spectral $\rho(\mathcal{L}_\omega)$ de la matrice $\mathcal{L}_\omega$ est égal à $\max(|\lambda_1|, |\lambda_2|, |\lambda_3|)$. Remarquons tout d'abord que $|1+\sqrt{2}\omega| > 1, \forall\omega \in]0, 2[$, et donc $|\lambda_1| > |\lambda_2|, \forall\omega \in]0, 2[$. Il ne reste donc plus qu'à comparer $|\lambda_1|$ et $|\lambda_3|$. Une rapide étude des fonctions $|\lambda_1|$ et $|\lambda_3|$ permet d'établir le graphe représentatif ci-contre.
On a donc :

$$\rho(\mathcal{L}_\omega) = |\lambda_3(\omega)| = \left|\frac{1-\omega}{1-\sqrt{2}\omega}\right| \text{ si } \omega \in]0, \sqrt{2}]$$

$$\rho(\mathcal{L}_\omega) = \lambda_1(\omega) = |1-\omega| \text{ si } \omega \in [\sqrt{2}, 2[.$$

4. La méthode est convergente si $\rho(\mathcal{L}_\omega) < 1$; Si $\omega \in [\sqrt{2}, 2[, \rho(\mathcal{L}_\omega) = \omega - 1 < 1$; si $\omega \in]0, \sqrt{2}[$,

$$\rho(\mathcal{L}_\omega) = \left|\frac{1-\omega}{1-\sqrt{2}\omega}\right| < 1$$

dès que $\frac{1-\omega}{\sqrt{2}\omega - 1} < 1$, c'est à dire $\omega > \frac{2}{1+\sqrt{2}}$.
Le minimum de $\rho(\mathcal{L}_\omega)$ est atteint pour $\omega_0 = 1$, on a alors $\rho(\mathcal{L}_\omega) = 0$.

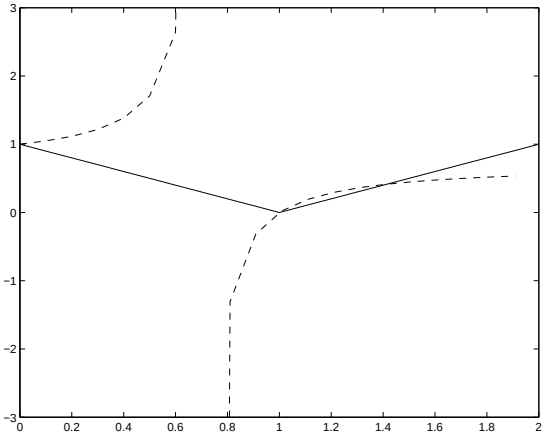

FIG. 1.7 – Graphe des valeurs propres λ_1 et λ_3

Exercice 41 page 51 (Méthode de la puissance pour calculer le rayon spectral de A)

1. Comme A est une matrice symétrique, A est diagonalisable dans $\mathbb{R}$. Soit $(f_1, \ldots, f_N) \in (\mathbb{R}^N)^N$ une base orthonormée de vecteurs propres de A associée aux valeurs propres $(\lambda_1, \ldots, \lambda_N) \in \mathbb{R}^N$. On décompose $x^{(0)}$ sur $(f_i)_{i=1,\ldots,N}$: $x^{(0)} = \sum_{i=1}^N \alpha_i f_i$. On a donc $Ax^{(0)} = \sum_{i=1}^N \lambda_i \alpha_i f_i$ et $A^n x^{(0)} = \sum_{i=1}^N \lambda_i^n \alpha_i f_i$.
   On en déduit :
$$\frac{x^{(n)}}{\lambda_N^n} = \sum_{i=1}^N \left(\frac{\lambda_i}{\lambda_N}\right)^n \alpha_i f_i.$$
   Comme $-\lambda_N$ n'est pas valeur propre,
$$\lim_{n\to+\infty} (\frac{\lambda_i}{\lambda_N})^n = 0 \text{ si } \lambda_i \neq \lambda_N. \tag{1.8.86}$$
   Soient $\lambda_1, \ldots, \lambda_p$ les valeurs propres différentes de λ_N, et $\lambda_{p+1}, \ldots, \lambda_N = \lambda_N$. On a donc
   $\lim_{n\to+\infty} \frac{x^{(n)}}{\lambda_N^n} = \sum_{i=p+1}^N \alpha_i f_i = x$, avec $Ax = \lambda_N x$.
   De plus, $x \neq 0$: en effet, $x^{(0)} \notin (Ker(A - \lambda_N Id))^\perp = Vect\{f_1, \ldots, f_p\}$, et donc il existe $i \in \{p+1, \ldots, N\}$ tel que $\alpha_i \neq 0$.
   Pour montrer (b), remarquons que :
$$\|x^{(n+1)}\| = \sum_{i=1}^N \lambda_i^{n+1} \alpha_i \text{ et } \|x^{(n)}\| = \sum_{i=1}^N \lambda_i^n \alpha_i$$
   car $(f_1, \ldots, f_N)$ est une base orthonormée. On a donc
$$\frac{\|x^{(n+1)}\|}{\|x^{(n)}\|} = \lambda_N^n \frac{\|\frac{x^{(n+1)}}{\lambda_N^{n+1}}\|}{\|\frac{x^{(n)}}{\lambda_N}\|} \to \lambda_N \frac{\|x\|}{\|x\|} = \lambda_N \text{ lorsque } n \to +\infty.$$

2. a) La méthode I s'écrit à partir de $x^{(0)}$ connu : $x^{(n+1)} = Bx^{(n)} + c$ pour $n \geq 1$, avec $c = (I - B)A^{-1}b$.
On a donc

$$\begin{aligned} x^{(n+1)} - x &= Bx^{(n)} + (Id - B)x - x \\ &= B(x^{(n)} - x). \end{aligned} \tag{1.8.87}$$

Si $y^{(n)} = x^{(n)} - x$, on a donc $y^{(n+1)} = By^{(n)}$, et d'après la question 1a) si $y^{(0)} \not\perp \mathrm{Ker}(B - \mu_N Id)$ où μ_N est la plus grande valeur propre de B, (avec $|\mu_N| = \rho(B) et - \mu_N$ non valeur propre), alors

$$\frac{\|y^{(n+1)}\|}{\|y^{(n)}\|} \longrightarrow \rho(B) \text{ lorsque } n \to +\infty,$$

c'est–à–dire

$$\frac{\|x^{(n+1)} - x\|}{\|x^{(n)} - x\|} \longrightarrow \rho(B) \text{ lorsque } n \to +\infty.$$

b) On applique maintenant 1a) à $y^{(n)} = x^{(n+1)} - x^{(n)}$ avec

$$y^{(0)} = x^{(1)} - x^{(0)} \text{ où } x^{(1)} = Ax^{(0)}.$$

On demande que $x^{(1)} - x^{(0)} \notin \mathrm{Ker}(B - \mu_N Id)^{\perp}$ comme en a), et on a bien $y^{(n+1)} = By^{(n)}$, donc $\frac{\|y^{(n+1)}\|}{\|y^{(n)}\|} \longrightarrow \rho(B)$ lorsque $n \to +\infty$.

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 :

$$\begin{cases} x \in \mathbb{R}^N \\ g(x) = 0. \end{cases} \tag{2.0.1}$$

Au Chapitre I on a étudié des méthodes de résolution du système (2.0.1) 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 (2.0.1) :
– les méthodes de point fixe : point fixe de contraction et point fixe de monotonie
– les méthodes de type Newton[1].

2.1 Les méthodes de point fixe

2.1.1 Point fixe de contraction

Soit $g \in C(\mathbb{R}^N, \mathbb{R}^N)$, on définit la fonction $f \in C(\mathbb{R}^N, \mathbb{R}^N)$ par $f(x) = x + g(x)$. On peut alors remarquer que $g(x) = 0$ si et seulement si $f(x) = x$. Résoudre le système non linéaire (2.0.1) revient donc à trouver un point fixe de f. Encore faut-il qu'un tel point fixe existe...

Théorème 2.1 (Point fixe) *Soit E un espace métrique complet, d la distance sur E, et $f : E \to E$ une fonction strictement contractante, c'est–à–dire telle qu'il existe $k \in]0, 1[$ tel que $d(f(x), f(y)) \leq kd(x, y)$ pour tout $x, y \in E$.*
Alors il existe un unique point fixe $\bar{x} \in E$ qui vérifie $f(\bar{x}) = \bar{x}$. De plus si $x^{(0)} \in E$, et $x^{(n+1)} = f(x^{(n)})$ $\forall n \geq 0$, alors $x^{(n)} \to \bar{x}$ quand $n + \infty$.

Démonstration :

Etape 1 : Existence de $\bar{x}$ et convergence de la suite

Soit $x^{(0)} \in E$ et $(x^{(n)})_{n \in \mathbb{N}}$ la suite définie par $x^{(n+1)} = f(x^{(n)})$ pour $n \geq 0$. On va montrer que :

1. $(x^{(n)})_n$ est de Cauchy (donc convergente car E est complet),

[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.

2. $\lim_{n\to+\infty} x^{(n)} = \bar{x}$ est point fixe de f.

Par hypothèse, on sait que pour tout $n \geq 1$,

$$d(x^{(n+1)}, x^{(n)}) = d(f(x^{(n)}), f(x^{(n-1)})) \leq k d(x^{(n)}, x^{(n-1)}).$$

Par récurrence sur n, on obtient que

$$d(x^{(n+1)}, x^{(n)}) \leq k^n d(x^{(1)}, x^{(0)}),\ \forall n \geq 0.$$

Soit $n \geq 0$ et $p \geq 1$, on a donc :

$$\begin{aligned} d(x^{(n+p)}, x^{(n)}) &\leq d(x^{(n+p)}, x^{(n+p-1)}) + \cdots + d(x^{(n+1)}, x^{(n)}) \\ &\leq \sum_{q=1}^{p} d(x^{(n+q)}, x^{(n+q-1)}) \\ &\leq \sum_{q=1}^{p} k^{n+q-1} d(x^{(1)}, x^{(0)}) \\ &\leq d(x^{(1)}, x^{(0)}) k^n (1 + k + \ldots + k^{p-1}) \\ &\leq d(x^{(1)}, x^{(0)}) \frac{k^n}{1-k} \longrightarrow 0 \text{ quand } n \to +\infty \text{ car } k < 1. \end{aligned}$$

La suite $(x^{(n)})_{n\in\mathbb{N}}$ est donc de Cauchy, *i.e.* :

$$\forall \varepsilon > 0,\ \ \exists n_\varepsilon \in \mathbb{N}\,;\ \ \forall n \geq n_\varepsilon,\ \ \forall\ \ p \geq 1\ \ d(x^{(n+p)}, x^{(n)}) \leq \varepsilon.$$

Comme E est complet, on a donc $x^{(n)} \longrightarrow \bar{x}$ dans E quand $n \to +\infty$.

Comme la fonction f est strictement contractante, elle est continue, donc on a aussi $f(x^{(n)}) \longrightarrow f(\bar{x})$ dans E quand $n + \infty$. En passant à la limite dans l'égalité $x^{(n+1)} = f(x^{(n)})$, on en déduit que $\bar{x} = f(\bar{x})$.

Etape 2 : Unicité

Soit $\bar{x}$ et $\bar{y}$ des points fixes de f, qui satisfont donc $\bar{x} = f(\bar{x})$ et $\bar{y} = f(\bar{y})$. Alors $d(f(\bar{x}), f(\bar{y})) = d(\bar{x}, \bar{y}) \leq k d(\bar{x}, \bar{y})$; comme $k < 1$, ceci est impossible sauf si $\bar{x} = \bar{y}$.

■

Remarque 2.2

1. *Sous les hypothèses du théorème 2.1, $d(x^{(n+1)}, \bar{x}) = d(f(x^{(n)}), f(\bar{x})) \leq k d(x^{(n)}, \bar{x})$; donc si $x^{(n)} \neq \bar{x}$ alors $\frac{d(x^{(n+1)}, \bar{x})}{d(x^{(n)}, \bar{x})} \leq k\ \ (< 1)$. La convergence est donc au moins linéaire (même si de fait, cette méthode converge en général assez lentement).*
2. *On peut généraliser le théorème du point fixe en remplaçant l'hypothèse "f strictement contractante" par " il existe $n > 0$ tel que $f^{(n)} = \underbrace{f \circ f \circ \ldots \circ f}_{n \text{ fois}}$ est strictement contractante " (reprendre la démonstration du théorème pour le vérifier).*

La question qui vient alors naturellement est : que faire si f n'est pas strictement contractante ? Soit $g \in C(\mathbb{R}^N, \mathbb{R}^N)$ telle que $f(x) = x + g(x)$. On aimerait déterminer les conditions sur g pour que f soit strictement contractante. Plus généralement si $\omega \neq 0$, on définit $f_\omega(x) = x + \omega g(x)$, et on remarque que x est solution du système (2.0.1) si et seulement si x est point fixe de $f_\omega(x)$.
On aimerait dans ce cas avoir des conditions pour que f_ω soit strictement contractante.

Théorème 2.3 (Point fixe de contraction avec relaxation)
On désigne par $|.|$ la norme euclidienne sur $\mathbb{R}^N$. Soit $g \in C(\mathbb{R}^N, \mathbb{R}^N)$ telle que

$$\exists \alpha > 0 \text{ tel que } (g(x) - g(y)) \cdot (x - y) \leq -\alpha |x - y|^2, \forall x, y \in \mathbb{R}^N, \tag{2.1.2}$$

$$\exists M > 0 \text{ tel que } |g(x) - g(y)| \leq M|x - y|, \ \forall x, y \in \mathbb{R}^N. \tag{2.1.3}$$

Alors la fonction f_ω 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 $x^{(n)} \longrightarrow \bar{x}$ quand $n + \infty$ avec $x^{(n+1)} = f_\omega(x^{(n)}) = x^{(n)} + \omega g(x^{(n+1)})$.

Remarque 2.4 *Le théorème 2.3 permet de montrer que sous les hypothèses* (2.1.2) *et* (2.1.3)*, et pour $\omega \in]0, \frac{2\alpha}{M^2}[$, on peut obtenir la solution de* (2.0.1) *en construisant la suite :*

$$\begin{cases} x^{(n+1)} = x^{(n)} + \omega g(x^{(n)}) \ \ n \geq 0, \\ x^{(0)} \in \mathbb{R}^N. \end{cases} \tag{2.1.4}$$

Or on peut aussi écrire cette suite de la manière suivante :

$$\begin{cases} \tilde{x}^{(n+1)} & = f(x^{(n)}), \ \ \forall n \geq 0 \\ x^{(n+1)} & = \omega \tilde{x}^{(n+1)} + (1 - \omega) x^{(n)}, \quad x^{(0)} \in \mathbb{R}^N. \end{cases} \tag{2.1.5}$$

En effet si $x^{(n+1)}$ est donné par la suite (2.1.5)*, alors $x^{(n+1)} = \omega \tilde{x}^{(n+1)} - (1-\omega)x^{(n)} = \omega f(x^{(n)}) + (1-\omega)x^{(n)} = \omega g(x^{(n)}) + x^{(n)}$. Le procédé de construction de la suite* (2.1.5) *est l'algorithme de relaxation sur f.*

Démonstration du théorème 2.3
Soit $0 < \omega < \dfrac{2\alpha}{M^2}$. On veut montrer que f est strictement contractante, c.à.d. qu'il existe $k < 1$ tel que $|f_\omega(x) - f_\omega(y)| \leq k|x - y| \ \forall (x, y) \in (\mathbb{R}^N)^2$. Soit $(x, y) \in (\mathbb{R}^N)^2$, alors, par définition de la norme euclidienne,

$$\begin{aligned} |f_\omega(x) - f_\omega(y)|^2 &= \big(x - y + \omega(g(x) - g(y))\big) \cdot \big(x - y + \omega(g(x) - g(y))\big) \\ &= |x - y|^2 + 2(x - y) \cdot (\omega(g(x) - g(y))) + \omega^2 |g(x) - g(y)|^2. \end{aligned}$$

Donc grâce aux hypothèses (2.1.2) et (2.1.3), on a : $|f_\omega(x) - f_\omega(y)|^2 \leq (1 - 2\omega\alpha + \omega^2 M^2) \ |x - y|^2$, et donc la fonction f_ω est strictement contractante si $1 - 2\omega\alpha + \omega^2 M^2 < 1$ ce qui est vérifié si $0 < \omega < \dfrac{2\alpha}{M^2}$. ■

Remarque 2.5 (Quelques rappels de calcul différentiel)
Soit $h \in C^2(\mathbb{R}^N, \mathbb{R})$. La fonction h est donc en particulier différentiable, c.à.d. que pour tout $x \in \mathbb{R}^N$, il existe $Dh(x) \in \mathcal{L}(\mathbb{R}^N, \mathbb{R})$ telle que $h(x + y) = h(x) + Dh(x)(y) + |y|\varepsilon(y)$ où $\varepsilon(y) \underset{y \to 0}{\to} 0$. On a dans ce cas, par définition du gradient, $Dh(x)(y) = \nabla h(x) \cdot y$ où $\nabla h(x) = (\partial_1 h(x), \cdots, \partial_N h(x))^t \in \mathbb{R}^N$ est le gradient de h au point x (on désigne par $\partial_i h$ la dérivée partielle de f par rapport à sa i-ème variable).
Comme on suppose $h \in C^2(\mathbb{R}^N, \mathbb{R})$, on a donc $g = \nabla h \in C^1(\mathbb{R}^N, \mathbb{R}^N)$, et g est continûment différentiable, c'est–à–dire

$$Dg(x) \in \mathcal{L}(\mathbb{R}^N, \mathbb{R}^N), \text{ et } g(x + y) = g(x) + Dg(x)(y) + |y|\varepsilon(y),$$

où $\varepsilon(y) \underset{y \to 0}{\to} 0$.
Comme $Dg(x) \in \mathcal{L}(\mathbb{R}^N, \mathbb{R}^N)$, on peut représenter $Dg(x)$ par une matrice de $\mathcal{M}_N(\mathbb{R})$, on confond alors l'application linéaire et la matrice qui la représente dans la base canonique, et on écrit par abus de notation $Dg(x) \in \mathcal{M}_N(\mathbb{R})$. On peut alors écrire, grâce à cet abus de notation, $Dg(x)(y) = Dg(x)y$ avec $(Dg(x)y)_i = \sum_{i,j=1,N} \partial^2_{i,j} h_j(x)$ où $\partial^2_{i,j} h = \partial_i(\partial_j h)(x)$.
Comme h est de classe C^2, la matrice $Dg(x)$ est symétrique. Pour $x \in \mathbb{R}^N$, on note $(\lambda_i(x))_{1 \leq i \leq N}$ les valeurs propres de $Dg(x)$, qui sont donc réelles.

La proposition suivante donne une condition suffisante pour qu'une fonction vérifie les hypothèses (2.1.2) et (2.1.3).

Proposition 2.6 *Soit $h \in C^2(\mathbb{R}^N, \mathbb{R})$, et $(\lambda_i)_{i=1,N}$ les valeurs propres de la matrice hessienne de h. On suppose qu'il existe des réels strictement positifs β et γ tels que $-\beta \leq \lambda_i(x) \leq -\gamma,\ \forall i \in \{1 \dots N\},\ \forall x \in \mathbb{R}^N$. (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* (2.1.2) *et* (2.1.3) *du théorème 2.3 avec $\alpha = \gamma$ et $M = \beta$.*

Démonstration de la proposition 2.6
Montrons d'abord que l'hypothèse (2.1.2) est vérifiée. Soit $(x, y) \in (\mathbb{R}^N)^2$, on veut montrer que $(g(x) - g(y)) \cdot (x - y) \leq -\gamma|x - y|^2$. On introduit pour cela la fonction $\varphi \in C^1(\mathbb{R}, \mathbb{R}^N)$ définie par :

$$\varphi(t) = g(x + t(y - x)).$$

On a donc $\varphi(1) - \varphi(0) = g(y) - g(x) = \int_0^1 \varphi'(t)dt$. Or $\varphi'(t) = Dg(x + t(y - x))(y - x)$. Donc $g(y) - g(x) = \int_0^1 Dg(x + t(y - x))(y - x)dt$. On en déduit que :

$$(g(y) - g(x)) \cdot (y - x) = \int_0^1 (Dg(x + t(y - x))(y - x) \cdot (y - x))\, dt.$$

Comme $\lambda_i(x) \in [-\beta, -\gamma]\ \forall i \in \{1, \dots, N\}$, on a donc $-\beta|y|^2 \leq Dg(z)y \cdot y \leq -\gamma|y|^2$. On a donc : $(g(y) - g(x)) \cdot (y - x) \leq \int_0^1 -\gamma|y - x|^2 dt = -\gamma|y - x|^2$ ce qui termine la démonstration de (2.1.2).
Montrons maintenant que l'hypothèse (2.1.3) est vérifiée. On veut montrer que $|g(y) - g(x)| \leq \beta|y - x|$. On peut écrire :

$$g(y) - g(x) = \int_0^1 Dg(x + t(y - x))(y - x)dt,$$

et donc

$$\begin{aligned} |g(y) - g(x)| &\leq \int_0^1 |Dg(x + t(y - x))(y - x)|dt \\ &\leq \int_0^1 |Dg(x + t(y - x))||y - x|dt, \end{aligned}$$

où $|.|$ est la norme sur $\mathcal{M}_N(\mathbb{R})$ induite par la norme euclidienne sur $\mathbb{R}^N$.
Or, comme $\lambda_i(x) \in [-\beta, -\gamma]$ pour tout $i = 1, \dots, N$, la matrice $-Dg(x + t(y - x))$ est symétrique définie positive. Et donc, d'après l'exercice 5 page 37,

$$|Dg(x + t(y - x)| = \rho(Dg(x + t(y - x)) \leq \beta.$$

On a donc ainsi montré que :

$$|g(y) - g(x)| \leq \beta|y - x|,$$

ce qui termine la démonstration.

■

Remarque 2.7 *Dans de nombreux cas, 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 $x = A^{-1}R(x)$. On peut donc appliquer l'algorithme de point fixe sur la fonction $f = A^{-1}R$, ce qui donne comme itération : $x^{(n+1)} = A^{-1}R(x^{(n)})$. Si on pratique un point fixe avec relaxation, avec paramètre de relaxation $\omega > 0$, alors l'itération s'écrit : $\tilde{x}^{(n+1)} = A^{-1}R(x^{(n)})$, $x^{(n+1)} = \omega\tilde{x}^{(n+1)} + (1 - \omega)x^{(n)}$.*

2.1.2 Point fixe de monotonie

Théorème 2.8 (Point fixe de monotonie)
Soient $A \in \mathcal{M}_N(\mathbb{R})$ et $R \in C(\mathbb{R}^N, \mathbb{R}^N)$. On suppose que :

1. $\forall x \in \mathbb{R}^N$, $Ax \geq 0 \Rightarrow x \geq 0$, *c'est-à-dire* $\big((Ax)_i \geq 0,\ \forall i = 1, \ldots, N\big) \Rightarrow \big(x_i \geq 0,\ \forall i = 1, \ldots, N\big)$.
2. *R est monotone, c.à.d. que si $x \geq y$ (composante par composante) alors $R(x) \geq R(y)$ (composante par composante).*
3. *0 est une sous-solution du problème, c'est-à-dire que $R(0) \geq 0$ 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 $Ax^{(n+1)} = R(x^{(n)})$. On a alors :

1. $0 \leq x^{(n)} \leq \tilde{x},\ \ \forall n \in \mathbb{N}$,
2. $x^{(n+1)} \geq x^{(n)},\ \ \forall n \in \mathbb{N}$,
3. *$x^{(n)} \longrightarrow \bar{x}$ quand $n \to +\infty$ et $A\bar{x} = R(\bar{x})$.*

Démonstration du théorème 2.8
Comme A est inversible la suite $(x^{(n)})_{n\in\mathbb{N}}$ vérifiant

$$\left\{ \begin{array}{l} x^{(0)} = 0, \\ Ax^{(n+1)} = R(x^{(n)}),\ \ n \geq 0 \end{array} \right.$$

est bien définie. On va montrer par récurrence sur n que $0 \leq x^{(n)} \leq \tilde{x}$ pour tout $n \geq 0$ et que $x^{(n)} \leq x^{(n+1)}$ pour tout $n \geq 0$.

1. Pour $n = 0$, on a $x^{(0)} = 0$ et donc $0 \leq x^{(0)} \leq \tilde{x}$ et $Ax^{(1)} = R(0) \geq 0$. On en déduit que $x^{(1)} \geq 0$ grâce aux hypothèses 1 et 3 et donc $x^{(1)} \geq x^{(0)} = 0$.
2. On suppose maintenant (hypothèse de récurrence) que $0 \leq x^{(p)} \leq \tilde{x}$ et $x^{(p)} \leq x^{(p+1)}$ pour tout $p \in \{0, \ldots, n-1\}$.
   On veut montrer que $0 \leq x^{(n)} \leq \tilde{x}$ et que $x^{(n)} \leq x^{(n+1)}$. Par hypothèse de récurrence pour $p = n-1$, on sait que $x^{(n)} \geq x^{(n-1)}$ et que $x^{(n-1)} \geq 0$. On a donc $x^{(n)} \geq 0$. Par hypothèse de récurrence, on a également que $x^{(n-1)} \leq \tilde{x}$ et grâce à l'hypothèse 2, on a donc $R(x^{(n-1)}) \leq R(\tilde{x})$. Par définition de la suite $(x^{(n)})_{n\in\mathbb{N}}$, on a $Ax^{(n)} = R(x^{(n-1)})$ et grâce à l'hypothèse 3, on sait que $A\tilde{x} \geq R(\tilde{x})$. On a donc : $A(\tilde{x} - x^{(n)}) \geq R(\tilde{x}) - R(x^{(n-1)}) \geq 0$. On en déduit alors (grâce à l'hypothèse 1) que $x^{(n)} \leq \tilde{x}$.
   De plus, comme $Ax^{(n)} = R(x^{(n-1)})$ et $Ax^{(n+1)} = R(x^{(n)})$, on a $A(x^{(n+1)} - x^{(n)}) = R(x^{(n)}) - R(x^{(n-1)}) \geq 0$ par l'hypothèse 2, et donc grâce à l'hypothèse 1, $x^{(n+1)} \geq x^{(n)}$.

On a donc ainsi montré (par récurrence) que

$$0 \leq x^{(n)} \leq \tilde{x},\ \ \forall n \geq 0$$

$$x^{(n)} \leq x^{(n+1)},\ \ \forall n \geq 0.$$

Ces inégalités s'entendent composante par composante, c.à.d. que si $x^{(n)} = (x_1^{(n)} \ldots x_N^{(n)})^t \in \mathbb{R}^N$ et $\tilde{x} = (\tilde{x}_1 \ldots \tilde{x}_N)^t \in \mathbb{R}^N$, alors $0 \leq x_i^{(n)} \leq \tilde{x}_i$ et $x_i^{(n)} \leq x_i^{(n+1)}$, $\forall i \in \{1, \ldots, N\}$, et $\forall n \geq 0$.

Soit $i \in \{1, \ldots, N\}$; la suite $(x_i^{(n)})_{n\in\mathbb{N}} \subset \mathbb{R}$ est croissante et majorée par $\tilde{x}_i$ donc il existe $\bar{x}_i \in \mathbb{R}$ tel que $\bar{x}_i = \lim_{n\to+\infty} x_i^{(n)}$. Si on pose $\bar{x} = (\bar{x}_1 \ldots \bar{x}_N)^t \in \mathbb{R}^N$, on a donc $x^{(n)} \longrightarrow \bar{x}$ quand $n \to +\infty$.

Enfin, comme $Ax^{(n+1)} = R(x^{(n)})$ et comme R est continue, on obtient par passage à la limite lorsque $n \to +\infty$ que $A\bar{x} = R(\bar{x})$ et que $0 \leq \bar{x} \leq \tilde{x}$.

■

L'hypothèse 1 du théorème 2.8 est souvent appelée "principe du maximum". Elle 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 23) et Δu sur $]0,1[\times]0,1[$ (voir page 27). Le principe du maximum est aussi caractérisé de la manière suivante (plus difficile à utiliser en pratique) :

Proposition 2.9 (CNS de monotonie) *L'hypothèse 1 du théorème 2.8 est vérifiée si et seulement si A inversible et A^{-1} a des coefficients ≥ 0.*

Démonstration :
Supposons d'abord que l'hypothèse 1 du théorème 2.8 est vérifiée et montrons que A inversible et que A^{-1} a des coefficients ≥ 0. Si x est tel que $Ax = 0$, alors $Ax \geq 0$ et donc, par hypothèse, $x \geq 0$. Mais on a aussi $Ax \leq 0$, soit $A(-x) \geq 0$ et donc par hypothèse, $x \leq 0$. On en déduit $x = 0$, ce qui prouve que A est inversible.

L'hypothèse 1 donne alors que $y \geq 0 \Rightarrow A^{-1}y \geq 0$. En prenant $y = e_1$ on obtient que la première colonne de A^{-1} est positive, puis en prenant $y = e_i$ on obtient que la i-ème colonne de A^{-1} est positive, pour $i = 2, \ldots, N$. Donc A^{-1} a tous ses coefficients positifs.

Supposons maintenant que A est inversible et que A^{-1} a des coefficients positifs. Soit $x \in \mathbb{R}^N$ tel que $Ax = y \geq 0$, alors $x = A^{-1}y \geq 0$. Donc A vérifie l'hypothèse 1.

■

Théorème 2.10 (Généralisation du précédent)
Soit $A \in \mathcal{M}_N(\mathbb{R})$, $R \in C^1(\mathbb{R}^N, \mathbb{R}^N)$, $R = (R_1, \ldots, R_N)^t$ tels que

1. *Pour tout $\beta \geq 0$ et pour tout $x \in \mathbb{R}^N$, $Ax + \beta x \geq 0 \Rightarrow x \geq 0$*
2. *$\dfrac{\partial R_i}{\partial x_j} \geq 0$, $\forall i, j$ t.q. $i \neq j$ (R_i est monotone croissante par rapport à la variable x_j si $j \neq i$) et $\exists \gamma > 0$, $-\gamma \leq \dfrac{\partial R_i}{\partial x_i} \leq 0, \forall x \in \mathbb{R}^N, \ \ \forall i \in \{1, \ldots, N\}$ (R_i est monotone décroissante par rapport à la variable x_i).*
3. *$0 \leq R(0)$ (0 est sous-solution) et $\exists \tilde{x} \geq 0$ tel que $A(\tilde{x}) \geq R(\tilde{x})$ ($\tilde{x}$ est sur-solution).*

Soient $x^{(0)} = 0$, $\beta \geq \gamma$, et $(x^{(n)})_{n\in\mathbb{N}}$ la suite définie par $Ax^{(n+1)} + \beta x^{(n+1)} = R(x^{(n)}) + \beta x^{(n)}$. Cette suite converge vers $\bar{x} \in \mathbb{R}^N$ et $A\bar{x} = R(\bar{x})$. De plus, $0 \leq x^{(n)} \leq \tilde{x} \ \ \forall n \in \mathbb{N}$ et $x^{(n)} \leq x^{(n+1)}, \ \ \forall n \in \mathbb{N}$.

Démonstration : On se ramène au théorème précédent avec $A+\beta Id$ au lieu de A et $R+\beta$ au lieu de R. ■

Remarque 2.11 (Point fixe de Brouwer) *On s'est intéressé ici uniquement à des théorèmes de point fixe "constructifs", i.e. qui donnent un algorithme pour le déterminer. Il existe aussi un théorème de point fixe dans $\mathbb{R}^N$ avec des hypothèses beaucoup plus générales (mais le théorème est non constructif), c'est le théorème de Brouwer[2] : si f est une fonction continue de la boule unité de $\mathbb{R}^N$ dans la boule unité, alors elle admet un point fixe dans la boule unité.*

2.1.3 Vitesse de convergence

Définition 2.12 (Vitesse de convergence) *Soit $(x^{(n)})_{n\in\mathbb{N}} \in \mathbb{R}^N$ et $\bar{x} \in \mathbb{R}^N$. On suppose que $x^{(n)} \to \bar{x}$ lorsque $n \to +\infty$, avec $x^{(n)} \neq \bar{x}$ pour tout $n \in \mathbb{N}$. On s'intéresse à la "vitesse de convergence" de la suite $(x^{(n)})_{n\in\mathbb{N}}$. On dit que :*

1. *la convergence est* **au moins linéaire** *s'il existe $\beta \in]0,1[$ et il existe $n_0 \in \mathbb{N}$ tels que si $n \geq n_0$ alors $\|x^{(n+1)} - \bar{x}\| \leq \beta \|x^{(n)} - \bar{x}\|$.*

[2] Luitzen Egbertus Jan Brouwer (1881-1966), mathématicien néerlandais.

2. *La convergence* **est linéaire** *si il existe* $\beta \in]0,1[$ *tel que*

$$\frac{\|x^{(n+1)} - \bar{x}\|}{\|x^{(n)} - \bar{x}\|} \longrightarrow \beta \text{ quand } n \to +\infty,$$

3. *La convergence est* **super linéaire** *si*

$$\frac{\|x^{(n+1)} - \bar{x}\|}{\|x^{(n)} - \bar{x}\|} \longrightarrow 0 \text{ quand } n \to +\infty,$$

4. *La convergence est* **au moins quadratique** *si il existe* $\beta \in]0,1[$ *et il existe* $n_0 \in \mathbb{N}$ *tels que si* $n \geq n_0$ *alors* $\|x^{(n+1)} - \bar{x}\| \leq \beta \|x^{(n)} - \bar{x}\|^2$,
5. *La convergence est* **quadratique** *si*

$$\exists \beta > 0 \quad \frac{\|x^{(n+1)} - \bar{x}\|}{\|x^{(n)} - \bar{x}\|^2} \longrightarrow \beta \text{ quand } n \to +\infty.$$

Remarque 2.13 *La convergence quadratique est évidemment plus "rapide" que la convergence linéaire.*

Proposition 2.14 *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*

$$\begin{array}{l} x^{(0)} \in \mathbb{R} \\ x^{(n+1)} = f(x^{(n)}). \end{array}$$

1. *Si on suppose que* $f'(\bar{x}) \neq 0$ *et* $|f'(\bar{x})| < 1$*, alors il existe* $\alpha > 0$ *tel que si* $x^{(0)} \in I_\alpha = [\bar{x} - \alpha, \bar{x} + \alpha]$ *alors* $x^{(n)} \to \bar{x}$ *lorsque* $n \to +\infty$*, et si* $x^{(n)} \neq \bar{x}$*, alors* $\dfrac{|x^{(n+1)} - \bar{x}|}{|x^{(n)} - \bar{x}|} \to |f'(\bar{x})| = \beta$*, où* $\beta \in]0,1[$*. 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* $x^{(n)} \to \bar{x}$ *quand* $n + \infty$*, et si* $x^{(n)} \neq \bar{x}$, $\forall n \in \mathbb{N}$ *alors*

$$\frac{|x^{(n+1)} - \bar{x}|}{|x^{(n)} - \bar{x}|^2} \to \beta = \frac{1}{2}|f''(\bar{x})|.$$

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 $x^{(n)} \to \bar{x}$. 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 $x^{(n)} \to \bar{x}$ 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 $x^{(n)} \to \bar{x}$ si $x^{(0)} \in I_\alpha$.

Cherchons maintenant la vitesse de convergence de la suite. Supposons que $f'(\bar{x}) \neq 0$ et $x^{(n)} \neq \bar{x}$ pour tout $n \in \mathbb{N}$. Comme $x^{(n+1)} = f(x^{(n)})$ et $\bar{x} = f(\bar{x})$, on a $|x^{(n+1)} - \bar{x}| = |f(x^{(n)}) - f(\bar{x})|$. Comme $f \in C^1(\mathbb{R}, \mathbb{R})$, il existe $\xi_n \in]x^{(n)}, \bar{x}[$ ou $]\bar{x}, x^{(n)}[$, tel que $f(x^{(n)}) - f(\bar{x}) = f'(\xi_n)(x^{(n)} - \bar{x})$. On a donc

$$\frac{|x^{(n+1)} - \bar{x}|}{|x^{(n)} - \bar{x}|} = |f'(\xi_n)| \longrightarrow |f'(\bar{x})| \text{ car } x^{(n)} \to \bar{x} \text{ et } f' \text{ est continue.}$$

On a donc une convergence linéaire.

2. Supposons maintenant que $f'(\bar{x}) = 0$ et $f \in C^2(\mathbb{R}, \mathbb{R})$. On sait déjà par ce qui précède qu'il existe $\alpha > 0$ tel que si $x^{(0)} \in I_\alpha$ alors $x^{(n)} \to\to \bar{x}$ lorsque $n \to +\infty$. On veut estimer la vitesse de convergence. On suppose pour cela que $x^{(n)} \neq \bar{x}$ pour tout $n \in \mathbb{N}$.
Comme $f \in C^2(\mathbb{R}, \mathbb{R})$, il existe $\xi_n \in]x^{(n)}, \bar{x}[$ tel que $f(x^{(n)}) - f(\bar{x}) = f'(\bar{x})(x^{(n)} - \bar{x}) + \frac{1}{2} f''(\xi_n)(x^{(n)} - \bar{x})^2$.
On a donc : $x^{(n+1)} - \bar{x} = \frac{1}{2} f''(\xi_n)(x^{(n)} - \bar{x})^2$ ce qui entraîne que

$$\frac{|x^{(n+1)} - \bar{x}|}{|x^{(n)} - \bar{x}|^2} = \frac{1}{2}|f''(\xi_n)| \longrightarrow \frac{1}{2}|f''(\bar{x})| \text{ quand } n \to +\infty.$$

La convergence est donc quadratique.

■

On étudie dans le paragraphe suivant la méthode de Newton pour la résolution d'un système non linéaire. 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 $(x^{(n)})_n \in \mathbb{R}^N$ qui converge vers $\bar{x}$ de manière quadratique. On pose

$$f(x) = x + h(x)g(x) \text{ avec } h \in C^2(\mathbb{R}, \mathbb{R}) \text{ tel que } h(x) \neq 0 \ \forall x \in \mathbb{R}.$$

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 2.14) $(x^{(n)})_{n\in\mathbb{N}}$ tel que $x^{(n)} \to \bar{x}$ de manière au moins quadratique. Or on a $f'(x) = 1 + h'(x)g(x) + g'(x)h(x)$ et donc $f'(\bar{x}) = 1 + g'(\bar{x})h(\bar{x})$. Il suffit donc de prendre h tel que $h(\bar{x}) = -\dfrac{1}{g'(\bar{x})}$. Ceci est possible si $g'(\bar{x}) \neq 0$.

En résumé, si $g \in C^3(\mathbb{R}, \mathbb{R})$ est telle que $g'(\overline{x}) \neq 0 \ \forall x \in \mathbb{R}$ et $g(\bar{x}) = 0$, on peut construire, pour x assez proche de $\bar{x}$, la fonction $f \in C^2(\mathbb{R}, \mathbb{R})$ définie par

$$f(x) = x - \frac{g(x)}{g'(x)}.$$

Grâce à la proposition 2.14, il existe $\alpha > 0$ tel que si $x^{(0)} \in I_\alpha$ alors la suite définie par $x^{(n+1)} = f(x^{(n)}) = x^{(n)} - \frac{g(x^{(n)})}{g'(x^{(n)})}$ converge vers $\bar{x}$ de manière au moins quadratique.

Remarquons que la construction de la suite de Newton s'écrit encore (dans le cas $N = 1$) $g'(x^{(n)})(x^{(n+1)} - x^{(n)}) = -g(x^{(n)})$ ou encore $g(x^{(n)}) + g'(x^{(n)})(x^{(n+1)} - x^{(n)}) = 0$.

2.2 Méthode de Newton

2.2.1 Construction et convergence de la méthode

On a vu ci-dessus comment se construit la méthode de Newton à partir du point fixe de monotonie 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 $\bar{x} \in \mathbb{R}^N$ tels que $g(\bar{x}) = 0$.
On cherche une méthode de construction d'une suite $(x^{(n)})_n \in \mathbb{R}^N$ qui converge vers $\bar{x}$ de manière quadratique. L'algorithme de Newton de construction d'une telle suite s'écrit :

$$\begin{cases} x^{(0)} \in \mathbb{R}^N \\ Dg(x^{(n)})(x^{(n+1)} - x^{(n)}) = -g(x^{(n)}), \ \forall n \geq 0. \end{cases} \tag{2.2.6}$$

(On rappelle que $Dg(x^{(n)}) \in \mathcal{M}_N(\mathbb{R})$ est la matrice représentant la différentielle de g en $x^{(n)}$.)

Pour chaque $n \in \mathbb{N}$, il faut donc effectuer les opérations suivantes :

1. Calcul de $Dg(x^{(n)})$,
2. Résolution du système linéaire $Dg(x^{(n)})(x^{(n+1)} - x^{(n)}) = -g(x^{(n)})$.

Remarque 2.15 *Si la fonction g dont on cherche un zéro est linéaire, i.e. si g est définie par $g(x) = Ax - b$ 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 $Ax = b$. En effet $Dg(x^{(n)}) = A$ et donc* (2.2.6) *s'écrit $Ax^{(n+1)} = b$.*

Pour assurer la convergence et la qualité de la méthode, on va chercher maintenant à répondre aux questions suivantes :

1. la suite $(x^{(n)})_n$ est–elle bien définie ? A–t–on $Dg(x^{(n)})$ inversible ?
2. A–t–on convergence $x^{(n)} \to \bar{x}$ quand $n + \infty$?
3. La convergence est-elle au moins quadratique ?

Théorème 2.16 (Convergence de la méthode de Newton, I) *Soient $g \in C^2(\mathbb{R}^N, \mathbb{R}^N)$ et $\bar{x} \in \mathbb{R}^N$ tels que $g(\bar{x}) = 0$. On munit $\mathbb{R}^N$ d'une norme $\|\cdot\|$. On suppose que $Dg(\bar{x})$ est inversible. Alors il existe $b > 0$, et $\beta > 0$ tels que*

1. *si $x^{(0)} \in B(\bar{x}, b) = \{x \in \mathbb{R}^N, \|x - \bar{x}\| < b\}$ alors la suite $(x^{(n)})_{n\in\mathbb{N}}$ est bien définie par* (2.2.6) *et $x^{(n)} \in B(\bar{x}, b)$ pour tout $n \in \mathbb{N}$,*
2. *si $x^{(0)} \in B(\bar{x}, b)$ et si la suite $(x^{(n)})_{n\in\mathbb{N}}$ est définie par* (2.2.6) *alors $x^{(n)} \to \bar{x}$ quand $n \to +\infty$,*
3. *si $x^{(0)} \in B(\bar{x}, b)$ et si la suite $(x^{(n)})_{n\in\mathbb{N}}$ est définie par* (2.2.6) *alors $\|x^{(n+1)} - \bar{x}\| \leq \beta\|x^{(n)} - \bar{x}\|^2 \; \forall n \in \mathbb{N}$.*

Pour démontrer ce théorème, on va commencer par démontrer le théorème suivant, qui utilise des hypothèses plus faibles mais pas très faciles à vérifier en pratique :

Théorème 2.17 (Convergence de la méthode de Newton, II)
Soient $g \in C^1(\mathbb{R}^N, \mathbb{R}^N)$ et $\bar{x} \in \mathbb{R}^N$ tels que $g(\bar{x}) = 0$. On munit $\mathbb{R}^N$ d'une norme $\|\cdot\|$ et $\mathcal{M}_N(\mathbb{R})$ de la norme induite. On suppose que $Dg(\bar{x})$ est inversible. On suppose de plus qu'il existe $a,\ a_1,\ a_2 \in \mathbb{R}_+^*$ tels que :

1. *si $x \in B(\bar{x}, a)$ alors $Dg(x)$ est inversible et $\|Dg(x))^{-1}\| \leq a_1$;*
2. *si $x, y \in B(\bar{x}, a)$ alors $\|g(y) - g(x) - Dg(x)(y - x)\| \leq a_2\|y - x\|^2$.*

Alors, si on pose : $b = \min\left(a, \dfrac{1}{a_1 a_2}\right) > 0$, $\beta = a_1 a_2$ et si $x^{(0)} \in B(\bar{x}, b)$, on a :

1. *$(x^{(n)})_{n\in\mathbb{N}}$ est bien définie par* (2.2.6),
2. *$x^{(n)} \to \bar{x}$ lorsque $n \to +\infty$,*
3. *$\|x^{(n+1)} - \bar{x}\| \leq \beta\|x^{(n)} - \bar{x}\|^2 \;\; \forall n \in \mathbb{N}$.*

Démonstration du théorème 2.17
Soit $x^{(0)} \in B(\bar{x}, b) \subset B(\bar{x}, a)$ où $b \leq a$. On va montrer par récurrence sur n que $x^{(n)} \in B(\bar{x}, b)\ \forall n \in \mathbb{N}$ (et que $(x^{(n)})_{n\in\mathbb{N}}$ est bien définie). L'hypothèse de récurrence est que $x^{(n)}$ est bien défini, et que $x^{(n)} \in B(\bar{x}, b)$. On veut montrer que $x^{(n+1)}$ est bien défini et $x^{(n+1)} \in B(\bar{x}, b)$.
Comme $b \leq a$, la matrice $Dg(x^{(n)})$ est inversible et $x^{(n+1)}$ est donc bien défini ; on a : $x^{(n+1)} - x^{(n)} = Dg(x^{(n)})^{-1}(-g(x^{(n)}))$. Pour montrer que $x^{(n+1)} \in B(\bar{x}, b)$ on va utiliser le fait que $b \leq \dfrac{1}{a_1 a_2}$.
Par hypothèse, on sait que si $x, y \in B(\bar{x}, a)$, on a

$$\|g(y) - g(x) - Dg(x)(y - x)\| \leq a_2\|y - x\|^2.$$

Prenons $y = \bar{x}$ et $x = x^{(n)} \in B(\bar{x}, a)$ dans l'inégalité ci-dessus. On obtient alors :

$$\|g(\bar{x}) - g(x^{(n)}) - Dg(x^{(n)})(\bar{x} - x^{(n)})\| \leq a_2 \|\bar{x} - x^{(n)}\|^2.$$

Comme $g(\bar{x}) = 0$ et par définition de $x^{(n+1)}$, on a donc :

$$\|Dg(x^{(n)})(x^{(n+1)} - x^{(n)}) - Dg(x^{(n)})(\bar{x} - x^{(n)})\| \leq a_2 \|\bar{x} - x^{(n)}\|^2,$$

et donc

$$\|Dg(x^{(n)})(x^{(n+1)} - \bar{x})\| \leq a_2 \|\bar{x} - x^{(n)}\|^2. \tag{2.2.7}$$

Or $x^{(n+1)} - \bar{x} = [Dg(x^{(n)})]^{-1}(Dg(x^{(n)}))(x^{(n+1)} - \bar{x})$, et donc

$$\|x^{(n+1)} - \bar{x}\| \leq \|Dg(x^{(n)})^{-1}\| \ \|Dg(x^{(n)})(x^{(n+1)} - \bar{x})\|.$$

En utilisant (2.2.7), les hypothèses 1 et 2 et le fait que $x^{(n)} \in B(\bar{x}, b)$, on a donc

$$\|x^{(n+1)} - \bar{x}\| \leq a_1 a_2 \|x^{(n)} - \bar{x}\|^2 < a_1 a_2 b^2. \tag{2.2.8}$$

Or $a_1 a_2 b^2 < b$ car $b \leq \dfrac{1}{a_1 a_2}$. Donc $x^{(n+1)} \in B(\bar{x}, b)$.
On a ainsi montré par récurrence que la suite $(x^{(n)})_{n \in \mathbb{N}}$ est bien définie et que $x^{(n)} \in B(\bar{x}, b)$ pour tout $n \geq 0$.
Pour montrer la convergence de la suite $(x^{(n)})_{n \in \mathbb{N}}$ vers $\bar{x}$, on repart de l'inégalité (2.2.8) :

$$a_1 a_2 \|x^{(n+1)} - \bar{x}\| \leq (a_1 a_2)^2 \|\bar{x} - x^{(n)}\|^2 = (a_1 a_2 \|x^{(n)} - \bar{x}\|)^2, \ \forall n \in \mathbb{N},$$

et donc par récurrence $a_1 a_2 \|x^{(n)} - \bar{x}\| \leq (a_1 a_2 \|x^{(0)} - \bar{x}\|)^{2^n} \quad \forall n \in \mathbb{N}$. Comme $x^{(0)} \in B(\bar{x}, b)$ et $b \leq \frac{1}{a_1 a_2}$, on a $(a_1 a_2 \|x^{(0)} - \bar{x}\|) < 1$ et donc $\|x^{(n)} - \bar{x}\| \to 0$ quand $n \to +\infty$.

La convergence est au moins quadratique car l'inégalité (2.2.8) s'écrit : $\|x^{(n+1)} - \bar{x}\| \leq \beta \|x^{(n)} - x\|^2$ avec $\beta = a_1 a_2$. ■

Démonstration du théorème 2.16
Soient $g \in C^2(\mathbb{R}^N, \mathbb{R}^N)$ et $\bar{x} \in \mathbb{N}$ tels que $g(\bar{x}) = 0$. Par hypothèse, $Dg(\bar{x})$ est inversible. Il suffit de démontrer (pour se ramener au théorème 2.17) qu'il existe $a, a_1, a_2 \in \mathbb{R}_+^*$ tels que

1. si $x \in B(\bar{x}, a)$ alors $Dg(x)$ est inversible et $\|(Dg(x))^{-1}\| \leq a_1$,
2. si $x, y \in B(\bar{x}, a)$ alors $\|g(y) - g(x) - Dg(x)(y - x)\| \leq a_2 \|y - x\|^2$.

Remarquons d'abord que $Dg(x) = Dg(\bar{x}) - Dg(\bar{x}) + Dg(x) = Dg(\bar{x})(Id + S)$ 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 $\|S\| \leq \dfrac{1}{2}$. 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 $\|Dg(x)^{-1}\| \leq \|(Dg(\overline{x}))^{-1}\| \|(Dg(x))^{-1} Dg(\overline{x})\| \leq 2\|(Dg(\overline{x}))^{-1}\|$.

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 si $x, y \in B(\bar{x}, a)$ alors $\|g(y) - g(x) - Dg(x)(y-x)\| \leq a_2\|y-x\|^2$.
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 qu'on veut majorer en 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

$$\begin{aligned} \|\varphi(1) - \varphi(0)| &= \|g(y) - g(x) - Dg(x)(y-x)\| \\ &\leq \int_0^1 \|Dg(x + t(y-x))(y-x) - Dg(x)(y-x)\|dt \\ &\leq \|y-x\| \int_0^1 \|Dg(x + t(y-x)) - Dg(x)\|dt. \end{aligned} \tag{2.2.9}$$

Pour majorer $\|Dg(x + t(y-x)) - Dg(x))\|$, on utilise alors le théorème des accroissements finis[3] (parfois aussi appelé "théorème de la moyenne") appliqué à Dg ; de l'inégalité (2.2.9), on tire donc que pour $x, y \in B(\bar{x}, a)$ et $t \in]0, 1[$:

$$\|Dg(x + t(y-x)) - Dg(x)\| \leq \qquad t\|y-x\| \sup_{c \in B(\bar{x}, a)} \|D(Dg)(c)\|_{\mathcal{L}(\mathbb{R}^N, \mathcal{M}_N(\mathbb{R}))}. \tag{2.2.10}$$

Comme $D(Dg) = D^2 g$ 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 (2.2.10) que :

$$\|Dg(x + t(y-x) - Dg(x))\| \leq a_2\|y-x\|,$$

et de l'inégalité (2.2.9) on déduit ensuite que

$$\|g(y) - g(x) - Dg(x)(y-x)\| \leq \int_0^1 a_2\|y-x\|dt \, \|y-x\| = a_2\|y-x\|^2.$$

On a donc ainsi démontré que g vérifie les hypothèses du théorème 2.17, ce qui termine la démonstration du théorème 2.16.

Remarque 2.18 *On ne sait pas bien estimer b dans le théorème 2.16, et ceci peut poser problème lors de l'implantation numérique : il faut choisir l'itéré initial $x^{(0)}$ "suffisamment proche" de $\overline{x}$ pour avoir convergence.*

[3] **Théorème des accroissements finis :** Soient E et F des espaces vectoriels normés, soient $h \in C^1(E, F)$ et $(x, y) \in E^2$. 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 : 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. $h(y) - h(x) = Dh(\xi)(y-x)$.

2.2.2 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 $x^{(0)}$ n'a pas été choisi "suffisamment proche" de $\overline{x}$, 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 $Dg(x^{(n)})$ à chaque itération, et d'autre part la factoriser pour résoudre le système linéaire $Dg(x^{(n)})(x^{(n+1)} - x^{(n)}) = -g(x^{(n)})$. (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 á 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 $\bar{x} \in \mathbb{R}$ tels que $g(\bar{x}) = 0$. On cherche à calculer $\bar{x}$. Si on le fait par la méthode de Newton, l'algorithme s'écrit :

$$\left\{ \begin{array}{l} x^{(0)} \in \mathbb{R}^N, \\ Dg(x^{(n)})(x^{(n+1)} - x^{(n)}) = -g(x^{(n)}), \ n \geq 0. \end{array} \right.$$

La méthode du "Faux quasi-Newton" (parfois appelée quasi-Newton) consiste à remplacer le calcul de la matrice jacobienne $Dg(x^{(n)})$ 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 $(x^{(n)})_{n\in\mathbb{N}}$ de la manière suivante :

$$\left\{ \begin{array}{l} x^{(0)} \in \mathbb{R}^N \\ Dg(x^{(n_i)})(x^{(n+1)} - x^{(n)}) = -g(x^{(n)}) \text{ si } n_i \leq n < n_{i+1}. \end{array} \right. \tag{2.2.11}$$

Avec cette méthode, on a moins de calculs et de factorisations de la matrice jacobienne $Dg(x)$ 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 : $g(x) = Ax + F_1(x) + F_2(x)$, avec $A \in \mathcal{M}_N(\mathbb{R})$ et $F_1, F_2 \in C^1(\mathbb{R}^N, \mathbb{R}^N)$. L'algorithme de Newton (2.2.6) s'écrit alors :

$$\left\{ \begin{array}{l} x^{(0)} \in \mathbb{R}^N \\ \big(A + DF_1(x^{(n)}) + DF_2(x^{(n)})\big)\,(x^{(n+1)} - x^{(n)}) = \\ \qquad\qquad -Ax^{(n)} - F_1(x^{(n)}) - F_2(x^{(n)}). \end{array} \right.$$

La méthode de Newton incomplet consiste à ne pas tenir compte de la jacobienne de F_2.

$$\left\{ \begin{array}{l} x^{(0)} \in \mathbb{R}^N \\ (A + DF_1(x^{(n)}))(x^{(n+1)} - x^{(n)}) = -Ax^{(n)} - F_1(x^{(n)}) - F_2(x^{(n)}). \end{array} \right. \tag{2.2.12}$$

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 $DF_2(x)$, 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 $A + DF_1(x^{(n)})$ soit "meilleure" que celle de la matrice $A + DF_1(x^{(n)}) + DF_2(x^{(n)})$; 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 $A + DF_1(x^{(n)})$ 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 $f(x,y) = (a,b)^t$ où $(a,b)^t \in \mathbb{R}^2$ 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 2.19 *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 $A(x^{(n+1)} - x^{(n)}) = -Ax^{(n)} - F_2(x^{(n)})$, elle s'écrit alors $Ax^{(n+1)} = -F_2(x^{(n)})$, ou encore $x^{(n+1)} = -A^{-1}F_2(x^{(n)})$ si A inversible. 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 $\overline{x} \in \mathbb{R}$ tel que $g(\overline{x}) = 0$ s'écrit :

$$\left\{ \begin{array}{l} x^{(0)} \in \mathbb{R} \\ g'(x^{(n)})(x^{(n+1)} - x^{(n)}) = -g(x^{(n)}), \quad \forall n \geq 0. \end{array} \right.$$

On aimerait simplifier le calcul de $g'(x^{(n)})$, c'est–à–dire remplacer $g'(x^{(n)})$ par une quantité "proche" sans calculer g'. Pour cela, on remplace la dérivée par un quotient différentiel. On obtient la méthode de la sécante :

$$\left\{ \begin{array}{l} x^{(0)}, x^{(1)} \in \mathbb{R} \\ \dfrac{g(x^{(n)}) - g(x^{(n-1)})}{x^{(n)} - x^{(n-1)}}(x^{(n+1)} - x^{(n)}) = -g(x^{(n)}) \quad n \geq 1. \end{array} \right. \tag{2.2.13}$$

Remarquons que dans la méthode de la sécante, $x^{(n+1)}$ dépend de $x^{(n)}$ et de $x^{(n-1)}$: 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 59 page 102) que si $g(\bar{x}) = 0$ et $g'(\bar{x}) \neq 0$, il existe $\alpha > 0$ tel que si $x^{(0)}, x^{(1)} \in [\bar{x} - \alpha.\bar{x} + \alpha] = I_\alpha$, la suite $(x^{(n)})_{n\in\mathbb{N}}$ construite par la méthode de la sécante (2.2.13) est bien définie, que $(x^{(n)})_{n\in\mathbb{N}} \subset I_\alpha$ et que $x^{(n)} \to \bar{x}$ quand $n \to +\infty$. De plus, la convergence est super linéaire, *i.e.* si $x^{(n)} \neq \bar{x}$ pour tout $n \in \mathbb{N}$, alors $\dfrac{x^{(n+1)} - \bar{x}}{x^{(n)} - \bar{x}} \to 0$ quand $n \to +\infty$. On peut même montrer (voir exercice 59 page 102) 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 $Dg(x^{(n)})$ dans la méthode de Newton (2.2.6), on va remplacer $Dg(x^{(n)})$ par $B^{(n)} \in \mathcal{M}_N(\mathbb{R})$ "proche de $Dg(x^{(n)})$". En s'inspirant de la méthode de la sécante en dimension 1, on cherche une matrice $B^{(n)}$ qui, $x^{(n)}$ et $x^{(n-1)}$ étant connus, vérifie la condition :

$$B^{(n)}(x^{(n)} - x^{(n-1)}) = g(x^{(n)}) - g(x^{(n-1)}) \tag{2.2.14}$$

Dans le cas où $N = 1$, cette condition détermine entièrement $B^{(n)}$: $B^{(n)} = \dfrac{g(x^{(n)}) - g(x^{(n-1)})}{x^{(n)} - x^{(n-1)}}$. Si $N > 1$, ceci ne permet pas de déterminer complètement $B^{(n)}$. Il y a plusieurs façons possibles de choisir $B^{(n)}$, 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[4].. Celle-ci consiste à choisir $B^{(n)}$ de la manière suivante : à $x^{(n)}$ et $x^{(n-1)}$ connus, on pose $\delta^{(n)} = x^{(n)} - x^{(n-1)}$ et $y^{(n)} = g(x^{(n)}) - g(x^{(n-1)})$; on suppose $B^{(n-1)} \in \mathcal{M}_N(\mathbb{R})$ connue, et on cherche $B^{(n)} \in \mathcal{M}_N(\mathbb{R})$ telle que

$$B^{(n)}\delta^{(n-1)} = y^{(n-1)} \tag{2.2.15}$$

(c'est la condition (2.2.14), qui ne suffit pas à déterminer $B^{(n)}$ de manière unique) et qui vérifie également :

$$B^{(n)}\xi = B^{(n-1)}\xi, \quad \forall \xi \in \mathbb{R}^N \text{ tel que } \xi \perp \delta^{(n)}. \tag{2.2.16}$$

Proposition 2.20 (Existence et unicité de la matrice de Broyden)
Soient $y^{(n)} \in \mathbb{R}^N$, $\delta^{(n)} \in \mathbb{R}^N$ et $B^{(n-1)} \in \mathcal{M}_N(\mathbb{R})$. Il existe une unique matrice $B^{(n)} \in \mathcal{M}_N(\mathbb{R})$ vérifiant (2.2.15) *et* (2.2.16)*; la matrice $B^{(n)}$ s'exprime en fonction de $y^{(n)}$, $\delta^{(n)}$ et $B^{(n-1)}$ de la manière suivante :*

$$B^{(n)} = B^{(n-1)} + \frac{y^{(n)} - B^{(n-1)}\delta^{(n)}}{\delta^{(n)} \cdot \delta^{(n)}}(\delta^{(n)})^t. \tag{2.2.17}$$

Démonstration : L'espace des vecteurs orthogonaux à $\delta^{(n)}$ est de dimension $N-1$. Soit $(\gamma_1, \ldots, \gamma_{N-1})$ une base de cet espace, alors $(\gamma_1, \ldots, \gamma_{N-1}, \delta^{(n)})$ est une base de $\mathbb{R}^N$ et si $B^{(n)}$ vérifie (2.2.15) et (2.2.16), les valeurs prises par l'application linéaire associée à $B^{(n)}$ sur chaque vecteur de base sont connues, ce qui détermine l'application linéaire et donc la matrice $B^{(n)}$ de manière unique. Soit $B^{(n)}$ définie par (2.2.17), on a :

$$B^{(n)}\delta^{(n)} = B^{(n-1)}\delta^{(n)} + \frac{y^{(n)} - B^{(n-1)}\delta^{(n)}}{\delta^{(n)} \cdot \delta^{(n)}}(\delta^{(n)})^t\delta^{(n)} = y^{(n)},$$

et donc $B^{(n)}$ vérifie (2.2.15). Soit $\xi \in \mathbb{R}^N$ tel que $\xi \perp \delta^{(n)}$, alors $\xi \cdot \delta^{(n)} = (\delta^{(n)})^t\xi = 0$ et donc

$$B^{(n)}\xi = B^{(n-1)}\xi + \frac{(y^{(n)} - B^{(n-1)}\delta^{(n)})}{\delta^{(n)} \cdot \delta^{(n)}}(\delta^{(n)})^t\xi = B^{(n-1)}\xi, \ \forall \xi \perp \delta^{(n)}.$$

■

L'algorithme de Broyden s'écrit donc :

$$\begin{cases} \text{Initialisation : } x^{(0)},\ x^{(1)} \in \mathbb{R}^N \ \ B_0 \in \mathcal{M}_N(\mathbb{R}) \\ \text{A } x^{(n)},\ x^{(n-1)} \text{ et } B^{(n-1)} \text{ connus, on pose} \\ \qquad \delta^{(n)} = x^{(n)} - x^{(n-1)} \text{ et } y^{(n)} = g(x^{(n)}) - g(x^{(n-1)}); \\ \text{Calcul de } B^{(n)} = B^{(n-1)} + \frac{y^{(n)} - B^{(n-1)}\delta^{(n)}}{\delta^{(n)} \cdot \delta^{(n)}}(\delta^{(n)})^t, \\ \text{Itération } n \text{ : résolution de : } B^{(n)}(x^{(n+1)} - x^{(n)}) = -g(x^{(n)}). \end{cases}$$

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 .

[4] C. G. Broyden, "A Class of Methods for Solving Nonlinear Simultaneous Equations." *Math. Comput.* 19, 577-593, 1965

2.3 Exercices

Exercice 43 (Calcul différentiel) *Suggestions en page 103, corrigé détaillé en page 105*

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 44 (Calcul différentiel, suite)

1. Soit $f \in C^2(\mathbb{R}^2, \mathbb{R})$ la fonction définie par $f(x_1, x_2) = ax_1 + bx_2 + cx_1x_2$, où a, b, et c sont trois réels fixés. Donner la définition et l'expression de $Df(x)$, $\nabla f(x)$,Df, $D^2f(x)$, $H_f(x)$.
2. Même question pour la fonction $f \in C^2(\mathbb{R}^3, \mathbb{R})$ définie par $f(x_1, x_2, x_3) = x_1^2 + x_1^2x_2 + x_2 \sin(x_3)$.

Exercice 45 (Méthode de monotonie) *Suggestions en page 103, corrigé détaillé en page 105.*

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 $u_1, \ldots, u_N$) :

$$\begin{array}{l} (Au)_i = \alpha_i f(u_i) + \lambda b_i \ \forall i \in \{1, \ldots, N\}, \\ u = (u_1, \ldots, u_N)^t \in \mathbb{R}^N, \end{array} \tag{2.3.18}$$

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

$$u \in \mathbb{R}^N, \ Au \geq 0 \Rightarrow u \geq 0. \tag{2.3.19}$$

On suppose qu'il existe $\mu > 0$ t.q. (2.3.18) 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 $(v^{(n)})_{n \in \mathbb{N}} \subset \mathbb{R}^N$ par $v^{(0)} = 0$ et, pour $n \geq 0$,

$$(Av^{(n+1)})_i = \alpha_i f(v_i^{(n)}) + \lambda b_i \ \forall i \in \{1, \ldots, N\}. \tag{2.3.20}$$

Montrer que la suite $(v^{(n)})_{n \in \mathbb{N}}$ est bien définie, convergente (dans $\mathbb{R}^N$) et que sa limite, notée $u^{(\lambda)}$, est solution de (2.3.18) (et vérifie $0 \leq u^{(\lambda)} \leq u^{(\mu)}$).

Exercice 46 (Point fixe amélioré) *Suggestions en page 103,*

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 :

$$\left\{ \begin{array}{l} x_0 \in \mathbb{R}, \\ \\ x_{n+1} = h(x_n), n \geq 0. \end{array} \right. \tag{2.3.21}$$

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 (2.3.21) 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 (2.3.21) 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 $(x_n)_{n\in\mathbb{N}}$ définie par (2.3.21) 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 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 (2.3.21) converge de manière cubique.

Exercice 47 (Nombre d'itérations fini pour Newton) *Corrigé détaillé en page 106*

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 $(x^{(n)})_{n\in\mathbb{N}}$ 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 $(x^{(n)})_{n\in\mathbb{N}}$ est bien définie.

1.2 Montrer que si $x^{(0)} \neq 0$, alors $x^{(n+1)} \neq x^{(n)}$ 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 $(x^{(n)})_{n\in\mathbb{N}}$ converge lorsque n tend vers l'infini et donner sa limite.

2. Soit $F : \mathbb{R}^N \to \mathbb{R}^N$ 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 48 (Newton et logarithme) *Suggestions en page 103*

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 49 (Méthode de Newton pour le calcul de l'inverse) *Corrigé en page 106*

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 :

$$\begin{cases} x^{(0)} \text{ donné,} \\ x^{(n+1)} = x^{(n)}(2 - ax^{(n)}). \end{cases} \tag{2.3.22}$$

   (b) Montrer que la suite $(x^{(n)})_{n\in\mathbb{N}}$ définie par (2.3.22) vérifie

$$\lim_{n\to+\infty} x^{(n)} = \begin{cases} \dfrac{1}{a} \text{ si } x^{(0)} \in]0, \dfrac{2}{a}[, \\ -\infty \text{ si } x^{(0)} \in]-\infty, 0[\cup]\dfrac{2}{a}, +\infty[\end{cases}$$

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 $B^{(n)}$ la suite ainsi définie.

   (d) Montrer que la suite $B^{(n)}$ définie dans la question précédente vérifie :

   $$Id - AB^{(n+1)} = (Id - AB^{(n)})^2.$$

   En déduire que la suite $(B^{(n)})_{n\in\mathbb{N}}$ converge vers A^{-1} si et seulement si $\rho(Id - AB^{(0)}) < 1$.

Exercice 50 (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 $A = P^{-1}\text{diag}(\lambda_1, \cdots, \lambda_N)P$.

   (i) Montrer que pour tout $k \in \mathbb{N}$, $X^{(k)}$ est bien définie et vaut $X^{(k)} = P^{-1}\text{diag}(\mu_1^{(k)}, \cdots, \mu_N^{(k)})P$ où $\mu_i^{(k)}$ est le $k^{ième}$ terme de la suite de Newton associé à la fonction f_{λ_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 51 (Valeurs propres et méthode de Newton)
Suggestions en page 103, corrigé détaillé en page 108

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 :

$$\begin{array}{l} Ax - \lambda x = 0, \\ x \cdot x = 1. \end{array} \tag{2.3.23}$$

Montrer que la méthode est localement convergente.

Exercice 52 (Modification de la méthode de Newton) *Suggestions en page 103.*

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 $n \geq 0$,

$$x^{(n+1)} = x^{(n)} - [Df(x^{(n)})^t Df(x^{(n)}) + \lambda Id]^{-1} Df(x^{(n)})^t f(x^{(n)}).$$

[Noter que, pour $\lambda = 0$, on retrouve la méthode de Newton.]

1. Montrer que la suite $(x^{(n)})_{n\in\mathbb{N}}$ 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 53 (Convergence de la méthode de Newton si $f'(\overline{x}) = 0$)
Suggestions en page 104, corrigé détaillé en page 109

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 54 (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 :

$$\left\{ \begin{array}{l} x_0 \in \mathbb{R}, \\ \\ x_{n+1} = h(x_n), n \geq 0. \end{array} \right. \tag{2.3.24}$$

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 (2.3.24) 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 (2.3.24) 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 55 (Variante de la méthode de Newton)
Corrigé détaillé en page 109

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 $(x^{(n)})_{n\in\mathbb{N}}$ par :

$$\begin{aligned} x^{(0)} &= x_0, \\ x^{(n+1)} &= x^{(n)} - \frac{f(x^{(n)})}{f'(y)}, \end{aligned} \tag{2.3.25}$$

où $y \in I$ est choisi arbitrairement.

1. Montrer par récurrence que la suite définie par (2.3.25) satisfait $x^{(n)} \in I$ pour tout $n \in \mathbb{N}$.
(On pourra remarquer que si $x^{(n+1)}$ est donné par (2.3.25) *alors* $x^{(n+1)} - x_0 = x^{(n)} - x_0 - \frac{f(x^{(n)}) - f(x_0)}{f'(y)} - \frac{f(x_0)}{f'(y)}$.*)*
2. Montrer que la suite $(x^{(n)})_{n\in\mathbb{N}}$ définie par (2.3.25) vérifie $|x^{(n)} - \bar{x}| \leq \frac{c}{2^n}$ et qu'elle converge vers $\bar{x}$ de manière au moins linéaire.
3. On remplace l'algorithme (2.3.25) par

$$\begin{aligned} x^{(0)} &= x_0, \\ x^{(n+1)} &= x^{(n)} - \frac{f(x^{(n)})}{f'(y^{(n)})}, \end{aligned} \tag{2.3.26}$$

où la suite $(y^{(n)})_{n\in\mathbb{N}}$ est une suite donnée d'éléments de I. Montrer que la suite $(x^{(n)})_{n\in\mathbb{N}}$ 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 (2.3.25) ou (2.3.26) peut-elle se généraliser, avec d'éventuelles modifications des hypothèses, à la dimension N ?

Exercice 56 (Méthode de Newton pour un système semi-linéaire) *Suggestions en page 104.*

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 $u_1,\dots,u_N$) :

$$\begin{array}{l}(Au)_i + \alpha_i f(u_i) = b_i \ \forall i \in \{1,\dots,N\},\\ u = (u_1,\dots,u_N)^t \in \mathbb{R}^N,\end{array} \tag{2.3.27}$$

où $A \in \mathcal{M}_N(\mathbb{R})$ est une matrice symétrique définie positive, $\alpha_i > 0$ pour tout $i \in \{1,\dots,N\}$ et $b_i \in \mathbb{R}$ pour tout $i \in \{1,\dots,N\}$.

On admet que (2.3.27) admet au moins une solution (ceci peut être démontré mais est difficile).

1. Montrer que (2.3.27) admet une unique solution.
2. Soit u la solution de (2.3.27). Montrer qu'il existe $a > 0$ t.q. la méthode de Newton pour approcher la solution de (2.3.27) converge lorsque le point de départ de la méthode, noté $u^{(0)}$, vérifie $|u - u^{(0)}| < a$.

Exercice 57 (Méthode de Steffensen)
Suggestions en page 104, corrigé détaillé en page 112

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 $f(x^{(n)} + f(x^{(n)})) \neq f(x^{(n)})$,

$$x^{(n+1)} = x^{(n)} - \frac{(f(x^{(n)}))^2}{f(x^{(n)} + f(x^{(n)})) - f(x^{(n)})}, \tag{2.3.28}$$

et si $f(x^{(n)} + f(x^{(n)})) = f(x^{(n)})$, $x^{(n+1)} = x^{(n)}$.

1. Montrer qu'il existe $\alpha > 0$ tel que si $x^{(n)} \in B(\overline{x},\alpha)$, alors $f(x^{(n)} + f(x^{(n)})) \neq f(x^{(n)})$ si $x^{(n)} \neq \overline{x}$. En déduire que si $x_0 \in B(\overline{x},\alpha)$, alors toute la suite $(x^{(n)})_{n\in\mathbb{N}}$ vérifie (2.3.28) pourvu que $x^{(n)} \neq \overline{x}$ 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

$$x^{(n+1)} - \overline{x} = a(x^{(n)})(x^{(n)} - \overline{x}), \text{ pour tout } n \in \mathbb{N} \text{ tel que } x^{(n)} \neq \overline{x}. \tag{2.3.29}$$

3. Montrer que la méthode est localement convergente en $\overline{x}$ et la convergence est au moins d'ordre 2.

Exercice 58 (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 :

$$x_{n+1} = x_n - \frac{g(x_n)}{g'(x_n)} - \frac{g''(x_n)[g(x_n)]^2}{2[g'(x_n)]^3}. \tag{2.3.30}$$

Montrer que la méthode (2.3.30) 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 59 (Méthode de la sécante) *Corrigé en page 2.5 page 114*

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.

2.4 Suggestions

Exercice 43 page 96 (Calcul différentiel)

1. Utiliser le fait que $Df(x)$ est une application linéaire et le théorème de Riesz. Appliquer ensuite la différentielle à un vecteur h bien choisi.
2. Mêmes idées...

Exercice 45 page 96 (Méthode de monotonie)

Pour montrer que la suite $(v^{(n)})_{n\in\mathbb{N}}$ 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 46 page 96 (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

$$|x_{n+1} - \overline{x}| = (x_n - \overline{x})(1 - \frac{g(x_n) - g(\overline{x})}{(x_n - \overline{x})g'(\varphi(x_n))}. \tag{2.4.31}$$

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 48 page 97 (Newton et logarithme)

Etudier les variations de la fonction φ définie par : $\varphi(x) = x - x\ln x$.

Exercice 51 page 98 (Valeurs propres et méthode de Newton)

Ecrire 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 52 page 99 (Modification de la méthode de Newton)

1. Remarquer que si $A \in \mathcal{M}_N(\mathbb{R})$ et $\lambda > 0$, alors $A^tA + \lambda Id$ 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 53 page 99 (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 56 page 100 (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 57 page 101 (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.

2.5 Corrigés

Exercice 43 page 96

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 $h_j^{(i)} = h\delta_{i,j}$, où $h > 0$ et $\delta_{i,j}$ désigne le symbole de Kronecker, i.e. $\delta_{i,j} = 1$ si $i = j$ et $\delta_{i,j} = 0$ sinon. En appliquant la définition de la différentielle avec $h^{(i)}$, on obtient :

$$f(x + h^{(i)}) - f(x) = Df(x)(h(i)) + \|h^{(i)}\|\varepsilon(h^{(i)}),$$

c'est–à–dire :

$$f(x_1, \ldots, x_{i-1}, x_i + h, x_{i-1}, \ldots, x_N) - f(x_1, \ldots, x_N) = (a(x))_i h + h\varepsilon(h^{(i)}).$$

En divisant par h et en faisant tendre h vers 0, on obtient alors que $(a(x))_i = \partial_i f(x)$.

2. Comme $f \in C^2(\mathbb{R}^N, \mathbb{R})$, on a $(\partial_i f(x) \in C^1(\mathbb{R}^N, \mathbb{R})$, et donc $\varphi \in C^1(\mathbb{R}^N, \mathbb{R}^N)$. Comme $D\varphi(x)$ est une application linéaire de $\mathbb{R}^N$ dans $\mathbb{R}^N$, il existe une matrice $A(x)$ carrée d'ordre N telle que $D\varphi(x)(y) = A(x)y$ pour tout $y \in \mathbb{R}^N$. Il reste à montrer que $(A(x))_{i,j} = \partial^2_{i,j} f(x)$. Soit $h^{(i)} \in \mathbb{R}^N$ défini à la question préceédente, pour $i,\ j = 1, \ldots, N$, on a

$$(D\varphi(x)(h^{(j)}))_i = (A(x)h^{(j)})_i = \sum_{k=1}^N a_{i,k}(x)h^{(j)})_k = ha_{i,j}(x).$$

Or par définition de la différentielle,

$$\varphi_i(x + h^{(j)}) - \varphi_i(x) = (D\varphi(x)(h^{(j)}))_i + \|h^{(j)}\|\varepsilon_i(h^{(j)}),$$

ce qui entraîne, en divisant par h et en faisant tendre h vers 0 : $\partial_j \varphi_i(x) = a_{i,j}(x)$. Or $\varphi_i(x) = \partial_i f(x)$, et donc $(A(x))_{i,j} = a_{i,j}(x) = \partial^2_{i,j} f(x)$.

Corrigé de l'exercice 45 page 96 (Méthode de monotonie)

Montrons que la suite $v^{(n)}$ est bien définie. Supposons $v^{(n)}$ connu ; alors $v^{(n+1)}$ est bien défini si le système

$$Av^{(n+1)} = d^{(n)},$$

où $d^{(x)}$ est défini par : $d_i^{(n)} = \alpha_i f(v_i^{(n)}) + \lambda b_i$ 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 $v^{(n+1)}$.
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 $v^{(n)} \to \bar{u}$ lorsque $n \to +\infty$, où $\bar{u}$ est tel que $A\bar{u} = R(\bar{u})$.

Corrigé de l'exercice 47 page 97 (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 :

$$x^{(n+1)} = x^{(n)} - \frac{f(x^{(n)})}{f'(x^{(n)})} = x^{(n)} - \frac{e^{x^{(n)}} - 1}{e^{x^{(n)}}}$$

est bien définie.

1.2 Par définition de la suite, on a $x^{(n+1)} - x^{(n)} = -\frac{e^{x^{(n)}}-1}{e^{x^{(n)}}} = 0$ ssi $x^{(n)} = 0$. Donc par récurrence sur n, si $x^{(0)} \neq 0$, on a $x^{(n+1)} \neq x^{(n)}$ 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 $(x^{(n)})_{n\in\mathbb{N}}$ 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 $x^{(n)}$ et $x^{(n+1)}$ deux itérés successifs donnés par la méthode de Newton, tels que $F(x^{(n)}) \neq 0$. On a donc :

$$DF(x^{(n)})(x^{(n+1)} - x^{(n)}) = -F(x^{(n)}), \tag{2.5.32}$$

et en particulier, $x^{(n+1)} \neq x^{(n)}$. Or, la condition de stricte convexité pour une fonction continûment différentiable entraï¿½ne que :

$$DF(x^{(n)})(x^{(n+1)} - x^{(n)}) < F(x^{(n+1)}) - F(x^{(n)}),$$

et donc, avec (2.5.32), $F(x^{(n+1)}) > 0$. On montre ainsi, par récurrence sur n, que si $F(x^{(0)}) \neq 0$, alors $F(x^{(n)}) > 0$ pour tout $n > 0$, ce qui montre que la méthode de Newton converge en un nombre fini d'opérations si et seulement si $F(x^{(0)}) = 0$.

Corrigé de l'exercice 49 page 97 (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 :

$$\begin{cases} x^{(0)} \text{ donné,} \\ x^{(n+1)} = x^{(n)}(2 - ax^{(n)}). \end{cases} \tag{2.5.33}$$

   (b) Soit $(x^{(n)})_{n\in\mathbb{N}}$ définie par (2.3.22). D'après le théorème du cours, on sait que la suite $(x^{(n)})_{n\in\mathbb{N}}$ converge de localement (de manière quadratique) dans un voisinage de $\frac{1}{a}$. On veut déterminer ici

l'intervalle de convergence précisément. On a $x^{(n+1)} = \varphi(x^{(n)})$ 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 :

$$\begin{array}{c|ccccc} x & & 0 & \frac{1}{a} & \frac{2}{a} & \\ \hline \varphi'(x) & & + & 0 & - & \\ \hline \varphi(x) & -\infty & \nearrow & \frac{1}{a} & \searrow & -\infty \end{array} \tag{2.5.34}$$

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 $x^{(n+1)} \geq x^{(n)}$ pour tout n, et donc la suite $(x^{(n)})_{n\in\mathbb{N}}$ 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 $\lim_{n\to+\infty} x^{(n)} = -\infty$. 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 $GL_N(\mathbb{R})(\mathbb{R})$ 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 $GL_N(\mathbb{R})(\mathbb{R})$ dans $GL_N(\mathbb{R})(\mathbb{R})$. Montrons qu'elle est dérivable. Soit $H \in GL_N(\mathbb{R})(\mathbb{R})$ 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 :

$$\begin{aligned} T(B + H) - T(B) &= \sum_{k=0}^{+\infty}(B^{-1}H)^k B^{-1} - B^{-1} \\ &= (Id + \sum_{k=1}^{+\infty}(-B^{-1}H)^k - Id)B^{-1} \\ &= \sum_{k=1}^{+\infty}(-B^{-1}H)^k B^{-1}. \end{aligned}$$

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

$$\frac{\|T(B+H) - T(B) - B^{-1}HB^{-1}}{\|H\|}\| \leq \|B^{-1}\|^3\|H\| \sum_{k=0}^{+\infty}(\|B^{-1}\|\|H\|)^k \to 0 \text{ lorsque } \|H\| \to 0.$$

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 :

$$\left\{ \begin{array}{l} B^0 \in GL_N(\mathbb{R})(\mathbb{R}), \\ Dg(B^n)(B^{n+1} - B^n) = -g(B^n). \end{array} \right.$$

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 :

$$\left\{ \begin{array}{l} B^0 \in GL_N(\mathbb{R})(\mathbb{R}), \\ -(B^{n+1} - B^n) = (Id - B^nA)B^n. \end{array} \right. \tag{2.5.35}$$

soit encore

$$\left\{ \begin{array}{l} B^0 \in GL_N(\mathbb{R})(\mathbb{R}), \\ B^{n+1} = 2B^n - B^nAB^n. \end{array} \right. \tag{2.5.36}$$

(d) Par définition, on a :

$$Id - AB^{n+1} = Id - A(2B^n - B^nAB^n) = Id - 2AB^n + AB^nAB^n.$$

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 $\rho(Id - AB^0) < 1$.

Corrigé de l'exercice 51 page 98 (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

$$F(y) = F(x, \lambda) = \left(\begin{array}{c} Ax - \lambda x \\ x \cdot x - 1 \end{array} \right),$$

et on a donc

$$DF(\overline{\lambda}, \overline{x})(z, \nu) = \left(\begin{array}{c} Az - \bar{\lambda}z - \nu\bar{x} \\ 2\bar{x} \cdot z \end{array} \right),$$

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 :

$$z \cdot A\bar{x} - \bar{\lambda}z \cdot \bar{x} - \nu\bar{x} \cdot \bar{x} = 0, \tag{2.5.37}$$

et comme $A\bar{x} = \bar{\lambda}\bar{x}$ et $\bar{x} \cdot \bar{x} = 1$, ceci entraîne que $\nu = 0$. En revenant à (2.5.37) on obtient alors que $Ax - \bar{\lambda}x = 0$, c.à.d. que $x \in Ker(A - \bar{\lambda}Id) = \mathbb{R}\bar{x}$ 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. Dnc, d'après le théorème du cours, la méthode de Newton est localement convergente.

Corrigé de l'exercice 53 page 99 (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 :

$$f'(x_n)(x_{n+1} - x_n) = -f(x_n) \tag{2.5.38}$$

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 (2.5.38) que

$$\begin{aligned} f''(\xi_n)(x_{n+1} - x_n) &= -\frac{1}{2}f''(\eta_n)(x_n - \overline{x}), \\ \text{soit } f''(\xi_n)(x_{n+1} - \overline{x}) &= (-\frac{1}{2}f''(\eta_n) + f''(\xi_n))(x_n - \overline{x}) \end{aligned}$$

$$\text{Donc } \frac{\|x_{n+1} - \overline{x}\|}{\|x_n - \overline{x}\|} = |1 - \frac{1}{2}\frac{f''(\eta_n)}{f''(\xi_n)}| \to \frac{1}{2} \text{ lorsque } n \to +\infty.$$

La méthode est donc d'ordre 1.
On peut obtenir une méthode d'ordre 2 en appliquant la méthode de Newton à f'.

Corrigé de l'exercice 55 page 100 (Variante de la méthode de Newton)

1. On a évidemment $x^{(0)} = x_0 \in I$. Supposons que $x^{(n)} \in I$ et montrons que $x^{(n+1)} \in I$. Par définition, on peut écrire :

$$x^{(n+1)} = x^{(n)} - \frac{f(x^{(n)}) - f(x_0)}{f'(y)} - \frac{f(x_0)}{f'(y)}.$$

Donc

$$x^{(n+1)} - x_0 = x^{(n)} - x_0 - \frac{f'(\xi_n)(x_n^{(n)} - x_0) - f(x_0)}{f'(y)}, \text{ où } \xi_n \in [x_0, x^{(n)}].$$

On en déduit que

$$x^{(n+1)} - x_0 = (1 - \frac{f'(\xi_n)}{f'(y)})(x_n^{(n)} - x_0) - \frac{f(x_0)}{f'(y)}.$$

Ceci entraîne :

$$\begin{aligned} |x^{(n+1)} - x_0| &= \frac{1}{|f'(y)|}|f'(\xi_n) - f'(y)||x^{(n)} - x_0| + \frac{|f(x_0)|}{f'(y)} \\ &\leq \lambda\frac{1}{2\lambda}c + \frac{c}{2\lambda}\lambda = c. \end{aligned}$$

Donc $x^{(n+1)} \in I$.

2. On a :

$$\begin{aligned} x^{(n+1)} - \bar{x} &= x^{(n)} - \bar{x} - \frac{f(x^{(n)}) - f(\bar{x})}{f'(y)} - \frac{f(\bar{x})}{f'(y)}. \\ \text{Donc} |x^{(n+1)} - \bar{x}| &\leq |x^{(n)} - \bar{x}||f'(y) - f'(\eta_n)|\frac{1}{|f'(y)|} \text{ où } \eta_n \in [\bar{x}, x^{(n)}]; \end{aligned}$$

Par hypothèse, on a donc

$$\begin{aligned} |x^{(n+1)} - \bar{x}| &\leq |x^{(n)} - \bar{x}|\frac{1}{2\lambda}\lambda \\ &\leq \frac{c}{2}|x^{(n)} - \bar{x}|. \end{aligned}$$

On en déduit par récurrence que

$$|x^{(n)} - \bar{x}| \leq \frac{c}{2^n}|x^{(0)} - \bar{x}|.$$

Ceci entraîne en particulier que

$$\begin{aligned} x^{(n)} &\to \bar{x} \\ n &\to +\infty. \end{aligned}$$

Il reste à montrer que la convergence est au moins linéaire. On a :

$$\frac{|x^{(n+1)} - \bar{x}|}{|x^{(n)} - \bar{x}|} = |f'(y) - f'(x^{(n)})|\frac{1}{|f'(y)|}$$

$$\text{Donc} \quad \frac{|x^{(n+1)} - \bar{x}|}{|x^{(n)} - \bar{x}|} \underset{n \to +\infty}{\to} |1 - \frac{f'(\bar{x})}{f'(y)}| = \beta \geq 0$$

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 $y^{(n)}$ ne change absolument rien à la preuve de la convergence de $x^{(n)}$ vers $\bar{x}$. Par contre, on a maintenant :

$$\begin{aligned} \frac{|x^{(n+1)} - \bar{x}|}{|x^{(n)} - \bar{x}|} &= |f'(y_n) - f'(\eta_n)|\frac{1}{|f'(y_n)|} \\ &= |1 - \frac{f'(\eta_n)}{f'(y_n)}| \end{aligned}$$

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. L'algorithme pour $N \geq 1$ se généralise en :

$$\left\{ \begin{array}{l} x^{(0)} = x_0 \\ x^{(n+1)} = x^{(n)} - (DF(y))^{-1} f(x^{(n)}). \end{array} \right.$$

On a donc

$$x^{(n+1)} - x_0 = x^{(n)} - x_0 - (DF(y))^{-1}(f(x^{(n)}) - f(x_0)) + (DF(y))^{-1} f(x_0).$$

Or $f(x^{(n)}) - f(x^{(0)}) = \varphi(1) - \varphi(0) = \displaystyle\int_0^1 \varphi'(t)dt$, où

$$\varphi(t) = f(tx^{(n)} + (1-t)x^{(0)})$$

et donc

$$\varphi'(t) = Df(tx^{(n)} + (1-t)x^{(0)})(x^{(n)} - x^{(0)}).$$

Donc

$$\begin{array}{l} x^{(n+1)} - x^{(0)} = \\ (x^{(n)} - x^{(0)}) \left(1 - (Df(y))^{-1} \displaystyle\int_0^1 Df(tx^n) + (1-t)x^{(0)})dt \right) \\ +(Df(y))^{-1} f(x_0) = \\ (x^{(n)} - x^{(0)})(Df(y))^{-1} \left(\displaystyle\int_0^1 \left(Df(y) - Df(tx^{(n)} + (1-t)x^{(0)} \right) dt \right) \\ +(Df(y))^{-1} f(x_0). \end{array}$$

On en déduit que :

$$\begin{array}{l} \|x^{(n+1)} - x^{(0)}\| \leq \\ \quad \|x^{(n)} - x^{(0)}\| \|(Df(y))^{-1}\| \displaystyle\int_0^1 \|Df(y) - Df(tx^{(n)} + (1-t)x^{(0)})\| dt \\ \quad +\|(Df(y))^{-1}\| \|f(x_0)\|. \end{array}$$

Si on suppose que $x^{(n)} \in I$, alors $tx^{(n)} + (1-t)x^{(0)} \in I$. 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 2.5.39 donne que

$$\begin{array}{rcl} \|x^{(n+1)} - x^{(0)}\| & \leq & \|x^{(n)} - x^{(0)}\| \lambda \dfrac{1}{2\lambda} + \lambda \dfrac{c}{2\lambda} \\ & \leq & c. \end{array}$$

ce qui prouve que $x^{(n+1)} \in I$.

On montre alors de la même manière que $x^{(n)}_{n \to \infty} \to \bar{x}$, (car $\|x^{(n+1)} - \bar{x}\| \leq \dfrac{1}{2}\|x^{(n)} - \bar{x}\|$).

Corrigé de l'exercice 57 page 101 (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 :

$$x_{n+1} - \overline{x} = a(x_n)(x_n - \overline{x})^2 \quad \text{si } x_n \neq \overline{x} \quad \text{et } x_0 \in B(\overline{x}, \alpha),$$

où a est une fonction continue. Par définition de la suite $(x_n)_{n\in\mathbb{N}}$, on a :

$$x_{n+1} - \overline{x} = x_n - \overline{x} - \frac{(f(x_n))^2}{f(x_n + f(x_r)) - f(x_n)}. \tag{2.5.39}$$

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 :

$$f(x_n + f(x_n)) - f(x_n) = f(x_n)\left[f(x_n)\int_0^1\int_0^t f''(x_n + sf(x_n))ds + f'(x_n)\right]. \tag{2.5.40}$$

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 (2.5.40) s'écrit alors :

$$f(x_n + f(x_n)) - f(x_n) = (f(x_n))^2 b(x_n) + f(x_n)f'(x_n). \tag{2.5.41}$$

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 (2.5.39) et (2.5.41) :

$$x_{n+1} - \overline{x} = x_n - \overline{x} - \frac{f(x_n)}{f(x_n)b(x_n) + f'(x_n))} \tag{2.5.42}$$

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 $c(x) \to \dfrac{1}{2} f''(\overline{x})$ lorsque $x \to \overline{x}$ et :

$$-f(x_n) = c(x)(\overline{x} - x_n)^2 + f'(x_n)(\overline{x} - x_n) \tag{2.5.43}$$

De (2.5.43) et(1.6.62), on obtient :

$$\begin{aligned} x_{n+1} - \overline{x} &= (x_n - \overline{x}) \left[1 + \frac{c(x_n)(x_n - \overline{x}) - f'(x_n)}{f(x_n)b(x_n) + f'(x_n)} \right] \\ &= \frac{(x_n - \overline{x})}{f(x_n)b(x_n) + f'(x_n)} \left(-c(x_n)(\overline{x} - x_n)^2 b(x_n) \right. \\ &\quad \left. -f'(x_n)(\overline{x} - x_n)b(x_n) + f'(x_n) + c(x_n)(x_n - \overline{x}) - f'(x_n) \right) \end{aligned}$$

On en déduit :

$$(x_{n+1} - \overline{x}) = (x_n - \overline{x})^2 a(x_n) \tag{2.5.44}$$

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
(7) $|a(x) - a(\overline{x})| \leq \varepsilon$.
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 (7), et $|x - \overline{x}| \leq \frac{1}{2(\beta+1)}$.
On déduit alors de (6) 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 (6), 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.

Corrigé de l'exercice 59 page 102 (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 :

$$\begin{aligned}(f(x_n) - f(x_{n-1})e_{n+1} &= e_n f(x_n) - e_n f(x_{n-1}) - f(x_n)e_n + f(x_n)e_{n-1} && (2.5.45)\\ &= -e_n f(x_{n-1}) + f(x_n)e_{n-1} && (2.5.46)\\ &= e_n e_{n-1}(\frac{f(x_n)}{e_n} - \frac{f(x_{n-1})}{e_{n-1}}). && (2.5.47)\end{aligned}$$

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

$$\mu'(x) = \frac{f'(x) - \mu(x)}{x - \overline{x}},\ \forall x > \overline{x}. \tag{2.5.48}$$

Or

$$\begin{aligned}\mu(x) &= \frac{1}{x - \overline{x}}(f(x) - f(\overline{x})) && (2.5.49)\\ &= \frac{1}{x - \overline{x}}(f(x) - (f(x) + (\overline{x} - x)f'(x) + \frac{1}{2}(\overline{x} - x)^2 f''(x)(\overline{x} - x)^3\varepsilon(x). && (2.5.50)\end{aligned}$$

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 (2.5.48) :

$$\mu'(x) = \frac{1}{2}f''(x) + (x - \overline{x})\varepsilon(x),\ \forall x > \overline{x}. \tag{2.5.51}$$

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 (2.5.51) et on obtient :

$$\lim_{x \to \overline{x}} \mu'(x) = \frac{1}{2}f''(\overline{x}). \tag{2.5.52}$$

(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 (2.5.52), 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 : $a_{n+1} = a_n a_{n-1} \geq M e_n M e_{n-1}$. Or par la question 2a, $e_n e_{n-1} = M_n e_{n+1} \leq M e_{n+1}$. On en déduit que la relation est encore vérifiée au rang $n+1$.

   (b) Par définition, $a_i = M e_i = M(x_i - \overline{x})$, 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

$$b_{n+1} = b_n + b_{n-1},\ \forall n \geq 1 \tag{2.5.53}$$

   L'ensemble de suites $(b_n)_{n\in\mathbb{N}}$ vérifiant (2.5.53) 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 (2.5.53) 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 (2.5.53), 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 petit que l'ordre de convergence de la méthode de Newton.

Chapitre 3

Optimisation

3.1 Définitions et rappels

3.1.1 Définition des problèmes d'optimisation

L'objectif de ce chapitre est de rechercher des minima ou des maxima d'une fonction $f \in C(\mathbb{R}^N, \mathbb{R})$ avec ou sans contrainte. Le problème d'optimisation sans contrainte s'écrit :

$$\left\{ \begin{array}{l} \text{Trouver } \bar{x} \in \mathbb{R}^N \text{ tel que :} \\ f(\bar{x}) \leq f(y), \quad \forall y \in \mathbb{R}^N. \end{array} \right. \tag{3.1.1}$$

Le problème d'optimisation avec contrainte s'écrit :

$$\left\{ \begin{array}{l} \text{Trouver } \bar{x} \in K \text{ tel que :} \\ f(\bar{x}) \leq f(y), \quad \forall y \in K. \end{array} \right. \tag{3.1.2}$$

où $K \subset \mathbb{R}^N$ et $K \neq \mathbb{R}^N$
Si $\bar{x}$ est solution du problème (3.1.1), on dit que $\bar{x} \in \arg\min_{\mathbb{R}^N} f$, et si $\bar{x}$ est solution du problème (3.1.2), on dit que $\bar{x} \in \arg\min_{K} f$.

3.1.2 Rappels et notations de calcul différentiel

Définition 3.1 *Soient E et F des espaces vectoriels normés, f une application de E dans F et $x \in E$. On dit que f est différentiable en x s'il existe $T \in \mathcal{L}(E, F)$ (où $\mathcal{L}(E, F)$ est l'ensemble des applications linéaires de E dans F) telle que $f(x + h) = f(x) + T(h) + \|h\|\varepsilon(h)$ avec $\varepsilon(h) \to 0$ quand $h \to 0$. L'application T est alors unique et on note $Df(x) = T \in \mathcal{L}(E, F)$.*

On peut remarquer qu'en dimension infinie, T dépend des normes associées à E et F. Voyons maintenant quelques cas particuliers d'espaces E et F :

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 $Df(x) \in \mathcal{L}(\mathbb{R}^N, \mathbb{R}^p)$, et il existe $A(x) \in \mathcal{M}_{p,N}(\mathbb{R})$ telle que $\underbrace{Df(x)(y)}_{\in \mathbb{R}^p} = \underbrace{Ay}_{\in \mathbb{R}^p}$, $\forall y \in \mathbb{R}^N$. On confond alors l'application linéaire $Df(x) \in \mathcal{L}(\mathbb{R}^N, \mathbb{R}^p)$ et la matrice $A(x) \in \mathcal{M}_{p,N}(\mathbb{R})$ qui la représente. On écrit donc :

$$A(x) = Df(x) = (a_{i,j})_{1 \leq i \leq p, 1 \leq j \leq N} \text{ où } a_{i,j} = \partial_j f_i(x),$$

∂_j désignant la dérivée partielle par rapport à la j-ème variable.

Exemple 3.2 *Prenons $N = 3$ et $p = 2$, notons $x = (x_1, x_2, x_3)^t$ et considérons $f : \mathbb{R}^3 \to \mathbb{R}^3$ définie par :*

$$f(x) = \begin{pmatrix} x_1^2 + x_2^3 + x_3^4 \\ 2x_1 - x_2 \end{pmatrix}$$

On vérifiera par le calcul (exercice) que pour $h == (h_1, h_2, h_3)^t$, on a :

$$Df(x)h = \begin{pmatrix} 2x_1h_1 + 3x_2^2h_2 + 4x_3^3h_3 \\ \\ 2h_1 - h_2 \end{pmatrix}$$

et donc, avec les notations précédentes,

$$A(x) = \begin{pmatrix} 2x_1 & 3x_2^2 & 4x_3^3 \\ \\ 2 & -1 & 0 \end{pmatrix}.$$

Cas où $E = \mathbb{R}^N$, $F = \mathbb{R}$ C'est 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 (avec l'abus de notation signalé dans le paragraphe précédent) $Df(x) \in \mathcal{M}_{1,N}(\mathbb{R})$, et on peut définir le gradient de f en x par $\nabla f(x) = (Df(x))^t \in \mathbb{R}^N$. Pour $(x, y) \in (\mathbb{R}^N)^2$, on a donc

$$Df(x)y = \sum_{j=i}^{N} \partial_j f(x) y_j = \nabla f(x) \cdot y \text{ où } \nabla f(x) = \begin{bmatrix} \partial_1 f(x) \\ \vdots \\ \\ \partial_N f(x) \end{bmatrix} \in \mathbb{R}^N.$$

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 $y \in E$, $Df(x)(y) = (\nabla f(x)|y)_E$.

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 $y \in E$, $D^2f(x)(y) \in \mathcal{L}(E, F)$; en particulier, 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 :

$$f \in C^2(\mathbb{R}^N, \mathbb{R}) \Leftrightarrow [f \in C^1(\mathbb{R}^N, \mathbb{R}) \text{ et } \nabla f \in C^1(\mathbb{R}^N, \mathbb{R}^N)].$$

Soit $g = \nabla f \in C^1(\mathbb{R}^N, \mathbb{R}^N)$, et $x \in \mathbb{R}^N$, alors $Dg(x) \in \mathcal{M}_N(\mathbb{R})$ et on peut définir la matrice hessienne de f, qu'on note H_f, par : $H_f(x) = Dg(x) = D(Df)(x) = (b_{i,j})_{i,j=1\ldots N} \in \mathcal{M}_N(\mathbb{R})$ où $b_{i,j} = \partial^2_{i,j} f(x)$ où $\partial^2_{i,j}$ désigne la dérivée partielle par rapport à la variable i de la dérivée partielle par rapport à la variable j. Notons que par définition, $Dg(x)$ est la matrice jacobienne de g en x.

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 soit un minimum global de f, c.à.d. :

$$\bar{x} \in E \text{ tel que } f(\bar{x}) \leq f(y) \ \forall y \in E, \tag{3.2.3}$$

ou un minimum local, c.à.d. :

$$\bar{x} \text{ tel que } \exists \alpha > 0 \;\; f(\bar{x}) \leq f(y) \;\; \forall y \in B(\bar{x}, \alpha). \tag{3.2.4}$$

Proposition 3.3 (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 (3.2.4) 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 $z \in E \setminus \{0\}$, alors 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 $\dfrac{\alpha}{\|z\|} > t > 0$, on a $Df(\bar{x})(z) + \varepsilon_z(t) \geq 0$. En faisant tendre t vers 0, on obtient que

$$Df(\bar{x})(z) \geq 0, \quad \forall z \in E.$$

On a aussi $Df(\bar{x})(-z) \geq 0 \;\; \forall z \in E$, et donc : $-Df(\bar{x})(z) \geq 0 \;\; \forall z \in E$.
On en conclut que

$$Df(\bar{x}) = 0.$$

Remarque 3.4 *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 3.5 (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

$$\forall A \in \mathbb{R}, \quad \exists R \in \mathbb{R}; \|x\| \geq R \Rightarrow f(x) \geq A. \tag{3.2.5}$$

On écrit (3.2.5) 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 3.6

1. *Le théorème est faux si E est de dimension infinie (i.e. si E est espace de Banach au lieu de $E = \mathbb{R}^N$), car si E est de dimension infinie, 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 $f(x) = x^2(x-1)(x+1)$.*

Définition 3.7 (Convexité) *Soit E un espace vectoriel et $f : E \to \mathbb{R}$. On dit que f est convexe si*

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) \text{ pour tout } (x,y) \in E^2 \text{ t.q. } x \neq y \text{ et } t \in [0,1].$$

On dit que f est strictement convexe si

$$f(tx + (1-t)y) < tf(x) + (1-t)f(y) \text{ pour tout } (x,y) \in E^2 \text{ t.q. } x \neq y \text{ et } t \in]0,1[.$$

Théorème 3.8 (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 $\bar{x} = \bar{\bar{x}}$.

Remarque 3.9 *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 3.5 et 3.8, on obtient le résultat d'existence et unicité suivant :

Théorème 3.10 (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 $f(\bar{x}) = \inf_{\mathbb{R}^N} f$.

Remarque 3.11 *Le théorème reste vrai (voir cours de maîtrise) 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 3.12 (1ère caractérisation de la convexité) *Soit E un espace vectoriel normé (sur $\mathbb{R}$) et $f \in C^1(E, \mathbb{R})$ alors :*

1. *f convexe si et seulement si $f(y) \geq f(x) + Df(x)(y-x)$, pour tout couple $(x,y) \in E^2$,*
2. *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 :

$$f(x + t(y-x)) - f(x) \leq t(f(y) - f(x)). \tag{3.2.6}$$

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 (3.2.6),

$$\varepsilon(t) + Df(x)(y-x) \leq f(y) - f(x), \quad \forall t \in]0,1[.$$

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 3.13 (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 3.3) que $Df(\bar{x}) = 0$ (la convexité est inutile).

($\Leftarrow$) Si f est convexe et différentiable, d'après la proposition 3.12, 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 $f(\bar{x}) = \inf_E f$.

Proposition 3.14 (2ème 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 3.12 :

$$f(y) \geq f(x) + \nabla f(x) \cdot (y - x). \tag{3.2.7}$$

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 :

$$f(y) - f(x) = \nabla f(x)(y-x) + \int_0^1 H_f(x + t(y-x))(y-x) \cdot (y-x)(1-t)dt. \tag{3.2.8}$$

Les inégalités (3.2.7) et (3.2.8) 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 :

$$\int_0^1 H_f(x + tz)z \cdot z(1-t)dt \geq 0 \quad \forall x, \forall z \in E. \tag{3.2.9}$$

En fixant $x \in E$, on écrit (3.2.9) 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 3.12 et montrer que : $f(y) \geq f(x) + \nabla f(x) \cdot (y - x)$ pour tout $(x, y) \in E^2$. Grâce à (3.2.8), 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 3.12. 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 car $f''(x) = 12x^2 \geq 0$, mais $f''(0) = 0$.

Cas d'une fonctionnelle quadratique Soient $A \in \mathcal{M}_N(\mathbb{R})$, $b \in \mathbb{R}^N$, et f la fonction de $\mathbb{R}^N$ dans $\mathbb{R}^N$ définie par $f(x) = \frac{1}{2}Ax \cdot x - b \cdot x$. Alors $f \in C^\infty(\mathbb{R}^N, \mathbb{R})$. Le calcul du gradient de f et de sa hessienne font l'objet de l'exercice 62 : on montre que

$$\nabla f(x) = \frac{1}{2}(Ax + A^t x) - b.$$

Donc si A est symétrique $\nabla f(x) = Ax - b$. Le calcul de la hessienne de f donne :

$$H_f(x) = D(\nabla f(x)) = \frac{1}{2}(A + A^t).$$

On en déduit que si A est symétrique, $H_f(x) = A$. On peut montrer en particulier (voir exercice 62) que si A est symétrique définie positive alors il existe un unique $\bar{x} \in \mathbb{R}^N$ tel que $f(\bar{x}) \leq f(x)$ pour tout $x \in \mathbb{R}^N$, et que ce $\bar{x}$ est aussi l'unique solution du système linéaire $Ax = b$.

3.3 Algorithmes d'optimisation sans contrainte

Soit $E = \mathbb{R}^N$ et $f \in C(E, \mathbb{R})$. On suppose qu'il existe $\bar{x} \in E$ tel que $f(\bar{x}) = \inf_E f$. On cherche à calculer $\bar{x}$ (si f est de classe C^1, on a nécessairement $\nabla f(\bar{x}) = 0$). On va donc maintenant développer des algorithmes (ou méthodes de calcul) du point $\bar{x}$ qui réalise le minimum de f.

3.3.1 Méthodes de descente

Définition 3.15 *Soient $f \in C(E, \mathbb{R})$ et $E = \mathbb{R}^N$.*

1. *Soit $x \in E$, on dit que $w \in E \setminus \{0\}$ est une direction de descente en x s'il existe $\rho_0 > 0$ tel que*

$$f(x + \rho w) \leq f(x) \ \forall \rho \in [0, \rho_0]$$

2. *Soit $x \in E$, on dit que $w \in E \setminus \{0\}$ est une direction de descente stricte en x si s'il existe $\rho_0 > 0$ tel que*

$$f(x + \rho w) < f(x) \ \forall \rho \in]0, \rho_0].$$

3. *Une "méthode de descente" pour la recherche de $\bar{x}$ tel que $f(\bar{x}) = \inf_E f$ consiste à construire une suite $(x_n)_n$ de la manière suivante :*
   (a) *Initialisation $x_0 \in E$;*
   (b) *Itération n : on suppose $x_0 \ldots x_n$ connus ($n \geq 0$) ;*

i. On cherche w_n direction de descente stricte de x_n

ii. On prend $x_{n+1} = x_n + \rho_n w_n$ avec $\rho_n > 0$ "bien choisi".

Proposition 3.16 *Soient $E = \mathbb{R}^N$, $f \in C^1(E, \mathbb{R})$, $x \in E$ et $w \in E \setminus \{0\}$; alors*

1. *si w direction de descente en x alors $w \cdot \nabla f(x) \leq 0$*
2. *si $\nabla f(x) \neq 0$ alors $w = -\nabla f(x)$ est une direction de descente stricte en x.*

Démonstration

1. Soit $w \in E \setminus \{0\}$ une direction de descente en x alors par définition,

$$\exists \rho_0 > 0 \text{ tel que } f(x + \rho w) \leq f(w), \ \ \forall \rho \in [0, \rho_0].$$

   Soit φ la fonction de $\mathbb{R}$ dans $\mathbb{R}$ définie par : $\varphi(\rho) = f(x + \rho w)$. On a $\varphi \in C^1(\mathbb{R}, \mathbb{R})$ et $\varphi'(\rho) = \nabla f(x + \rho w) \cdot w$. Comme w est une direction de descente, on peut écrire : $\varphi(\rho) \leq \varphi(0), \forall \rho \in [0, \rho_0]$, et donc

$$\forall \rho \in]0, \rho_0[, \quad \frac{\varphi(\rho) - \varphi(0)}{\rho} \leq 0;$$

   en passant à la limite lorsque ρ tend vers 0, on déduit que $\varphi'(0) \leq 0$, c.à.d. $\nabla f(x) \cdot w \leq 0$.

2. Soit $w = -\nabla f(x) \neq 0$. On veut montrer qu'il existe $\rho_0 > 0$ tel que si $\rho \in]0, \rho_0]$ alors $f(x + \rho w) < f(x)$ ou encore que $\varphi(\rho) < \varphi(0)$ où φ est la fonction définie en 1 ci-dessus. On a : $\varphi'(0) = \nabla f(x) \cdot w = -|\nabla f(x)|^2 < 0$. Comme φ' est continue, il existe $\rho_0 > 0$ tel que si $\rho \in [0, \rho_0]$ alors $\varphi'(\rho) < 0$. Si $\rho \in]0, \rho_0]$ alors $\varphi(\rho) - \varphi(0) = \int_0^\rho \varphi'(t)dt < 0$, et on a donc bien $\varphi(\rho) < \varphi(0)$ pour tout $\rho \in]0, \rho_0]$, ce qui prouve que w est une direction de descente stricte en x.

■

Algorithme du gradient à pas fixe Soient $f \in C^1(E, \mathbb{R})$ et $E = \mathbb{R}^N$. On se donne $\rho > 0$.

$$\begin{cases} \text{Initialisation :} & x_0 \in E, \\ \text{Itération } n : & x_n \text{ connu, } (n \geq 0) \\ & w_n = -\nabla f(x_n), \\ & x_{n+1} = x_n + \rho w_n. \end{cases} \tag{3.3.10}$$

Théorème 3.17 (Convergence du gradient à pas fixe) *Soient $E = \mathbb{R}^N$ et $f \in C^1(E, \mathbb{R})$ On suppose que :*

1. *$\exists \alpha > 0$ tel que $(\nabla f(x) - \nabla f(y)) \cdot (x - y) \geq \alpha \|x - y\|^2, \forall (x, y) \in E^2$,*
2. *$\exists M > 0$ tel que $\|\nabla f(x) - \nabla f(y)\| \leq M \|x - y\|, \forall (x, y) \in E^2$,*

alors :

1. *f est strictement convexe,*
2. *$f(x) \to +\infty$ quand $|x| \to +\infty$,*
3. *il existe un et un seul $\bar{x} \in E$ tel que $f(\bar{x}) = \inf_E f$ (conséquence de 1. et 2.),*
4. *si $0 < \rho < \dfrac{2\alpha}{M^2}$ alors la suite $(x_n)_{n \in \mathbb{N}}$ construite par* (3.3.10) *converge vers $\bar{x}$ lorsque $n \to +\infty$.*

La démonstration de ce théorème fait l'objet de l'exercice 63.

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 :

$$\left\{ \begin{array}{ll} \text{Initialisation :} & x_0 \in \mathbb{R}^N. \\ \text{Itération } n : & x_n \text{ connu.} \\ & \text{On calcule } w_n = -\nabla f(x_n). \\ & \text{On choisit } \rho_n \geq 0 \text{ tel que} \\ & \qquad f(x_n + \rho_n w_n) \leq f(x_n + \rho w_n) \ \forall \rho \geq 0. \\ & \text{On pose } x_{n+1} = x_n + \rho_n w_n. \end{array} \right. \tag{3.3.11}$$

Les questions auxquelles on doit répondre pour s'assurer du bien fondé de ce nouvel algorithme sont les suivantes :

1. Existe–t–il ρ_n tel que $f(x_n + \rho_n w_n) \leq f(x_n + \rho w_n), \forall \rho \geq 0$?
2. Comment calcule–t'on ρ_n ?
3. La suite $(x_n)_{n\in\mathbb{N}}$ construite par l'algorithme converge–t–elle ?

La réponse aux questions 1. et 3. est apportée par le théorème suivant :

Théorème 3.18 (Convergence du gradient à pas optimal)
Soit $f \in C^1(\mathbb{R}^N, \mathbb{R})$ telle que $f(x) \to +\infty$ quand $|x| \to +\infty$. Alors :

1. *La suite $(x_n)_{n\in\mathbb{N}}$ est bien définie par (3.3.11). On choisit $\rho_n > 0$ tel que $f(x_n + \rho_n w_n) \leq f(x_n + \rho w_n)$ $\forall \rho \geq 0$ (ρ_n existe mais n'est pas nécessairement unique).*
2. *La suite $(x_n)_{n\in\mathbb{N}}$ est bornée et si $(x_{n_k})_{k\in\mathbb{N}}$ est une sous suite convergente, i.e. $x_{n_k} \to x$ lorsque $k \to +\infty$, on a nécessairement $\nabla f(x) = 0$. De plus si f est convexe on a $f(x) = \inf_{\mathbb{R}^N} f$*
3. *Si f est strictement convexe on a alors $x_n \to \bar{x}$ quand $n \to +\infty$, avec $f(\bar{x}) = \inf_{\mathbb{R}^N} f$*

La démonstration de ce théorème fait l'objet de l'exercice 65. On en donne ici les idées principales.

1. On utilise l'hypothèse $f(x) \to +\infty$ quand $|x| \to +\infty$ pour montrer que la suite $(x_n)_{n\in\mathbb{N}}$ construite par (3.3.11) existe : en effet, à x_n connu,

   1er cas : si $\nabla f(x_n) = 0$, alors $x_{n+1} = x_n$ et donc $x_p = x_n \ \forall p \geq n$,

   2ème cas : si $\nabla f(x_n) \neq 0$, alors $w_n = \nabla f(x_n)$ est une direction de descente stricte.

   Dans ce deuxième cas, il existe donc ρ_0 tel que

   $$f(x_n + \rho w_n) < f(x_n), \ \forall \rho \in]0, \rho_0]. \tag{3.3.12}$$

   De plus, comme $w_n \neq 0$, $|x_n + \rho w_n| \to +\infty$ quand $\rho \to +\infty$ et donc $f(x_n + \rho w_n) \longrightarrow +\infty$ quand $\rho \to +\infty$. Il existe donc $M > 0$ tel que si $\rho > M$ alors $f(x_n + \rho w_n) \geq f(x_n)$. On a donc :

   $$\inf_{\rho\in\mathbb{R}_+^*} f(x_n + \rho w_n) = \inf_{\rho\in[0,M]} f(x_n + \rho w_n).$$

   Comme $[0, M]$ est compact, il existe $\rho_n \in [0, M]$ tel que $f(x_n + \rho_n w_n) = \inf_{\rho\in[0,M]} f(x_n + \rho w_n)$. De plus on a grâce à (3.3.12) que $\rho_n > 0$.
2. Le point 2. découle du fait que la suite $(f(x_n))_{n\in\mathbb{N}}$ est décroissante, donc la suite $(x_n)_{n\in\mathbb{N}}$ est bornée (car $f(x) \to +\infty$ quand $|x| \to +\infty$). On montre ensuite que si $x_{n_k} \to x$ lorsque $k \to +\infty$ alors $\nabla f(\bar{x}) = 0$ (ceci est plus difficile, les étapes sont détaillées dans l'exercice 65).

Reste la question du calcul de ρ_n. Soit φ la fonction de $\mathbb{R}_+$ dans $\mathbb{R}$ définie par : $\varphi(\rho) = f(x_n + \rho w_n)$. Comme $\rho_n > 0$ et $\varphi(\rho_n) \leq \varphi(\rho)$ pour tout $\rho \in \mathbb{R}_+$, on a nécessairement $\varphi'(\rho_n) = \nabla f(x_n + \rho_n w_n) \cdot w_n = 0$. Considérons le cas d'une fonctionnelle quadratique, *i.e.* $f(x) = \frac{1}{2} Ax \cdot x - b \cdot x$, A étant une matrice symétrique définie positive. Alors $\nabla f(x_n) = Ax_n - b$, et donc $\nabla f(x_n + \rho_n w_n) \cdot w_n = (Ax_n + \rho_n A w_n - b) \cdot w_n = 0$. On a ainsi dans ce cas une expression explicite de ρ_n :

$$\rho_n = \frac{(b - Ax_n) \cdot w_n}{Aw_n \cdot w_n},$$

(en effet, $Aw_n \cdot w_n \neq 0$ car A est symétrique définie positive).
Dans le cas d'une fonction f générale, on n'a pas en général de formule explicite pour ρ_n. 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 ρ_n optimal.

Algorithme du gradient à pas variable Dans ce nouvel algorithme, on ne prend pas forcément le paramètre optimal pour ρ, mais on lui permet d'être variable d'une itération à l'autre. L'algorithme s'écrit :

$$\left\{ \begin{array}{ll} \text{Initialisation :} & x_0 \in \mathbb{R}^N. \\ \text{Itération :} & \text{On suppose } x_n \text{ connu ; soit } w_n = -\nabla f(x_n) \text{ où : } w_n \neq 0 \\ & (\text{si } w_n = 0 \text{ l'algorithme s'arrête}). \\ & \text{On prend } \rho_n > 0 \text{ tel que } f(x_n + \rho_n w_n) < f(x_n). \\ & \text{On pose } x_{n+1} = x_n + \rho_n w_n. \end{array} \right. \tag{3.3.13}$$

Théorème 3.19 (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 $(x_n)_{n\in\mathbb{N}}$ par (3.3.13).*
2. *La suite $(x_n)_{n\in\mathbb{N}}$ est bornée. Si $x_{n_k} \to x$ quand $k + \infty$ et si $\nabla f(x_{n_k}) \to 0$ quand $n \to +\infty$ alors $\nabla f(x) = 0$. Si de plus f est convexe on a $f(x) = \inf_{\mathbb{R}^N} f$*
3. *Si $\nabla f(x_n) \to 0$ quand $n \to +\infty$ et si f est strictement convexe alors $x_n \to \bar{x}$ et $f(\bar{x}) = \inf_{\mathbb{R}^N} f$.*

Démonstration : Elle est facile à partir de la démonstration du théorème précédent : reprendre en l'adaptant l'exercice 65.

3.3.2 Algorithmes du gradient conjugué

La méthode du gradient conjugué a été découverte en 1952 par Hestenes et Steifel pour la minimisation de fonctionnelles quadratiques, c'est-à -dire de fonctionnelles de la forme

$$f(x) = \frac{1}{2} Ax \cdot x - b \cdot x,$$

où $A \in \mathcal{M}_N(\mathbb{R})$ est une matrice symétrique définie positive et $b \in \mathbb{R}^N$. On rappelle (voir section (3.2.2) et exercice (62)) que $f(\bar{x}) = \inf_{\mathbb{R}^N} f \Leftrightarrow A\bar{x} = b$.

Définition 3.20 (Vecteurs conjugués) *Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice symétrique définie positive,*

1. *Deux vecteurs v et w de $\mathbb{R}^N \setminus \{0\}$ sont dits A-conjugués si $Av \cdot w = w \cdot Av = 0$.*
2. *Une famille $(w^{(1)}, \ldots, w^{(p)})$ de $\mathbb{R}^N \setminus \{0\}$ est dite A-conjuguée si $w^{(i)} \cdot Aw^{(j)} = 0$ pour tout couple $(i, j) \in \{1, \ldots, p\}^2$ tel que $i \neq j$.*

Proposition 3.21 *Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice symétrique définie positive, $(w^{(1)}, \dots, w^{(p)})$ une famille de $\mathbb{R}^N$, alors :*

1. *si la famille $(w^{(1)}, \dots, w^{(p)})$ est A-conjuguée alors elle est libre ;*
2. *dans le cas où $p = N$, si la famille $(w^{(1)}, \dots, w^{(N)})$ est A-conjuguée alors c'est une base de $\mathbb{R}^N$.*

Démonstration : Le point 2. est immédiat dès qu'on a démontré le point 1. Supposons donc que $(w^{(1)}, \dots, w^{(p)})$ est une famille A-conjuguée, *i.e.* $w^{(i)} \neq 0$, $\forall i$ et $w^{(i)} \cdot Aw^{(j)} = 0$ si $i \neq j$; soit $(\alpha_i)_{i=1,\dots,p} \subset \mathbb{R}$, supposons que $\sum_{i=1}^p \alpha_i w^{(i)} = 0$, on a donc $\sum_{i=1}^p \alpha_i w^{(i)} \cdot Aw^{(j)} = 0$ et donc $\alpha_j w^{(j)} \cdot Aw^{(j)} = 0$. Or $w^{(j)} \cdot Aw^{(j)} \neq 0$ car $w^{(j)} \neq 0$ et A est symétrique définie positive. On en déduit que $\alpha_j = 0$ pour $j = 1, \dots, p$. La famille $(w^{(1)}, \dots, w^{(p)})$ est donc libre.

Proposition 3.22 *Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice symétrique définie positive, $b \in \mathbb{R}^N$ et f une fonction définie de $\mathbb{R}^N$ dans $\mathbb{R}^N$ par $f(x) = \frac{1}{2}Ax \cdot x - b \cdot x$. On suppose que la suite $(x^{(n)})_n$ est définie par :*

Initialisation $x^{(0)} \in \mathbb{R}^N$
Itération n $x^{(n+1)} = x^{(n)} + \rho_n w^{(n)}$ *où*
1) $w^{(n)} \neq 0$ *est une direction de descente stricte en* $x^{(n)}$
2) ρ_n *est optimal dans la direction* $w^{(n)}$.

Si la famille $(w^{(0)}, \dots, w^{(N-1)})$ est une famille A-conjuguée alors $x^{(N)} = \bar{x}$ avec $A\bar{x} = b$.

Démonstration Soit $w^{(n)}$ direction de descente stricte en $x^{(n)}$ et ρ_n optimal dans la direction $w^{(n)}$; alors $\rho_n > 0$ et $\nabla f(x^{(n+1)}) \cdot w^{(n)} = 0$, c'est-à-dire

$$(Ax^{(n+1)} - b) \cdot w^{(n)} = 0 \tag{3.3.14}$$

On va montrer que

$$(Ax^{(N)} - b) \cdot w^{(p)} = 0, \ \forall p \in \{0, \dots, N-1\}.$$

Comme $(w^{(0)}, \dots, w^{(N-1)})$ est une base de $\mathbb{R}^N$, on en déduit alors que $Ax^{(N)} = b$, c'est-à-dire $x^{(N)} = \bar{x}$. Remarquons d'abord grâce à (3.3.14) que $(Ax^{(N)} - b) \cdot w^{(N-1)} = 0$. Soit maintenant $p < N - 1$. On a :

$$Ax^{(N)} - b = A(x^{(N-1)} + \rho_{N-1} w^{(N-1)}) - b = Ax^{(N-1)} - b + \rho_{N-1} Aw^{(N-1)}.$$

On a donc en itérant,

$$Ax^{(N)} - b = Ax^{(p+1)} - b + \rho_{N-1} Aw^{(N-1)} + \dots + \rho_{p+1} Aw^{(p+1)}, \ \forall p \geq 1$$

. On en déduit que

$$(Ax^{(N)} - b) \cdot w^{(p)} = (Ax^{(p+1)} - b) \cdot w^{(p)} + \sum_{j=p+1}^{N-1} (\rho_i Aw_j \cdot w^{(p)}).$$

Comme les directions w_i sont conjuguées, on a donc $(Ax^{(N)} - b) \cdot w^{(p)} = 0$ pour tout $p = 0 \dots, N-1$ et donc $Ax^{(N)} = b$. ∎

Le résultat précédent suggère de rechercher une méthode de minimisation de la fonction quadratique f selon le principe suivant : Pour $x^{(0)} \dots x^{(n)}$ connus, $w^{(0)}, \dots, w^{(n-1)}$ connus, on cherche $w^{(n)}$ tel que :

1. $w^{(n)}$ soit une direction de descente stricte en $x^{(n)}$,
2. $w^{(n)}$ soit A-conjugué avec $w^{(p)}$ pour tout $p < n$.

Si on arrive à trouver $w^{(n)}$ on prend alors $x^{(n+1)} = x^{(n)} + \rho_n w^{(n)}$ avec ρ_n optimal dans la direction $w^{(n)}$. La propriété précédente donne $x^{(N)} = \bar{x}$ avec $A\bar{x} = b$.

Définition 3.23 (Méthode du gradient conjugué) *Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice symétrique définie positive, $b \in \mathbb{R}^N$ et $f(x) = \frac{1}{2}Ax \cdot x - b \cdot x$.*

$$\left\{\begin{array}{l}
\textbf{Initialisation} \\
\text{Soit } x^{(0)} \in \mathbb{R}^N, \text{ et soit } r^{(0)} = b - Ax^{(0)} = -\nabla f(x^{(0)}). \\
\quad 1) \text{ Si } r^{(0)} = 0, \text{ alors } Ax^{(0)} = b \text{ et donc } x^{(0)} = \bar{x}, \\
\quad \text{auquel cas l'algorithme s'arrête.} \\
\quad 2) \text{ Si } r^{(0)} \neq 0, \text{ alors on pose } w^{(0)} = r^{(0)}, \text{ et on choisit } \rho_0 \text{ optimal} \\
\quad \text{dans la direction } w^{(0)}. \\
\quad \text{On pose alors } x^{(1)} = x^{(0)} + \rho_0 w^{(0)}. \\
\\
\textbf{Itération } 1 \leq n \leq N-1 : \\
\text{On suppose } x^{(0)}, \ldots, x^{(n)} \text{ et } w^{(0)}, \ldots, w^{(n-1)} \text{ connus et on pose} \\
\quad r^{(n)} = b - Ax^{(n)}. \\
\quad 1) \text{ Si } r^{(n)} = 0 \text{ on a } Ax^{(n)} = b \text{ donc } x^{(n)} = \bar{x} \\
\quad \text{auquel cas l'algorithme s'arrête.} \\
\quad 2) \text{ Si } r^{(n)} \neq 0, \text{ alors on pose } w^{(n)} = r^{(n)} + \lambda_{n-1} w^{(n-1)} \\
\quad \text{avec } \lambda_{n-1} \text{ tel que } \quad w^{(n)} \cdot Aw^{(n-1)} = 0, \\
\quad \text{et on choisit } \rho_n \text{ optimal dans la direction } w^{(n)} ; \\
\quad \text{On pose alors } x^{(n+1)} = x^{(n)} + \rho_n w^{(n)}.
\end{array}\right. \tag{3.3.15}$$

Théorème 3.24 *Soit A une symétrique définie positive, $A \in \mathcal{M}_N(\mathbb{R})$, $b \in \mathbb{R}^N$ et $f(x) = \frac{1}{2}Ax \cdot x - b \cdot x$ alors* (3.3.15) *définit une suite $(x^{(n)})_{n=0,\ldots,p}$ avec $p \leq N$ telle que $x^{(N)} = \bar{x}$ avec $A\bar{x} = b$.*

Démonstration
Initialisation Si $r^{(0)} = 0$, alors $Ax^{(0)} = b$ et donc $x^{(0)} = \bar{x}$ auquel cas $p = 0$. Si $r^{(0)} \neq 0$, comme $w^{(0)} = r^{(0)} = b - Ax^{(0)} = -\nabla f(x^{(0)})$, $w^{(0)}$ est une direction de descente stricte ; il existe donc ρ_0 qui minimise la fonction φ définie de $\mathbb{R}$ dans $\mathbb{R}$ par $\varphi(\rho) = f(x^{(0)} + \rho w^{(0)})$. La valeur de ρ_0 est obtenue en demandant que $\varphi'(\rho) = 0$, ce qui donne : $\rho_0 = \dfrac{r^{(0)} \cdot w^{(0)}}{Aw^{(0)} \cdot w^{(0)}}$. L'élément $x^{(1)} = x^{(0)} + \rho_0 w^{(0)}$ est donc bien défini. Notons que $r^{(1)} = Ax^{(1)} - b = r^{(0)} - \rho_0 Aw^{(0)}$, et donc $r^{(1)} \cdot w^{(0)} = 0$.
Itération n
On suppose $x^{(0)}, \ldots, x^{(n)}$ et $w^{(0)}, \ldots, w^{(n)}$ connus, et on pose $r^{(n)} = b - Ax^{(n)}$.
Si $r^{(n)} = 0$ alors $Ax^{(n)} = b$ et donc $x^{(n)} = \bar{x}$ auquel cas l'algorithme s'arrête et $p = n$.
Si $r^{(n)} \neq 0$, on pose $w^{(n)} = r^{(n)} + \lambda_{n-1} w^{(n-1)}$. Comme $w^{(n-1)} \neq 0$, on peut choisir λ_{n-1} tel que $w^{(n)} \cdot Aw^{(n-1)} = 0$, c.à.d. $(r^{(n)} + \lambda_{n-1} w^{(n-1)}) \cdot Aw^{(n-1)} = 0$, en prenant

$$\lambda_{n-1} = -\frac{r^{(n)} \cdot Aw^{(n-1)}}{w^{(n-1)} \cdot Aw^{(n-1)}}.$$

Montrons maintenant que $w^{(n)}$ est une direction de descente stricte en $x^{(n)}$. On a :

$$\begin{aligned}
w^{(n)} \cdot (-\nabla f(x^{(n)})) &= (r^{(n)} + \lambda_{n-1} w^{(n-1)}) \cdot (-\nabla f(x^{(n)})) \\
&= (-\nabla f(x^{(n)}) + \lambda_{n-1} w_{n-1}) \cdot (-\nabla f(x_n)) \\
&= |\nabla f(x^{(n)})|^2 - \lambda_{n-1} w^{(n-1)} \cdot \nabla f(x^{(n)}).
\end{aligned}$$

Or $w^{(n-1)} \cdot \nabla f(x^{(n)}) = 0$ car ρ_{n-1} est le paramètre de descente optimal en $x^{(n-1)}$ dans la direction $w^{(n-1)}$, on a donc :

$$-w^{(n)} \cdot \nabla f(x^{(n)}) = |\nabla f(x^{(n)})|^2 = |r^{(n)}|^2 > 0$$

ceci donne que $w^{(n)}$ est une direction de descente stricte en $x^{(n)}$ On peut choisir $\rho_n > 0$ optimal en $x^{(n)}$ dans la direction $w^{(n)}$, et le calcul de ρ_n (similaire à celui de l'étape d'initialisation) donne

$$\rho_n = \frac{r^{(n)} \cdot w^{(n)}}{Aw^{(n)} \cdot w^{(n)}}. \tag{3.3.16}$$

On peut donc bien définir $x^{(n+1)} = x^{(n)} + \rho_n w^{(n)}$. Remarquons que ce choix de ρ_n entraîne que

$$r^{(n)} \cdot w^{(n-1)} = 0. \tag{3.3.17}$$

Pour pouvoir appliquer la proposition 3.22, il reste à montrer que la famille $w^{(0)}, \ldots, w^{(n)}$ est A-conjuguée. Ceci est l'objet de la proposition 3.26 qui suit. Grâce à cette proposition, on obtient que si $r^{(n)} \neq 0$, $n = 0, \ldots, N-1$, la famille $(w^{(0)}, \ldots, w^{(N-1)})$ est donc A-conjuguée, et $w^{(n)}$ est une direction de descente stricte en $x^{(n)}$ pour tout $n \leq N-1$. On en déduit par la proposition 3.22 que $x^{(N)} = \bar{x}$. ■

La démonstration de la proposition 3.26 que l'on vient d'utiliser se fait par récurrence, et nécessite les petits résultats préliminaires énoncés dans le lemme suivant :

Lemme 3.25 *Sous les hypothèses et notations de la définition 3.23, on a :*

$$\rho_n = \frac{r^{(n)} \cdot r^{(n)}}{w^{(n)} \cdot Aw^{(n)}}, \tag{3.3.18}$$

$$r^{(n)} = r^{(n-1)} + \rho_{n-1} Aw^{(n-1)}, \tag{3.3.19}$$

$$r^{(n)} \cdot r^{(n-1)} = 0, \tag{3.3.20}$$

$$\lambda_{n-1} = \frac{r^{(n)} \cdot r^{(n)}}{r^{(n-1)} \cdot r^{(n-1)}}, \tag{3.3.21}$$

Démonstration :
1. Comme ρ_n est le paramètre optimal dans la direction $w^{(n)}$, on sait (voir (3.3.16)) que

$$\rho_n = \frac{r^{(n)} \cdot w^{(n)}}{Aw^{(n)} \cdot w^{(n)}}.$$

Or par définition, $w^{(n)} = r^{(n)} + \lambda_{n-1} w^{(n-1)}$, et donc $w^{(n)} \cdot r^{(n)} = r^{(n)} \cdot r^{(n)} + \lambda_{n-1} w^{(n-1)} \cdot r^{(n)}$. Il ne reste plus à remarquer que $w^{(n-1)} \cdot r^{(n)} = 0$ en raison de l'optimalité de ρ_{n-1} (voir (3.3.17)). On en déduit que

$$\rho_n = \frac{r^{(n)} \cdot r^{(n)}}{w^{(n)} \cdot Aw^{(n)}}.$$

2. Par définition, $x^{(n)} = x^{(n-1)} + \rho_{n-1} w^{(n-1)}$, donc $Ax^{(n)} = Ax^{(n-1)} + \rho_{n-1} Aw^{(n-1)}$, ce qui entraîne $r^{(n)} = r^{(n-1)} + \rho_{n-1} Aw^{(n-1)}$.

3. Par définition, et grâce à (3.3.19), on a :

$$r^{(n)} \cdot r^{(n-1)} = r^{(n-1)} \cdot r^{(n-1)} + \rho_{n-1} Aw^{(n-1)} \cdot r^{(n-1)}.$$

Or $w^{(n-1)} = r^{(n-1)} + \lambda_{n-1} w^{(n-2)}$, et donc $r^{(n-1)} = w^{(n-1)} - \lambda_{n-1} w^{(n-2)}$. On en déduit que

$$r^{(n)} \cdot r^{(n-1)} = r^{(n-1)} \cdot r^{(n-1)} - \rho_{n-1} Aw^{(n-1)} \cdot w^{(n-1)} - \rho_{n-1} \lambda_{n-1} Aw^{(n-1)} \cdot w^{(n-2)}.$$

Or $Aw^{(n-1)} \cdot w^{(n-2)} = 0$ et par (3.3.18), on a $r^{(n-1)} \cdot r^{(n-1)} - \rho_{n-1} Aw^{(n-1)} \cdot w^{(n-1)} = 0$.

4. Par définition,

$$\lambda_{n-1} = -\frac{r^{(n)} \cdot Aw^{(n-1)}}{w^{(n-1)} \cdot Aw^{(n-1)}}.$$

Or par (3.3.19), on a :

$$Aw^{(n-1)} = \frac{1}{\rho_{n-1}}(r^{(n-1)} - r^{(n)}).$$

On conclut grâce à (3.3.20) et (3.3.18). ■

Proposition 3.26 *Sous les hypothèses et notations de la définition 3.23, soit $n \in \mathbb{N}$ tel que $1 \leq n \leq N$, si $r^{(q)} \neq 0$ pour $0 \leq q \leq n$, les propriétés suivantes sont vérifiées :*

1. $r^{(n)} \cdot w^{(q)} = 0$, $\forall q = 0, \ldots, n-1$,
2. *Vect*$(r^{(0)}, \ldots, r^{(n)})$ = *Vect*$(r^{(0)}, \ldots, A^n r^{(0)})$,
3. *Vect*$(w^{(0)}, \ldots, w^{(n)})$ = *Vect*$(r^{(0)}, \ldots, A^n r^{(0)})$,
4. $w^{(n)} \cdot Aw^{(q)} = 0$, $\forall q = 0, \ldots, n-1$,
5. $r^{(n)} \cdot r^{(q)} = 0$, $\forall q = 0, \ldots, n-1$,

où $Vect(w^{(0)}, \ldots, w^{(n)})$ désigne l'espace vectoriel engendré par les vecteurs $w^{(0)}, \ldots, w^{(n)}$. En particulier, la famille $(w^{(0)}, \ldots, w^{(N-1)})$ est A-conjuguée.
L'espace Vect$(r^{(0)}, \ldots, A^n r^{(0)})$ est appelé espace de Krylov.

Démonstration :

On démontre les propriétés 1. à 5 par récurrence.
Etudions tout d'abord le cas $n = 1$. Remarquons que $r^{(1)} \cdot w^{(0)} = 0$ en vertu de (3.3.17) (on rappelle que cette propriété découle du choix optimal de ρ_0).
On a grâce à (3.3.19) :

$$r^{(1)} = r^{(0)} - \rho_0 Aw^{(0)} = r^{(0)} - \rho_0 Ar^{(0)},$$

car $w^{(0)} = r^{(0)}$. On a donc $Vect(r^{(0)}, r^{(1)}) = Vect(r^{(0)}, Ar^{(0)})$.
De plus, comme $w^{(0)} = r^{(0)}$, et $w^{(1)} = r^{(1)} + \lambda_1 w^{(0)}$, on a

$$Vect(r^{(0)}, r^{(1)}) = Vect(w^{(0)}, w^{(1)}).$$

On en déduit que 2. et 3. sont vraies pour $n = 1$.
Enfin, on a bien $w^{(1)} \cdot Aw^{(0)} = 0$ car $w^{(0)}$ et $w^{(1)}$ sont conjuguées, et $r^{(0)} \cdot r^{(1)} = 0$ en vertu de (3.3.20).

On a ainsi montré que les propriétés 1. à 5. sont vérifiées au rang $n = 1$. Supposons maintenant que ces propriétés soient vérifiées jusqu'au rang n, et démontrons qu'elles le sont encore au rang $n + 1$.
1. En vertu de (3.3.19), et par les hypothèses de récurrence 1. et 4., on a :

$$r^{(n+1)} \cdot w^{(q)} = r^{(n)} \cdot w^{(q)} - \rho_n Aw^{(n)} \cdot w^{(q)} = 0, \ \forall q \leq n-1.$$

De plus, (3.3.20) entraîne $r^{(n+1)} \cdot w^{(n)} = 0$
2. Montrons que $Vect(r^{(0)}, r^{(1)} \ldots, r^{(n+1)}) = Vect(r^{(0)}, Ar^{(0)}, \ldots, A^{(n+1)} r^{(0)})$. Pour ce faire, commençons par remarquer que

$$r^{(n+1)} \in Vect(r^{(0)}, Ar^{(0)} \ldots, A^{(n+1)} r^{(0)}).$$

En effet, en vertu de (3.3.19), on a : $r^{(n+1)} = r^{(n)} - \rho_n Aw^{(n)}$, et par hypothèse de récurrence, on a

$$r^{(n)} \in Vect(r^{(0)}, Ar^{(0)} \ldots, A^n r^{(0)}), \text{ et } w^{(n)} \in Vect(r^{(0)}, Ar^{(0)} \ldots, A^n r^{(0)}).$$

Montrons maintenant que $A^{n+1}r^{(0)} \in Vect(r^{(0)}, r^{(1)} \dots, r^{(n+1)})$. Comme $r^{(n+1)} \in Vect(r^{(0)}, Ar^{(0)} \dots, A^{(n+1)}r^{(0)})$, il existe une famille $(\alpha_k)_{k=0,\dots,n+1}$ telle que

$$r^{(n+1)} = \sum_{k=0}^{n+1} \alpha_k A^k r(0) = \sum_{k=0}^{n} \alpha_k A^k r(0) + \alpha_{n+1} A^{n+1} r^{(0)}.$$

Or grâce à la propriété 1. on sait que $r^{(n+1)} \cdot w^{(q)} = 0$, $\forall q \leq n$, et donc $r^{(n+1)} \notin Vect(w^{(0)}, w^{(1)} \dots, w^{(n)})$. On a donc $\alpha_{n+1} \neq 0$, et on peut donc écrire

$$A^{n+1}r^{(0)} = \frac{1}{\alpha_{n+1}}(r^{(n+1)} - \sum_{k=0}^{n} \alpha_k A^k r^{(0)}) \in Vect(r^{(0)}, r^{(1)} \dots, r^{(n+1)}),$$

par hypothèse de récurrence.
3. Montrons maintenant que

$$Vect(w^{(0)}, w^{(1)} \dots, w^{(n+1)}) = Vect(r^{(0)}, Ar^{(0)} \dots, A^{n+1}r^{(0)}).$$

On a : $w^{(n+1)} = r^{(n+1)} + \lambda_n w^{(n)}$. Or on vient de montrer que

$$r^{(n+1)} \in Vect(r^{(0)}, Ar^{(0)} \dots, A^{n+1}r^{(0)}),$$

et par hypothèse de récurrence, $w^{(n)} \in Vect(r^{(0)}, Ar^{(0)} \dots, A^{n}r^{(0)})$. On a donc bien $w^{(n+1)} \in Vect(r^{(0)}, Ar^{(0)} \dots, A^{n+1}r^{(0)})$. Montrons que réciproquement, $A^{n+1}r^{(0)}) \in Vect(w^{(0)}, w^{(1)} \dots, w^{(n+1)})$. On a montré en 2. que

$$A^{n+1}r^{(0)} = \frac{1}{\alpha_{n+1}}(r^{(n+1)} - \sum_{k=0}^{n} \alpha_k A^k r^{(0)}).$$

Or $r^{(n+1)} = w^{(n+1)} - \lambda_n w^{(n)} \in Vect(w^{(0)}, w^{(1)} \dots, w^{(n+1)})$, et

$$\sum_{k=0}^{n} \alpha_k A^k r^{(0)}) \in Vect(r^{(0)}, r^{(1)} \dots, r^{(n)}) = Vect(w^{(0)}, w^{(1)} \dots, w^{(n)}),$$

par hypothèse de récurrence. On en déduit que

$$A^{n+1}r^{(0)} \in Vect(w^{(0)}, w^{(1)} \dots, w^{(n)}).$$

4. On veut maintenant montrer que $w^{(n+1)} \cdot Aw^{(q)} = 0$, $\forall q \leq n$. Pour $q = n$, cette propriété est vérifiée en raison du choix de $w^{(n+1)}$ (conjuguée avec $w^{(n)}$). Pour $q < n$, on calcule :

$$w^{(n+1)} \cdot Aw^{(q)} = r^{(n+1)} \cdot Aw^{(q)} + \lambda_n w^{(n)} \cdot Aw^{(q)}. \tag{3.3.22}$$

Or $w^{(n)} \cdot Aw^{(q)} = 0$ pour tout $q \leq n-1$ par hypothèse de récurrence. De plus, toujours par hypothèse de récurrence, $w^{(q)} \in Vect(r^{(0)}, Ar^{(0)} \dots, A^{q}r^{(0)})$, et donc

$$Aw^{(q)} \in Vect(r^{(0)}, Ar^{(0)} \dots, A^{q+1}r^{(0)}) = Vect(w^{(0)}, w^{(1)} \dots, w^{(q+1)}).$$

On a montré en 1. que $r^{(n+1)} \cdot w^{(k)} = 0$ pour tout $k \leq n$, on a donc $r^{(n+1)} \cdot Aw^{(q)} = 0$, et en reportant dans (3.3.22), on obtient donc que $w^{(n+1)} \cdot Aw^{(q)} = 0$ pour tout $q \leq n$.
5. Il reste à montrer que $r^{(n+1)} \cdot r^{(q)} = 0$ pour tout $q \leq n$. Pour $q = n$, on l'a démontré dans le lemme 3.25. Pour $q \leq n-1$, on a

$$r^{(n+1)} \cdot r^{(q)} = (r^{(n)} - \lambda_n Aw^{(n)}) \cdot r^{(q)} = r^{(n)} \cdot r^{(q)} - \lambda_n Aw^{(n)} \cdot r^{(q)}.$$

Or $r^{(n)} \cdot r^{(q)} = 0$ par hypothèse de récurrence, et $Aw^{(n)} \cdot r^{(q)} = w^{(n)} \cdot Ar^{(q)}$; or $Ar^{(q)} \in Vect(r^{(0)}, \dots, r^{(q)}$ et $w^{(n)} \cdot r^{(k)} = 0$ pour tout $k \leq n-1$ par hypothèse de récurrence 1. On en déduit que $r^{(n+1)} \cdot r^{(q)} = 0$.
Ceci termine la démonstration de la proposition (3.26).

■

Remarque 3.27 (Gradient conjugué préconditionné)
1. On a vu que $\lambda_{n-1} = \dfrac{r^{(n)} \cdot r^{(n)}}{r^{(n-1)} \cdot r^{(n-1)}}$ *et que* $\rho_n = \dfrac{r^{(n)} \cdot r^{(n)}}{w^{(n)} \cdot Aw^{(n)}}$.
On peut calculer le nombre d'opérations nécessaires pour calculer $\bar{x}$ *(c.à.d. pour calculer* $x^{(N)}$*, sauf dans le cas miraculeux où* $x^{(N)} = \bar{x}$ *pour* $n < N$*) 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 n'est pas intéressante comme méthode directe car elle demande 12 fois plus d'opérations que Choleski.*

2. On peut alors se demander si la méthode est intéressante comme méthode itérative, c.à.d. si on peut espérer que $x^{(n)}$ *soit "proche de* $\bar{x}$*" pour "*$n \ll N$*". 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* $\bar{x}$*. 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".*

On calcule 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* $y = L^t x$*, alors le système* $Ax = b$ *peut aussi s'écrire* $L^{-1}A(L^t)^{-1}y = L^{-1}b$*, et le système* $(L^t)^{-1}y = x$ *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,*

$$Bx \cdot x = L^{-1}A(L^t)^{-1}x \cdot x = A(L^t)^{-1}x \cdot (L^t)^{-1}x,$$

et donc $Bx \cdot x > 0$ *si* $x \neq 0$*. 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*

$$f(y) = \frac{1}{2}By \cdot y - L^{-1}b \cdot y.$$

On en déduit l'expression de la suite $(y^{(n)})_{n\in\mathbb{N}}$ *et donc* $(x^{(n)})_{n\in\mathbb{N}}$*.*
On peut alors montrer (voir exercice 70) que l'algorithme du gradient conjugué préconditionné ainsi obtenu peut s'écrire directement pour la suite $(x^{(n)})_{n\in\mathbb{N}}$ *, de la manière suivante :*

Itération n *On pose* $r^{(n)} = b - Ax^{(n)}$*,*
on calcule $s^{(n)}$ *solution de* $LL^t s^{(n)} = r^{(n)}$*.*
On pose alors $\lambda_{n-1} = \dfrac{s^{(n)} \cdot r^{(n)}}{s^{(n-1)} \cdot r^{(n-1)}}$ *et* $w^{(n)} = s^{(n)} + \lambda_{n-1}w^{(n-1)}$*.*
Le paramètre optimal ρ_n *a pour expression :* $\rho_n = \dfrac{s^{(n)} \cdot r^{(n)}}{Aw^{(n)} \cdot w^{(n)}}$*, et on pose alors* $x^{(n+1)} = x^{(n)} + \rho_n w^{(n)}$*.*
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.*

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 (3.3.15), mais on adapte le calcul de λ_{n-1} et ρ_n.

Itération n :
A $x^{(0)}, \ldots, x^{(n)}$ et $w^{(0)}, \ldots, w^{(n-1)}$ connus, on calcule $r^{(n)} = -\nabla f(x^{(n)})$.
Si $r^{(n)} = 0$ alors $Ax^{(n)} = b$ et donc $x^{(n)} = \bar{x}$ auquel cas l'algorithme s'arrête.
Si $r^{(n)} \neq 0$, on pose $w^{(n)} = r^{(n)} + \lambda_{n-1} w^{(n-1)}$ où λ_{n-1} peut être choisi de différentes manières :

1ère méthode (Fletcher–Reeves)

$$\lambda_{n-1} = \frac{r^{(n)} \cdot r^{(n)}}{r^{(n-1)} \cdot r^{(n-1)}},$$

2ème méthode (Polak–Ribière)

$$\lambda_{n-1} = \frac{(r^{(n)} - r^{(n-1)}) \cdot r^{(n)}}{r^{(n-1)} \cdot r^{(n-1)}}.$$

On pose alors $x^{(n+1)} = x^{(n)} + \rho_n w^{(n)}$, où ρ_n est choisi, si possible, optimal dans la direction $w^{(n)}$.
La démonstration de la convergence de l'algorithme de Polak–Ribière fait l'objet de l'exercice 71 page 154.

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 n'existe pas et il vaut donc mieux la réserver au 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 :

$$f(x) = \inf_{\mathbb{R}^N} f \Rightarrow g(x) = 0.$$

Si de plus f est convexe alors on a $g(x) = 0 \Rightarrow f(x) = \inf_{\mathbb{R}^N} f$. 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 $H_f(x)$ est la matrice hessienne de f en x. La méthode de Newton s'écrit dans ce cas :

$$\left\{ \begin{array}{ll} \text{Initialisation} & x^{(0)} \in \mathbb{R}^N, \\ \text{Itération } n & H_f(x^{(n)})(x^{(n-1)} - x^{(n)}) = -\nabla f(x^{(n)}). \end{array} \right. \tag{3.3.23}$$

Remarque 3.28 *La méthode de Newton pour minimiser une fonction f convexe est une méthode de descente. En effet, si $H_f(x_n)$ est inversible, on a $x^{(n+1)} - x^{(n)} = [H_f(x^{(n)})]^{-1}(-\nabla f(x^{(n)}))$ soit encore $x^{(n+1)} = x^{(n)} + \rho_n w^{(n)}$ où $\rho_n = 1$ et $w^{(n)} = [H_f(x^{(n)})]^{-1}(-\nabla f(x^{(n)}))$. Si f est convexe, H_f est une matrice symétrique positive (déjà vu). Comme on suppose $H_f(x^{(n)})$ inversible par hypothèse, la matrice $H_f(x^{(n)})$ est donc symétrique définie positive.*
Donc $w^{(n)}$ est alors une direction de descente stricte si $w^{(n)} \neq 0$ (donc $\nabla f(x^{(n)}) \neq 0$). On en déduit que

$$-w^{(n)} \cdot \nabla f(x^{(n)}) = [H_f(x^{(n)})]^{-1} \nabla f(x^{(n)}) \cdot \nabla f(x^{(n)}) > 0$$

ce qui est une condition suffisante pour que $w^{(n)}$ soit une direction de descente stricte.
La méthode de Newton est donc une méthode de descente avec $w^{(n)} = -H_f(x^{(n)})(\nabla f(x^{(n)}))$ et $\rho_n = 1$.

On peut aussi remarquer, en vertu du théorème 2.16 page 90, que si $f \in C^3(\mathbb{R}^N, \mathbb{R})$, si $\bar{x}$ est tel que $\nabla f(\bar{x}) = 0$ et si $H_f(\bar{x}) = D(\nabla f)(\bar{x})$ est inversible alors il existe $\varepsilon > 0$ tel que si $x_0 \in B(\bar{x}, \varepsilon)$, alors la suite $(x^{(n)})_n$ est bien définie par (3.3.23) et $x^{(n)} \to \bar{x}$ lorsque $n \to +\infty$. De plus, d'après la proposition 2.14, il existe $\beta > 0$ tel que $|x^{(n+1)} - \bar{x}| \leq \beta |x^{(n)} - \bar{x}|^2$ pour tout $n \in \mathbb{N}$.

Remarque 3.29 (Sur l'implantation numérique) *La convergence de la méthode de Newton est très rapide, mais nécessite en revanche le calcul de $H_f(x)$, 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 3.30 *Soient $f \in C^1(\mathbb{R}^N, \mathbb{R})$, $x \in \mathbb{R}^N$ tel que $\nabla f(x) \neq 0$, et soit $B \in \mathcal{M}_N(\mathbb{R})$ une matrice symétrique définie positive ; alors $w = -B\nabla f(x)$ est une direction de descente stricte en x.*

Démontration On a : $w \cdot \nabla f(x) = -B\nabla f(x) \cdot \nabla f(x) < 0$ car B est symétrique définie positive et $\nabla f(x) \neq 0$ donc w est une direction de descente stricte en x. En effet, soit φ la fonction de $\mathbb{R}$ dans $\mathbb{R}$ définie par $\varphi(\rho) = f(x+\rho w)$. Il est clair que $\varphi \in C^1(\mathbb{R}, \mathbb{R})$, $\varphi'(\rho) = \nabla f(x + \rho w) \cdot w$ et $\varphi'(0) = \nabla f(x) \cdot w < 0$. Donc $\exists \rho_0 > 0$ tel que $\varphi'(\rho) < 0$ si $\rho \in]0, \rho_0[$. Par le théorème des accroissements finis, $\varphi(\rho) < \varphi(0)\ \forall \rho \in]0, \rho_0[$ 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 $x^{(n)}$ le vecteur $w^{(n)} = -(B^{(n)})^{-1}(\nabla f(x^{(n)}))$ où la matrice $B^{(n)}$ est censée approcher $H_f(x^{(n)})$ (sans calculer la dérivée seconde de f). On suppose $x^{(n)}, x^{(n-1)}$ et $B^{(n-1)}$ connus. Voyons comment on peut déterminer $B^{(n)}$. On peut demander par exemple que la condition suivante soit satisfaite :

$$\nabla f(x^{(n)}) - \nabla f(x^{(n-1)}) = B^{(n)}(x^{(n)} - x^{(n-1)}). \tag{3.3.24}$$

Ceci est un système à N équations et $N \times N$ inconnues, et ne permet donc pas de déterminer entièrement la matrice $B^{(n)}$ si $N > 1$. Voici un moyen possible pour déterminer entièrement $B^{(n)}$, dû à Broyden. On pose $s^{(n)} = x^{(n)} - x^{(n-1)}$, on suppose que $s^{(n)} \neq 0$, et on pose $y^{(n)} = \nabla f(x^{(n)}) - \nabla f(x_{n-1})$. On choisit alors $B^{(n)}$ telle que :

$$\begin{cases} B^{(n)}s^{(n)} = y^{(n)} \\ B^{(n)}s = B^{(n-1)}s,\ \forall s \perp s^{(n)} \end{cases} \tag{3.3.25}$$

On a exactement le nombre de conditions qu'il faut avec (3.3.25) pour déterminer entièrement $B^{(n)}$. Ceci suggère la méthode suivante :

Initialisation Soient $x^{(0)} \in \mathbb{R}^N$ et $B^{(0)}$ une matrice symétrique définie positive. On pose $w^{(0)} = (B^{(0)})^{-1}(-\nabla f(x^{(0)}))$; alors $w^{(0)}$ est une direction de descente stricte sauf si $\nabla f(x^{(0)}) = 0$.
On pose alors $x^{(1)} = x^{(0)} + \rho^{(0)}w^{(0)}$, où $\rho^{(0)}$ est optimal dans la direction $w^{(0)}$.

Itération n On suppose $x^{(n)}$, $x^{(n-1)}$ et $B^{(n-1)}$ connus, $(n \geq 1)$, et on calcule $B^{(n-1)}$ par (3.3.25). On pose $w^{(n)} = -(B^{(n)})^{-1}(\nabla f(x^{(n)}))$. On choisit $\rho^{(n)}$ optimal en $x^{(n)}$ dans la direction $w^{(n)}$, et on pose $x^{(n+1)} = x^{(n)} + \rho^{(n)}w^{(n)}$.

Le problème avec cet algorithme est que si la matrice est $B^{(n-1)}$ symétrique définie positive, la matrice $B^{(n)}$ ne l'est pas forcément, et donc $w^{(n)}$ n'est pas forcément une direction de descente stricte. On va donc modifier cet algorithme dans ce qui suit.

Méthode de BFGS La méthode DFGS (de Broyden[1], Fletcher[2], Goldfarb[3] et Shanno[4]) cherche à construire $B^{(n)}$ proche de $B^{(n-1)}$, telle que $B^{(n)}$ vérifie (3.3.24) et telle que si $B^{(n-1)}$ est symétrique définie positive alors $B^{(n)}$ 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.

[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

On suppose $x^{(n)}$, $x^{(n-1)}$, $B^{(n-1)}$ connus, et on définit

$$\mathcal{C}_n = \{B \in \mathcal{M}_N(\mathbb{R}) | B \text{ symétrique, vérifiant (3.3.24)}\},$$

qui est une partie de $\mathcal{M}_N(\mathbb{R})$ convexe fermée non vide. On choisit alors $B^{(n)} = P_{\mathcal{C}_n} B^{(n-1)}$ où $P_{\mathcal{C}_n}$ désigne la projection orthogonale sur $\mathcal{C}_n$. La matrice $B^{(n)}$ ainsi définie existe et est unique ; elle est symétrique d'après le choix de $\mathcal{C}_n$. On peut aussi montrer que si $B^{(n-1)}$ symétrique définie positive alors $B^{(n)}$ l'est aussi.
Avec un choix convenable de la norme sur $\mathcal{M}_N(\mathbb{R})$, on obtient le choix suivant de $B^{(n)}$ si $s^{(n)} \neq 0$ et $\nabla f(x^{(n)}) \neq 0$ (sinon l'algorithme s'arrête) :

$$B^{(n)} = B^{(n-1)} + \frac{y^{(n)}(y^{(n)})^t}{(s^{(n)})^t \cdot y^{(n)}} - \frac{B^{(n-1)} s^{(n)} (s^{(n)})^t B^{(n-1)}}{(s^{(n)})^t B^{(n-1)} s^{(n)}}. \tag{3.3.26}$$

L'algorithme obtenu est l'algorithme de BFGS.

Algorithme de BFGS

$$\left\{ \begin{array}{ll} \underline{\text{Initialisation}} & \text{On choisit } x^{(0)} \in \mathbb{R}^N \text{ et} \\ & B^{(0)} \text{ symétrique définie positive} \\ & (\text{ par exemple } B^{(0)} = Id) \text{ et on pose} \\ & w^{(0)} = -B^{(0)} \nabla f(x^{(0)}) \\ & \text{si } \nabla f(x^{(0)}) \neq 0, \text{ on choisit } \rho^{(0)} \text{ optimal} \\ & \text{dans la direction } w^{(0)}, \text{ et donc} \\ & w^{(0)} \text{ est une direction de descente stricte.} \\ & \text{On pose } x^{(1)} = x^{(0)} + \rho^{(0)} w^{(0)}. \\ \underline{\text{Itération } n} & \text{A } x^{(n)}, \ x^{(n-1)} \text{ et } B_{n-1} \text{ connus } \ (n \geq 1) \\ & \text{On suppose} \\ & s^{(n)} = x^{(n)} - x^{(n-1)} \ \ y^{(n)} = \nabla f(x) - \nabla f(x^{(n-1)}) \\ & \text{si } s^{(n)} \neq 0 \text{ et } \nabla f(x^{(n)}) \neq 0, \\ & \text{on choisit } B^{(n)} \text{ vérifiant (3.3.26)} \\ & \text{On calcule } w^{(n)} = -(B^{(n)})^{-1} (\nabla f(x^{(n)})) \\ & \text{(direction de descente stricte en } x^{(n)}). \\ & \text{On calcule } \rho^{(n)} \text{ optimal dans la direction } w^{(n)} \\ & \text{et on pose } x^{(n+1)} = x^{(n)} + \rho^{(n)} w^{(n)}. \end{array} \right. \tag{3.3.27}$$

On donne ici sans démonstration le théorème de convergence suivant :

Théorème 3.31 (Fletcher, 1976) *Soit $f \in C^2(\mathbb{R}^N, \mathbb{R})$ telle que $f(x) \to +\infty$ quand $|x| \to +\infty$. On suppose de plus que f est strictement convexe (donc il existe un unique $\bar{x} \in \mathbb{R}^N$ tel que $f(\bar{x}) = \inf_{\mathbb{R}^N} f$) et on suppose que la matrice hessienne $H_f(\bar{x})$ est symétrique définie positive.*
Alors si $x^{(0)} \in \mathbb{R}^N$ et si $B^{(0)}$ est symétrique définie positive, l'algorithme BFGS définit bien une suite $x^{(n)}$ et on a $x^{(n)} \to \bar{x}$ quand $n \to +\infty$
De plus, si $x^{(n)} \neq \bar{x}$ pour tout n, la convergence est super linéaire i.e.

$$\left| \frac{x^{(n+1)} - \bar{x}}{x^{(n)} - \bar{x}} \right| \to 0 \text{ quand } n \to +\infty.$$

Pour éviter la résolution d'un système linéaire dans BFGS, on peut choisir de travailler sur $(B^{(n)})^{-1}$ au lieu de $B^{(n)}$.

$$\left\{\begin{array}{l}\textbf{Initialisation } \text{Soit } x^{(0)} \in \mathbb{R}^N \text{ et } K^{(0)} \text{ symétrique définie positive}\\ \text{telle que } \rho_0 \text{ soit optimal dans la direction } -K^{(0)}\nabla f(x^{(0)}) = w^{(0)}\\ x^{(1)} = x^{(0)} + \rho_0 w^{(0)}\\ \textbf{Itération } n : \text{A } x^{(n)}, x^{(n-1)}, K^{(n-1)} \text{ connus, } n \geq 1,\\ \text{on pose } s^{(n)} = x^{(n)} - x^{(n-1)}, y^{(n)} = \nabla f(x^{(n)}) - \nabla f(x^{(n-1)})\\ \text{et } K^{(n)} = P_{\mathcal{C}_n} K^{(n-1)}.\\ \text{On calcule } w^{(n)} = -K^{(n)}\nabla f(x^{(n)}) \text{ et on choisit } \rho_n\\ \text{optimal dans la direction } w^{(n)}.\\ \text{On pose alors } x^{(n+1)} = x^{(n)} + \rho_n w^{(n)}.\end{array}\right. \tag{3.3.28}$$

Remarquons que le calcul de la projection de $P_{\mathcal{C}_n} K^{(n-1)}$ peut s'effectuer avec la formule (3.3.26) où on a remplacé $B^{(n-1)}$ par $K^{(n-1)}$. Malheureusement, on obtient expérimentalement une convergence nettement moins bonne pour l'algorithme de quasi-Newton modifié (3.3.28) que pour l'algorithme de BFGS (3.3.26).

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 $\nabla f(x^{(n)})$. Leur convergence est linéaire (donc lente).

Méthode de gradient conjugué : Si f est quadratique (c.à.d. $f(x) = \frac{1}{2}Ax \cdot x - b \cdot x$ 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 $H_f(x^{(n)})$ (et de $\nabla f(^{(n)})$). Si on peut calculer $H_f(x^{(n)})$, cette méthode est parfaite.

Méthode de quasi Newton : L'avantage de la méthode de quasi Newton est qu'on ne calcule que $\nabla f(x^{(n)})$ et pas $H_f(x^{(n)})$). La convergence est super linéaire. Par rapport à une méthode de gradient où on calcule $w^{(n)} = -\nabla f(x^{(n)})$, la méthode BFGS nécessite une résolution de système linéaire : $w^{(n)} = (B^{(n)})^{-1}(-\nabla f(x^{(n)}))$.

Quasi–Newton modifié :

Pour éviter la résolution de système linéaire dans BFGS, on peut choisir de travailler sur $(B^{(n)})^{-1}$ au lieu de $B^{(n)}$, pour obtenir l'algorithme de quasi Newton (3.3.28). Cependant, on perd alors en vitesse de convergence.

Comment faire si on ne veut (ou peut) pas calculer $\nabla f(x^{(n)})$? On peut utiliser des "méthodes sans gradient", c.à.d. qu'on choisit *a priori* les directions $w^{(n)}$. Ceci peut se faire soit par un choix déterministe, soit par un choix stochastique.

Un choix déterministe possible est de calculer $x^{(n)}$ en résolvant N problèmes de minimisation en une dimension d'espace. Pour chaque direction $i = 1, \ldots, N$, on prend $w^{(n,i)} = e_i$, où 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 :

$$f(x_1^{(n)}, x_2^{(n)}, \ldots, \theta, \ldots, x_N^{(n)}) \leq f(x_1^{(n)}, x_2^{(n)}, \ldots, t, \ldots, x_N^{(n)}), \forall t \in \mathbb{R}.$$

Remarquons que si f est quadratique, on retrouve la méthode de Gauss Seidel.

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 :

$$\begin{cases} \bar{u} \in K \\ f(\bar{u}) = \inf\limits_K f \end{cases} \tag{3.4.29}$$

Ce problème est un problème de minimisation avec contrainte (ou "sous contrainte") au sens où l'on cherche u qui minimise f en astreignant u a être dans K. Voyons quelques exemples de ces contraintes (définies par l'ensemble K), qu'on va expliciter à l'aide des p fonctions continues, $g_i \in C(E, \mathbb{R})$ $i = 1 \dots p$.

1. **Contraintes égalités.** On pose $K = \{x \in E,\, g_i(x) = 0\ i = 1 \dots p\}$. On verra plus loin que le problème de minimisation de f peut alors être résolu grâce au théorème des multiplicateurs de Lagrange (voir théorème 3.38).
2. **Contraintes inégalités.** On pose $K = \{x \in E,\, g_i(x) \leq 0\ i = 1 \dots, p\}$. On verra plus loin que le problème de minimisation de f peut alors être résolu grâce au théorème de Kuhn–Tucker (voir théorème 3.42).
   - *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 ùn 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 $f(x) = \frac{1}{2} Ax \cdot x - b \cdot x$, et les fonctions g_i sont affines, alors on dit qu'on a affaire ùn problème de "programmation quadratique".
3. **Programmation convexe.** Dans le cas où f est convexe et K est convexe, on dit qu'on a affaire ùn problème de "programmation convexe".

3.4.2 Existence – Unicité – Conditions d'optimalité simple

Théorème 3.32 (Existence) *Soit $E = \mathbb{R}^N$ et $f \in C(E, \mathbb{R})$.*

1. *Si K est un sous-ensemble fermé borné de E, alors il existe $\bar{x} \in K$ tel que $f(\bar{x}) = \inf\limits_K f$.*
2. *Si K est un sous-ensemble fermé de E, et si f est croissante à l'infini, c'est–à–dire que $f(x) \to +\infty$ quand $|x| \to +\infty$, alors $\exists \bar{x} \in K$ tel que $f(\bar{x}) = \inf\limits_K f$*

Démonstration

1. Si K est un sous-ensemble fermé borné de E, comme f est continue, elle atteint ses bornes sur K, d'où l'existence de $\bar{x}$.
2. Si f est croissante à l'infini, alors il existe $R > 0$ tel que si $\|x\| > R$ alors $f(x) > f(0)$; donc $\inf\limits_K f = \inf\limits_{K \cap B_R} f$, où B_R désigne la boule de centre 0 et de rayon R. L'ensemble $K \cap B_R$ est compact, car intersection d'un fermé et d'un compact. Donc, par ce qui précède, il existe $\bar{x} \in K$ tel que $f(\bar{x}) = \inf\limits_{K \cap B_R} f = \inf\limits_{B_R} f$. ∎

Théorème 3.33 (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 $\bar{x}$ de K tel que $f(\bar{x}) = \inf\limits_K f$.*

Démonstration
Supposons que $\bar{x}$ et $\bar{\bar{x}}$ soient deux solutions du problème (3.4.29), avec $\bar{x} \neq \bar{\bar{x}}$
Alors $f(\frac{1}{2}\bar{x} + \frac{1}{2}\bar{\bar{x}}) < \frac{1}{2}f(\bar{x}) + \frac{1}{2}f(\bar{\bar{x}}) = \inf_K f$. On aboutit donc à une contradiction. ■

Des théorèmes d'existence 3.32 et d'unicité 3.33 on déduit immédiatement le théorème d'existence et d'unicité suivant :

Théorème 3.34 (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 $f(x) \Rightarrow +\infty$ quand $\|x\| \to +\infty$, alors il existe un unique élément $\bar{x}$ de K solution du problème de minimisation* (3.4.29)*, i.e. tel que $f(\bar{x}) = \inf_K f$*

Remarque 3.35 *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 3.36 (Condition simple d'optimalité) *Soient $E = \mathbb{R}^N$, $f \in C(E, \mathbb{R})$ et $\bar{x} \in K$ tel que $f(\bar{x}) = \inf_K f$. On suppose que f est différentiable en $\bar{x}$*

1. *Si $\bar{x} \in \overset{\circ}{K}$ alors $\nabla f(\bar{x}) = 0$.*
2. *Si K est convexe, alors $\nabla f(\bar{x}) \cdot (x - \bar{x}) \geq 0$ pour tout $x \in K$.*

Démonstration

1. Si $\bar{x} \in \overset{\circ}{K}$, alors il existe $\varepsilon > 0$ tel que $B(\bar{x}, \varepsilon) \subset K$ et $f(\bar{x}) \leq f(x)\ \forall x \in B(\bar{x}, \varepsilon)$. Alors on a déjà vu (voir preuve de la Proposition 3.3 page 120) que ceci implique $\nabla f(\bar{x}) = 0$.
2. Soit $x \in K$. Comme $\bar{x}$ réalise le minimum de f sur K, on a : $f(\bar{x} + t(x - \bar{x})) = f(tx + (1-t)\bar{x}) \geq f(\bar{x})$ pour tout $t \in]0, 1]$, par convexité de K. On en déduit que

$$\frac{f(\bar{x} + t(x - \bar{x})) - f(\bar{x})}{t} \geq 0 \text{ pour tout } t \in]0, 1].$$

   En passant à la limite lorsque t tend vers 0 dans cette dernière inégalité, on obtient : $\nabla f(\bar{x}) \cdot (x - \bar{x}) \geq 0$.
■

3.4.3 Conditions d'optimalité dans le cas de contraintes égalité

Dans tout ce paragraphe, on considèrera les hypothèses et notations suivantes :

$$\begin{array}{l} f \in C(\mathbb{R}^N, \mathbb{R}),\ \ g_i \in C^1(\mathbb{R}^N, \mathbb{R}),\ \ i = 1 \dots p\,; \\ K = \{u \in \mathbb{R}^N,\ \ g_i(u) = 0\ \ \forall i = 1 \dots p\}\,; \\ g = (g_1, \dots, g_p)^t \in C^1(\mathbb{R}^N, \mathbb{R}^p) \end{array} \tag{3.4.30}$$

Remarque 3.37 (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 rang$(Dg(u)) = \dim(Im(Dg(u))) \leq p$. On rappelle de plus que

$$Dg(u) = \begin{pmatrix} \frac{\partial g_1}{\partial x_1}, & \dots, & \frac{\partial g_1}{\partial x_N} \\ \dots, & \ddots, & \dots \\ \frac{\partial g_p}{\partial x_1}, & \dots, & \frac{\partial g_p}{\partial x_N} \end{pmatrix},$$

et que rang $(Dg(u)) \leq \min(N,p)$. *De plus, si rang* $(Dg(u)) = p$, *alors les vecteurs* $(Dg_i(u))_{i=1\ldots p}$ *sont linéairement indépendants dans* $\mathbb{R}^N$.

Théorème 3.38 (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 :*

$$\text{il existe } (\lambda_1, \ldots, \lambda_p)^t \in \mathbb{R}^p \; tel que \nabla f(\bar{u}) + \sum_{i=1}^{p} \lambda_i \nabla g_i(\bar{u}) = 0.$$

(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.

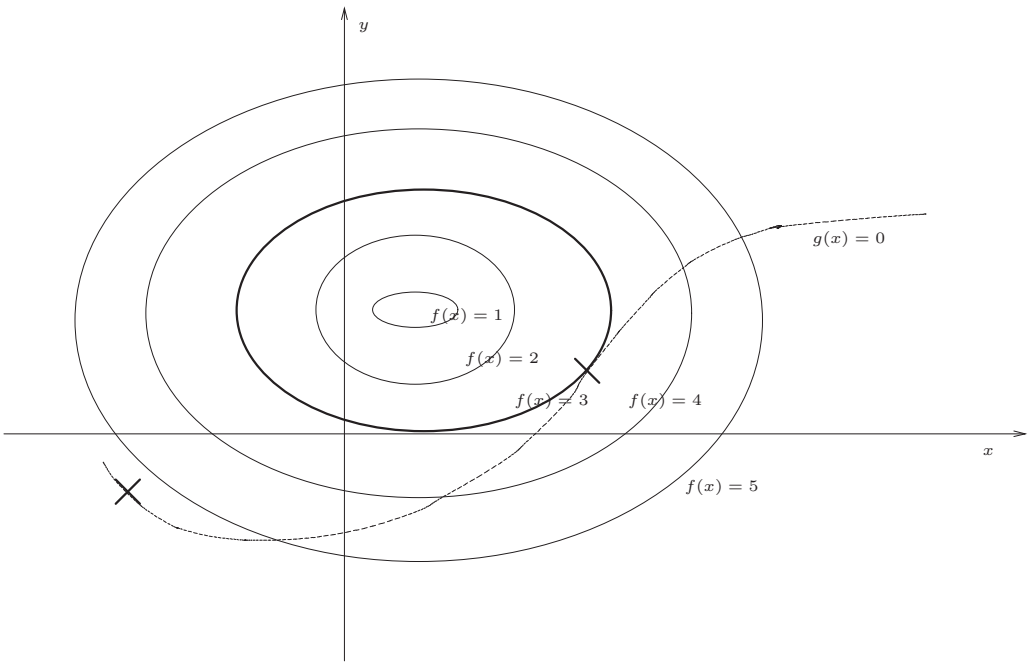

FIG. 3.1 – Interprétation géométrique des multiplicateurs de Lagrange

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 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

$$\begin{cases} (\bar{w}, \bar{v}) \in K \\ \bar{f}(\bar{w}, \bar{v}) \leq f(w, v)\ \ \forall (w, v) \in \overline{K} \end{cases} \tag{3.4.31}$$

D'autre part, le théorème des fonctions implicites (voir note de bas de page 141) 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 (3.4.31) 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 3.36,

$$D\psi(\bar{w}) = 0. \tag{3.4.32}$$

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 (3.4.32) que

$$D_1\bar{f}(\bar{w}, \phi((\bar{w}))w - [D_2\bar{g}(\bar{w}, \phi((\bar{w}))]^{-1} D_1\bar{g}(\bar{w}, \phi((\bar{w}))w = 0,\ \text{pour tout } w \in G. \tag{3.4.33}$$

De plus, comme $D_2\bar{g}(\bar{w}, \phi((\bar{w}))]^{-1} D_2\bar{g}(\bar{w}, \phi((\bar{w})) = Id$, on a :

$$D_2\bar{f}(\bar{w}, \phi((\bar{w}))z - D_2\bar{f}(\bar{w}, \phi((\bar{w}))D_2\bar{g}(\bar{w}, \phi((\bar{w}))]^{-1} D_2\bar{g}(\bar{w}, \phi((\bar{w}))z = 0,\ \forall z \in F. \tag{3.4.34}$$

[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 $D\phi(x) = -[D_2h(x, \phi(x))]^{-1} \cdot D_1h(x, \phi(x))$.

Soit $x \in \mathbb{R}^N$, et $(z,w) \in F \times G$ tel que $x = z + w$. En additionant (3.4.33) et (3.4.34), et en notant $\Lambda = -D_2\bar{f}(\bar{w}, \phi((\bar{w}))D_2\bar{g}(\bar{w}, \phi((\bar{w}))]^{-1}$, on obtient :

$$Df(\bar{u})x + \Lambda Dg(\bar{u})x = 0,$$

ce qui donne, en transposant : $Df(\bar{u}) + \sum_{i=1}^{p} \lambda_i \nabla g_i(\bar{u}) = 0$, avec $\Lambda = (\lambda_1, \ldots, \lambda_p)$.

■

Remarque 3.39 (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 $K = \{u \in \mathbb{R}^N, \quad g_i(u) = 0, \quad i = 1, \ldots, p\}$.*
Le problème qu'on cherche à résoudre est le problème de minimisation (3.4.29) *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 (3.4.29) *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*

$$\begin{cases} \dfrac{\partial f}{\partial x_j}(\bar{u}) + \displaystyle\sum_{i=1}^{p} \lambda_i \frac{\partial g_i}{\partial x_j} = 0, j = 1, \ldots N, \\ g_i(\bar{u}) = 0, \quad i = 1, \ldots, p. \end{cases} \tag{3.4.35}$$

Le système (3.4.35) *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 3.40 *On vient de montrer que si $\bar{x}$ solution de* (3.4.29) *et $Im(Dg(\bar{x})) = \mathbb{R}^p$, alors $\bar{x}$ solution de* (3.4.35)*. Par contre, si $\bar{x}$ est solution de* (3.4.35)*, ceci n'entraîne pas que $\bar{x}$ est solution de* (3.4.29)*.*

Des exemples d'application du théorème des multiplicateurs de Lagrange sont donnés dans les exercices 76 page 157 et 77 page 157.

3.4.4 Contraintes inégalités

Soit $f \in C(\mathbb{R}^N, \mathbb{R})$ et $g_i \in C^1(\mathbb{R}^N, \mathbb{R}) \quad i = 1, \ldots, p$, on considère maintenant un ensemble K de la forme : $K = \{x \in \mathbb{R}^N, \quad g_i(x) \leq 0 \quad \forall i = 1 \ldots p\}$, et on cherche à résoudre le problème de minimisation (3.4.29) qui sécrit :

$$\begin{cases} \bar{x} \in K \\ f(\bar{x}) \leq f(x), \quad \forall x \in K. \end{cases}$$

Remarque 3.41 *Soit $\bar{x}$ une solution de* (3.4.29) *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 $f(\bar{x}) \leq f(x) \quad \forall x \in B(\bar{x}, \varepsilon)$. On est alors ramené à un problème de minimisation sans contrainte, et si i 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 (3.4.29).

Théorème 3.42 (Kuhn–Tucker) *Soit $f \in C(\mathbb{R}^N, \mathbb{R})$, soit $g_i \in C^1(\mathbb{R}^N, \mathbb{R})$, pour $i = 1, \ldots, p$, et soit $K = \{x \in \mathbb{R}^N, \quad g_i(x) \leq 0 \quad \forall i = 1 \ldots p\}$. On suppose qu'il existe $\bar{x}$ solution de* (3.4.29)*, et on pose $I(\bar{x}) = \{i \in \{1, \ldots, p\}; |g_i(\bar{x}) = 0\}$. On suppose que f est différentiable en $\bar{x}$ et que la famille (de $\mathbb{R}^N$) $\{\nabla g_i(\bar{x}), i \in I(\bar{x})\}$ est libre. . Alors il existe une famille $(\lambda_i)_{i \in I(\bar{x})} \subset \mathbb{R}_+$ telle que*

$$\nabla f(\bar{x}) + \sum_{i \in I(\bar{x})} \lambda_i \nabla g_i(\bar{x}) = 0.$$

Remarque 3.43 *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* $\{h(x) = 0\} = \{h(x) \leq)\} \cap \{-h(x) \leq 0\}$. *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 :*

$$\begin{cases} \nabla f(\bar{x}) + \sum_{i=1}^{p} \lambda_i \nabla g_i(\bar{x}) = 0, \\ \lambda_i g_i(\bar{x}) = 0, \quad \forall i = 1, \ldots, p, \\ g_i(\bar{x}) \leq 0, \quad \forall i = 1, \ldots, p, \\ \lambda_i \geq 0, \quad \forall i = 1, \ldots, p. \end{cases}$$

$i = 1 \ldots p \quad g_i(\bar{x}) \leq 0 \quad i = 1 \ldots p$
$\lambda_i \geq 0$

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 3.44 (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. *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 (3.4.29).

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_n connu $\qquad x_{n+1} = p_K(x_n - \rho \nabla f(x_n))$

où p_K est la projection sur K définie par la proposition 3.44.

Lemme 3.45 *Soit* $(x_n)_n$ *construite par l'algorithme (GPFK). On suppose que* $x_n \to x$ *quand* $n + \infty$. *Alors* x *est solution de* (3.4.29).

Démonstration :

Soit $p_K : \mathbb{R}^N \to K \subset \mathbb{R}^N$ la projection sur K définie par la proposition 3.44. Alors p_K est continue. Donc si $x_n \to x$ quand $n \to +\infty$ alors $x = p_K(x - \rho \nabla f(x))$ et $x \in K$ (car $x_n \in K$ et K est fermé).
La caractérisation de $p_K(x - \rho \nabla f(x))$ donnée dans la proposition 3.44 donne alors :

$(x - \rho\nabla f(x) - x/x - y) \geq 0$ pour tout $y \in K$, et comme $\rho > 0$, ceci entraîne $(\nabla f(x)/x - y)$ 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 3.46 (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* (3.4.29)*,*
2. *si $0 < \rho < \dfrac{2\alpha}{M^2}$, la suite (x_n) 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 (3.4.29).
2. On pose, pour $x \in \mathbb{R}^N$, $h(x) = p_K(x - \rho\nabla f(x))$. On a donc $x_{n+1} = h(x_n)$. Pour montrer que la suite $(x_n)_{n\in\mathbb{N}}$ converge, il suffit donc de montrer que h est strictement contractante dès que

$$0 < \rho < \frac{2\alpha}{M^2}. \tag{3.5.36}$$

Grâce au lemme 3.47 démontré plus loin, 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 si la condition (3.5.36) est vérifiée (voir théorème 3.17 page 125 et exercice 63 page 149), et plus précisément :

$$|\bar{h}(x) - \bar{h}(y)| \leq (1 - 2\alpha\rho + M^2\rho^2)|x - y|^2.$$

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 dès que $0 < \frac{2\alpha}{M^2}$. La suite $(x_n)_{n\in\mathbb{N}}$ converge donc bien vers $x = \bar{x}$

Lemme 3.47 (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 3.44, 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

$$\begin{aligned} \|p_K(x) - p_K(y)\|^2 = &\ (p_K(x) - x + x - y + y - p_K(y)|p_K(x) - p_K(y)) \\ = &\ (p_K(x) - x|p_K(x) - p_K(y))_E + (x - y|p_K(x) - p_K(y)) + \\ &\ (y - p_K(y)|p_K(x) - p_K(y)). \end{aligned}$$

Or $(p_K(x) - x|p_K(x) - p_K(y)) \geq 0$ et $(y - p_K(y)|p_K(x) - p_K(y))$, d'où :

$$\|p_K(x) - p_K(y)\| \leq (x - y|p_K(x) - p_K(y)),$$

et donc, grâce à l'inégalité de Cauchy-Schwarz,

$$\|p_K(x) - p_K(y)\| \leq \|x - y\| \, \|p_K(x) - p_K(y)\| \leq \|x - y\|.$$

■

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_n connu
$w_n = -\nabla f(x_n)$; calculer ρ_n optimal dans la direction w_n
$x_{n+1} = p_K(x_n + \rho_n w_n)$

La démonstration de convergence de cet algorithme se déduit de celle de l'algorithme à pas fixe.

Remarque 3.48 *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 :
1er cas On suppose ici que $K = C^+ = \{x \in \mathbb{R}^N, \ x = (x_1, \ldots, x_n)^t \ x_i \geq 0 \ \forall i\}$.

Si $y \in \mathbb{R}^N \ y = (y_+ 1 \ldots y_N)^t$, on peut montrer (exercice 3.6 page 159) que

$$(p_K(y))_i = y_i^+ = \max(y_i, 0), \quad \forall i \in \{1, \ldots, N\}$$

2ème cas 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 :

$$\left\{ \begin{array}{l} f \in \mathcal{C}^1(\mathbb{R}^N, \mathbb{R}), \\ g_i \in \mathcal{C}^1(\mathbb{R}^N, \mathbb{R}), \\ K = \{x \in \mathbb{R}^N, \ g_i(x) \leq 0 \ \ i = 1, \ldots, p\}, \text{ et } K \text{ est non vide.} \end{array} \right. \tag{3.5.37}$$

On définit un problème "primal" comme étant le problème de minimisation d'origine, c'est–à–dire

$$\begin{cases} \bar{x} \in K, \\ f(\bar{x}) \leq f(x), \text{ pour tout } x \in K, \end{cases} \tag{3.5.38}$$

On définit le "lagrangien" comme étant la fonction L définie de $\mathbb{R}^N \times \mathbb{R}^p$ dans $\mathbb{R}$ par :

$$L(x,\lambda) = f(x) + \lambda \cdot g(x) = f(x) + \sum_{i=1}^{p} \lambda_i g_i(x), \tag{3.5.39}$$

avec $g(x) = (g_1(x), \ldots, g_p(x))^t$ et $\lambda = (\lambda_1(x), \ldots, \lambda_p(x))^t$.
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 3.49 *Le théorème de Kuhn-Tucker entraîne que si $\bar{x}$ est solution du problème primal* (3.5.38) *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 :

$$M(\lambda) = \inf_{x \in \mathbb{R}^N} L(x,\lambda), \text{ pour tout } \lambda \in \mathbb{R}^p. \tag{3.5.40}$$

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 :

$$\begin{cases} x \in \mathbb{R}^N \\ L(x,\lambda) \leq L(y,\lambda) \text{ pour tout } x \in \mathbb{R}^N, \end{cases} \tag{3.5.41}$$

Lemme 3.50 *L'application M de $\mathbb{R}^p$ dans $\mathbb{R}$ définie par* (3.5.40) *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 :

$$\begin{cases} \mu \in C^+, \\ M(\mu) \geq M(\lambda) \ \ \forall \lambda \in C^+. \end{cases} \tag{3.5.42}$$

Définition 3.51 *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*

$$L(x, \lambda) \leq L(x, \mu) \leq L(y, \mu) \text{ pour tout } y \in \mathbb{R} \text{ et pour tout } \lambda \in C^+.$$

Proposition 3.52 *Sous les hypothèses* (3.5.37)*, soit* L *définie par* $L(x, \lambda) = f(x) + \lambda g(x)$ *et* $(x, \mu) \in \mathbb{R}^N \times C^+$ *un point selle de* L *sur* $\mathbb{R}^N \times C^+$.
alors

1. $\bar{x}$ *est solution du problème* (3.5.38)*,*
2. μ *est solution de* (3.5.42)*,*
3. $\bar{x}$ *est solution du problème* (3.5.41) *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 (3.5.42), et si $\bar{x}$ solution de (3.5.41) avec $\lambda = \mu$, alors $(\bar{x}, \mu)$ est un point selle de L, et donc $\bar{x}$ est solution de (3.5.38).

De ces résultats découle l'idée de base des méthodes de dualité : on cherche μ solution de (3.5.42). On obtient ensuite une solution $\bar{x}$ du problème (3.5.38), en cherchant $\bar{x}$ comme solution du problème (3.5.41) avec $\lambda = \mu$ (qui est un problème de minimisation sans contraintes). La recherche de la solution μ du problème dual (3.5.42) 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 143) pour résoudre de manière itérative le problème dual (3.5.42). 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 3.44 page 143). L'algorithme (GPFK) pour la recherche de μ s'écrit donc :
Initialisation : $\mu_0 \in C_+$
Itération : $\mu_{n+1} = p_{C_+}(\mu_n + \rho \nabla M(\mu_n))$
Pour définir complètement l'algorithme d'Uzawa, il reste à préciser les points suivants :

1. Calcul de $\nabla M(\mu_n)$,
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 3.6 page 159) : pour $\lambda \in \mathbb{R}^N$, 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 3.53 *Sous les hypothèses* (3.5.37)*, on suppose que pour tout* $\lambda \in \mathbb{R}^N$*, le problème* (3.5.41) *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 (3.5.41). On peut dont maintenant donner le détail de l'itération générale de l'algorithme d'Uzawa :

Itération de l'algorithme d'Uzawa. Soit $\mu_n \in C^+$ connu ;

1. On cherche $x_n \in \mathbb{R}^N$ solution de $\left\{ \begin{array}{l} x_n \in \mathbb{R}^N, \\ L(x_n, \mu_n) \leq L(x, \mu_n), \ \forall x \in \mathbb{R}^N \end{array} \right.$ (On a donc $x_n = x_{\mu_n}$)
2. On calcule $\nabla M(\mu_n) = g(x_n)$

3. $\overline{\mu}_{n+1} = \mu_n + \rho \nabla M(\mu_n) = \mu_n + \rho g(x_n) = ((\overline{\mu}_{n+1})_1, \ldots, (\overline{\mu}_{n+1})_p)^t$
4. $\mu_{n+1} = p_{C^+}(\overline{\mu}_{n+1})$, c'est–à-dire $\mu_{n+1} = ((\mu_{n+1})_1, \ldots, (\mu_{n+1})_p)^t$ avec $(\mu_{n+1})_i = \max(0, (\overline{\mu}_{n+1})_i)$ pour tout $i = 1, \ldots, p$.

On a alors le résultat suivant de convergence de l'algorithme :

Proposition 3.54 (Convergence de l'algorithme d'Uzawa) *Sous les hypothèses* (3.5.37)*, on suppose de plus que :*

1. *il existe* $\alpha > 0$ *tel que* $(\nabla f(x) - \nabla f(y)) \cdot (x - y) \geq \alpha |x - y|^2$ *pour tout* $(x, y) \in (\mathbb{R}^N)^2$,
2. *il existe* $M_f > 0$ $|\nabla f(x) - \nabla f(y)| \leq M_f |x - y|$ *pour tout* $(x, y) \in (\mathbb{R}^N)^2$,
3. *pour tout* $\lambda \in C^+$*, il existe un unique* $x_\lambda \in \mathbb{R}^N$ *tel que* $L(x_\lambda, \lambda) \leq L(x, \lambda)$ *pour tout* $x \in \mathbb{R}^N$.

Alors si $0 < \rho < \dfrac{2\alpha}{{M_f}^2}$*, la suite* $((x_n, \mu_n))_n \in \mathbb{R}^N \times C^+$ *donnée par l'algorithme d'Uzawa vérifie :*

1. $x_n \to \bar{x}$ *quand* $n \to +\infty$*, où* $\bar{x}$ *est la solution du problème* (3.5.38)*,*
2. $(\mu_n)_{n \in \mathbb{N}}$ *est bornée.*

Remarque 3.55 (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* (3.5.41) *et existence et unicité de la solution* $\bar{x}$ *du problème* (3.5.38)*.)*

3.6 Exercices

Exercice 60 (Convexité et continuité) *Suggestions en page 161.*

1. Soit $f : \mathbb{R} \to \mathbb{R}$. On suppose que f est 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.

Exercice 61 (Maximisation)
Suggestions en page 161

Soit E un espace vectoriel normé et $f : E \to \mathbb{R}$.

1. Donner une condition suffisante d'existence de $\bar{x} \in E$ tel que $x = \sup_{x \in E} f(x)$.
2. Donner une condition suffisante d'unicité de $\bar{x} \in E$ tel que $x = \sup_{x \in E} f(x)$.
3. Donner une condition suffisante d'existence et unicité de $\bar{x} \in E$ tel que $x = \sup_{x \in E} f(x)$.

Exercice 62 (Minimisation d'une fonctionnelle quadratique) *Suggestions en page 3.7. Corrigé détaillé en page 164*

Soient $A \in \mathcal{M}_N(\mathbb{R})$, $b \in \mathbb{R}^N$, et f la fonction de $\mathbb{R}^N$ dans $\mathbb{R}$ définie par $f(x) = \frac{1}{2}Ax \cdot x - b \cdot x$.

1. Montrer que $f \in C^\infty(\mathbb{R}^N, \mathbb{R})$ et calculer le gradient et la matrice hessienne de f en tout point.

2. Montrer que si A est symétrique définie positive alors il existe un unique $\bar{x} \in \mathbb{R}^N$ qui minimise f, et que ce $\bar{x}$ est l'unique solution du système linéaire $Ax = b$.

Exercice 63 (Convergence de l'algorithme du gradient à pas fixe)
Suggestions en page 162, corrigé détaillé en page 164

Soit $f \in C^1(\mathbb{R}^N, \mathbb{R})$ ($N \geq 1$). On suppose que f vérifie :

$$\exists \alpha > 0 \text{ t.q. } (\nabla f(x) - \nabla f(y)) \cdot (x - y) \geq \alpha |x - y|^2, \ \forall x, y \in \mathbb{R}^N, \tag{3.6.43}$$

$$\exists M > 0 \text{ t.q. } |\nabla f(x) - \nabla f(y)| \leq M |x - y|, \ \forall x, y \in \mathbb{R}^N. \tag{3.6.44}$$

1. Montrer que
$$f(y) - f(x) \geq \nabla f(x) \cdot (y - x) + \frac{\alpha}{2} |y - x|^2, \ \forall x, y \in \mathbb{R}^N.$$

2. Montrer que f est strictement convexe et que $f(x) \to \infty$ quand $|x| \to \infty$. En déduire qu'il existe un et un seul $\overline{x} \in \mathbb{R}^N$ t.q. $f(\overline{x}) \leq f(x)$ pour tout $x \in \mathbb{R}^N$.
3. Soient $\rho \in]0, (2\alpha/M^2)[$ et $x_0 \in \mathbb{R}^N$. Montrer que la suite $(x_n)_{n\in\mathbb{N}}$ définie par $x_{n+1} = x_n - \rho\nabla f(x_n)$ (pour $n \in \mathbb{N}$) converge vers $\overline{x}$.

Exercice 64 (Mise en oeuvre de GPF et GPO) .

On considère la fonction $f : \mathbb{R}^2 \to \mathbb{R}$ définie par $f(x_1, x_2) = 2x_1^2 - x_1x_2 - 3x_1 - 2x_2 + 4$.

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 les deux premiers itérés donnés par l'algorithme du gradient à pas fixe (GPF) et du gradient à pas optimal (GPO), en partant de $(x_1^{(0)}, x_2^{(0)}) = (0, 0)$.

Exercice 65 (Convergence de l'algorithme du gradient à pas optimal) *Suggestions en page 162. Corrigé détaillé en page 165*

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 $(x_n)_{n\in\mathbb{N}}$ de l'algorithme du gradient à pas optimal.) On suppose x_n connu ($n \in N$). On pose $w_n = -\nabla f(x_n)$. Si $w_n = 0$, on pose $x_{n+1} = x_n$. Si $w_n \neq 0$, montrer qu'il existe $\overline{\rho} > 0$ t.q. $f(x_n+\overline{\rho}w_n) \leq f(x_n+\rho w_n)$ pour tout $\rho \geq 0$. On choisit alors un $\rho_n > 0$ t.q. $f(x_n+\rho_n w_n) \leq f(x_n+\rho w_n)$ pour tout $\rho \geq 0$ et on pose $x_{n+1} = x_n + \rho_n w_n$.
   On considère, dans les questions suivantes, la suite $(x_n)_{n\in\mathbb{N}}$ ainsi construite.
4. Montrer que (avec R et M donnés aux questions précédentes)
   (a) la suite $(f(x_n))_{n\in\mathbb{N}}$ est une suite convergente,
   (b) $x_n \in B_R$ pour tout $n \in \mathbb{N}$,
   (c) $f(x_n + \rho w_n) \leq f(x_n) - \rho|w_n|^2 + (\rho^2/2)M|w_n|^2$ pour tout $\rho \in [0, 1/|w_n|]$.
   (d) $f(x_{n+1}) \leq f(x_n) - |w_n|^2/(2M)$, si $|w_n| \leq M$.
   (e) $-f(x_{n+1}) + f(x_n) \geq |w_n|^2/(2\overline{M})$, avec $\overline{M} = \sup(M, \tilde{M})$, $\tilde{M} = \sup\{|\nabla f(x)|, x \in B_R\}$.
5. Montrer que $\nabla f(x_n) \to 0$ (quand $n \to \infty$) 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 $x_n \to \overline{x}$ quand $n \to \infty$.

Exercice 66 (Algorithme du gradient à pas optimal)

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 ρ_n en fonction de A et de x_n.] A quelle condition suffisante cet algorithme converge-t-il ?

Exercice 67 (Fonction non croissante à l'infini) *Suggestions en page 162.*

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 $T(x) = \sup\{\rho \geq 0;\ [x, x - \rho\nabla f(x)] \subset A\}$. 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 $\rho_k \in [0, T(x_k)]$ t.q. $f(x_k - \rho_k \nabla f(x_k)) = \min\{f(x_k - \rho\nabla f(x_k)), 0 \leq \rho \leq T(x_k)\}$ (La fonction T est définie à la question 2) et on pose $x_{k+1} = x_k - \rho_k \nabla f(x_k)$.

   (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 ρ_k, t.q. $f(x_k - \rho_k\nabla f(x_k) = \min\{f(x_k - \rho\nabla f(x_k)), 0 \leq \rho \leq T(x_k)\}$).

   (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 $(x_{k_n})_{n\in\mathbb{N}}$ et $x \in A$ t.q. $x_{k_n} \to x$, quand $n \to \infty$, 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 68 (Méthode de relaxation) *Corrigé détaillé en page 168*

Soit f une fonction continûment différentiable de $E = \mathbb{R}^N$ dans $\mathbb{R}$ vérifiant l'hypothèse (3.6.43) :
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 :

$$
\left\{
\begin{array}{ll}
\text{Initialisation :} & x^{(0)} \in E, \\
\text{Itération } n : & x^{(n)} \text{ connu, } (n \geq 0) \\
& \text{Calculer } x_1^{(n+1)} \text{ tel que, pour tout } \xi \in \mathbb{R}, \\
& f(x_1^{(n+1)}, x_2^{(n)}, x_3^{(n)}, \ldots, x_N^{(n)}) \leq f(\xi, x_2^{(n)}, x_3^{(n)}, \ldots, x_N^{(n)}), \\
& \text{Calculer } x_2^{(n+1)} \text{ tel que, pour tout } \xi \in \mathbb{R}, \\
& f(x_1^{(n+1)}, x_2^{(n+1)}, x_3^{(n)}, \ldots, x_N^{(n)}) \leq f(x_1^{(n+1)}, \xi, x_3^{(n)}, \ldots, x_N^{(n)}), \\
& \qquad \ldots \\
& \text{Calculer } x_k^{(n+1)} \text{ tel que, pour tout } \xi \in \mathbb{R}, \\
& f(x_1^{(n+1)}, \ldots, x_{k-1}^{(n+1)}, x_k^{(n+1)}, x_{(k+1)}^{(n)}, \ldots, x_N^{(n)}) \\
& \qquad \leq f(x_1^{(n+1)}, \ldots, x_{k-1}^{(n+1)}, \xi, x_{(k+1)}^{(n)}, \ldots, x_N^{(n)}), \\
& \qquad \ldots \\
& \text{Calculer } x_N^{(n+1)} \text{ tel que, pour tout } \xi \in \mathbb{R}, \\
& f(x_1^{(n+1)}, x_2^{(n+1)}, \ldots, x_{N-1}^{(n+1)}, x_N^{(n+1)}) \leq f(x_1^{(n+1)}, \ldots, x_{N-1}^{(n+1)}, \xi).
\end{array}
\right. \tag{3.6.45}
$$

2. Pour $n \in \mathbb{N}$ et $1 \leq k \leq N$, soit $\varphi_k^{(n+1)}$ la fonction de $\mathbb{R}$ dans $\mathbb{R}$ définie par :

$$\varphi_k^{(n+1)}(s) = f(x_1^{(n+1)}, \ldots, x_{k-1}^{(n+1)}, s, x_{(k+1)}^{(n)}, \ldots, x_N^{(n)}).$$

Montrer qu'il existe un unique élément $\overline{s} \in \mathbb{R}$ tel que

$$\varphi_k^{(n+1)}(\overline{s}) = \inf_{s \in \mathbb{R}} \varphi_k^{(n+1)}(s).$$

En déduire que la suite $(x^{(n)})_{n \in \mathbb{N}}$ construite par l'algorithme (3.6.45) 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 $(x^{(n)})_{n \in \mathbb{N}}$ la suite définie par l'algorithme (3.6.45). Pour $n \geq 0$, on définit $x^{(n+1,0)} = x^{(n)} = (x_1^{(n)}, \ldots, x_N^{(n)})^t$, et pour $1 \leq k \leq N$, $x^{(n+1,k)} = (x_1^{(n+1)}, \ldots, x_k^{(n+1)}, x_{k+1}^{(n)}, \ldots, x_N^{(n)})^t$ (de sorte que $x^{(n+1,N)} = x^{(n+1)}$).

(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 $(x^{(n)})_{n \in \mathbb{N}}$ vérifie

$$f(x^{(n)}) - f(x^{(n+1)}) \geq \frac{\alpha}{2} \|x^{(n)} - x^{(n+1)}\|^2.$$

En déduire que $\lim_{n \to +\infty} \|x^{(n)} - x^{(n+1)}\| = 0$ et que, pour $1 \leq k \leq N$, $\lim_{n \to +\infty} \|x^{(n+1,k)} - x^{(n+1)}\| = 0$.

4. Montrer que

$$\|x^{(n+1)} - \overline{x}\| \leq \frac{1}{\alpha} \left(\sum_{k=1}^{N} |\partial_k f(x^{(n+1)})|^2 \right)^{\frac{1}{2}}.$$

5. Montrer que les suites $(x^{(n)})_{n\in\mathbb{N}}$, et $(x^{(n+1,k)})_{n\in\mathbb{N}}$, pour $k = 1, \ldots, N$, sont bornées.
Montrer que

$$|\partial_k f(x^{(n+1)})| \to 0 \text{ lorsque } n \to +\infty.$$

(On rappelle que $\partial_k f(x^{(n+1,k)}) = 0$.)
Conclure quant à la convergence de la suite$(x^{(n)})_{n\in\mathbb{N}}$ 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 (3.6.45) 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 (3.6.45) appliqué à g ne converge pas vers $\overline{x}$. Quelle est l'hypothèse mise en défaut ici ?

Exercice 69 (Gradient conjugué pour une matrice non symétrique)

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

$$Ax = b;. \tag{3.6.46}$$

1. Montrer que $x \in \mathbb{R}^N$ est solution de (1.6.60) si et seulement si x est solution de

$$Mx = A^tb;. \tag{3.6.47}$$

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 (3.6.47), 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 70 (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 $y^{(0)}, \ldots, y^{(n)}$ et $\tilde{w}^{(0)}, \ldots, \tilde{w}^{(n-1)}$ connus, et on pose : $\tilde{r}^{(n)} = \tilde{b} - By^{(n)}$. Si $\tilde{r}^{(n)} \neq 0$, on calcule : $\tilde{w}^{(n)} = \tilde{r}^{(n)} + \lambda_{n-1}\tilde{w}^{(n-1)}$ avec $\lambda_{n-1} = \frac{\tilde{r}^{(n)} \cdot \tilde{r}^{(n)}}{\tilde{r}^{(n-1)} \cdot \tilde{r}^{(n-1)}}$ et on pose alors : $y^{(n+1)} = y^{(n)} + \rho_n \tilde{w}^{(n)}$ avec $\rho_n = \frac{\tilde{r}^{(n)} \cdot \tilde{r}^{(n)}}{\tilde{w}^{(n)} \cdot B\tilde{w}^{(n)}}$,

3. En utilisant le cours, justifier que la famille $y^{(n)}$ ainsi construite est finie. A quoi est égale sa dernière valeur ?

Pour $n \in \mathbb{N}$, on pose : $x^{(n)} = L^{-t}y^{(n)}$ (avec $L^{-t} = (L^{-1})^t = (L^t)^{-1}$), $r^{(n)} = b - Ax^{(n)}$, $w^{(n)} = L^{-t}\tilde{w}^{(n)}$ et $s^{(n)} = (LL^t)^{-1}r^{(n)}$.

4. Soit $n > 0$ fixé. Montrer que :

$$(a) \quad \lambda_{n-1} = \frac{s^{(n)} \cdot r^{(n)}}{s^{(n-1)} \cdot r^{(n-1)}}, \qquad (b) \quad \rho_n = \frac{s^{(n)} \cdot r^{(n)}}{w^{(n)} \cdot Aw^{(n)}},$$

$$(c) \quad w^{(n)} = s^{(n)} + \lambda_n w^{(n-1)}, \qquad (d) \quad x^{(n+1)} = x^{(n)} + \rho_n w^{(n)}.$$

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 71 (Méthode de Polak-Ribière) *Suggestions en page 162, corrigé en page 170*

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 $\alpha > 0$, $\beta \geq \alpha$ t.q. $\alpha|y|^2 \leq H(x)y \cdot y \leq \beta|y|^2$ 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. $x^{(n)}$, $w^{(n-1)}$ connus ($n \geq 1$). On pose $g^{(n)} = -\nabla f(x^{(n)})$. Si $g^{(n)} = 0$, l'algorithme s'arrête (on a $x^{(n)} = \overline{x}$). Si $g^{(n)} \neq 0$, on pose $\lambda_{n-1} = [g^{(n)} \cdot (g^{(n)} - g^{(n-1)})]/[g^{(n-1)} \cdot g^{(n-1)}]$, $w^{(n)} = g^{(n)} + \lambda_{n-1} w^{(n-1)}$ et $x^{(n+1)} = x^{(n)} + \rho_n w^{(n)}$ avec ρ_n "optimal" dans la direction w_n. (Noter que ρ_n existe bien.)

   On suppose dans la suite que $g^{(n)} \neq 0$ pour tout $n \in \mathbb{N}$.
2. Montrer (par récurrence sur n) que $g^{(n+1)} \cdot w^{(n)} = 0$ et $g^{(n)} \cdot g^{(n)} = g^{(n)} \cdot w^{(n)}$, pour tout $n \in \mathbb{N}$.
3. On pose

$$J^{(n)} = \int_0^1 H(x^{(n)} + \theta \rho_n w^{(n)}) d\theta.$$

   Montrer que $g^{(n+1)} = g^{(n)} + \rho_n J^{(n)} w^{(n)}$ et que $\rho_n = (-g^{(n)} \cdot w^{(n)})/(J^{(n)} w^{(n)} \cdot w^{(n)})$ (pour tout $n \in \mathbb{N}$).
4. Montrer que $|w^{(n)}| \leq (1 + \beta/\alpha)|g^{(n)}|$ pour tout $n \in \mathbb{N}$. [Utiliser, pour $n \geq 1$, la question précédente et la formule donnant λ_{n-1}.]
5. Montrer que $x^{(n)} \to \overline{x}$ quand $n \to \infty$.

Exercice 72 (Algorithme de quasi Newton)
Corrigé détaillé en page 173

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. $x^{(n)}$ connu ($n \geq 0$. On pose $x^{(n+1)} = x^{(n)} - \rho_n K^{(n)} g^{(n)}$ avec $g^{(n)} = \nabla f(x^{(n)})$, $K^{(n)}$ une matrice symétrique définie positive à déterminer et ρ_n "optimal" dans la direction $w^{(n)} = -K^{(n)} g^{(n)}$. (Noter que ρ_n existe bien.)

Partie 1. Calcul de ρ_n. On suppose que $g^{(n)} \neq 0$.

1. Montrer que $w^{(n)}$ est une direction de descente stricte en $x^{(n)}$ et calculer la valeur de ρ_n (en fonction de $K^{(n)}$ et $g^{(n)}$).
2. On suppose que, pour un certain $n \in \mathbb{N}$, on a $K^{(n)} = (H(x^{(n)}))^{-1}$ (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 $\rho_n = 1$.
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 $K^{(0)} = Id$ et

$$K^{(n+1)} = K^{(n)} + \frac{s^{(n)}(s^{(n)})^t}{s^{(n)} \cdot y^{(n)}} - \frac{(K^{(n)}y^{(n)})(K^{(n)}(y^{(n)})^t}{K^{(n)}y^{(n)} \cdot y^{(n)}},\ n \geq 0, \tag{3.6.48}$$

avec $s^{(n)} = x^{(n+1)} - x^{(n)}$ et $y^{(n)} = g^{(n+1)} - g^{(n)} = As^{(n)}$.

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 $s^{(0)}, \ldots, s^{(n-1)}$ sont des vecteurs A-conjugués et non-nuls et que $K^{(0)}, \ldots, K^{(n)}$ 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 $g^{(n)} \neq 0$. Montrer que $s^{(n)} \neq 0$ (cf. Partie I) et que, pour $i < n$,

   $$s^{(n)} \cdot As^{(i)} = 0 \Leftrightarrow g^{(n)} \cdot s^{(i)} = 0.$$

   Montrer que $g^{(n)} \cdot s^{(i)} = 0$ pour $i < n$. [On pourra remarquer que $g^{(i+1)} \cdot s^{(i)} = g^{(i+1)} \cdot w^{(i)} = 0$ et $(g^{(n)} - g^{(i+1)}) \cdot s^{(i)} = 0$ par l'hypothèse de conjugaison de $s^{(0)}, \ldots, s^{(n-1)}$.] En déduire que $s^{(0)}, \ldots, s^{(n)}$ sont des vecteurs A-conjugués et non-nuls.

   (b) Montrer que $K^{(n+1)}$ est symétrique.

   (c) Montrer que $K^{(n+1)}As^{(i)} = s^{(i)}$ si $0 \leq i \leq n$.

   (d) Montrer que, pour tout $x \in \mathbb{R}^N$, on a

   $$K^{(n+1)}x \cdot x = \frac{(K^{(n)}x \cdot x)(K^{(n)}y^{(n)} \cdot y^{(n)}) - (K^{(n)}y^{(n)} \cdot x)^2}{K^{(n)}y^{(n)} \cdot y^{(n)}} + \frac{(s^{(n)} \cdot x)^2}{As^{(n)} \cdot s^{(n)}}.$$

   En déduire que $K^{(n+1)}$ 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 $g^{(n)} \neq 0$ si $0 \leq n \leq N-1$. Montrer (par récurrence sur n, avec la question précédente) que $s^{(0)}, \ldots, s^{(N-1)}$ sont des vecteurs A-conjugués et non-nuls et que $K^{(N)}As^{(i)} = s^{(i)}$ si $i < N$. En déduire que $K^{(N)} = A^{-1}$, $\rho_N = 1$ et $x^{(N+1)} = A^{-1}b = \overline{x}$.

Exercice 73 (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 $d \in \mathbb{R}^P$. Pour $x \in \mathbb{R}^N$, on pose

$$J(x) = (f(x) - d) \cdot C(f(x) - d).$$

On cherche à minimiser J.

I *Propriétés d'existence et d'unicité*

(a) Montrer que J est bornée inférieurement.

(b) Pour C et d donnés, donner trois exemples de fonctions f pour lesquels les fonctionnelles J associées sont telles que l'on ait :

i. existence et unicité de $\bar{x} \in \mathbb{R}^N$ qui réalise le minimum de J, pour le premier exemple.

ii. existence et non unicité de $\bar{x} \in \mathbb{R}^N$ qui réalise le minimum de J, pour le second exemple.

iii. non existence de $\bar{x} \in \mathbb{R}^N$ qui réalise le minimum de J, pour le troisième exemple.

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 $Df(x)$, $D_2f(x)$ (pour $x \in \mathbb{R}^N$), ainsi que Df et D_2f ? Montrer que pour tout $x \in \mathbb{R}^N$, il existe $M(x) \in \mathcal{M}_{P,N}(\mathbb{R})$, où $\mathcal{M}_{P,N}(\mathbb{R})$ désigne l'ensemble des matrices réelles à P lignes et N colonnes, telle que $Df(x)(y) = M(x)y$ pour tout $y \in \mathbb{R}^N$.

(b) Pour $x \in \mathbb{R}^N$, calculer $\nabla J(x)$.

(c) Pour $x \in \mathbb{R}^N$, calculer la matrice hessienne de J en x (qu'on notera $H(x)$). On suppose maintenant que M ne dépend pas de x ; montrer que dans ce cas $H(x) = 2M(x)^t CM(x)$.

III *Algorithmes d'optimisation*

Dans toute cette question, on suppose qu'il existe un unique $\bar{x} \in \mathbb{R}^N$ 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 $(x_n)_{n\in\mathbb{N}}$ 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_n 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 $n \in \mathbb{N}$, la minimisation de la fonctionnelle J par celle de la fonctionnelle J_n, définie de $\mathbb{R}^N$ dans $\mathbb{R}$, et obtenue à partir de J en effectuant un développement limité au premier ordre de $f(x)$ en x_n, c.à.d.

$$J_n(x) = (f(x_n) + Df(x_n)(x - x_n) - d) \cdot C(f(x_n) + Df(x_n)(x - x_n) - d).$$

i. Soit $n \geq 0$, $x_n \in \mathbb{R}^N$ connu, $M_n = M(x_n) \in \mathcal{M}_{P,N}(\mathbb{R})$, et $x \in \mathbb{R}^N$. On pose $h = x - x_n$. Montrer que

$$J_n(x) = J(x_n) + M_n^t CM_n h \cdot h + 2M_n^t C(f(x_n) - d) \cdot h.$$

ii. Montrer que la recherche du minimum de J_n 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 74 (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 $n \in \mathbb{N}$, on définit la fonction f_n par $f_n(x) = f(x) + n\psi(x)$.

1. Montrer qu'il existe au moins un élément $\bar{x}_n \in \mathbb{R}^N$ tel que $f_n(\bar{x}_n) = \inf_{x\in\mathbb{R}^N} f_n(x)$, 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 $n \in \mathbb{N}$,
$$f(\bar{x}_n) \leq f_n(\bar{x}_n) \leq f(\bar{x}_K).$$
3. En déduire qu'il existe une sous-suite $(\bar{x}_{n_k})_{k\in\mathbb{N}}$ et $y \in K$ tels que $\bar{x}_{n_k} \to y$ lorsque $k \to +\infty$.
4. Montrer que $y = \bar{x}_K$. En déduire que toute la suite $(\bar{x}_n)_{n\in\mathbb{N}}$ converge vers $\bar{x}_K$.
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 75 (Sur l'existence et l'unicité) *Corrigé en page 175*

Etudier l'existence et l'unicité des solutions du problème (3.4.29), 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 :

$$\begin{array}{llll} (i) & K = \{|x| \leq 1\}\,; & (ii) & K = \{|x| = 1\} \\ (iii) & K = \{|x| \geq 1\}\,; & (iv) & K = \{|x| > 1\}. \end{array} \tag{3.6.49}$$

Exercice 76 (Aire maximale d'un rectangle à périmètre donné)
Corrigé en page 176

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 (3.4.29), 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

$$\begin{cases} \bar{x} = (\bar{x}_1, \bar{x}_2)^t \in \tilde{K} \\ f(\bar{x}_1, \bar{x}_2) \leq f(x_1, x_2), \quad \forall\, (x_1, x_2)^t \in \tilde{K}, \end{cases} \tag{3.6.50}$$

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 (3.6.50) alors $x = (1/2, 1/2)$. En déduire que le problème (3.6.50) admet une unique solution donnée par $\bar{x} = (1/2, 1/2)$.

Exercice 77 (Fonctionnelle quadratique) *Suggestions en page 163, corrigé en page 176*

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 $K = \{x \in \mathbb{R}^N, \quad g(x) = 0\}$ et on cherche à résoudre le problème de minimisation (3.4.29).

1. Montrer que l'ensemble K est non vide, fermé et convexe. En déduire que le problème (3.4.29) admet une unique solution.

2. Montrer que si $\bar{x}$ est solution de (3.4.29), alors il existe $\lambda \in \mathbb{R}$ tel que $y = (\bar{x}, \lambda)^t$ soit l'unique solution du système :

$$\left[\begin{array}{c|c} A & d \\ \hline d^t & 0 \end{array}\right] \left[\begin{array}{c} \bar{x} \\ \hline \lambda \end{array}\right] = \left[\begin{array}{c} b \\ \hline c \end{array}\right] \tag{3.6.51}$$

Exercice 78 (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 $a = (a_1, \ldots, a_N) \in \mathbb{R}^N$, $a \neq 0$. Pour $x = (x_1, \ldots, x_N) \in \mathbb{R}^N$, on pose : $f(x) = \sum_{i=1}^N |x_i - a_i|^2$, $g(x) = \sum_{i=1}^N |x_i|^2$. 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 $b \in \mathbb{R}^N$. Pour $v \in \mathbb{R}^N$, on pose $f(v) = (1/2)Av \cdot v - b \cdot v$ et $g(v) = Bv \cdot v$. Peut-on appliquer le théorème de Lagrange et quelle condition donne-t-il sur u si $f(u) = \min\{f(v), v \in K\}$ avec $K = \{v \in \mathbb{R}^N; g(v) = 1\}$?

Exercice 79 (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, 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 :

$$u \in U, \ J(u) \leq J(v), \ \forall v \in U. \tag{3.6.52}$$

2. Soit $u \in U$, montrer que u est solution de (3.6.52) si et seulement si $(Au - b) \cdot (v - u) + j(v) - j(u) \geq 0$, pour tout $v \in U$.

Exercice 80

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 81 (Application simple du théorème de Kuhn-Tucker) *Corrigé en page 177*

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 (3.4.29) et appliquer le théorème de Kuhn-Tucker pour la détermination de cette solution.

Exercice 82 (Exemple d'opérateur de projection)

Correction en page 177

1.Soit $K = C^+ = \{x \in \mathbb{R}^N,\ x = (x_1, \ldots, x_n)^t,\ x_i \geq 0,\ \forall i = 1, \ldots, N\}$.

(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 $K = \{x = (x_1, \ldots, x_N)^t; \alpha_i \leq \beta_i, i = 1, \ldots, N\}$.

1. Montrer que K est un convexe fermé non vide.

2. Soit p_K l'opérateur de projection définie à la proposition 3.44 page 143. Montrer que pour tout $y \in \mathbb{R}^N$, on a :

$$(p_K(y))_i = \max(\alpha_i, \min(y_i, \beta_i)), \quad \forall i = 1, \ldots, N$$

Exercice 83 (Méthode de relaxation avec Newton problèmes sans contrainte)

On considère le problème :

$$\left\{ \begin{array}{l} \overline{x} \in K, \\ f(\overline{x}) \leq f(x), \forall x \in K, \end{array} \right. \tag{3.6.53}$$

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 :

$$\left\{ \begin{array}{ll} \text{Initialisation :} & x^{(0)} \in E, \\ \text{Itération } n : & x^{(n)} \text{ connu, } (n \geq 0) \\ & \text{Calculer } x_1^{(n+1)} \in [a_1, b_1] \text{ tel que :} \\ & f(x_1^{(n+1)}, x_2^{(n)}, x_3^{(n)}, \ldots, x_N^{(n)}) \leq f(\xi, x_2^{(n)}, x_3^{(n)}, \ldots, x_N^{(n)}), \text{ pour tout } \xi \in [a_1, b_1], \\ & \text{Calculer } x_2^{(n+1)} \in [a_2, b_2] \text{ tel que :} \\ & f(x_1^{(n+1)}, x_2^{(n+1)}, x_3^{(n)}, \ldots, x_N^{(n)}) \leq f(x_1^{(n+1)}, \xi, x_3^{(n)}, \ldots, x_N^{(n)}), \\ & \qquad \text{pour tout } \xi \in [a_2, b_2], \\ & \qquad \ldots \\ & \text{Calculer } x_k^{(n+1)} \in [a_k, b_k], \text{ tel que :} \\ & f(x_1^{(n+1)}, \ldots, x_{k-1}^{(n+1)}, x_k^{(n+1)}, x_{(k+1)}^{(n)}, \ldots, x_N^{(n)}) \\ & \qquad \leq f(x_1^{(n+1)}, \ldots, x_{k-1}^{(n+1)}, \xi, x_{(k+1)}^{(n)}, \ldots, x_N^{(n)}), \text{ pour tout } \xi \in [a_k, b_k], \\ & \qquad \ldots \\ & \text{Calculer } x_N^{(n+1)} \in [a_N, b_N] \text{ tel que :} \\ & f(x_1^{(n+1)}, x_2^{(n+1)}, \ldots, x_{N-1}^{(n+1)}, x_N^{(n+1)}) \leq f(x_1^{(n+1)}, \ldots, x_{N-1}^{(n+1)}, \xi), \\ & \qquad \text{pour tout } \xi \in [a_N, b_N]. \end{array} \right. \tag{3.6.54}$$

Montrer que la suite $x^{(n)}$ construite par l'algorithme (3.6.54) 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 $\overline{x}$:

$$\left\{ \begin{array}{ll} \text{Initialisation :} & x^{(0)} \in E, \\ \text{Itération } n : & x^{(n)} \text{ connu, } (n \geq 0) \\ & \text{Calculer } x_1^{(n+1)} \geq 2 - x_2^{(n)} \text{ tel que :} \\ & f(x_1^{(n+1)}, x_2^{(n)}) \leq f(\xi, x_2^{(n)}), \text{ pour tout } \xi \geq 2 - x_2^{(n)}, \\ & \text{Calculer } x_2^{(n+1)} \geq 2 - x_1^{(n)} \text{ tel que :} \\ & f(x_1^{(n+1)}, x_2^{(n+1)}) \leq f(x_1^{(n+1)}, \xi), \text{ pour tout } \xi \geq 2 - x_1^{(n)}. \end{array} \right. \tag{3.6.55}$$

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.

Exercice 84 (Convergence de l'algorithme d'Uzawa)

Soient $N, p \in \mathbb{N}^\star$. Soit $f \in C^1(\mathbb{R}^N, \mathbb{R})$ ($N \geq 1$) t.q.

$$\exists \alpha > 0, \ (\nabla f(x) - \nabla f(y)) \cdot (x - y) \geq \alpha |x - y|^2, \ \forall x, y \in \mathbb{R}^N.$$

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 :

$$x \in D, \quad f(x) \leq f(y), \ \forall y \in D. \tag{3.6.56}$$

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 (3.6.56).

   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 :

   $$x \in \mathbb{R}^N, \quad L(x, u) \leq L(y, u), \ \forall y \in \mathbb{R}^N. \tag{3.6.57}$$

   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 :

   $$L(\overline{x}, v) \leq L(\overline{x}, \overline{u}) \leq L(y, \overline{u}), \ \forall (y, v) \in \mathbb{R}^N \times \mathcal{C}^+. \tag{3.6.58}$$

   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 :

   $$u \in \mathcal{C}^+, \quad M(u) \geq M(v), \ \forall v \in \mathcal{C}^+. \tag{3.6.59}$$

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 (3.6.56) et x_u est l'unique solution de (3.6.57)) et que u est solution de (3.6.59). [On pourra commencer par montrer, en utilisant la première inégalité, que $x \in D$ et $u \cdot (Cx - d) = 0$.]

   Montrer que $\nabla f(\overline{x}) + C^t u = 0$ 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 $w = P_{\mathcal{C}^+} v \Longleftrightarrow ((v-w)\cdot(w-z) \geq 0, \forall z \in \mathcal{C}^+)$.]

6. Déduire des questions 2, 4 et 5 que le problème (3.6.59) admet au moins une solution.

7. Montrer que l'algorithme du gradient à pas fixe avec projection pour trouver la solution de (3.6.59) 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 (3.6.58)) 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. $\nabla f(\overline{x}) + C^t\overline{u} = 0$.

3.7 Suggestions

Exercice 60 page 149 (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 $E = C([0,1], \mathbb{R})\ldots$

Exercice 61 page 149 (Minimisation d'une fonctionnelle quadratique)

Appliquer les théorèmes du cours à $-f$.

Exercice 62 page 149 (Minimisation d'une fonctionnelle quadratique)

1. Calculer la différentielle de f en formant la différence $f(x+h)-f(x)$ et en utilisant la définition. Calculer la hessienne en formant la différence $\nabla f(x+h)-\nabla f(x)$.
2. Utiliser le cours...

Exercice 63 page 149 (Algorithme du gradient à pas fixe)

1. Introduire la fonction φ définie (comme d'habitude...) par $\varphi(t) = f(tx+(1-t)y)$, intégrer entre 0 et 1 et utiliser l'hypothèse (3.3.15) sur $\nabla f(x+t(y-x))-\nabla f(x)$.

2. Utiliser le cours pour la stricte convexité et l'existence et l'unicité de $\bar{x}$, et la question 1 pour montrer que $f(x)\to+\infty$ lorsque $|x|\to+\infty$.

3. Montrer grâce aux hypothèses (3.3.15) et (3.3.16) que $|x_{n+1}-\bar{x}|^2 < |x_n-\bar{x}|^2(1-2\alpha\rho+M^2\rho^2)$.

Exercice 65 page 150 (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 $\varphi(\rho) = f(x_n+\rho w_n)$.

4. a. Montrer que f est minorée et remarquer que la suite $(f(x_n))_{n\in\mathbb{N}}$ 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 ρ_n est optimal et le résultat de 4.c.
4.e. Etudier le polynôme du 2nd degré en ρ défini par : $P_n(\rho) = f(x_n)-\rho|w_n|^2+\frac{1}{2}M|w_n|^2\rho^2$ dans les cas où $|w_n|\le M$ (fait la quesiton 4.c) puis dans le cas $|w_n|\ge M$.

5. utiliser l'inégalité prouvée en 4.e. pour montrer que $|w_n|\to 0$ 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 67 page 151 (Cas où f n'est pas croissante à l'infini)

S'inspirer des techniques utilisées aux exercices 63 et 65 (il faut impérativement les avoir fait avant...).

Exercice 71 page 154 (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 concurrence, utiliser le fait que si $w_n\cdot\nabla f(x_n)<0$ alors w_n est une direction de descente stricte de f en x_n, et que si ρ_n est optimal alors $\nabla f(x_n+\rho_n w_n)=0$.
3. Utiliser la fonction φ définie par $\varphi(\theta) = \nabla f(x_n+\theta\rho_n w_n)$.
4. C'est du calcul...
5. Montrer d'abord que $-g_n w_n \le -\gamma|w_n||g_n|$. Montrer ensuite (en utilisant la bonne vieille fonction φ définie par $\varphi(t) = f(x_n+t\rho_n)$, que $g_n\to 0$ lorsque $n\to+\infty$.

Exercice 77 page 157 (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 $\bar{x}$ est solution de (3.4.29), alors $y = (\bar{x}, \lambda)^t$ est solution du système (3.6.51), et montrer ensuite que le système (3.6.51) admet une unique solution.

3.8 Corrigés

Corrigé de l'exercice 62 page 149 (Minimisation d'une fonctionnelle quadratique)

1. Puisque $f(x) = \frac{1}{2}Ax \cdot x - b \cdot x$, $f \in C^\infty(\mathbb{R}^N, \mathbb{R})$. Calculons le gradient de f :

$$\begin{aligned} f(x+h) &= \frac{1}{2}A(x+h)\cdot(x+h) - b\cdot(x+h) \\ &= \frac{1}{2}Ax\cdot x + \frac{1}{2}Ax\cdot h + \frac{1}{2}Ah\cdot x + \frac{1}{2}Ah\cdot h - b\cdot x - b\cdot h \\ &= f(x) + \frac{1}{2}(Ax\cdot h + Ah\cdot x) - b\cdot h + \frac{1}{2}Ah\cdot h \\ &= f(x) + \frac{1}{2}(Ax + A^t x)\cdot h - b\cdot h + \frac{1}{2}Ah\cdot h. \end{aligned}$$

Et comme $\|Ah\cdot h\| \leq \|A\|_2 \|h\|^2$, on a :

$$\nabla f(x) = \frac{1}{2}(Ax + A^t x) - b. \tag{3.8.60}$$

Si A est symétrique $\nabla f(x) = Ax - b$. Calculons maintenant la hessienne de f. D'après (3.8.60), on a :

$$\nabla f(x+h) = \frac{1}{2}(A(x+h) + A^t(x+h)) - b = \nabla f(x) + \frac{1}{2}(Ah + A^t h)$$

et donc $H_f(x) = D(\nabla f(x)) = \frac{1}{2}(A + A^t)$. On en déduit que si A est symétrique, $H_f(x) = A$.
2. Si A est symétrique définie positive, alors f est strictement convexe. De plus, si A est symétrique définie positive, alors $f(x) \to +\infty$ quand $|x| \to +\infty$. En effet,

$$\begin{aligned} Ah\cdot h &\geq \alpha|h|^2 \text{ où } \alpha \text{ est la plus petite valeur propre de } A, \text{ et } \alpha > 0 \\ f(h) &\geq \frac{\alpha}{2}\|h\|^2 - \|b\|\|h\|; \text{ or } \|bh\| \leq \|b\|\ \|h\| \\ f(h) &\geq \|h\|\left(\frac{\alpha\|h\|}{2} - b\right) \longrightarrow \infty \text{ quand } h \to +\infty \end{aligned}$$

On en déduit l'existence et l'unicité de $\bar{x}$ qui minimise f. On a aussi :

$$\nabla f(\bar{x}) = 0 \Leftrightarrow f(\bar{x}) = \inf_{\mathbb{R}^N} f$$

Par la question 1. $\bar{x}$ est donc l'unique solution du système $A\bar{x} = b$.

Corrigé de l'exercice 63 page 149 (Convergence de l'algorithme du gradient à pas fixe)

1. Soit φ la fonction définie de $\mathbb{R}$ dans $\mathbb{R}^N$ par : $\varphi(t) = f(x + t(y-x))$. Alors $\varphi(1) - \varphi(0) = \int_0^1 \nabla f(x + t(y-x))\cdot(y-x)dt$, et donc :

$$f(y) - f(x) = \int_0^1 \nabla f(x + t(y-x))\cdot(y-x)dt.$$

On a donc :

$$f(y) - f(x) - \nabla f(x)\cdot(y-x) = \int_0^1 (\nabla f(x + t(y-x))\cdot(y-x) - \nabla f(x)\cdot(y-x))dt,$$

c'est–à–dire :

$$f(y) - f(x) - \nabla f(x)\cdot(y-x) = \int_0^1 \underbrace{(\nabla f(x+t(y-x)) - \nabla f(x))}_{\geq \alpha t(y-x)^2}\cdot(y-x)dt.$$

Grâce à la première hypothèse sur f, ceci entraîne :

$$f(y) - f(x) - \nabla f(x)\cdot(y-x) \geq \alpha\int_0^1 t|y-x|^2dt = \frac{\alpha}{2}|y-x|^2 > 0 \text{ si } y \neq x. \tag{3.8.61}$$

2. On déduit de la question 1 que f est strictement convexe. En effet, grâce à la question 1, pour tout $(x,y) \in E^2$, $f(y) > f(x) + \nabla f(x)\cdot(y-x)$; et d'après la première caractérisation de la convexité, voir proposition 3.11 p.47, on en déduit que f est strictement convexe.
   Montrons maintenant que $f(y) \to +\infty$ quand $|y| \to +\infty$.
   On écrit (3.8.61) pour $x = 0$: $f(y) \geq f(0) + \nabla f(0)\cdot y + \dfrac{\alpha}{2}|y|^2$.
   Comme $\nabla f(0)\cdot y \geq -|\nabla f(0)|(y)$, on a donc
   $f(y) \geq f(0) - |\nabla f(0)|\,|y| + \dfrac{\alpha}{2}|y|^2,$ et donc :
   $f(y) \geq f(0) + |y|\left(\dfrac{\alpha}{2}|y| - |\nabla f(0)|\right) \to +\infty$ quand $|y| + \infty$.
3. On pose $h(x) = x - \rho\nabla f(x)$. L'algorithme du gradient à pas fixe est un algorithme de point fixe pour h.

$$x_{n+1} = x_n - \rho\nabla f(x_n) = h(x_n).$$

Grâce au théorème 2.3 page 84, on sait que h est strictement contractante si $0 < \rho < \dfrac{2\alpha}{M^2}$.
Donc $x_n \to \bar{x}$ unique point fixe de h, c'est–à–dire $\bar{x} = h(\bar{x}) = \bar{x} - \rho\nabla f(\bar{x})$. Ceci entraîne

$$\nabla f(\bar{x}) = 0 \text{ donc } f(\bar{x}) = \inf_E f \text{ car } f \text{ est convexe .}$$

Corrigé de l'exercice 65 page 150 (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 $\|H\|$ atteint son $\max$ sur B_{R+1} qui est un fermé borné de $\mathbb{R}^N$. Soit $M = \max_{x\in B_{R+1}} \|H(x)\|$, on a $H(x)y\cdot y \leq My\cdot y \leq M|y|^2$.
3. Soit $w_n = -\nabla f(x_n)$.
   Si $w_n = 0$, on pose $x_{n+1} = x_n$.
   Si $w_n \neq 0$, montrons qu'il existe $\bar{\rho} > 0$ tel que

$$f(x_n + \bar{\rho}w_n) \leq f(x_n + \rho w_n) \quad \forall \rho > 0.$$

   On sait que $f(x) \to +\infty$ lorsque $|x| \to +\infty$.
   Soit $\varphi : \mathbb{R}_+ \to \mathbb{R}$ définie par $\varphi(\rho) = f(x_n + \rho w_n)$. On a $\varphi(0) = f(x_n)$ et $\varphi(\rho) = f(x_n + \rho w_n) \to +\infty$ lorsque $\rho \to +\infty$.
   En effet si $\rho \to +\infty$, on a $|x_n + \rho w_n| \to +\infty$. Donc φ étant continue, φ admet un minimum, atteint en $\bar{\rho}$, et donc $\exists \bar{\rho} \in \mathbb{R}_+$; $f(x_n + \bar{\rho}w) \leq f(x_n + \rho w_n) \quad \forall \rho > 0$.

4. a) Montrons que la suite $(f(x_n))_{n\in\mathbb{N}}$ est convergente. La suite $(f(x_n))_{n\in\mathbb{N}}$ vérifie

$$f(x_{n+1}) \leq f(x_n).$$

De plus $f(x) \to +\infty$ lorsque $|x| \to +\infty$ donc f est bornée inférieurement. On en conclut que la suite $(f(x_n))_{n\in\mathbb{N}}$ est convergente.

b) Montrons que $x_n \in B_R \ \forall n \in \mathbb{N}$. On sait que si $x \notin B_R$ alors $f(x) > f(x_0)$. Or la suite $(f(x_n))_{n\in\mathbb{R}}$ est décroissante donc $f(x_n) \leq f(x_0) \ \forall n$, donc $x_n \in B_R, \ \forall n \in \mathbb{N}$.

c) Montrons que $f(x_n + \rho w_n) \leq f(x_n) - \rho|w_n|^2 + \frac{\rho^2}{2}M|w_n|^2, \ \forall \rho \in [0, \frac{1}{|w_n|}]$. Soit φ définie de $\mathbb{R}_+$ dans $\mathbb{R}$ par $\varphi(\rho) = f(x_n + \rho w_n)$. 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 $\varphi'(\rho) = \nabla f(x_n + \rho w_n) \cdot w_n$ et $\varphi''(\rho) = H(x_n + \rho w_n)w_n \cdot w_n$. Donc

$$\varphi(\rho) = \underbrace{\varphi(0)}_{0} + \rho \underbrace{\nabla f(x_n)}_{-w_n} \cdot w_n + \frac{\rho^2}{2}H(x_n + \tilde{\rho} w_n)w_n \cdot w_n.$$

Si $\rho \in [0, \frac{1}{|w_n|}]$ on a

$$\begin{aligned} |x_n + \tilde{\rho} w_n| &\leq |x_n| + \frac{1}{|w_n|}|w_n| \\ &\leq R + 1, \end{aligned}$$

donc $x_n + \tilde{\rho} w_n \in B_{R+1}$ et par la question 2,

$$H(x_n + \tilde{\rho} w_n)w_n \cdot w_n \leq M|w_n|^2.$$

On a donc bien

$$\varphi(\rho) = f(x_n + \rho w_n) \leq f(x_n) - \rho|w_n|^2 + \frac{\rho^2}{2}M|w_n|^2.$$

d) Montrons que $f(x_{n+1}) \leq f(x_n) - \frac{|w_n|^2}{2M}$ si $|w_n| \leq M$.

Comme le choix de ρ_n est optimal, on a

$$f(x_{n+1}) = f(x_n + \rho_n w_n) \leq f(x_n + \rho w_n), \quad \forall \rho \in \mathbb{R}_+.$$

donc en particulier

$$f(x_{n+1}) \leq f(x_n + \rho w_n), \quad \forall \rho \in [0, \frac{1}{|w_n|}].$$

En utilisant la question précédente, on obtient

$$f(x_{n+1}) \leq f(x_n) - \rho|w_n|^2 + \frac{\rho^2}{2}M|w_n|^2 = \varphi(\rho), \quad \forall \rho \in [0, \frac{1}{|w_n|}]. \tag{3.8.62}$$

Or la fonction φ atteint son minimum pour

$$-|w_n|^2 + \rho M|w_n|^2 = 0$$

c'est–à–dire $\rho M = 1$ ou encore $\rho = \frac{1}{M}$ ce qui est possible si $\frac{1}{|w_n|} \geq \frac{1}{M}$ (puisque 3.8.62 est vraie si $\rho \leq \frac{1}{|w_n|}$).

Comme on a supposé $|w_n| \leq M$, on a donc

$$f(x_{n+1}) \leq f(x_n) - \frac{|w_n|^2}{M} + \frac{|w_n|^2}{2M} = f(x_n) - \frac{|w_n|^2}{2M}.$$

e) Montrons que $-f(x_{n+1}) + f(x_n) \geq \frac{|w_n|^2}{2\bar{M}}$ où $\bar{M} = \sup(M, \tilde{M})$ avec $\tilde{M} = \sup\{|\nabla f(x)|,\ x \in C_R\}$.

On sait par la question précédente que si

$$|w_n| \leq M, \text{ on a } -f(x_{n+1}) - f(x_n) \geq \frac{|w_n|^2}{2M}.$$

Montrons que si $|w_n| \geq M$, alors $-f(x_{n+1}) + f(x_n) \geq \frac{|w_n|^2}{2\tilde{M}}$. On aura alors le résultat souhaité.

On a

$$f(x_{n+1}) \leq f(x_n) - \rho|w_n|^2 + \frac{\rho^2}{2}M|w_n|^2,\ \ \forall \rho \in [0, \frac{1}{|w_n|}].$$

Donc

$$f(x_{n+1}) \leq \min_{[0, \frac{1}{|w_n|}]} \underbrace{[f(x_n) - \rho|w_n|^2 + \frac{\rho^2}{2}M|w_n|^2]}_{P_n(\rho)}$$

– 1er cas si $|w_n| \leq M$, on a calculé ce $\min$ à la question c).

– si $|w_n| \geq M$, la fonction $P_n(\rho)$ est décroissante sur $[0, \frac{1}{|w_n|}]$ et le minimum est donc atteint pour $\rho = \frac{1}{|w_n|}$.

$$\begin{aligned} \text{Or } P_n\left(\frac{1}{|w_n|}\right) &= f(x_n) - |w_n| + \frac{M}{2} \leq f(x_n) - \frac{|w_n|}{2} \\ &\leq f(x_n) - \frac{|w_n|^2}{2\tilde{M}}. \end{aligned}$$

5. Montrons que $\nabla f(x_n) \to 0$ lorsque $n \to +\infty$. On a montré que $\forall n$, $|w_n|^2 \leq 2\bar{M}(f(x_n) - f(x_{n+1}))$. Or la suite $(f(x_n))_{n\in\mathbb{N}}$ est convergente. Donc $|w_n| \to 0$ lorsque $n \to +\infty$ et $w_n = \nabla f(x_n)$ ce qui prouve le résultat.

   La suite $(x_n)_{n\in\mathbb{N}}$ est bornée donc $\exists (n_k)_{k\in\mathbb{N}}$ et $\tilde{x} \in \mathbb{R}^N$; $x_{n_k} \to x$ 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 $\exists\,!\ \bar{x} \in \mathbb{R}^N$ tel que $\nabla f(\bar{x}) = 0$. Montrons que $f(\bar{x}) \leq f(x)\ \forall x \in \mathbb{R}^N$ et que $x_n \to \bar{x}$ quand $n \to +\infty$. 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 $\bar{x}$, et donc $f(\bar{x}) \leq f(x)$ pour tout $x \in \mathbb{R}^N$.

   Montrons maintenant que la suite $(x_n)_{n\in\mathbb{N}}$ converge vers $\bar{x}$. En raison de l'hypotèse d'unicité, on a forcément $\tilde{x} = \bar{x}$, et on sait qu'on a convergence d'une sous-suite de $(x_n)_{n\in\mathbb{N}}$ vers $\bar{x}$ par la question 5. Il reste donc à montrer que c'est toute la suite qui converge. Supposons qu'elle ne converge pas ; alors

$$\exists \varepsilon > 0;\ \forall k \in \mathbb{N},\ \exists n_k \geq k \ \text{ et } |x_{n_k} - \bar{x}| > \varepsilon \tag{3.8.63}$$

   Mais d'après la question 5), on peut extraire de la suite $(x_{n_k})_{k\in\gamma}$ une sous–suite qui converge, ce qui contredit (3.8.63). Donc la suite $(x_n)_{n\in\mathbb{N}}$ converge.

Corrigé de l'exercice 68 page 151 (Méthode de relaxation)

1. On vu à l'exercice 63, questions 1 et 2, que si f vérifie l'hypothèse (3.6.43) 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 (3.6.43) 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 les résultats de l'exercice 63, questions 1 et 2 au cas $N = 1$, on en déduit l'existence et unicité de $\bar{s}$ tel que

$$\varphi_k^{(n+1)}(\overline{s}) = \inf_{s \in \mathbb{R}} \varphi_k^{(n+1)}(s).$$

Comme l'algorithme (3.6.45) procède à N minimisations de ce type à chaque itération, on en déduit que la suite $(x^{(n)})_{n\in\mathbb{N}}$ construite par cet algorithme est bien définie.

3.(a) Par définition, $x_k^{(n+1)}$ réalise le minimum de la fonction $\varphi_k^{(n+1)}$ sur $\mathbb{R}$. Comme de plus, $\varphi_k^{(n+1)} \in C^1(\mathbb{R},\mathbb{R})$, on a donc $(\varphi_k^{(n+1)})'(x_k^{(n+1)}) = 0$. Or $(\varphi_k^{(n+1)})'(x_k^{(n+1)}) = \partial_k f(x^{(n+1,k)})$, et donc $\partial_k f(x^{(n+1,k)}) = 0$.
D'après la question 2 de l'exercice 63, 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 $x^{(n+1,k-1)} - x^{(n+1,k)} = -x_k^{(n+1)} e_k$ 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 $(x^{(n)})_{n\in\mathbb{N}}$, on a :

$$f(x^{(n)}) - f(x^{(n+1)}) = \sum_{k=1}^{N} f(x^{(n+1,k-1)}) - f(x^{(n+1,k)}).$$

Par la question précédente, on a donc :

$$f(x^{(n)}) - f(x^{(n+1)}) \geq \frac{\alpha}{2} \sum_{k=1}^{N} |x^{(n+1,k-1))} - x^{(n+1,k)}|^2.$$

Or $x^{(n+1,k-1))} - x^{(n+1,k)} = -x_k^{(n+1)} e_k$, et $(e_k)_{k\in Ndim}$ est une base orthonormée. On peut donc écrire que

$$\begin{aligned} \sum_{k=1}^{N} |x^{(n+1,k-1))} - x^{(n+1,k)}|^2 &= \sum_{k=1}^{N} |(x_k^{(n)} - x_k^{(n+1)}) e_k|^2 \\ &= |\sum_{k=1}^{N} (x_k^{(n)} - x_k^{(n+1)}) e_k|^2 \\ &= |\sum_{k=1}^{N} (x^{(n+1,k-1))} - x^{(n+1,k)})|^2 \\ &= |x^{(n)} - x^{(n+1)}|^2. \end{aligned}$$

On en déduit que

$$f(x^{(n)}) - f(x^{(n+1)}) \geq \frac{\alpha}{2}|x^{(n)} - x^{(n+1)}|^2.$$

La suite $(f(x^{(n)}))_{n\in\mathbb{N}}$ est bornée inférieurement par $f(\bar{x})$; l'inégalité précédente montre qu'elle est décroissante, donc elle converge. On a donc $f(x^{(n)}) - f(x^{(n+1)} \to 0$ lorsque $n \to +\infty$, et donc par l'inégalité précédente,

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

De plus, pour $1 \leq k \leq N$,

$$\begin{aligned}
|x^{(n+1,k)} - x^{(n+1)}|^2 &= \sum_{\ell=k}^{N} |(x_\ell^{(n)} - x_\ell^{(n+1)})e_\ell|^2 \\
&= |\sum_{\ell=k}^{N} (x_\ell^{(n)} - x_\ell^{(n+1)})e_\ell|^2 \\
&= |\sum_{\ell=k}^{N} (x^{(n+1,\ell-1))} - x^{(n+1,\ell)})|^2 \\
&\leq |x^{(n)} - x^{(n+1)}|^2.
\end{aligned}$$

d'où l'on déduit que $\lim_{n\to+\infty} |x^{(n+1,k)} - x^{(n+1)}| = 0$.

4. En prenant $x = \bar{x}$ et $y = x^{(n+1)}$ dans l'hypothèse (3.6.43) et en remarquant que, puisque $\bar{x}$ réalise le minimum de f, on a $\nabla f(\bar{x}) = 0$, on obtient :

$$(-\nabla f(x^{(n+1)}) \cdot (\bar{x} - x^{(n+1)}) \geq \alpha|\bar{x} - x^{(n+1)}|^2,$$

et donc, par l'inégalité de Cauchy Schwarz :

$$|x^{(n+1)} - \overline{x}| \leq \frac{1}{\alpha}\left(\sum_{k=1}^{N} |\partial_k f(x^{(n+1)})|^2\right)^{\frac{1}{2}}.$$

5. Par les questions 1 et 2 de l'exercice 63, 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 $(f(x_n))_{n\in\mathbb{N}}$ étant décroissante, on a $f(x_n)) \leq f(x_0)$ pour tout n, et donc $|x_n| \leq R$ pour tout n. Par la question 3(b), on sait que pour tout $k \geq 1$, $\lim_{n\to+\infty} |x^{(n+1,k)} - x^{(n+1)}| = 0$, ce qui prouve que les suites $(x^{(n+1,k)})_{n\in\mathbb{N}}$, pour $k = 1, \ldots, N$, sont également bornées.

Comme $\lim_{n\to+\infty} |x^{(n+1,k)} - x^{(n+1)}| = 0$, on a pour tout $\eta > 0$, l'existence de $N_\eta \in \mathbb{N}$ tel que $|x^{(n+1,k)} - x^{(n+1)}| < \eta$ 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 $n \geq N_\eta$: $|\partial_k f(x^{(n+1,k)}) - \partial_k f(x^{(n+1)})| \leq \epsilon$, ce qui démontre que :

$$|\partial_k f(x^{(n+1)})| \to 0 \text{ lorsque } n \to +\infty.$$

On en conclut par le résultat de la question 4 que $x^{(n)} \to \bar{x}$ lorsque $n \to +\infty$.

6. On a vu a l'exercice 62 que dans ce cas, $\nabla f(x) = \frac{1}{2}(A + A^t)x - b$. L'algorithme 3.6.45 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 (3.6.45) appliqué à g est stationnaire. La suite ne converge donc pas vers $\overline{x}$. La fonction g n'est pas différentiable sur la droite $x_1 = x_2$.

Corrigé de l'exercice 71 page 154 (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 :

$$f(y) - f(x) = \varphi'(0) + \int_0^1 (1-t)\varphi''(t)dt. \tag{3.8.64}$$

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 3.8.64 que

$$f(y) \geq f(x) + \nabla f(x) \cdot (y-x) + \frac{\alpha}{2}|y-x|^2. \tag{3.8.65}$$

L'inégalité 3.8.65 entraîne la stricte convexité de f et sa croissance à l'infini (voir démonstration de la convergence du gradient à pas fixe, exercice 27).
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 $g^{(n+1)} \cdot w^{(n)} = 0$ et $g^{(n)} \cdot g^{(n)} = g^{(n)} \cdot w^{(n)}$ pout tout $n \in \mathbb{N}$.
Pour $n = 0$, 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 $g^{(n)} \cdot w^{(n-1)} = 0$ et $g^{(n-1)} \cdot g^{(n-1)} = g^{(n-1)} \cdot w^{(n-1)}$, et montrons que $g^{(n+1)} \cdot w^{(n)} = 0$ et $g^{(n)} \cdot g^{(n)} = 0$.
Par définition, on a :

$$w^{(n)} = g^{(n)} + \lambda_{n-1} w^{(n-1)}, \text{ donc}$$

$$w^{(n)} \cdot g^{(n)} = g^{(n)} \cdot g^{(n)} + \lambda_{n-1} w^{(n-1)} \cdot g^{(n)} = g^{(n)} \cdot g^{(n)}$$

par hypothèse de récurrence. On déduit de cette égalité que $w^{(n)} \cdot g^{(n)} > 0$ (car $g^{(n)} \neq 0$) et donc $w^{(n)}$ est une direction de descente stricte en $x^{(n)}$. On a donc $\nabla f(x^{(n+1)}) \cdot w^{(n)} = 0$, et finalement $g^{(n+1)} \cdot w^{(n)} = 0$.

3. Par définition, $g^{(n)} = -\nabla f(x^{(n)})$; or on veut calculer $g^{(n+1)} - g^{(n)} = -\nabla f(x^{(n+1)}) + \nabla f(x^{(n)})$. Soit φ la fonction de $\mathbb{R}$ dans $\mathbb{R}$ définie par :

$$\varphi(t) = -\nabla f(x^{(n)} + t(x^{(n+1)} - x^{(n)})).$$

On a donc :

$$\begin{aligned} \varphi(1) - \varphi(0) &= g^{(n+1)} - g^{(n)} \\ &= \int_0^1 \varphi'(t)dt. \end{aligned}$$

Calculons φ' : $\varphi'(t) = H(x^{(n)} + t(x^{(n+1)} - x^{(n)}))(x^{(n+1)} - x^{(n)})$. Et comme $x^{(n+1)} = x^{(n)} + \rho_n w^{(n)}$, on a donc :

$$g^{(n+1)} - g^{(n)} = \rho_n J_n w^{(n)}. \tag{3.8.66}$$

De plus, comme $g^{(n+1)} \cdot w^{(n)} = 0$ (question 1), on obtient par (3.8.66) que

$$\rho_n = \frac{g^{(n)} \cdot w^{(n)}}{J_n w^{(n)} \cdot w^{(n)}}$$

(car $J_n w^{(n)} \cdot w^{(n)} \neq 0$, puisque J_n est symétrique définie positive).

4. Par définition, on a $w^{(n)} = g^{(n)} + \lambda_{n-1} w^{(n-1)}$, et donc

$$|w^{(n)}| \leq |g^{(n)}| + |\lambda_{n-1}||w^{(n-1)}|. \tag{3.8.67}$$

Toujours par définition, on a :

$$\lambda_{n-1} = \frac{g^{(n)} \cdot (g^{(n)} - g^{(n-1)})}{g^{(n-1)} \cdot g^{(n-1)}}.$$

Donc, par la question 3, on a :

$$\lambda_{n-1} = \frac{\rho_n g^{(n)} \cdot J^{(n-1)} w^{(n-1)}}{g^{(n-1)} \cdot g^{(n-1)}}.$$

En utilisant la question 2 et à nouveau la question 3, on a donc :

$$\lambda_{n-1} = -\frac{J^{(n-1)} w^{(n-1)} \cdot g^{(n)}}{J^{(n-1)} w^{(n-1)} \cdot w^{(n-1)}},$$

et donc

$$\lambda_{n-1} = \frac{|J^{(n-1)} w^{(n-1)} \cdot g^{(n)}|}{J^{(n-1)} w^{(n-1)} \cdot w^{(n-1)}},$$

car $J^{(n-1)}$ est symétrique définie positive.
De plus, en utilisant les hypothèses sur H, on vérifie facilement que

$$\alpha|x|^2 \leq J^{(n)} x \cdot x \leq \beta|x|^2 \ \ \forall x \in \mathbb{R}^N.$$

On en déduit que

$$\lambda_{n-1} \leq \frac{|J^{(n-1)} w^{(n-1)} \cdot g^{(n)}|}{\alpha|w^{(n-1)}|^2}.$$

On utilise alors l'inégalité de Cauchy–Schwarz :

$$\begin{aligned} |J^{(n-1)} w^{(n-1)} \cdot g^{(n)}| &\leq \|J^{(n-1)}\|_2 \ |w^{(n-1)}| \ |g^{(n-1)}| \\ &\leq \beta|w^{(n-1)}| \ |g^{(n-1)}|. \end{aligned}$$

On obtient donc que

$$\lambda_{n-1} \leq \frac{\beta}{\alpha}\frac{|g^{(n-1)}|}{|w^{(n-1)|}},$$

ce qui donne bien grâce à (3.8.67) :

$$|w^{(n)}| \leq |g^{(n)}|(1+\frac{\beta}{\alpha}).$$

5. • Montrons d'abord que la suite $(f(x^{(n)}))_{n\in\mathbb{N}}$ converge. Comme $f(x^{(n+1)}) = f(x^{(n)} + \rho_n w^{(n)}) \leq f(x^{(n)} + \rho w^{(n)})\ \forall \rho \geq 0$, on a donc en particulier $f(x^{(n+1)}) \leq f(x^{(n)})$. La suite $(f(x^{(n)}))_{n\in\mathbb{N}}$ est donc décroissante. De plus, elle est minorée par $f(\bar{x})$. Donc elle converge, vers une certaine limite $\ell \in \mathbb{R}$, lorsque n tend vers $+\infty$.

• La suite $(x^{(n)})_{n\in\mathbb{N}}$ est bornée : en effet, comme f est croissante à l'infini, il existe $R > 0$ tel que si $|x| > R$ alors $f(x) \geq f(x^{(0)})$. Or $f(x^{(n)}) \geq f(x^{(0)})$ pout tout $n \in \mathbb{N}$, et donc la suite $(x^{(n)})_{n\in\mathbb{N}}$ est incluse dans la boule de rayon R.

• Montrons que $\nabla f(x^{(n)}) \to 0$ lorsque $n \to +\infty$.
On a, par définition de $x^{(n+1)}$,

$$f(x^{(n+1)}) \leq f(x^{(n)} + \rho w^{(n)}),\quad \forall \rho \geq 0.$$

En introduisant la fonction φ définie de $\mathbb{R}$ dans $\mathbb{R}$ par $\varphi(t) = f(x^{(n)} + t\rho w^{(n)})$, on montre facilement (les calculs sont les mêmes que ceux de la question 1) que

$$f(x^{(n)} + \rho w^{(n)}) = f(x^{(n)}) + \rho\nabla f(x^{(n)})\cdot w^{(n)} + \rho^2 \int_0^1 H(x^{(n)} + t\rho w^{(n)})w^{(n)}\cdot w^{(n)}(1-t)dt,$$

pour tout $\rho \geq 0$. Grâce à l'hypothèse sur H, on en déduit que

$$f(x^{(n+1)}) \leq f(x^{(n)}) + \rho\nabla f(x^{(n)})\cdot w^{(n)} + \frac{\beta}{2}\rho^2|w^{(n)}|^2,\ \forall \rho \geq 0.$$

Comme $\nabla f(x^{(n)})\cdot w^{(n)} = -g^{(n)}\cdot w^{(n)} = -|g^{(n)}|^2$ (question 2) et comme $|w^{(n)}| \leq |g^{(n)}|(1+\frac{\beta}{\alpha})$ (question 4), on en déduit que :

$$f(x^{(n+1)}) \leq f(x^{(n)}) - \rho|g^{(n)}|^2 + \rho^2\gamma|g^{(n)}|^2 = \psi_n(\rho),\ \forall \rho \geq 0,$$

où $\gamma = \frac{\beta^2}{2} + (1+\frac{\beta}{\alpha})^2$. La fonction ψ_n est un polynôme de degré 2 en ρ, qui atteint son minimum lorsque $\psi_n'(\rho) = 0$, *i.e.* pour $\rho = \dfrac{1}{2\gamma}$. On a donc, pour $\rho = \dfrac{1}{2\gamma}$,

$$f(x^{(n+1)}) \leq f(x^{(n)}) - \frac{1}{4\gamma}|g^{(n)}|^2,$$

d'où on déduit que

$$|g^{(n)}|^2 \leq 4\gamma(f(x^{(n)}) - f(x^{(n+1)}) \underset{n\to+\infty}{\to} 0$$

On a donc $\nabla f(x^{(n)}) \to 0$ lorsque $n \to +\infty$.

• La suite $(x^{(n)})_{n\in\mathbb{N}}$ étant bornée, il existe une sous–suite qui converge vers $x \in \mathbb{R}^N$, comme $\nabla f(x^{(n)}) \to 0$ et comme
$nablaf$ est continue, on a $\nabla f(x) = 0$. Par unicité du minimum (f est croissante à l'infini et strictement convexe) on a donc $x = \bar{x}$.
Enfin on conclut à la convergence de toute la suite par un argument classique (voir question 6 de l'exercice 65 page 150).

Corrigé de l'exercice 72 page 154 (Algorithme de quasi Newton)

Partie 1

1. Par définition de $w^{(n)}$, on a :

$$w^{(n)} \cdot \nabla f(x^{(n)}) = -K^{(n)} \nabla f(x^{(n)}) \cdot \nabla f(x^{(n)}) < 0$$

car K est symétrique définie positive.
Comme ρ_n est le paramètre optimal dans la direction $w^{(n)}$, on a $\nabla f(x^{(n)} + \rho_n w^{(n)}) \cdot w^{(n)} = 0$, et donc $Ax^{(n)} \cdot w^{(n)} + \rho_n A w^{(n)} \cdot w^{(n)} = b \cdot w^{(n)}$; on en déduit que

$$\rho_n = -\frac{g^{(n)} \cdot w^{(n)}}{Aw^{(n)} \cdot w^{(n)}}.$$

Comme $w^{(n)} = -K^{(n)} g^{(n)}$, ceci s'écrit encore :

$$\rho_n = \frac{g^{(n)} \cdot K^{(n)} g^{(n)}}{AK^{(n)} g^{(n)} \cdot K^{(n)} g^{(n)}}.$$

2. Si $K^{(n)} = A^{-1}$, la formule précédente donne immédiatement $\rho_n = 1$.
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 $Ax^{(n)} = b$, 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 $g^{(n)} \neq 0$. Par définition, on a $s^{(n)} = x^{(n+1)} - x^{(n)} = -\rho_n K^{(n)} g^{(n)}$, avec $\rho_n > 0$. Comme $K^{(n)}$ est symétrique définie positive elle est donc inversible ; donc comme $g^{(n)} \neq 0$, on a $K^{(n)} g^{(n)} \neq 0$ et donc $s^{(n)} \neq 0$.
Soit $i < n$, par définition de $s^{(n)}$, on a :

$$s^{(n)} \cdot As^{(i)} = -\rho_n K^{(n)} g^{(n)} \cdot As^{(i)}.$$

Comme $K^{(n)}$ est symétrique,

$$s^{(n)} \cdot As^{(i)} = -\rho_n g^{(n)} \cdot K^{(n)} As^{(i)}.$$

Par hypothèse, on a $K^{(n)} As^{(i)} = s^{(i)}$ pour $i < n$, donc on a bien que si $i < n$

$$s^{(n)} \cdot As^{(i)} = 0 \Leftrightarrow g^{(n)} \cdot s^{(i)} = 0.$$

Montrons maintenant que $g^{(n)} \cdot s^{(i)} = 0$ 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

$$\begin{aligned}(g^{(n)} - g^{(i+1)}) \cdot s^{(i)} &= (Ax^{(n)} - Ax^{(i+1)}) \cdot s^{(i)} \\ &= \sum_{k=i+1}^{n-1} (Ax^{(k+1)} - Ax^{(k)}) \cdot s^{(i)} \\ &= \sum_{k=i+1}^{n-1} As^{(k)} \cdot s^{(i)}, \\ &= 0\end{aligned}$$

Par hypothèse de A–conjugaison de la famille $(s^{(i)})_{i=1,k-1}$ on déduit alors facilement des deux égalités précédentes que $g^{(n)} \cdot s^{(i)} = 0$. Comme on a montré que $g^{(n)} \cdot s^{(i)} = 0$ si et seulement si $s^{(n)} \cdot As^{(i)} = 0$, on en conclut que la famille $(s^{(i)})_{i=1,\ldots,n}$ est $A-$conjuguée, et que les vecteurs $s^{(i)}$ sont non nuls.

2. Montrons que $K^{(n+1)}$ est symétrique. On a :

$$(K^{(n+1)})^t = (K^{(n)})^t + \frac{(s^{(n)}(s^{(n)})^t)^t}{s^{(n)} \cdot y^{(n)}} - \frac{[(K^{(n)}y^{(n)})(K^{(n)}y^{(n)})^t]^t}{K^{(n)}y^{(n)} \cdot y^{(n)}} = K^{(n+1)},$$

car $K^{(n)}$ est symétrique.

3. Montrons que $K^{(n+1)}As^{(i)} = s^{(i)}$ si $0 \leq i \leq n$. On a :

$$K^{(n+1)}As^{(i)} = K^{(n)}As^{(i)} + \frac{s^{(n)}(s^{(n)})^t}{s^{(n)} \cdot y^{(n)}}As^{(i)} - \frac{(K^{(n)}y^{(n)})(K^{(n)}(y^{(n)})^t}{K^{(n)}y^{(n)} \cdot y^{(n)}}As^{(i)}. \qquad (3.8.68)$$

– Considérons d'abord le cas $i < n$. On a

$$s^{(n)}(s^{(n)})^t As^{(i)} = s^{(n)}[(s^{(n)})^t As^{(i)}] = s^{(n)}[s^{(n)} \cdot As^{(i)}] = 0$$

car $s^{(n)} \cdot As^{(i)} = 0$ si $i < n$. De plus, comme $K^{(n)}$ est symétrique, on a :

$$(K^{(n)}y^{(n)})(K^{(n)}y^{(n)})^t As^{(i)} = K^{(n)}y^{(n)}(y^{(n)})^t K^{(n)}As^{(i)}.$$

Or par la question (c), on a $K^{(n)}As^{(i)} = s^{(i)}$ si $0 \leq i \leq n$. De plus, par définition, $y^{(n)} = As^{(n)}$. On en déduit que

$$(K^{(n)}y^{(n)})(K^{(n)}y^{(n)})^t As^{(i)} = K^{(n)}y^{(n)}(As^{(n)})^t s^{(i)} = K^{(n)}y^{(n)}(s^{(n)})^t As^{(i)} = 0$$

puisque on a montré en (a) que les vecteurs $s^{(0)}, \ldots, s^{(n)}$ sont A-conjugués. On déduit alors de (3.8.68) que

$$K^{(n+1)}As^{(i)} = K^{(n)}As^{(i)} = s^{(i)}.$$

– Considérons maintenant le cas $i = n$. On a

$$K^{(n+1)}As^{(n)} = K^{(n)}As^{(n)} + \frac{s^{(n)}(s^{(n)})^t}{s^{(n)} \cdot y^{(n)}}As^{(n)} - \frac{(K^{(n)}y^{(n)})(K^{(n)}(y^{(n)})^t}{K^{(n)}y^{(n)} \cdot y^{(n)}}As^{(n)},$$

et comme $y^{(n)} = As^{(n)}$,, ceci entraîne que

$$K^{(n+1)}As^{(n)} = K^{(n)}As^{(n)} + s^{(n)} - K^{(n)}y^{(n)} = s^{(n)}.$$

4. Pour $x \in \mathbb{R}^N$, calculons $K^{(n+1)}x \cdot x$:

$$K^{(n+1)}x \cdot x = K^{(n)}x \cdot x + \frac{s^{(n)}(s^{(n)})^t}{s^{(n)} \cdot y^{(n)}}x \cdot x - \frac{(K^{(n)}y^{(n)})(K^{(n)}y^{(n)})^t}{K^{(n)}y^{(n)} \cdot y^{(n)}}x \cdot x.$$

Or $s^{(n)}(s^{(n)})^t x \cdot x = s^{(n)}(s^{(n)} \cdot x) \cdot x = (s^{(n)} \cdot x)^2$, et de même, $(K^{(n)}y^{(n)})(K^{(n)}y^{(n)})^t x \cdot x = (K^{(n)}y^{(n)} \cdot x)^2$. On en déduit que

$$K^{(n+1)}x \cdot x = K^{(n)}x \cdot x + \frac{(s^{(n)} \cdot x)^2}{s^{(n)} \cdot y^{(n)}} - \frac{(K^{(n)}y^{(n)} \cdot x)^2}{K^{(n)}y^{(n)} \cdot y^{(n)}}.$$

En remarquant que $y^{(n)} = As^{(n)}$, et en réduisant au même dénominateur, on obtient alors que

$$K^{(n+1)}x \cdot x = \frac{(K^{(n)}x \cdot x)(K^{(n)}y^{(n)} \cdot y^{(n)}) - (K^{(n)}y^{(n)} \cdot x)^2}{(K^{(n)}y^{(n)} \cdot y^{(n)})} + \frac{(s^{(n)} \cdot x)^2}{As^{(n)} \cdot s^{(n)}}.$$

Montrons maintenant que $K^{(n+1)}$ est symétrique définie positive. Comme $K^{(n)}$ est symétrique définie positive, on a grâce à l'inégalité de Cauchy-Schwarz que $(K^{(n)}y^{(n)} \cdot x)^2 \leq (K^{(n)}x \cdot x)(K^{(n)}y^{(n)})$ avec égalité si et seulement si x et $y^{(n)}$ sont colinéaires. Si x n'est pas colinéaire à $y^{(n)}$, on a donc donc clairement

$$K^{(n+1)}x \cdot x > 0.$$

Si maintenant x est colinéaire à $y^{(n)}$, *i.e.* $x = \alpha y^{(n)}$ avec $\alpha \in \mathbb{R}_+^*$, on a, grâce au fait que $y^{(n)} = As^{(n)}$,

$$\frac{(s^{(n)} \cdot x)^2}{As^{(n)} \cdot s^{(n)}} = \alpha^2 \frac{(s^{(n)} \cdot As^{(n)})^2}{As^{(n)} \cdot s^{(n)}} > 0, \text{ et donc } K^{(n+1)}x \cdot x > 0.$$

On en déduit que $K^{(n+1)}$ est symétrique définie positive.

5. On suppose que $g^{(n)} \neq 0$ si $0 \leq n \leq N-1$. On prend comme hypothèse de récurrence que les vecteurs $s^{(0)}, \ldots, s^{(n-1)}$ sont A-conjugués et non-nuls, que $K^{(j)}As^{(i)} = s^{(i)}$ 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 $K^{(N)}As^{(i)} = s^{(i)}$ pour $i = 0$ à $N-1$. Or les vecteurs $s^{(0)}, \ldots, s^{(n-1)}$ sont A-conjugués et non-nuls : ils forment donc une base. On en déduit que $K^{(N)}A = Id$, ce qui prouve que $K^{(N)} = A^{-1}$, et donc, par définition de $x^{(N+1)}$, que $x^{(N+1)} = A^{-1}b = \overline{x}$.

Exercice 75 page 157 (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 3.34 page 139. 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 (3.4.29) est donc $\bar{x} = 0$.

(ii) L'ensemble $K = \{|x| = 1\}$ est fermé borné mais non convexe. Le théorème d'existence 3.32 page 138 s'applique donc, mais pas le théorème d'unicité 3.33 page 138. De fait, on peut remarquer que $K = \{-1, 1\}$, et donc $\{f(x), x \in K\} = \{1\}$. Il existe donc deux solutions du problème (3.4.29) : $\bar{x}_1 = 1$ 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 3.34 page 139 : il existe un unique $\bar{x}_1 \in \mathbb{R}$ et un unique $\bar{x}_2 \in \mathbb{R}$ solution de (3.4.29) 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 3.32 page 138 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 76 page 157 (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 :

$$\left\{ \begin{array}{l} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in K \\ f(\bar{x}_1, \bar{x}_2) \leq f(x_1, x_2) \quad \forall (x_1, x_2) \in K \end{array} \right. \tag{3.8.69}$$

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 (3.8.69) ou (3.6.50). L'ensemble $\tilde{K}$ est un convexe ferme borné, la fonction f est continue, et donc par le théorème 3.32 page 138, il existe au moins une solution du problème (3.6.50) (ou (3.8.69)).

3. Calculons ∇g : $\nabla g(x) = (1, 1)^t$, donc rang $Dg(x, y) = 1$. Par le théorème de Lagrange, si $x = (x_1, x_2)^t$ est solution de (3.8.69), 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 :

$$-\bar{y} + \lambda = 0 \; - \bar{x} + \lambda = 0 \quad \bar{x} + \bar{y} = 1.$$

On a donc

$$\bar{x} = \bar{y} = \frac{1}{2}.$$

Exercice 77 page 157 (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 3.32 et 3.33 il existe un unique $\bar{x}$ solution de (3.4.29).

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' :

$$\left\{ \begin{array}{l} A\bar{x} - b + \lambda d = 0, \\ d \cdot \bar{x} = c \end{array} \right. , \tag{3.8.70}$$

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 (3.8.70) et et $\bar{x}$ est solution de (3.4.29).

Exercice 81 page 158 (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 3.34 page 139, il y a donc existence et unicité de la solution du problème (3.4.29). 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 3.6 page 159 (Exemple d'opérateur de projection)

2. Soit p_K l'opérateur de projection définie à la proposition 3.44 page 143, 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 $y_i > \beta_i$,

$$(p_K(y))_i = \max(\alpha_i, \min(y_i, \beta_i)) \text{ pour tout } i = 1, \ldots, N.$$

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) :

$$\left\{ \begin{array}{l} x'(t) = f(x(t), t) \;\; t > 0, \\ x(0) = \bar{x}_0. \end{array} \right. \tag{4.1.1}$$

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 (4.1.1). Prenons par exemple l'équation du second ordre décrivant le comportement de l'amortisseur d'une voiture :

$$\left\{ \begin{array}{l} my'' + cy' + ky = 0, \\ y(0) = \bar{x}_0, \\ y'(0) = 0. \end{array} \right. \tag{4.1.2}$$

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 $x = (x_1, x_2)^t$:

$$\left\{ \begin{array}{l} x'(t) = f(x(t), t), \\ x(0) = (\bar{x}_0, 0)^t, \end{array} \right. \quad \text{avec } f(x,t) = \left(\begin{array}{c} x_2, \\ -\dfrac{1}{m}(cx_2 + kx_1) \end{array} \right). \tag{4.1.3}$$

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 (4.1.1), c'est–à–dire que x est solution de (4.1.1) sur $[0, T_M[$, et que s'il existe $\alpha > 0$ et $y \in C^2([0, \alpha[, \mathbb{R}^N)$ solution de (4.1.1) 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) *En fait, pour avoir existence et unicité d'une solution maximale de* (4.1.1)*, on peut affaiblir l'hypothèse $f \in C^1(\mathbb{R}^N \times \mathbb{R}, \mathbb{R}^N)$ en $f \in C(\mathbb{R}^N \times \mathbb{R}, \mathbb{R}^N)$ qui soit "lipschitzienne sur les bornés", c'est–à–dire qui vérifie :*

$$\begin{array}{l} \forall A > 0, \exists M_A \in \mathbb{R}_+ \text{ tel que } \forall t \in [0, T[, \forall (x,y) \in B_A \times B_A, \\ |f(x,t) - f(y,t)| \leq M_A |x - y|. \end{array} \tag{4.1.4}$$

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 (4.1.4)*, alors qu'elle n'est évidemment pas forcément globalement lipschitzienne (prendre $f(x) = x^2$ pour s'en convaincre). De même la propriété* (4.1.4) *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 :*

$$\begin{cases} \dfrac{dx}{dt}(t) = x^2(t) \\ x(0) = 1 \end{cases}$$

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 $x(t) = \frac{1}{1-t}$ $t \in [0, 1[$.

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* (4.1.1) *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 (4.1.1). L'objet de ce chapitre est de présenter des méthodes pour obtenir des solutions (numériques) approchées de la solution de (4.1.1). Plus précisément, on adopte les notations et hypothèses suivantes :

Notations et hypothèses :

$$\left\{ \begin{array}{l} \text{Soit } f \text{ vérifiant l'hypothèse (4.1.4))} \\ \text{ et soit } x \text{ solution maximale de (4.1.1) (définie sur } [0, T_M[), \\ \text{ on se donne } T \in]0, T_M[, \text{ on cherche à calculer } x \text{ sur } [0,T], \\ \text{ où } x \in C^1([0,T], \mathbb{R}^N) \text{ est solution de (4.1.1).} \\ \text{On se donne une discrétisation de } [0,T], \text{ i.e. } n \in \mathbb{N} \text{ et} \\ (t_0, t_1, \ldots, t_n) \in \mathbb{R}^{n+1} \text{ tels que } 0 < t_0 < t_1 < \ldots < t_n = T. \\ \text{ On pose } h_k = t_{k+1} - t_k, \ \forall k = 0, \ldots, n-1, \\ \text{ et } h = \max\{h_0, \ldots, h_{n-1}\}. \text{ Pour } k = 1, \ldots n, \text{ on cherche } x_k \\ \text{ valeur approchée de } x(t_k) = \bar{x}_k, \\ \text{ et on appelle } e_k = \bar{x}_k - x_k \text{ l'erreur de discrétisation.} \end{array} \right. \quad (4.1.5)$$

[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 $t \in [0,T]$.

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* (4.1.5)*, on appelle schéma à un pas pour la résolution numérique de* (4.1.1)*, un algorithme de construction des valeurs* $(x_k)_{k=1,n}$ *qui s'écrit sous la forme suivante :*

$$\begin{cases} x_0 \text{ donné (approximation de } \bar{x}_0) \\ \dfrac{x_{k+1} - x_k}{h_k} = \phi(x_k, t_k, h_k), \quad k = 0, \ldots n-1, \end{cases} \tag{4.1.6}$$

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 (4.1.6), 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 (4.1.6) avec la fonction ϕ très simple suivante :
$$\phi(x_k, t_k, h_k) = f(x_k, t_k). \tag{4.1.7}$$

2. Schéma Euler implicite

$$\begin{cases} x_0 \text{ donné} \\ \dfrac{x_{k+1} - x_k}{h_k} = f(x_{k+1}, t_{k+1}). \quad k = 0, \ldots n-1, \end{cases} \tag{4.1.8}$$

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 (4.1.6) (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 :

$$D_1 f(y,t)z \cdot z \leq 0 \quad \forall y \in \mathbb{R}^N, \ \forall z \in \mathbb{R}^N, \quad \forall t \geq 0, \tag{4.1.9}$$

alors x_{k+1} calculé par (4.1.8) est bien défini en fonction de x_k, t_k, et h_k. On peut donc bien écrire le schéma (4.1.8) sous la forme (4.1.6) 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 (4.1.9), ce schéma entre donc bien dans le cadre des schémas (4.1.6) 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* (4.1.5) *et on étudie le schéma* (4.1.6)*.*

1. *Pour $k = 0, \dots, n$, on définit l'erreur de consistance du schéma* (4.1.6) *en t_k par :*

$$R_k = \frac{\overline{x}_{k+1} - \overline{x}_k}{h_k} - \phi(\overline{x}_k, t_k, h_k). \tag{4.2.10}$$

2. *Le schéma est consistant si*

$$\max\{|R_k|, k = 0 \dots n-1\} \to 0 \quad \textit{lorsque } h \to 0. \tag{4.2.11}$$

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 $|R_k| \leq Ch^p$, $\forall k = 1, \dots, n-1$.*

Donnons maintenant une condition nécessaire sur ϕ pour que le schéma (4.1.6) soit consistant.

Proposition 4.6 (Caractérisation de la consistance) *Sous les hypothèses et notations* (4.1.5)*, 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* (4.1.6) *est consistant.*

Démonstration Comme $x \in C^1([0,T], \mathbb{R}^N)$ est la solution exacte de (4.1.1), 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 (4.1.1) est dans $C^p(\mathbb{R}_+, \mathbb{R}^N)$.

Définition 4.7 (Stabilité) *Sous les hypothèses* (4.1.5)*, on dit que le schéma* (4.1.6) *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* (4.1.5)*.*

1. *Le schéma* (4.1.6) *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* (4.1.5)*, on dit que le schéma* (4.1.6) *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*

$$\begin{array}{l} x_{k+1} = x_k + h_k\phi(t_k, x_k, h_k), \\ y_{k+1} = y_k + h_k\phi(t_k, y_k, h_k) + \varepsilon_k, \end{array} \quad \text{pour } k = 0, \dots, n-1, \tag{4.2.12}$$

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 90 page 190.

Théorème 4.10 (Convergence) *Sous les hypothèses et notations* (4.1.5)*, on suppose que le schéma* (4.1.6) *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 4.2.10. 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* (4.1.5)*, une condition suffisante pour que le schéma* (4.1.6) *soit stable par rapport aux erreurs est que*

$$\begin{array}{l} \exists h^* > 0, \exists M > 0; \forall (x, y) \in \mathbb{R}^N \times \mathbb{R}^N, \forall h < h^*, \ \forall t \in [0, T], \\ |\phi(x, t, h) - \phi(y, t, h)| \leq M|x - y|. \end{array} \tag{4.2.13}$$

La démonstration de cette proposition est laissée en exercice (exercice 90 page 190).

4.3 Théorème général de convergence

Théorème 4.12 *On se place sous les hypothèses et notations* (4.1.5)*.*

1. *On suppose que le schéma* (4.1.6) *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*

$$\begin{array}{l} \forall (y,z) \in B_A \times B_A,\ \ \forall t \in [0,T],\ \ \forall h \in [0,h^*], \\ |\phi(y,t,h) - \phi(z,t,h)| \leq M_A |y-z|, \end{array} \tag{4.3.14}$$

où B_A désigne la boule de rayon A. (Noter que cette hypothèse sur ϕ est semblable à l'hypothèse (4.1.4) *"Lipschitz sur les bornés" faite sur f dans la remarque 4.1 page 178).*

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 $k = 0, \ldots n$, 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 $k = 0, \ldots, n$. (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 (4.1.1), 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 (4.3.14),
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 :

$$\left\{ \begin{array}{l} |e_k| \leq \alpha_k h^p + \beta_k |e_0|, \\ x_k \in B_{2A}, \end{array} \right. , \tag{4.3.15}$$

$$\text{avec } \alpha_k = Ce^{t_k M_{2A}}\,(1+h_0)\ldots(1+h_{k-1}) \text{ et } \beta_k = e^{t_k M_{2A}}. \tag{4.3.16}$$

Si on suppose (4.3.15) vraie, on peut terminer la démonstration du théorème : en effet pour $x \geq 0$, on a $1+x \leq e^x$, et donc :

$$(1+h_0)(1+h_1)\ldots(1+h_{k-1}) \leq e^{h_0+h_1+\cdots h_{k-1}} = e^{t_k} \leq e^T.$$

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 (4.3.15) et (4.3.16) que

$$\begin{array}{rl} |e_k| & \leq Ce^{T(M_{2A}+1)}h^p + e^{TM_{2A}}|e_0| \\ & \leq K(h^p + |e_0|) \text{ avec } K = \max(Ce^{T(M_{2A}+1)}, e^{T(M_2 A)}), \end{array}$$

et que $x_k \in B_{2A}$. Il ne reste donc plus qu'à démontrer (4.3.15) par récurrence sur k.
– Pour $k = 0$, les formules (4.3.16) 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 $\varepsilon e^{TM_2 A} \leq \frac{A}{2}$ et donc $\varepsilon \leq \frac{A}{2}$. On en déduit que $x_0 \in B_{2A}$.

– Supposons maintenant que les relations (4.3.15) et (4.3.16) sont vraies jusqu'au rang k et démontrons qu'elles le sont encore au rang $k+1$.
Par définition du schéma (4.1.6) et de l'erreur de consistance (4.2.10), 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

$$|e_{k+1}| \leq |e_k| + h_k|\phi(\bar{x}_k, t_k, h_k) - \phi(x_k, t_k, h_k)| + h_k|R_k|. \tag{4.3.17}$$

Comme $x_k \in B_{2A}$ et $\bar{x}_k \in B_A$, en utilisant la propriété (4.3.14) 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 (4.1.6) est supposé consistant d'ordre p, on a $|R_k| \leq Ch^p$. On peut donc déduire de (4.3.17) que

$$|e_{k+1}| \leq |e_k|(1 + M_{2A}h_k) + h_k Ch^p,$$

et, en utilisant l'hypothèse de récurrence (4.3.15) :

$$|e_{k+1}| \leq (1 + h_k M_{2A})(\alpha_k h^p + \beta_k|e_0|) + h_k Ch^p.$$

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 (4.3.15) 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 $k = 1, \dots n$. 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 90 page 190) que si le schéma (4.1.6) *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 $k = 0, \dots, n$.*

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 (4.1.1) *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 (4.1.5) et on étudie le schéma (4.1.6). On donne quatre exemples de schémas de la forme (4.1.6) :

Exemple 1 Euler explicite On rappelle que le schéma s'écrit (voir (4.1.7)) :

$$\frac{x_{k+1} - x_k}{h_k} = f(x_k, t_k),$$

On a donc $\phi(x_k, t_k, h_k) = f(x_k, t_k)$.
On peut montrer (voir exercice 89 page 190) 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 :

$$\frac{x_{k+1} - x_k}{h_k} = f\left(x_k + \frac{h_k}{2} f(x_k, t_k), t_k + \frac{h_k}{2}\right) = \phi(x_k, t_k, h_k) \tag{4.4.18}$$

– 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

$$\frac{x_{k+1} - x_k}{h_k} = \frac{1}{2} f(x_k, t_k) + \frac{1}{2}[f(x_k + h_k f(x_k, t_k), t_{k+1})]. \tag{4.4.19}$$

– 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 (4.1.9) donnée plus haut est vérifiée, le schéma d'Euler implicite (4.1.8) 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 :

$$\begin{array}{l} x_{k+1} = x_k - hx_k = (1 - h)x_k \text{ et donc} \\ x_k = (1 - h)^k, \quad \forall k = 0, \ldots, n, \text{ avec } nh = T. \end{array} \tag{4.5.20}$$

(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

$$x_n = (1 - h)^{T/h} \to e^{-T} \text{ quand } h \to 0.$$

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 (4.5.20) 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 (4.1.8), 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 (4.1.1) 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

$$x_n = (1+h)^n \to e^T \text{ c.à.d. lorsque } n \to +\infty.$$

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 (4.1.8), 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 (4.1.6), si pour tout $k = 0 \dots n-1$, x_k étant donné, il existe x_{k+1} qui satisfait :

$$\frac{x_{k+1} - x_k}{h_k} = f(x_{k+1}, t_{k+1}), \;\; k = 0, \dots, \; n-1.$$

On va montrer dans le théorème suivant que ceci est le cas si la condition (4.1.9) qu'on rappelle ici est vérifiée :

$$D_1 f(y,t)z \cdot z \leq 0, \;\; \forall y, z \in \mathbb{R}^N, \;\; \forall t \in [0,T].$$

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* (4.1.5) *et* (4.1.9)*. Alors*

1. $(x_k)_{k=0\dots n}$ *est bien définie par* (4.1.8)*,*
2. $|e_k| \leq |e_0| + h \displaystyle\int_0^{t_k} |x''(s)| ds, \;\; \forall k = 0, \dots, \; n.$

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 (4.1.9), on déduit que :

$$(f(y,t) - f(z,t), (y-z)) \leq 0, \; \forall y, z \in \mathbb{R}^N, \; \forall t \in [0,T]. \tag{4.6.21}$$

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

$$F(h,y) = 0. \tag{4.6.22}$$

Pour $h = 0$, cette équation admet évidemment une unique solution $y = x_k$. Soit $I = \{\bar{h} \in \mathbb{R}_+^*$ t.q. (4.6.22) 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 (4.6.22).

Supposons que $\sup I = H < +\infty$. Montrons d'abord que H est atteint. Soit $(h_n)_{n \in \mathbb{N}} \subset I$ telle que $h_n \to H$ lorsque $n \to +\infty$, alors la suite $(y_n)_{n \in \mathbb{N}}$ définie par $y_n = x_k + h_n f(y_n, t_k + h_n)$ est bornée : en effet,

$$y_n = x_k + h_n(f(y_n, t_k + h_n) - f(0, t_k + h)) + h_n f(0, t_k + h),$$

en prenant le produit scalaire des deux membres de cette égalité avec y_n et en utilisant (4.6.21) et l'inégalité de Cauchy-Schwarz, on obtient que :

$$|y_n| \leq |x_k| + H|f(0, t_k + h)|.$$

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 (4.6.22). On a bien $F(H,Y) = 0$, et $D_2 F(H,Y) = Id - HD_1 f(Y, t_k + H)$ est inversible grâce à l'hypothèse (4.1.9). Donc il existe un voisinage de (H,Y) sur lequel (4.6.22) 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 (4.1.8) 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 (4.1.9) ceci entraîne (par (4.6.21)) 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* (4.1.8) *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 85 (Condition de Lipschitz et unicité) *Corrigé en page 197*

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 :
$$\begin{aligned} &y'(t) = \varphi_a(y(t)), \ t \in [0, +\infty[\\ &y(0) = 0. \end{aligned} \tag{4.7.23}$$

Montrer que si φ_a est lipschitzienne sur les bornés alors le problème de Cauchy (4.7.23) admet une solution unique, et que si φ_a n'est pas lipschitzienne sur les bornés alors le problème de Cauchy (4.7.23) admet au moins deux solutions.

Exercice 86 (Fonctions lipschitziennes sur les bornés)

Les fonctions suivantes sont elles lipschitziennes sur les bornés ?

1.
$$\begin{array}{rl}\varphi_1 : & \mathbb{R} \to \mathbb{R} \\ & x \mapsto \min(x^2, \sqrt{(x^2+1)})\end{array}$$

2.
$$\begin{array}{rl}\varphi_2 : & \mathbb{R}^2 \to \mathbb{R}^2 \\ & (x,y) \mapsto (x^2 - xy, |y + 2xy|)\end{array}$$

3.
$$\begin{array}{rl}\varphi_3 : & \mathbb{R}^2_+ \to \mathbb{R}^2_+ \\ & (x,y) \mapsto (\sqrt{x+y}, x^2+y^2)\end{array}$$

Exercice 87 (Loi de Malthus) *Corrigé en page 197*

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 88 (Histoire de sardines) *Corrigé en page 197*

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 89 (Consistance et ordre des schémas) *Corrigé en page 198*

On reprend les hypothèses et notations (4.1.5).

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 (4.1.7)) 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 90 (Stabilité par rapport aux erreurs et convergence) *Corrigé donné en page 200*

On se place sous les hypothèses et notations (4.1.5) page 179, et on considère le schéma (4.1.6) page 180 pour la résolution numérique de l'équation différentielle (4.1.1) page 178.

1. Montrer que si le schéma (4.1.6) est stable par rapport aux erreurs au sens de la définition 4.9 page 182, et qu'il est consistant d'ordre p au sens de la définition 4.5 page 181, 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

$$\exists h^* > 0, \exists M > 0;\ \forall (x,y) \in \mathbb{R}^N \times \mathbb{R}^N, \forall h < h^*,\ \forall t \in [0,T],$$
$$|\phi(x,t,h) - \phi(y,t,h)| \leq M|x-y|,$$

alors le schéma est stable par rapport aux erreurs.

Exercice 91 (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 $[0, T_M[$) :

$$\left\{ \begin{array}{l} \dfrac{dx}{dt}(t) = f(x(t),t),\ \ t > 0, \\ x(0) = \bar{x}_0. \end{array} \right. \tag{E}$$

On se donne $T \in]0, T_M[$, et une discrétisation de $[0,T]$, définie par $n \in \mathbb{N}$ et $(t_0, t_1, \dots, t_n) \in \mathbb{R}^{n+1}$ tels que $0 = t_0 < t_1 < \dots < t_n = T$. 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 92 (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 :

$$\begin{array}{l} \overline{x} \in \mathbb{R}^N, \\ f(\overline{x}) \leq f(x),\ \forall x \in \mathbb{R}^N, \end{array} \tag{4.7.24}$$

$$\begin{array}{l} \dfrac{dx}{dt}(t) = -\nabla f(x(t)),\ t \in \mathbb{R}^+, \\ x(0) = x_0. \end{array} \tag{4.7.25}$$

1. Montrer que l'algorithme du gradient à pas fixe (de pas noté ρ) pour trouver la solution de (4.7.24) (avec point de départ x_0) est le schéma d'Euler explicite pour la résolution approchée de (4.7.25) (avec pas de temps ρ).
2. Montrer qu'il existe un unique $\overline{x}$ solution de (4.7.24).
3. Montrer que (4.7.25) admet une et une seule solution sur $\mathbb{R}_+$ et que cette solution converge vers $\overline{x}$ (solution de (4.7.24)) 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 93 (Méthode de Taylor) *Corrigé en page 200*

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 (4.1.1), 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

$$\begin{array}{l} f^{(0)} = f, \\ f^{(m+1)} = \left(\partial_1 f^{(m)}\right) f + \partial_2 f^{(m)}, \text{ pour } m \geq 0. \end{array} \tag{4.7.26}$$

1. Montrer que pour tout $m \in \mathbb{N}$, la solution x du problème de Cauchy (4.1.1) 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

$$\left\{ \begin{array}{l} x_0 = \bar{x}_0, \\ x_{k+1} = x_k + h\psi_p(x_k, t_k, h), \text{ pour } k = 1, \ldots, n. \end{array} \right. \tag{4.7.27}$$

3. Montrer que dans le cas $p = 1$, le système (4.7.27) 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 (4.7.27) 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 94 (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 :

$$y'(t) = f(y(t)),\ t \in \mathbb{R}_+, \tag{4.7.28}$$
$$y(0) = y_0. \tag{4.7.29}$$

Question 1.

1.1 Soit $T \in \overline{\mathbb{R}}_+$; on suppose que $y \in C^1([0,T[,\mathbb{R})$ est solution de (4.7.28)-(4.7.29). 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 (4.7.28)-(4.7.29) 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 (4.7.28)-(4.7.29) par le schéma d'Euler implicite de pas $h \in \mathbb{R}_+^*$, qui s'écrit :

$$y_{n+1} = y_n + hf(y_{n+1}), n \in \mathbb{N}. \tag{4.7.30}$$

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 $(y_n)_{n\in\mathbb{N}}$ solution du schéma d'Euler implicite (4.7.30) telle que $y_n \in]0,1[$ pour tout $n \in \mathbb{N}$.

3.2 Soit $(y_n)_{n\in\mathbb{N}}$ 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 :

$$y_n \leq C\Big(\frac{1}{1+h\beta}\Big)^n, \forall n \in \mathbb{N}.$$

Exercice 95 (Méthodes semi-implicite et explicite) *Corrigé en page 202*

On s'intéresse dans cet exercice au système différentiel :

$$\begin{cases} x_1'(t) = -x_1(t) - x_1(t)x_2(t), \\ x_2'(t) = -\dfrac{x_2(t)}{x_1(t)}, \end{cases} \qquad t > 0, \tag{4.7.31}$$

avec les conditions initiales

$$x_1(0) = a,\ x_2(0) = b, \tag{4.7.32}$$

où a et b appartiennent à l'intervalle $]0,1[$.

1. On pose $x = (x_1, x_2)^t$. Montrer que le système (4.7.31)-(4.7.32) s'écrit

$$\begin{cases} x'(t) = f(x(t)),\ t > 0, \\ x(0) = (a, b)^t, \end{cases} \tag{4.7.33}$$

avec $f \in C^1((\mathbb{R}_+^*)^2, \mathbb{R}^2)$.

2. Les questions suivantes sont facultatives : elles permettent de montrer que le système (4.7.33) 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 (4.7.33) (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 (4.7.33) appartenant à $C^1([0, \beta[, (\mathbb{R}_+^*)^2)$.

   (c) Montrer que le système (4.7.33) 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 (4.7.31)-(4.7.32) : soit k le pas de discrétisation, choisi tel que $0 < k < \frac{1}{2}$.

$$\begin{cases} \dfrac{x_1^{(n+1)} - x_1^{(n)}}{k} = -x_1^{(n)} - x_1^{(n)} x_2^{(n+1)}, \\ \dfrac{x_2^{(n+1)} - x_2^{(n)}}{k} = -\dfrac{x_2^{(n+1)}}{x_1^{(n)}}, \\ x_1^{(0)} = a,\ x_2^{(0)} = b. \end{cases} \tag{4.7.34}$$

   (a) Montrer par récurrence sur n que les suites $(x_1^{(n)})_{n\in\mathbb{N}}$ et $(x_2^{(n)})_{n\in\mathbb{N}}$ données par (4.7.34) sont bien définies, décroissantes et strictement positives.

   (b) Montrer que le schéma numérique (4.7.34) s'écrit sous la forme

$$\frac{x^{(n+1)} - x^{(n)}}{k} = \phi(x^{(n)}, k), \tag{4.7.35}$$

   avec $x^{(n)} = (x_1^{(n)}, x_2^{(n)})^t$, $\phi \in C^\infty((\mathbb{R}_+^*)^2 \times \mathbb{R}_+, \mathbb{R}^2)$ et $\phi(x, 0) = f(x)$.

   (c) (Consistance)
   Soit $T > 0$. Pour $n \in \mathbb{N}$, on note $t_n = nk$. Montrer qu'il existe $C(T) \in \mathbb{R}_+$ tel que

$$\frac{x(t_{n+1}) - x(t_n)}{k} = \phi(x(t_n), k) + R_k^{(n)},\ \text{pour tout } n \text{ tel que } nk \leq T, \tag{4.7.36}$$

   avec $|R_k^{(n)}| \leq C(T)k$.

   (d) (Stabilité)
   Soit $T > 0$.

   (i) Montrer que $x_1^{(n)} \geq (1 - k - kb)^{\frac{T}{k}}$ 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

$$\begin{cases} a(T) \leq x_1^{(n)} \leq a, \\ b(T) \leq x_2^{(n)} \leq b, \end{cases} \quad \text{pour tout } n \text{ tel que } nk \leq T. \tag{4.7.37}$$

(e) (Convergence)
Soit $T > 0$. Montrer qu'il existe $D(T) \in \mathbb{R}_+$ tel que

$$|x^{(n)} - x(t_n)| \leq D(T)k, \text{ pour tout } n \text{ tel que } nk \leq T. \tag{4.7.38}$$

En déduire la convergence du schéma (4.7.34).

(f) On remplace maintenant le schéma (4.7.34) par le schéma d'Euler explicite pour le système (4.7.33). Ecrire ce schéma. Montrer que pour tout pas de discrétisation $k > 0$, il existe des valeurs de n telles que $x_1^{(n)} \leq 0$ ou $x_2^{(n)} \leq 0$. (On pourra montrer que si $x_1^{(n)} > 0$ et $x_2^{(n)} > 0$ pour tout $n \in \mathbb{N}$, alors $x_1^{(n)}$ tend vers 0 lorsque n tend vers $+\infty$, et donc qu'il existe n tel que $x_2^{(n)} \leq 0$, ce qui contredit l'hypothèse). Commenter.

Exercice 96

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 :

$$\forall (y,z) \in (\mathbb{R}^n)^2, \ (f(y,t) - f(z,t), y - z) \leq 0. \tag{4.7.39}$$

On considère le système différentiel :

$$y'(t) = f(y(t),t) \, \forall t \in [0,T[, \tag{4.7.40}$$

$$y(0) = y^{(0)}. \tag{4.7.41}$$

1. Montrer que pour tout $y \in \mathbb{R}^n$ et $t \in [0,T[$, on a :

$$(f(y,t),y) \leq \frac{1}{2}(\|f(0,t)\|^2 + \|y\|^2). \tag{4.7.42}$$

En déduire qu'il existe une unique solution $y \in C^1([0,T[, \mathbb{R}^n)$ vérifiant (4.7.40)-(4.7.41).

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 $k = 0, \ldots, N$, on note $t_k = kh$, et on se propose d'étudier l'algorithme suivant, où $0 \leq \theta \leq 1$.

$$y_0 \in \mathbb{R}^n \text{ est donné} \tag{4.7.43}$$

$$y_{k,1} = y_k + \theta h f(y_{k,1}, t_k + \theta h), \text{ pour } k = 0, \ldots, N-1, \tag{4.7.44}$$

$$y_{k+1} = y_k + h f(y_{k,1}, t_k + \theta h) \text{ pour } k = 0, \ldots, N-1, \tag{4.7.45}$$

2. Montrer qu'il existe une unique solution $(y_k)_{k=0,\ldots,N} \subset \mathbb{R}^n$ de (4.7.43)-(4.7.44)-(4.7.45).

Pour $k = 0, \ldots, N-1$, on pose $y(t_k) = \overline{y}_k$, où y est la solution exacte de (4.7.40)-(4.7.41), $t_{k,1} = t_k + \theta h$, on définit $\tilde{y}_{k,1}$ par :

$$\tilde{y}_{k,1} = \overline{y}_k + \theta h f(\tilde{y}_{k,1}, t_{k,1}), \tag{4.7.46}$$

et on définit l'erreur de consistance R_k du schéma (4.7.43)-(4.7.44)-(4.7.45) au point t_k par :

$$R_k = \frac{\overline{y}_{k+1} - \overline{y}_k}{h} - f(\tilde{y}_{k,1}, t_{k,1}) \tag{4.7.47}$$

3. Pour $k = 0, \ldots, N$, on pose $\overline{y}_{k,1} = y(t_{k,1})$, et, pour $k = 0, \ldots, N-1$ on pose :

$$\tilde{R}_k = \frac{1}{h}(\overline{y}_{k,1} - \overline{y}_k) - \theta f(\overline{y}_{k,1}, t_{k,1}). \tag{4.7.48}$$

Montrer que pour tout $k = 0, \ldots, N-1$:

$$\tilde{y}_{k,1} - \overline{y}_{k,1} = \theta h\big(f(\tilde{y}_{k,1}, t_{k,1}) - f(\overline{y}_{k,1}, t_{k,1})\big) + h\tilde{R}_k, \tag{4.7.49}$$

En déduire qu'il existe C_1 ne dépendant que de y et de T t.q. : $\|\tilde{y}_{k,1} - \overline{y}_{k,1}\| \leq C_1 h^2$.

4. Montrer qu'il existe C_2 ne dépendant que de f, y et T t.q.

$$\|\overline{y}_{k+1} - \overline{y}_k - hf(\overline{y}_k, t_{k,1}) - hR_k\| \leq C_2 h^3, \; \forall k = 0, \ldots, N-1. \tag{4.7.50}$$

5. Déduire des questions précédentes qu'il existe C_3 ne dépendant que de y, f et T t.q. :

$$\|R_k\| \leq C_3((\theta - \frac{1}{2})h + h^2) \tag{4.7.51}$$

et en déduire l'ordre du schéma (4.7.43)-(4.7.44)-(4.7.45).

6. Montrer que pour tout $k = 1, \ldots, N$, on a :

$$\big(\overline{y}_k - y_k, f(y_{k,1}, t_{k,1}) - f(\tilde{y}_{k,1}, t_{k,1})\big) \leq -\theta h \|f(y_{k,1}, t_{k,1}) - f(\tilde{y}_{k,1}, t_{k,1})\|^2. \tag{4.7.52}$$

7. Montrer que pour tout $k = 0, \ldots, N$, on a :

$$\|e_{k+1} - hR_k\|^2 = \|e_k\|^2 + 2h(f(y_{k,1}, t_{k,1}) - f(\tilde{y}_{k,1}, t_{k,1}), e_k) + h^2 \|f(y_{k,1}, t_{k,1}) - f(\tilde{y}_{k,1}, t_{k,1})\|^2. \tag{4.7.53}$$

8. Montrer que si $\theta \geq \frac{1}{2}$, on a :

$$\|e_k\| \leq \|e_0\| + C_3(h^2 + (\theta - \frac{1}{2})h), \; \forall k = 1, \ldots, N. \tag{4.7.54}$$

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 :

$$z_0 \in \mathbb{R}^n \text{ donné} \tag{4.7.55}$$

$$z_{k,1} = z_k + \theta h f(z_{k,1}, t_{k,1}), \text{ pour } k = 0, \ldots, N-1, \tag{4.7.56}$$

$$z_{k+1} = z_k + +\varepsilon_k + hf(z_{k,1}, t_{k,1}) \text{ pour } k = 0, \ldots, N-1, \tag{4.7.57}$$

En s'inspirant des questions 6 et 7, montrer que si $\theta \geq \frac{1}{2}$, on a :

$$\|y_{k+1} - z_{k+1} + \varepsilon_k\|^2 \leq \|y_k - z_k\|^2, \tag{4.7.58}$$

et en déduire que

$$\|y_k - z_k\| \leq \|y_0 - z_0\| + \sum_{i=0}^{k-1} \|\varepsilon_i\|. \tag{4.7.59}$$

4.8 Corrigés

Corrigé de l'exercice 85 page 189 (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 (4.7.23) admet une unique solution qui est la solution constante et égale à zéro.
Si φ_a est lipschitzienne sur les bornés, i.e. si $a \in]0, 1[$, la fonction nulle est encore solution du problème (4.7.23), 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 $a \in]0, 1[$.)

Corrigé de l'exercice 87 page 190 (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.

Corrigé de l'exercice 88 page 190 (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 :

$$\begin{cases} x(t) = \dfrac{1}{\left(10^{-6} - \frac{10^{-4}}{4}\right)e^{-4t} + \frac{10^{-4}}{4}} \quad 0 \leq t \leq 30 \\ x_{30} = x(30), \\ x(t) = \dfrac{1}{\left(\frac{1}{x_{30}} - \frac{10^{-4}}{3.9}\right)e^{-3.9t} + \frac{10^{-4}}{3.9}} \qquad t \geq 30. \end{cases}$$

On en conclut que $\lim_{t\to+\infty} x(t) = 3.9\ 10^4$.

Corrigé de l'exercice 89 page 190 (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 182.

2. Pour les schémas d'Euler amélioré et de Heun, le théorème 4.12 page 182 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 :

$$\left| \frac{\bar{x}_{k+1} - \bar{x}_k}{h_k} - x'(t_k + \frac{h_k}{2}) \right| \leq C h^2 \tag{4.8.60}$$

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

$$|\frac{1}{2}(x'(t_{k+1}) + x'(t_k)) - x'(t_k + \frac{h_k}{2})| \leq C h^2.$$

En utilisant à nouveau (4.8.60), on en déduit que l'erreur de consistance est d'ordre 2 (à condition que x soit trois fois dérivable . . .).

Corrigé de l'exercice 90 page 190 (Stabilité par rapport aux erreurs et convergence)

1. Par définition du schéma (4.1.6) et de l'erreur de consistance (4.2.10), on a :

$$\begin{array}{ll} x_{k+1} & = x_k + h_k\phi(x_k, t_k, h_k) \\ \bar{x}_{k+1} & = \bar{x}_k + h_k\phi(\bar{x}_k, t_k, h_k) + h_k R_k. \end{array}$$

Comme le schéma (4.1.6) est supposé stable par rapport aux données, on a en prenant $y_k = \bar{x}_k$ et $\varepsilon_k = h_k R_k$ dans (4.2.12) page 182 :

$$e_{k+1} \leq K(|x_0 - \bar{x}_0| + \sum_{i=0}^{k-1} |h_i R_i|) \text{ pour tout } k = 0, \ldots, n-1.$$

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 $(x_k)_{k=0,\ldots,n-1}$ et $(y_k)_{k=0,\ldots,n-1}$ vérifiant (4.2.12), c'est–à–dire :

$$\begin{array}{l} x_{k+1} = x_k + h_k\phi(x_k, t_k, h_k), \\ y_{k+1} = y_k + h_k\phi(y_k, t_k, h_k) + \varepsilon_k, \end{array} \quad \text{pour } k = 0, \ldots, n-1,$$

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 (4.1.6) est stable par rapport aux erreurs.

Corrigé de l'exercice 93 page 191 (Méthode de Taylor)

1. Soit x solution du problème de Cauchy (4.1.1). 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

$$x^{(p+1)}(t) = f^{(p)}(x(t), t) \text{ pour } p = 0, \ldots, m,$$

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 (4.7.27) s'écrit :

$$\begin{cases} x_0 = \bar{x}_0, \\ x_{k+1} = x_k + hf(x_k, t_k), \text{ pour } k = 1, \ldots, n. \end{cases}$$

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

$$x_k = \left(\sum_{j=0}^{p} \frac{h^j}{j!} \right)^k \text{ pour } k = 1, \ldots, \ell,$$

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 (4.1.1) 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

$$\begin{aligned} x_k &= \left(\sum_{j=0}^{p} \frac{h^j}{j!} \right)^k \\ &= \left(e^h - R(h) \right)^k, \end{aligned}$$

avec $0 < R(h) < e^h \frac{h^{p+1}}{(p+1)!}$. On a donc

$$\begin{aligned} x_k &= e^k h \left(1 - \frac{R(h)}{e^h} \right)^k \\ &= e^k h (1 - a)^k, \end{aligned}$$

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 182 (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 $k = 0, \ldots, n+1$.

Corrigé de l'exercice 4.7 page 193 (Méthodes semi-implicite et explicite)

1.

$$\left\{\begin{array}{l} \dfrac{x_1^{(n+1)} - x_1^{(n)}}{k} = -x_1^{(n)} - x_1^{(n)} x_2^{(n+1)}, \\ \dfrac{x_2^{(n+1)} - x_2^{(n)}}{k} = -\dfrac{x_2^{(n+1)}}{x_1^{(n)}}, \\ x_1^{(0)} = a,\ x_2^{(0)} = b. \end{array}\right. \tag{4.8.61}$$

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

$$x_2^{(n+1)} = \frac{1}{1 + \frac{k}{x_1^{(n)}}} x_2^{(n)}, \tag{4.8.62}$$

et donc en utilisant l'hypothèse de récurrence, on obtient que $x_2^{(n+1)} < x_2^{(n)}$ et $0 < x_2^{(n+1)} < b$.
De plus

$$x_1^{(n+1)} = x_1^{(n)} - k x_1^{(n)} - k x_1^{(n)} x_2^{(n+1)} = x_1^{(n)}(1 - k - k x_2^{(n+1)}), \tag{4.8.63}$$

et donc grâce au fait que $0 < x_2^{(n+1)} < b$ (et donc $1 - k - k x_2^{(n+1)} > 1 - k - kb$, et à l'hypothèse de récurrence, on déduit que $x_1^{(n+1)} < x_1^{(n)}$ et $0 < x_1^{(n+1)} < a$.

2. Après calcul, on obtient que le schéma numérique (4.7.34) s'écrit sous la forme

$$\frac{x^{(n+1)} - x^{(n)}}{k} = \phi(x^{(n)}, k), \tag{4.8.64}$$

avec $x^{(n)} = (x_1^{(n)}, x_2^{(n)})^t$, et où $\phi \in C^\infty((\mathbb{R}_+^*)^2 \times \mathbb{R}_+, \mathbb{R}^2)$ est définie par

$$\phi(x,k) = \begin{pmatrix} -x_1(1 + \dfrac{x_1 x_2}{x_1 + k}) \\ -\dfrac{x_2}{x_1 + k} \end{pmatrix}, \tag{4.8.65}$$

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 $(x_1^{(n)}, x_2^{(n)}) \in (\mathbb{R}_+^*)^2$ et $k > 0$, le couple $(x_1^{(n+1)}, x_2^{(n+1)})$ est bien défini par (4.7.34) 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 $|R_k^{(n)}| \leq C(T)k$.

4. (Stabilité)

   (i) Soit $T > 0$. De (4.8.63) et du fait que $0 < x_2^{(n)} < b$ on déduit que

   $$x_1^{(n+1)} \geq x_1^{(n)}(1 - k - kb),$$

   et donc par récurrence sur n que

   $$x_1^{(n)} \geq x_1^{(0)}(1 - k - kb)^n,$$

   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 $x_1^{(n)} \geq (1 - k - kb)^{\frac{T}{k}}$.

   (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 $a(T) \leq x_1^{(n)} \leq a$, 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 (4.8.62) que

   $$x_2^{(n+1)} \geq \frac{1}{1 + \frac{k}{a(T)}} x_2^{(n)},$$

et donc que

$$x_2^{(n)} \geq \Big(\frac{1}{1+\frac{k}{a(T)}}\Big)^{\frac{T}{k}} x_2^{(0)},$$

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 $b(T) \leq x_2^{(n)} \leq b$, 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 : $x^{(n)} \in [a(T), a] \times [b(T), b]$.
   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 (4.7.34) par le schéma d'Euler explicite. Celui s'écrit :

$$\begin{cases} \dfrac{x_1^{(n+1)} - x_1^{(n)}}{k} = -x_1^{(n)} - x_1^{(n)} x_2^{(n)}, \\ \dfrac{x_2^{(n+1)} - x_2^{(n)}}{k} = -\dfrac{x_2^{(n)}}{x_1^{(n)}}, \\ x_1^{(0)} = a,\ x_2^{(0)} = b. \end{cases} \tag{4.8.66}$$

Supposons $x_1^{(n)} > 0$ et $x_2^{(n)} > 0$ pour tout n. La première équation de (4.7.34) donne alors que

$$\frac{x_1^{(n+1)} - x_1^{(n)}}{k} = -x_1^{(n)},$$

et donc $x_1^{(n+1)} < (1-k)x_1^{(n)}$. On en déduit par récurrence que $x_1^{(n)} < (1-k)^n a \to 0$ 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 $x_1^{(n)} \leq k$. Or pour cette valeur de n,

$$x_2^{(n+1)} = x_2^{(n)}(1 - \frac{k}{x_1^{(n)}}) \leq 0,$$

ce qui contredit l'hypothèse $x_2^{(n)} > 0$ 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...)