UFR de Mathématiques

Université Claude Bernard - Lyon I
43, Boulevard du 11 Novembre 1918
69622, Villeurbanne cedex, France

Cours rédigé par

Francis FILBET

Introduction à l'analyse numérique

Spécialité : Licence en Mathématiques.

Mathématiques Pures et Mathématiques Appliquées – Lyon

Table des matières

Chapitre 1

Les systèmes linéaires

1 Quelques exemples de systèmes linéaires

1.1 Algorithme de Google

Pour classer les réponses à une requête donnée, Google utilise deux facteurs :

- le **score de pertinence** qui mesure l'adéquation entre la requête et le contenu de la page web
- l'**indice de popularité** qui ne dépend pas de la consultation de la page par les internautes mais du nombre de liens qui pointent sur cette page à partir d'autres pages web.

Nous nous intéressons ici seulement au calcul de l'indice de popularité des pages web. Pour cela, nous considérons qu'il y a N pages web au total ; en pratique N est de l'ordre de 5 à 6 milliards. Nous calculons la probabilité $p(A)$ de consulter la page A, en sachant que nous arrivons directement à la page A dans $100\,(1-d)\,\%$ des cas et indirectement c'est-à-dire en utilisant un lien trouvé sur une autre page web dans $100\,d\,\%$ des cas où d est un réel compris entre zéro et un.

Nous appelons indice de popularité de A le nombre $x_A = N \times p(A)$. Il vérifie alors l'équation suivante

$$x_A = (1-d) \, + \, d \sum_{i=1}^{k} \frac{x_{T_i}}{N(T_i)},$$

où les T_i, $1 \leq i \leq k$ sont les pages qui ont un lien qui pointe vers la page A et $N(T_i)$ est le nombre de liens de la page T_i.

Pour simplifier la compréhension, nous considérons le cas de quatre pages A, B, C et D où la page A pointe vers les pages B, C et D, la page B pointe vers A, C et D, la page C pointe vers A, B et D et la page D pointe vers A, B et C, ce qui peut être résumé par le schéma de la Figure 1.1.

Nous pouvons alors écrire un système linéaire $A\,x = b$ de taille 4×4 lié au cas décrit dans la Figure 1.1, où l'inconnue est le vecteur $x = (x_A, x_B, x_C, x_D)^T$, tandis que la matrice A et le

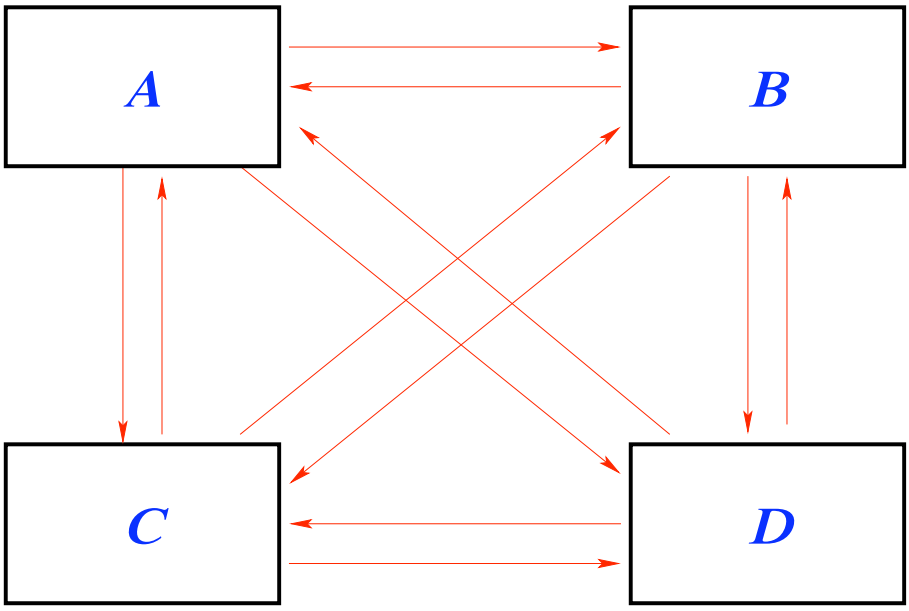

FIG. 1.1 – Exemple de liens entre les pages A, B, C et D.

vecteur b sont donnés par

$$A = \begin{pmatrix} 1 & a & a & a \\ a & 1 & a & a \\ a & a & 1 & a \\ a & a & a & 1 \end{pmatrix}, \quad b = (1-d) \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix},$$

avec $a = -d/3$.

Ce système n'est en général pas facile à résoudre (ici nous trouvons bien sûr que $x = (1,1,1,1)^T$), imaginons ce qu'il en est lorsque que nous tenons compte de l'ensemble des pages web et de l'ensemble des requêtes. Il faut donc résoudre un système linéaire avec un grand nombre d'inconnues et pratiquement de manière instantanée.

1.2 L'équation de la chaleur

Nous examinons maintenant un autre problème issu de la physique. Imaginons une chambre dans laquelle nous appliquons une source de chaleur aux bords et au centre de la pièce.

La chaleur va se répandre à l'intérieur de la pièce en suivant une dynamique décrite par une équation aux dérivées partielles. Nous renvoyons le lecteur au Chapitre 7 pour plus de détails sur

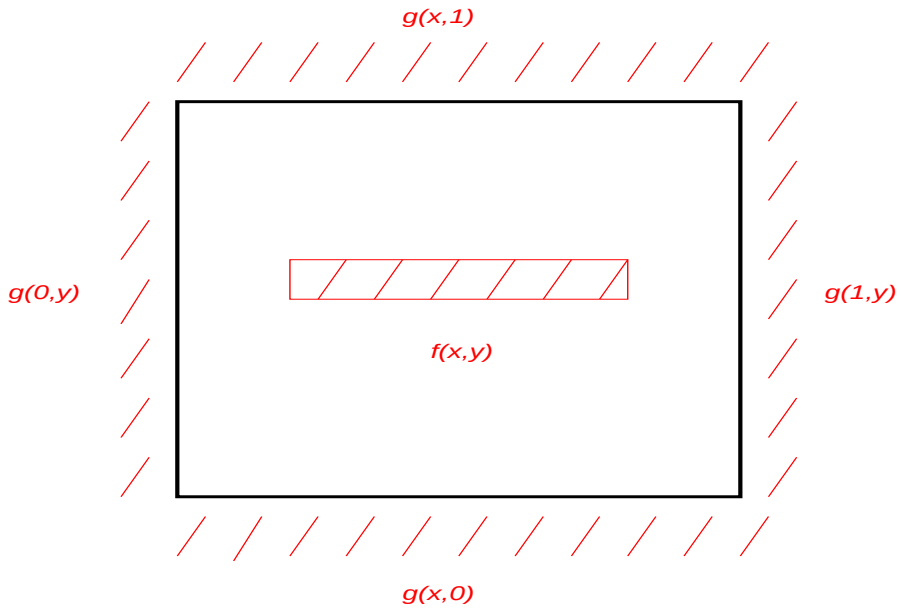

FIG. 1.2 – Étude de l'équation de la chaleur dans une pièce en dimension deux

l'étude numérique de ces équations. Ici, nous considérons l'équation suivante (appelée équation de la chaleur linéaire) :

$$\frac{\partial u}{\partial t} - \Delta u = f, \quad t \in I\!R^+ \quad (x,y) \in \Omega,$$

où f désigne la puissance surfacique et

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

Si la température aux bords est maintenue constante, la distribution de chaleur dans la pièce converge vers un état stationnaire. Dans ce cas, la dérivée temporelle dans l'équation précédente disparaît, et nous nous retrouvons avec l'équation de Laplace :

$$-\Delta u = f, \quad (x,y) \in \Omega.$$

La résolution de cette équation représente un problème aux bords typique : la solution dépend fortement de la condition imposée aux bords du domaine $\Gamma = \partial\Omega$

$$u = g, \quad (x,y) \in \Gamma,$$

où g désigne la source de chaleur au bord (par exemple un radiateur attaché au mur).

Prenons alors $\Omega = (0,1) \times (0,1)$ et recouvrons le domaine Ω par une grille formée de parallèles aux axes. Nous considérons alors un point P de la grille, dont les voisins sont notés

E (Est), O, (Ouest), N (Nord) et S (Sud). À l'aide d'une formule de Taylor, nous écrivons en chaque point voisin de P et pour une fonction u

$$\left\{\begin{array}{rcl} u(E) &=& u(P) + h\,\frac{\partial u}{\partial x}(P) + h^2\,\frac{\partial^2 u}{\partial x^2}(P) + O(h^3), \\ u(O) &=& u(P) - h\,\frac{\partial u}{\partial x}(P) + h^2\,\frac{\partial^2 u}{\partial x^2}(P) + O(h^3), \\ u(N) &=& u(P) + h\,\frac{\partial u}{\partial y}(P) + h^2\,\frac{\partial^2 u}{\partial y^2}(P) + O(h^3), \\ u(S) &=& u(P) - h\,\frac{\partial u}{\partial y}(P) + h^2\,\frac{\partial^2 u}{\partial y^2}(P) + O(h^3). \end{array}\right. \tag{1.1}$$

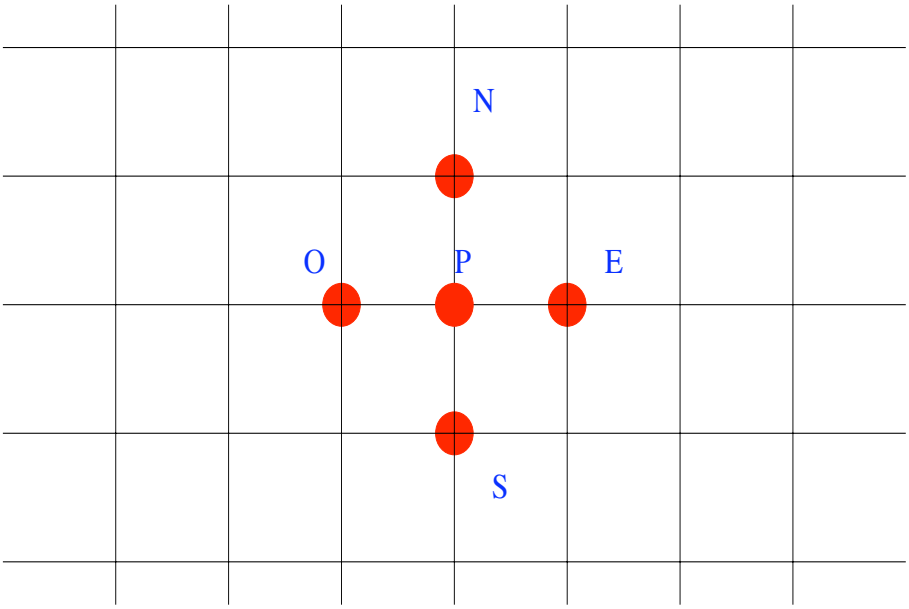

FIG. 1.3 – Localisation des points voisins d'un point arbitraire P où nous effectuons un développement de Taylor.

Nous avons choisi pour simplifier les points de la grille équidistants $(x_i, y_j) = (i\,h, j\,h)$, avec $h = 1/(n+1)$ et en notant $u_{i,j}$ l'inconnue approchant la solution u de l'équation de Laplace aux points (x_i, y_j), nous obtenons le système linéaire de n^2 équations en écrivant $P = (x_i, y_j)$ et $E = (x_{i-1}, y_j)$, $N = (x_i, y_{j+1})$,... Puis, en sommant les quatre égalités de (1.1) tout en négligeant les termes d'ordre supérieur à trois (c'est-à-dire les terme $O(h^3)$), nous obtenons le système suivant pour l'approximation $u_{i,j}$

$$\frac{4\,u_{i,j} - u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1}}{h^2} = f(x_i, y_j) =: f_{i,j}$$

avec également n^2 inconnues en interprétant la condition aux limites $u = g$ sur Γ selon

$$u_{0,j} \;=\; g(0, y_j), \quad u_{n+1,j} \;=\; g(1, y_j), \quad j = 0, ..., n+1$$

et

$$u_{i,0} \;=\; g(x_i, 0), \quad u_{i,n+1} \;=\; g(x_i, 0), \quad i = 0, ..., n+1.$$

Nous verrons au Chapitre 7 comment justifier rigoureusement que nous obtenons bien une approximation de la solution du problème de l'équation de Laplace mais ce n'est pas cela qui nous préoccupe pour l'instant.

Ainsi, nous aboutissons à la résolution d'un système linéaire de grande taille :

$$A\,x \;=\; b,$$

avec

$$A = \begin{pmatrix} D & -I_n & 0 & \dots & 0 \\ -I_n & D & -I_n & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -I_n \\ 0 & \dots & 0 & -I_n & D \end{pmatrix}$$

et D est une matrice carrée de taille $n \times n$

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

et x est le vecteur des inconnues et b la donnée

$$x = \begin{pmatrix} u_{1,1} \\ u_{2,1} \\ \vdots \\ u_{2,1} \\ u_{2,2} \\ \vdots \\ u_{n,n} \end{pmatrix}, \quad b = \begin{pmatrix} h^2\, f_{1,1} + g(0,h) + g(h,0) \\ h^2\, f_{2,1} + g(2\,h,0) \\ \vdots \\ h^2\, f_{1,2} + g(0,2\,h) \\ h^2\, f_{2,2} \\ \vdots \\ h^2\, f_{n,n} + g(1,1-h) + g(1-h,1) \end{pmatrix}.$$

Dans la suite nous présentons quelques rappels sur les matrices. Puis, nous proposons deux types de méthodes de résolution : les méthodes directes qui sont exactes (méthodes d'élimination de Gauss, décomposition LU et factorisation de Choleski) et les méthodes itératives (algorithmes de Jacobi et Gauss-Seidel) qui forment une suite d'approximations de la solution. Nous proposons enfin un complément détaillant d'autres méthodes classiques.

2 Rappels sur les matrices

Avant de s'intéresser à la résolution numérique de systèmes linéaires, nous présentons quelques rappels d'algèbre linéaire.

Soit u un vecteur à n composantes, notées $(u_i)_{1\leq i\leq n}$, à valeur dans un corps $\mathbb{K}$ (par exemple le corps des réels $\mathbb{R}$). Nous noterons l'adjoint du vecteur colonne u, le vecteur ligne $u^\star$ de $\mathbb{K}^n$ tel que $u^\star = (\overline{u}_1, ..., \overline{u}_n)$ et le transposé de u, le vecteur ligne $u^T = (u_1, ..., u_n)$. Nous rappelons d'abord le produit matrice $\times$ vecteur : soit A une matrice à m lignes et n colonnes et u un vecteur de $\mathbb{K}^n$, nous définissons $v = A\,u \in \mathbb{K}^m$ le vecteur

$$v_i = \sum_{j=1}^{n} a_{i,j}\, u_j, \quad i = 1, ..., m$$

et pour B un matrice à n lignes et p colonnes, nous avons $C = AB$ avec

$$c_{i,j} = \sum_{k=1}^{n} a_{i,k}\, b_{k,j}, \quad i = 1, ..., m, \quad j = 1, ..., p.$$

Nous avons aussi la notion de matrice adjointe et transposée

Définition 2.1 *Soit $A \in \mathcal{M}_{m,n}(\mathbb{K})$, c'est-à-dire une matrice de $\mathbb{K}$ à m lignes et n colonnes. Nous appelons matrice adjointe de A la matrice $A^\star$, à n lignes et m colonnes, donnée par $A^\star = (a^\star_{i,j})_{\substack{1\leq i\leq n\\ 1\leq j\leq m}}$ et*

$$a^\star_{i,j} = \overline{a}_{j,i}, \quad 1 \leq i \leq n, \quad 1 \leq j \leq m.$$

Nous appelons matrice transposée de A la matrice A^T, à n lignes et m colonnes, donnée par $A^T = (a^T_{i,j})_{\substack{1\leq i\leq n\\ 1\leq j\leq m}}$ et

$$a^T_{i,j} = a_{j,i}, \quad 1 \leq i \leq n, \quad 1 \leq j \leq m.$$

Dans la base canonique de $\mathbb{K}^n$, nous définissons le produit scalaire entre deux vecteurs u et $v \in \mathbb{K}^n$, le scalaire de $\mathbb{K}$ donné par

$$(u, v) = \sum_{i=1}^{n} u_i \, v_i^*.$$

2.1 Cas des matrices carrées

Considérons maintenant le cas particulier de matrices carrées lorsque le nombre de lignes est égal au nombre de colonnes. Nous rappelons les définitions suivantes

Définition 2.2 *Soit $A \in \mathcal{M}_{n,n}(\mathbb{K})$ une matrice carrée d'ordre n.*

- *A est une matrice normale si et seulement si $A^\star\, A \;=\; A\, A^\star$.*
- *A est une matrice unitaire si et seulement si $A^\star\, A \;=\; A\, A^\star \;=\; I_n$, où I_n désigne la matrice identité d'ordre n, la matrice est alors inversible. Dans le cas où $\mathbb{K} = \mathbb{R}$, nous parlons de matrice orthogonale et $A^T\, A = A\, A^T = I_n$, c'est-à-dire $A^T = A^{-1}$.*
- *A est une matrice hermitienne si et seulement si $A^\star \;=\; A$. Dans le cas où $\mathbb{K} = \mathbb{R}$, nous parlons de matrice symétrique et nous avons $A^T \;=\; A$.*

Proposition 2.1 *Toute matrice hermitienne est une matrice normale.*

Démonstration. Soit $A \in \mathcal{M}_{n,n}(\mathbb{K})$ telle que $A^\star = A$. Nous avons alors

$$A\, A^\star \;=\; A\, A \;=\; A^\star\, A.$$

□

Pour la suite, nous rappelons qu'une valeur propre de A est donnée par $\lambda \in \mathbb{K}$ telle que

$$det(A \;-\; \lambda\, I_n) \;=\; 0.$$

Ainsi, il existe au moins un vecteur v non nul (dit *vecteur propre*) vérifiant

$$A\, v \;=\; \lambda\, v.$$

Définition 2.3 *Soit $A \in \mathcal{M}_{n,n}(\mathbb{K})$; nous appelons spectre de A l'ensemble des valeurs propres de A*

$$Sp(A) = \{\lambda_i \in \mathbb{K}; 1 \leq i \leq n; \exists\, v_i \in \mathbb{K}^n : v_i \neq 0 \quad A\, v_i = \lambda_i\, v_i\}\,.$$

Nous appelons rayon spectral de A le nombre réel positif $\rho(A)$ tel que

$$\rho(A) = \max_{1\leq i\leq n} \{|\lambda_i|; \quad \lambda_i \in Sp(A)\}\,.$$

Comme nous l'avons vu en début de chapitre, l'objectif de cette partie est de mettre au point des algorithmes de résolution numérique pour un système de la forme

$$A\, x = b, \tag{2.2}$$

où $A \in \mathcal{M}_{n,n}(\mathbb{K})$ et $b \in \mathbb{R}^n$ sont donnés et $x \in \mathbb{R}^n$ est l'inconnue. Nous donnons d'abord une condition nécessaire et suffisante sur la matrice A pour que ce système admette une solution unique.

Définition 2.4 *Une matrice carrée $A \in \mathcal{M}_{n,n}(\mathbb{K})$ d'ordre n est dite inversible ou régulière ou encore non singulière, si et seulement si il existe une matrice B d'ordre n telle que*

$$A\, B = B\, A = I_n.$$

Dans ce cas, la matrice B est unique et est appelée la matrice inverse de A, et est notée A^{-1}.

Nous rappelons alors la définition du determinant d'une matrice carrée

Définition 2.5 *Soit $A \in \mathcal{M}_{n,n}(\mathbb{K})$, le determinant de la matrice A, noté $det(A)$, est donné par*

$$det(A) = \sum_{\sigma\in S_n} sign(\sigma) a_{1,\sigma(1)} \ldots a_{n,\sigma(n)},$$

où la somme porte sur l'ensemble des permutations de n, c'est-à-dire $n!$ permutations et $sign(\sigma) = \pm 1$ selon que le nombre de transpositions (permutation de deux éléments) de σ est pair ($sign(\sigma) = 1$) ou impair ($sign(\sigma = -1)$).

Le système linéaire (2.2) admet alors une unique solution si et seulement si la matrice A est inversible et la solution est donnée par $x = A^{-1}\, b$. Le plus souvent le calcul de A^{-1} est long et fastidieux, même pour un ordinateur, c'est pourquoi nous avons recours à des algorithmes de résolution exacte (nous parlons alors d'une méthode directe) ou des méthodes d'approximation de la solution (nous disons une méthode itérative).

Avant de présenter de tels algorithmes, nous énonçons quelques propriétés des matrices inversibles, ce qui nous permettra de s'assurer que le problème à résoudre admet bien une solution.

Théorème 2.1 (Théorème des matrices inversibles) *Soit $A \in \mathcal{M}_{n,n}(\mathbb{K})$ une matrice carrée d'ordre n à coefficients dans un corps $\mathbb{K}$ (par exemple le corps des réels $\mathbb{R}$). Les propositions suivantes sont équivalentes :*

- *A est inversible,*
- *A possède n pivots,*
- *$det(A) \neq 0$, (déterminant non nul),*
- *le rang de A est égal à n,*
- *le système homogène $A\,x = 0$ a pour unique solution $x = 0$,*
- *pour tout b dans $\mathbb{K}^n$, le système linéaire $A\,x = b$ a exactement une solution,*

Nous avons aussi un résultat permettant de calculer la solution d'un système linéaire, cet algorithme est dû à Cramer.

Théorème 2.2 *Soit $A \in \mathcal{M}_{n,n}(\mathbb{K})$ une matrice carrée d'ordre n inversible. Alors, la solution du système $A\,x = b$ est donnée par*

$$x_i = det(A_i)/det(A), \quad i = 1, \ldots, n$$

où A_i est la matrice A à laquelle nous remplaçons le $i$ème colonne par le vecteur b.

Cette méthode bien que très élégante est très coûteuse. Elle nécessite en effet plus de $n^2\,n!$ opérations, elle n'est donc jamais utilisée en pratique sauf en dimension $n = 2$. Avant de proposer des solutions alternatives, voyons un premier théorème qui permet d'obtenir une représentation triangulaire d'une matrice carrée par changement de base.

Théorème 2.3 (Théorème de Shur) *Soit $A \in \mathcal{M}_{n,n}(\mathbb{K})$ une matrice quelconque. Alors, il existe une matrice unitaire $U \in \mathcal{M}_{n,n}(\mathbb{K})$ (c'est-à-dire $U^*\,U = I_n$) telle que*

$$T = U^*\,A\,U,$$

où T est une matrice triangulaire dont la diagonale est composée par l'ensemble des valeurs propres de A.

Démonstration. Nous renvoyons à [3][Théorème 1.2-1] pour une preuve complète. □

Nous avons alors le corollaire suivant particulièrement important pour les applications pratiques

Corollaire 2.1 *Soit $A \in \mathcal{M}_{n,n}(\mathbb{K})$ une matrice hermitienne. Alors, il existe une matrice unitaire $U \in \mathcal{M}_{n,n}(\mathbb{K})$ c'est-à-dire $U^*\,U = I_n$, telle que*

$$D = U^*\,A\,U,$$

où D est une matrice diagonale, composée par l'ensemble des valeurs propres de A.

Nous rappelons finalement quelques propriétés qu'il est bon de connaître pour la suite.

Proposition 2.2 *Soient A, $B \in \mathcal{M}_{n,n}(\mathbb{K})$ inversibles, nous avons alors*
- *pour tout $\alpha \in \mathbb{K}$, $(\alpha\, A)^{-1} = \frac{1}{\alpha} A^{-1}$*
- $(A\, B)^{-1} = B^{-1}\, A^{-1}$
- $(AB)^T = B^T\, A^T$.
- $(A^T)^{-1} = (A^{-1})^T$

De plus, le determinant $det(A)$ vérifie les propriétés suivantes
- *pour tout $\alpha \in \mathbb{K}$, $det(\alpha\, A) = \alpha^n det(A)$.*
- *la valeur du determinant est inchangée si une ligne (resp. colonne) multipliée par un scalaire est ajoutée à une autre ligne (resp. colonne) ;*
- *le determinant d'une matrice triangulaire est égal au produit des termes diagonaux ;*
- *si deux lignes (resp colonnes) sont interchangées, le determinant est mutiplié par -1 ;*
- $det(A\, B) = det(A)\, det(B)$

La conséquence de ces propriétés est que l'ensemble des matrices carrées d'ordre n inversibles forme un groupe, appelé le groupe linéaire et noté habituellement Gl_n. En général, "presque toutes" les matrices sont inversibles. Sur le corps $\mathbb{K}$, cela peut être formulé de façon plus précise : l'ensemble des matrices non inversibles, considéré comme sous-ensemble de $\mathcal{M}_{n,n}(\mathbb{K})$, est un ensemble négligeable, c'est-à-dire de mesure de Lebesgue nulle. Intuitivement, cela signifie que si vous choisissez au hasard une matrice carrée à coefficients réels, la probabilité pour qu'elle soit non inversible est égale à zéro. La raison est que des matrices non inversibles peuvent être considérées comme racines d'une fonction polynôme donnée par le déterminant.

2.2 Cas des matrices triangulaires

Le plus souvent dans les applications, nous aurons affaire à des matrice creuses, c'est-à-dire avec beaucoup de coefficients nuls. En particulier, nous avons

Définition 2.6 *Soit $A \in \mathcal{M}_{n,n}(\mathbb{K})$ une matrice carrée d'ordre n.*
- *A est une matrice triangulaire inférieure si et seulement si $A = (a_{i,j})_{1 \leq i,j \leq n}$ avec*

$$a_{i,j} = 0, \quad 1 \leq i < j \leq n.$$

- *A est une matrice triangulaire supérieure lorsque*

$$a_{i,j} = 0, \quad 1 \leq j < i \leq n.$$

- *A est une matrice diagonale dès lors que $a_{i,j} = 0$ pour $i \neq j$.*

Ces matrices particulières jouent un rôle important en analyse numérique car elles sont facilement inversibles ou du moins, nous pouvons facilement trouver la solution $x \in \mathbb{K}^n$ du système linéaire $A\, x = b$ lorsque A est une matrice triangulaire. En effet, considérons le cas

d'une matrice triangulaire supérieure, alors la solution x se calcule par un algorithme de remontée (nous observons qu'une matrice triangulaire inversible a tous ses éléments diagonaux non nuls)

$$x_n = \frac{b_n}{a_{n,n}}.$$

Puis, pour tout $i = n-1, n-2, ..., 1$

$$x_i = \frac{b_i \; - \; \sum_{j=i+1}^{n} a_{i,j}\, x_j}{a_{i,i}}.$$

Remarque 2.1 *La plupart des méthodes directes qui calculent la solution exacte vont consister à se ramener à cette situation particulière d'une matrice triangulaire.*

Nous rappelons ensuite un résultat qui sera utile pour la suite à propos des matrices triangulaires

Lemme 2.1 *Soit $L \in \mathcal{M}_{n,n}(\mathbb{K})$ une matrice triangulaire inférieure inversible (tous les éléments diagonaux sont non nuls). Alors, L^{-1} est aussi une matrice triangulaire inférieure.*
Soit U une matrice triangulaire supérieure inversible (tous les éléments diagonaux sont non nuls). Alors, U^{-1} est aussi une matrice triangulaire supérieure.

Démonstration. Nous prouvons seulement la première assertion, la deuxième suit la même idée. Soit A, l'inverse de la matrice L, nous avons alors pour tout $(i,j) \in \{1, \ldots, n\}^2$

$$\delta_{i,j} = \sum_{k=1}^{n} a_{i,k}\, l_{k,j} \; = \; \sum_{k=1}^{j} a_{i,k}\, l_{k,j},$$

où $\delta_{i,j}$ est le symbole de Kronecker

$$\delta_{i,j} = \begin{cases} 1 & \text{si } i = j, \\ \\ 0 & \text{sinon.} \end{cases}$$

Ainsi, pour $j = 1$ et $i > 1$, nous avons

$$0 \; = \; a_{i,1}\, l_{1,1} \; \Rightarrow \; a_{i,1} \; = \; 0;$$

pour $j = 2$ et $i > 2$, nous avons

$$0 \; = \; a_{i,1}\, l_{1,2} \; + \; a_{i,2}\, l_{2,2} \; = \; a_{i,2}\, l_{2,2} \; \Rightarrow \; a_{i,2} \; = \; 0.$$

Par récurrence, nous prouvons que pour tout $j > i$, nous avons $a_{i,j} = 0$. En conclusion, la matrice $A = L^{-1}$ est bien triangulaire inférieure. □

2.3 Autres propriétés et résultats

L'ensemble $\mathcal{M}_{n,n}(\mathbb{K})$ peut être considéré comme étant un espace vectoriel muni d'une norme $\|.\|$. Pour cela, il nous faut définir correctement la norme d'une matrice.

Définition 2.7 *Nous appelons norme matricielle toute application $\|.\|$ de $\mathcal{M}_{n,n}(\mathbb{K})$ à valeur dans $\mathbb{R}^+$ qui vérifie les propriétés suivantes*

– *pour toute matrice $A \in \mathcal{M}_{n,n}(\mathbb{K})$,*

$$\|A\| = 0 \Rightarrow A = 0_{\mathbb{K}^{n\times n}},$$

– *pour toute matrice $A \in \mathcal{M}_{n,n}(\mathbb{K})$, et pour tout $\alpha \in \mathbb{K}$,*

$$\|\alpha A\| = |\alpha| \, \|A\|,$$

– *pour toutes matrices $A,\ B \in \mathcal{M}_{n,n}(\mathbb{K})$ (inégalité triangulaire),*

$$\|A + B\| \leq \|A\| + \|B\|,$$

– *pour toutes matrices $A,\ B \in \mathcal{M}_{n,n}(\mathbb{K})$,*

$$\|A\,B\| \leq \|A\| \, . \, \|B\|.$$

Notons bien que pour définir une norme matricielle, cela requiert une condition supplémentaire par rapport à la définition d'une norme vectorielle. Il n'est donc pas évident *a priori* de pouvoir construire une telle application seulement à partir d'une norme vectorielle. Cependant, nous avons la proposition suivante assurant l'existence d'une telle application.

Proposition 2.3 *Sur $\mathcal{M}_{n,n}(\mathbb{K})$ l'ensemble des normes matricielles est non vide.*

Démonstration. Il suffit de montrer que la norme de Froebenus $\|.\|_F$ définie par

$$\|A\|_F = \left(\sum_{i,j=1}^{n} |a_{i,j}|^2 \right)^{1/2}$$

est bien une norme matricielle. Les trois premières propriétés sont connues puisque $\|.\|_F$ est une norme vectorielle de $\mathbb{K}^{n^2}$. Il reste à démontrer que pour toute $A,\ B \in \mathcal{M}_{n,n}(\mathbb{K})$

$$\|A\,B\|_F \leq \|A\|_F \, \|B\|_F.$$

En effet,

$$\|A\,B\|_F^2 = \|C\|_F^2 = \sum_{i,j=1}^{n} c_{i,j}^2 = \sum_{i,j=1}^{n} \left(\sum_{k=1}^{n} a_{i,k}\, b_{k,j} \right)^2 .$$

En utilisant l'inégalité de Cauchy-Schwartz, nous avons

$$\begin{aligned} \|A\,B\|_F^2 &\leq \sum_{i,j=1}^{n} \left(\sum_{k=1}^{n} a_{i,k}^2 \right) \left(\sum_{k=1}^{n} b_{k,j}^2 \right) = \left(\sum_{i,k=1}^{n} a_{i,k}^2 \right) \left(\sum_{k,j=1}^{n} b_{k,j}^2 \right) \\ &= \|A\|_F^2 \, \|B\|_F^2. \end{aligned}$$

□

Voyons maintenant comment nous pouvons définir une norme matricielle à partir d'une norme vectorielle quelconque.

Définition 2.8 *Soit $\|.\|$ une norme vectorielle sur $\mathbb{K}^n$. Nous définissons la norme matricielle $\|.\|$ subordonnée à la norme vectorielle $\|.\|$ comme étant l'application donnée par*

$$A \in \mathcal{M}_{n,n}(\mathbb{K}) \longrightarrow \|A\| := \sup_{v \in \mathbb{K}^n} \frac{\|A\,v\|}{\|v\|}.$$

Nous vérifions facilement que cette application définit bien une norme matricielle.

Pour $1 \leq p \leq \infty$, nous savons que l'application qui à $v \in \mathbb{K}^n$ fait correspondre le réel positif

$$\|v\|_p = \left(\sum_{i=1}^{n} |v_i|^p \right)^{1/p}$$

ou

$$\|v\|_\infty = \max_{1 \leq i \leq n} |v_i|,$$

pour $p = \infty$, est une norme vectorielle sur $\mathbb{K}^n$. Nous posons alors pour $A \in \mathcal{M}_{n,n}(\mathbb{K})$

$$\|A\|_p = \sup_{v \in \mathbb{K}^n} \frac{\|A\,v\|_p}{\|v\|_p},$$

qui est une norme matricielle subordonnée à la norme vectorielle $\|.\|_p$. En général nous ne pouvons pas calculer $\|A\|_p$ directement en fonction de $(a_{i,j})_{1 \leq i,j \leq n}$ sauf pour $p = 1$ et $p = \infty$.

Proposition 2.4 *Soit $A \in \mathcal{M}_{n,n}(\mathbb{K})$ une matrice quelconque. Alors, nous avons*

$$\|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^{n} |a_{i,j}|$$

et

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

Démonstration. Soit $v \in I\!K^n$. D'une part, nous avons

$$\begin{aligned} \|A\,v\|_1 &= \sum_{i=1}^{n} \left| \sum_{j=1}^{n} a_{i,j}\, v_j \right| \leq \sum_{i=1}^{n} \sum_{j=1}^{n} |a_{i,j}|\,|v_j|, \\ &\leq \sum_{j=1}^{n} \left(\sum_{i=1}^{n} |a_{i,j}| \right) |v_j| \leq \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{i,j}| \|v\|_1. \end{aligned}$$

D'où

$$\|A\|_1 \leq \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{i,j}|.$$

D'autre part, nous montrons qu'il existe $w \in I\!K^n$ unitaire $\|w\|_1$ tel que

$$\|A\,w\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{i,j}|.$$

En effet, nous choisissons $j_0 \in \{1, ..., n\}$ tel que

$$\sum_{i=1}^{n} |a_{i,j_0}| = \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{i,j_0}|,$$

puis nous prenons $w \in I\!K^n$ tel que $w_i = 0$ pour $i \neq j_0$ et $w_{j_0} = 1$. Alors,

$$\|A\,w\|_1 = \sum_{i=1}^{n} |(A\,w)_i| = \sum_{i=1}^{n} |a_{i,j_0}| = \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{i,j_0}|.$$

En utilisant une méthode analogue, nous montrons le résultat pour la norme matricielle $\|.\|_\infty$. □

Dans le cas particulier de la norme subordonnée à la norme euclidienne de $I\!K^n$ c'est-à-dire définie par : $A \in \mathcal{M}_{n,n}(I\!K)$

$$\|A\|_2 = \sup_{v\in I\!K^n} \frac{\|A\,v\|_2}{\|v\|_2},$$

nous avons le résultat suivant.

Proposition 2.5 *Pour $A \in \mathcal{M}_{n,n}(I\!K)$, nous avons*

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

Démonstration. Prenons $I\!K = \mathbb{C}$ ou $I\!R$, nous avons alors

$$\begin{aligned} \|A\|_2^2 &= \sup_{v\in I\!K^n} \frac{\|A\,v\|_2^2}{\|v\|_2^2} = \sup_{v\in I\!K^n} \frac{(A\,v)^\star\,(A\,v)}{v^\star\,v}, \\ &= \sup_{v\in I\!K^n} \frac{v^\star\,A^\star\,A\,v}{v^\star\,v}. \end{aligned}$$

Or, nous vérifions que

$$(A^\star\, A)^\star \;=\; A^\star\, (A^\star)^\star \;=\; A^\star\, A.$$

Ainsi, la matrice $A^\star\, A$ est hermitienne, elle est donc diagonalisable : il existe $U \in \mathcal{M}_{n,n}(\mathbb{K})$ unitaire, c'est-à-dire $U^\star\, U = I_n$ telle que

$$U\, A^\star\, A\, U \;=\; diag(\mu_k).$$

Les valeurs $(\mu_k)_{1\leq k\leq n}$ sont les valeurs singulières de la matrice A, c'est-à-dire les valeurs propres de la matrice hermitienne $A^*\, A$. Notons que $\mu_k \geq 0$ puisque de la relation $A^*\, Ap_k = \mu_k\, p_k$, nous déduisons que

$$(A\, p_k)^*\, Ap_k = \mu_k\, p_k^*\, p_k,$$

c'est-à-dire $\mu_k = \|A\, p_k\|_2^2/\|p_k\|_2^2 \;\geq\; 0$. Il vient ensuite

$$\sup_{v\in\mathbb{K}^n} \frac{v^\star\, A^\star\, A\, v}{v^\star\, v} \;=\; \sup_{v\in\mathbb{K}^n} \frac{v^\star\, U^\star\, U\, A^\star\, A\, U^\star\, U\, v}{v^\star\, U^\star\, U\, v}.$$

Puis, par changement de variable $w = U\, v$, nous avons

$$\|A\|_2^2 \;=\; \sup_{w\in\mathbb{K}^n} \frac{w^\star\, U\, A^\star\, A\, U^\star\, w}{w^\star\, w} \;=\; \sup_{w\in\mathbb{K}^n} \frac{\sum_{k=1}^n \mu_k |w_k|^2}{\sum_{k=1}^n |w_k|^2},$$

où les μ_k sont les valeurs propres de $A^\star\, A$ et en prenant le vecteur de la base cannonique ne comprenant que des 0 excepté la k-ème composante qui correspond à la plus grande valeur propre μ_k en module, nous obtenons $\|A\|_2^2 = \rho(A\, A^\star)$. □

2.4 Conditionnement de matrices

Considérons la matrice de Hilbert donnée par

$$a_{i,j} \;=\; \frac{1}{i+j-1}, \quad i,j = 1,\ldots,n.$$

Pour $n = 4$, nous avons par exemple

$$\begin{pmatrix} 1 & 1/2 & 1/3 & 1/4 \\ 1/2 & 1/3 & 1/4 & 1/5 \\ 1/3 & 1/4 & 1/5 & 1/6 \\ 1/4 & 1/5 & 1/6 & 1/7 \end{pmatrix}.$$

Pour un vecteur donné b tel que

$$b^T \;=\; \left(\frac{25}{12}, \frac{77}{60}, \frac{57}{60}, \frac{319}{420}\right) \;\simeq\; (2.0833, 1.2833, 0.9500, 0.7599)\,.$$

Nous calculons alors $x = A^{-1}\,b$ et vérifions facilement que

$$x^T \;=\; (1, 1, 1, 1)\,.$$

Ensuite, si nous changeons légèrement la source $\tilde{b}$

$$\tilde{b}^T \;=\; (2.1, 1.3, 1.0, 0.8)\,,$$

la solution $\tilde{x}$ du système $\tilde{x} = A^{-1}\tilde{b}$ est donnée par

$$\tilde{x}^T \;=\; (5.6, -48, 114, -70)\,,$$

ce qui est très loin de la solution x.

Essayons de comprendre ce phénomène de manière plus globale : nous constatons qu'une petite perturbation de la donnée b induit une grande modification de la solution x. Plus généralement, soient A une matrice réelle d'ordre n inversible et b un vecteur de $\mathbb{K}^n$; nous notons par x la solution du système $A\,x = b$ donnée par $x = A^{-1}\,b$. Pour une perturbation δb du vecteur b nous notons $x + \delta x$ la solution de

$$A\,(x \,+\, \delta x) \;=\; b \,+\, \delta b.$$

Pour une norme vectorielle $\|.\|$ de $\mathbb{K}^n$, nous cherchons à contrôler l'erreur relative $\|\delta u\| \,/\, \|u\|$ en fonction de la norme de l'erreur relative $\|\delta b\| \,/\, \|b\|$ et de la norme matricielle subordonnée $\|A\|$.

Par linéarité de A, nous avons d'une part puisque A est inversible

$$A\,\delta x \;=\; \delta b \quad \Rightarrow \|\delta x\| \;\leq\; \|A^{-1}\|\,\|\delta b\|$$

et d'autre part

$$A\,x \;=\; b \quad \Rightarrow \|b\| \;\leq\; \|A\|\,\|x\|,$$

ou encore de manière équivalente

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

Ainsi, nous obtenons l'estimation suivante

$$\frac{\|\delta x\|}{\|x\|} \;\leq\; \|A^{-1}\|\,\|A\|\,\frac{\|\delta b\|}{\|b\|}.$$

et proposons la définition suivante

Définition 2.9 *Nous appelons conditionnement de la matrice A relativement à la norme matricielle $\|.\|$ subordonnée à la norme vectorielle $\|.\|$, le nombre*

$$cond(A) = \|A\| \, \|A^{-1}\|.$$

Vérifions sur l'exemple précédent la véracité de la définition du conditionnement, nous avons pour la norme $\|.\|_1$,

$$cond_1(A) \; = \; \|A\|_1 \, \|A^{-1}\|_1 \; \simeq \; 28\,375 \; > \; 1.$$

ou pour la norme $\|.\|_2$

$$cond_2(A) \; = \; \|A\|_2 \, \|A^{-1}\|_2 \; \simeq \; 15\,514 \; > \; 1.$$

Les valeurs du conditionnement de la matrice A semble être assez élevées indépendamment de la norme choisie.

Nous concluons cette partie en donnant quelques propriétés sur le conditionnement d'une matrice A

Proposition 2.6 *Soient $A \in \mathcal{M}_{n,n}(\mathbb{K})$ une matrice carrée inversible et $\|.\|$ une norme matricielle subordonnée à la norme vectorielle $\|.\|$. Alors, nous avons*

(i) $cond(A^{-1}) = cond(A)$,
(ii) soit $\alpha \in \mathbb{K}$, $cond(\alpha \, A) = cond(A)$,
(iii) $cond(I_n) = 1$,
(iv) $cond(A) \geq 1$.

Démonstration. Les propositions (i) et (ii) son évidentes. Pour (iii) il suffit de remarquer que pour une norme subordonnée $\|I_n\| = 1$ et donc

$$cond(I_n) \; = \; \|I_n\|^2 \; = \; 1.$$

Pour (iv) nous avons d'une part en utilisant (iii),

$$cond(A \, A^{-1}) \; = \; cond(I_n) \; = \; 1.$$

D'autre part,

$$cond(A \, A^{-1}) \; = \; \|A \, A^{-1}\| \, \|(A \, A^{-1})^{-1}\| \; \leq \; (\|A\| \, \|A^{-1}\|)^2 \; = \; cond(A)^2.$$

Ainsi, nous obtenons le résultat

$$cond(A) \; \geq \; 1.$$

□

Nous voyons bien que le conditionnement sert à mesurer la sensibilité du système aux perturbations de b et de A.

Définition 2.10 *Soit $A \in \mathcal{M}_{n,n}(\mathbb{K})$ une matrice inversible. Nous dirons que*
(i) un système linéaire est bien conditionné, si $cond(A)$ n'est pas trop grand par rapport à un (qui est le conditionnement de l'identité) ;
(ii) un système linéaire est mal conditionné, si $cond(A)$ est grand par rapport à un.

L'inconvénient de la définition du conditionnement est qu'il fait apparaître $\|A^{-1}\|$ qui n'est pas facile à calculer d'autant plus que nous ne connaissons pas la forme explicite de la matrice A^{-1}. Dans le cas particulier d'une matrice normale et pour la norme matricielle $\|.\|_2$ nous avons néanmoins le résultat suivant.

Proposition 2.7 *Soit A une matrice hermitienne ($A^\star = A$). Alors, nous avons*

$$\|A\|_2 \;=\; \rho(A).$$

De plus, si A est une matrice hermitienne inversible. Alors, nous avons

$$cond_2(A) \;=\; \frac{\max\{|\lambda_i|, i = 1, .., n\}}{\min\{|\lambda_i|, i = 1, .., n\}}.$$

Démonstration. Puisque A est hermitienne, nous pouvons appliquer le Corollaire 2.1 et montrons qu'il existe une matrice U telle que $U\,U^\star = I_n$ et

$$U^\star\,A\,U \;=\; diag(\lambda_i),$$

où $(\lambda_i)_i$ sont les valeurs propres de A, nous avons

$$\|A\|_2^2 \;=\; \sup_{v \in \mathbb{K}^n} \frac{\|A\,v\|_2^2}{\|v\|_2^2} \;=\; \sup_{v \in \mathbb{K}^n} \frac{(A\,v)^\star\,(A\,v)}{v^\star\,v}.$$

Or,

$$\begin{aligned} v^\star\,A^\star\,A\,v &= v^\star\,U\,U^\star\,A^\star\,U\,U^\star\,A\,U\,U^\star\,v, \\ &= (U^\star\,v)^\star\,(diag(\lambda_i))^\star\,(diag(\lambda_i))\,U^\star\,v \end{aligned}$$

et $U^\star\,A\,U \;=\; (U^\star\,A\,U)^\star$. Nous posons alors $w \;=\; U^\star\,v$ et obtenons

$$v^\star\,A^\star\,A\,v \;=\; w^\star diag(\overline{\lambda}_i)\,diag(\lambda_i)\,w.$$

Ainsi, nous avons montré le premier résultat

$$\|A\|_2^2 \;=\; \sup_{w \in \mathbb{K}^n} \frac{w^\star diag(|\lambda_i|^2)\,w}{w^\star\,w}$$

D'une part, nous avons pour tout $w \in \mathbb{K}^n$

$$\frac{w^\star diag(|\lambda_i|^2)\,w}{w^\star\,w} \;\leq\; \rho(A)^2.$$

D'autre part, en choisissant le vecteur de la base canonique $w_{j_0} = e_{j_0}$ où $j_0 \in \{1, \ldots, n\}$ tel que $|\lambda_{j_0}| = \rho(A)$, nous avons

$$|\lambda_{j_0}|^2 = \frac{e_{j_0}^\star \bar{\lambda}_{j_0} \lambda_{j_0} e_{j_0}}{e_{j_0}^\star e_{j_0}} \leq \sup_{w \in I\!K^n} \frac{w^\star diag(|\lambda_i|^2)\, w}{w^\star\, w}$$

et donc en regroupant les deux denières inégalités

$$\|A\|_2^2 = \rho(A)^2.$$

Nous supposons maintenant que A est une matrice hermitienne inversible. Pour calculer $cond_2(A)$, il nous suffit de remarquer que $\lambda_i \neq 0$ et $1/\lambda_i$ est valeur propre de A^{-1} et donc en appliquant le résultat précédent à A^{-1} (qui est également une matrice hermitienne), nous avons

$$\|A^{-1}\|_2 = \rho(A^{-1}) = \frac{1}{\min\{|\lambda_i|, i = 1, ..., n\}},$$

d'où nous déduisons

$$cond_2(A) = \|A\|_2 \, \|A^{-1}\|_2 = \frac{\max\{|\lambda_i|, i = 1, .., n\}}{\min\{|\lambda_i|, i = 1, .., n\}}.$$

□

Préconditionemment d'un système linéaire

Pour remedier au problème du mauvais conditionnement d'une matrice, nous pouvons appliquer une méthode de préconditionnement. En effet, en vue de résoudre $Ax = b$ nous multiplions ce système d'équation à gauche par une matrice inversible P

$$P\,A\,x = P\,b,$$

avec P choisie pour que $P\,A$ soit bien conditionnée (dans le cas le plus favorable, nous aurions $P = A^{-1}$). Cependant, il n'y a pas de méthode standard pour trouver la matrice P, le plus souvent nous chercherons une matrice à la fois facile à inverser et "assez proche" de A^{-1}.

La résolution d'un système linéaire algébrique est au cœur de la plupart des calculs en analyse numérique. Ici, nous décrivons les algorithmes de résolution les plus populaires qui sont appliqués à des systèmes généraux. Nous considérons le système algébrique réel

$$A\,x = b, \tag{2.3}$$

où A est une matrice non singulière de dimension n. Nous distinguerons deux types de méthodes. D'une part, *les méthodes directes* où nous calculons exactement la solution $x = A^{-1}\,b$ (en évitant bien sûr de calculer A^{-1}). D'autre part *les méthodes itératives* où nous calculons une solution approchée.

3 Méthodes directes

Des méthodes directes permettent de calculer la solution exacte du problème (2.3) en un nombre fini d'étapes (en l'absence d'erreur d'arrondi). La méthode directe la plus classique est la Méthode d'Élimination de Gauss (GEM en anglais), qui consiste à décomposer A comme le produit LU où L et U sont respectivement une matrice triangulaire inférieure et une matrice triangulaire supérieure.

Cette méthode fut nommée d'après le mathématicien C. F. Gauss, mais il semble qu'elle fût déjà connue des Chinois depuis au moins le premier siècle de notre ère. Elle est référencée dans l'important livre chinois "Les neuf chapitres sur l'art du calcul" (rédigé sous la dynastie Han), dont elle constitue le huitième chapitre, sous le titre de la "disposition rectangulaire". La méthode est présentée au moyen de dix-huit exercices [2]. C. F. Gauss a pour la première fois formalisé la méthode que nous allons voir en toute généralité, dans un mémoire de 1810 sur l'orbite de l'astéroïde nommé Pallas. Notons cependant que plus tard, il se trouva confronté à un problème de triangulation de la région de Hanovre, qui impliquait 26 triangles ; même Gauss, qui pourtant excellait dans l'art du calcul, ne pouvait pas résoudre à la main des systèmes de plusieurs dizaines d'équations. Il mit alors au point un algorithme de calcul approché, qui fut redécouvert par P. L. von Seidel. L'algorithme de Gauss-Seidel est utilisé de nos jours pour de très grands systèmes. Nous présenterons cet algorithme itératif dans la dernière partie du chapitre.

3.1 Méthode de Gauss avec et sans pivot

L'élimination de Gauss ou l'élimination de Gauss-Jordan est un algorithme de l'algèbre linéaire pour déterminer les solutions d'un système d'équations linéaires, pour déterminer le rang d'une matrice ou pour calculer l'inverse d'une matrice carrée inversible.

Méthode de Gauss sans pivot.

Commençons par donner un exemple de la méthode de Gauss qui consiste à remplacer le système initial par un système tri-diagonal. Nous considérons le système d'équations suivant :

$$\begin{cases} x_1 + 2\,x_2 + 2\,x_3 = 2, \\ x_1 + 3\,x_2 + 3\,x_3 = 2, \\ 3\,x_1 + 7\,x_2 + 8\,x_3 = 8. \end{cases}$$

Nous établissons la matrice correspondante

$$A^{(1)} = \begin{pmatrix} 1 & 2 & 2 \\ 1 & 3 & 3 \\ 3 & 7 & 8 \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, \quad b^{(1)} = \begin{pmatrix} 2 \\ 2 \\ 8 \end{pmatrix}$$

et nous appliquons la première étape de Gauss. Ainsi, nous ajoutons un multiple de la première ligne aux deux autres lignes pour obtenir des zéros : nous faisons la différence entre la deuxième ligne et la première, puis la différence entre la troisième et trois fois la première, il vient alors

$$\begin{cases} x_1 + 2\,x_2 + 2\,x_3 = 2, \\ \quad + x_2 + x_3 = 0, \\ \quad + x_2 + 2\,x_3 = 2, \end{cases}$$

ou encore

$$A^{(2)} = \begin{pmatrix} 1 & 2 & 2 \\ 0 & 1 & 1 \\ 0 & 1 & 2 \end{pmatrix}, \quad b^{(2)} = \begin{pmatrix} 2 \\ 0 \\ 2 \end{pmatrix}.$$

Ensuite, la deuxième ligne est multipliée par -1 et nous ajoutons le résultat à la troisième :

$$\begin{cases} x_1 + 2x_2 + 2x_3 = 2, \\ \quad x_2 + x_3 = 0, \\ \quad + x_3 = 2, \end{cases}$$

ou encore

$$A^{(3)} = \begin{pmatrix} 1 & 2 & 2 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \quad b^{(3)} = \begin{pmatrix} 2 \\ 0 \\ 2 \end{pmatrix}.$$

Nous pouvons alors résoudre ce système de mainère exacte et la solution du système est donnée par

$$x = \begin{pmatrix} 2 \\ -2 \\ 2 \end{pmatrix}.$$

Le principe de la méthode consiste donc à se ramener par des opérations simples (combinaisons linéaires) à un système triangulaire équivalent qui sera donc facile à résoudre. Avant de décrire précisément l'algorithme, nous introduisons une forme de matrice bien particulière et donnons quelques unes de ses propriétés utiles pour la suite.

Lemme 3.1 *Soit $B^{(k)} \in \mathcal{M}_{n,n}(\mathbb{K})$ une matrice de la forme*

$$B^{(k)} := \begin{pmatrix} 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & b_{k+1}^{(k)} & 0 & \dots & 0 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \dots & 0 & b_n^{(k)} & 0 & \dots & 0 \end{pmatrix}.$$

Alors, nous avons

(i) $B^{(k)}\, B^{(l)} = 0$, pour tout $1 \leq k \leq l \leq n$,

(ii) pour $L^{(k)} = I_n - B^{(k)}$, nous avons $L^{(k)}$ est inversible et $[L^{(k)}]^{-1} = I_n + B^{(k)}$,

(iii) $L^{(k)}\, L^{(l)} = I_n - (B^{(k)} + B^{(l)})$, pour tout $1 \leq k \leq l \leq n$.

Démonstration. Soient $B^{(k)}$ et $B^{(l)} \in \mathcal{M}_{n,n}(\mathbb{K})$ données par la représentation ci-dessus, nous avons pour $1 \leq i, j \leq n$,

$$\left(B^{(k)}\, B^{(l)}\right)_{i,j} = \sum_{m=1}^{n} \left(B^{(k)}\right)_{i,m} \left(B^{(l)}\right)_{m,j}.$$

Or, pour $m \neq k$, nous savons que $\left(B^{(k)}\right)_{i,m} = 0$ et donc

$$\left(B^{(k)}\, B^{(l)}\right)_{i,j} = \left(B^{(k)}\right)_{i,k} \left(B^{(l)}\right)_{k,j}.$$

Ensuite, lorsque $j \neq l$, nous avons $\left(B^{(l)}\right)_{k,j} = 0$ et donc le produit $\left(B^{(k)}\, B^{(l)}\right)_{i,j}$ est nul.

D'autre part, lorsque $j = l$, nous avons $\left(B^{(k)}\, B^{(l)}\right)_{i,l} = \left(B^{(k)}\right)_{i,k} \left(B^{(l)}\right)_{k,l}$. Puisque $k \leq l$, le coefficient $\left(B^{(l)}\right)_{k,l}$ est également nul et donc le produit $\left(B^{(k)}\, B^{(l)}\right)_{i,l}$ est aussi égal à zéro.

Nous obtenons donc le résultat

$$B^{(k)}\, B^{(l)} = 0, \quad 1 \leq k \leq l \leq n.$$

Le reste de la preuve devient alors évident. D'une part nous vérifions que

$$(I_n - B^{(k)})\,(I_n + B^{(k)}) \;=\; I_n - [B^{(k)}]^2 \;=\; I_n$$

et donc $\left(L^{(k)}\right)^{-1} = (I_n - B^{(k)})^{-1} = (I_n + B^{(k)})$. Ensuite, nous vérifions que pour $1 \leq k \leq l \leq n$

$$L^{(k)}\, L^{(l)} \;=\; I_n - (B^{(k)} + B^{(l)}).$$

□

Voyons maintenant comment étendre la méthode de Gauss-Jordan pour un système quelconque. Nous nous intéressons au système linéaire

$$A\, x \;=\; b,$$

où A est une matrice carrée inversible de taille $n \times n$ et $b \in I\!K^n$. Pour commencer, posons $A^{(1)} = A$ et $b^{(1)} = b$. Ainsi, à partir de ce changement de notation le système s'écrit

$$\left\{\begin{array}{lcl}
a^{(1)}_{1,1}\, x_1 \;+\; a^{(1)}_{1,2}\, x_2 + \quad \ldots \quad + a^{(1)}_{1,n}\, x_n & = & b^{(1)}_1, \\
a^{(1)}_{2,1}\, x_1 \;+\; a^{(1)}_{2,2}\, x_2 + \quad \ldots \quad + a^{(1)}_{2,n}\, x_n & = & b^{(1)}_2, \\
\vdots & & \vdots \\
a^{(1)}_{n,1}\, x_1 \;+\; a^{(1)}_{n,2}\, x_2 + \quad \ldots \quad + a^{(1)}_{n,n}\, x_n & = & b^{(1)}_n.
\end{array}\right.$$

Nous supposons que $a^{(1)}_{1,1} \neq 0$, et nous appelons alors ce coefficient le premier *pivot*. Ensuite, pour $i = 2, .., n$ nous remplaçons simplement la ligne i par une combinaison linéaire des lignes 1 et i de manière à faire apparaître des zéros sur la première colonne. Pour cela pour $i = 2, .., n$ nous posons

$$\alpha^{(1)}_i = \frac{a^{(1)}_{i,1}}{a^{(1)}_{1,1}}$$

et nous appliquons la transformation

$$l_i \longrightarrow l_i \;-\; \alpha^{(1)}_i\, l_1,$$

ou encore

$$\left\{\begin{array}{lll} a_{1,1}^{(1)}\,x_1 & +a_{1,2}^{(1)}\,x_2+ \quad \ldots \quad +a_{1,n}^{(1)}\,x_n & = \; b_1^{(1)}, \\ 0\,x_1 & +(a_{2,2}^{(1)}-\alpha_2^{(1)}\,a_{1,2}^{(1)})\,x_2+ \quad \ldots \quad +(a_{2,n}^{(1)}-\alpha_2^{(1)}\,a_{1,n}^{(1)})\,x_n & = \; b_2^{(1)}-\alpha_2^{(1)}\,b_1^{(1)}, \\ \vdots & & \\ 0\,x_1 & +(a_{n,2}^{(1)}-\alpha_n^{(1)}\,a_{1,2}^{(1)})\,x_2+ \quad \ldots \quad +(a_{n,n}^{(1)}-\alpha_n^{(1)}\,a_{1,n}^{(1)})\,x_n & = \; b_n^{(1)}-\alpha_n^{(1)}\,b_1^{(1)}. \end{array}\right.$$

Nous obtenons alors un système de la forme $A^{(2)}\,x \,=\, b^{(2)}$, c'est-à-dire

$$\left\{\begin{array}{lll} a_{1,1}^{(1)}\,x_1+ & a_{1,2}^{(1)}\,x_2+ \quad \ldots \quad +a_{1,n}^{(1)}\,x_n & = \; b_1^{(1)}, \\ & a_{2,2}^{(2)}\,x_2+ \quad \ldots \quad +a_{2,n}^{(2)}\,x_n & = \; b_2^{(2)}, \\ & \vdots & \\ & a_{n,2}^{(2)}\,x_2+ \quad \ldots \quad +a_{n,n}^{(2)}\,x_n & = \; b_n^{(2)}, \end{array}\right.$$

avec

$$\left\{\begin{array}{l} \alpha_i^{(1)}=\dfrac{a_{i,1}^{(1)}}{a_{1,1}^{(1)}}, \quad i=2,..,n; \\ a_{i,j}^{(2)}=a_{i,j}^{(1)}-\alpha_i^{(1)}\,a_{1,j}^{(1)}, \quad i=2,..,n;\ j=2,..,n; \\ b_i^{(2)}=b_i^{(1)}-\alpha_i^{(1)}\,b_1^{(1)}, \quad i=2,..,n, \end{array}\right.$$

tandis que la première ligne est inchangée, c'est-à-dire $a_{1,j}^{(2)}=a_{1,j}^{(1)}$ pour tout $1\leq j\leq n$. Ce qui d'un point de vue matriciel revient à faire le produit suivant

$$A^{(2)}=L^{(1)}\,A^{(1)}, \quad b^{(2)}=L^{(1)}\,b^{(1)},$$

où $L^{(1)}=I_n-B^{(1)}$ et $B^{(1)}$ est donnée par

$$B^{(1)}:=\begin{pmatrix} 0 & 0 & \ldots & \ldots & 0 \\ \alpha_2^{(1)} & 0 & & & \vdots \\ \vdots & \vdots & & & \vdots \\ \alpha_n^{(1)} & 0 & \ldots & \ldots & 0 \end{pmatrix}.$$

Nous vérifions aussi le nombre d'opérations de cette première étape. La transformation de $A^{(1)}x = b^{(1)}$ en le système $A^{(2)}\,x = b^{(2)}$ nécessite

- $\star$ $(n-1)$ divisions,
- $\star$ $n\,(n-1) = (n-1)^2 + (n-1)$ multiplications,
- $\star$ $n\,(n-1) = (n-1)^2 + (n-1)$ additions.

Puis, nous raisonnons par récurrence. Supposons qu'à l'étape $(k-1)$, nous avons construit le système $A^{(k)}\,x \,=\, b^{(k)}$ qui s'écrit sous la forme

$$\left\{\begin{array}{lllllll} a_{1,1}^{(1)}\,x_1 + & a_{1,2}^{(1)}\,x_2 + & \ldots & \ldots & + a_{1,n}^{(1)}\,x_n & = & b_1^{(1)}, \\ & a_{2,2}^{(2)}\,x_2 + & \ldots & \ldots & + a_{2,n}^{(2)}\,x_n & = & b_2^{(2)}, \\ & & \ddots & \vdots & \vdots & \vdots & \\ & & a_{k,k}^{(k)}\,x_k + & \ldots & + a_{n,n}^{(k)}\,x_n & = & b_k^{(k)}, \\ & & a_{n,k}^{(k)}\,x_k + & \ldots & + a_{n,n}^{(k)}\,x_n & = & b_n^{(k)} \end{array}\right. .$$

Nous supposons que $a_{k,k}^{(k)} \neq 0$, ce coefficient est aussi appelé un pivot. Alors, les k premières lignes restent inchangées tandis que pour $i \geq k+1$,

$$\left\{\begin{array}{l} \alpha_i^{(k)} = \dfrac{a_{i,k}^{(k)}}{a_{k,k}^{(k)}}, \quad i \,=\, k+1,..,n; \\ a_{i,j}^{(k+1)} = a_{i,j}^{(k)} \,-\, \alpha_i^{(k)}\,a_{k,j}^{(k)}, \quad i \,=\, k+1,..,n; \quad j = k+1,..,n; \\ b_i^{(k+1)} = b_i^{(k)} \,-\, \alpha_i^{(k)}\,b_k^{(k)}, \quad i = k+1,..,n, \end{array}\right.$$

et comme précédemment les k premières lignes restent inchangées.

Encore une fois, d'un point de vue matriciel cela revient à faire le produit suivant

$$A^{(k+1)} = L^{(k)}\,A^{(k)}, \quad b^{(k+1)} = L^{(k)}\,b^{(k)},$$

où $L^{(k)} = I_n - B^{(k)}$ et $B^{(k)}$ est donnée par

$$B^{(k)} := \begin{pmatrix} 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & \alpha_{k+1}^{(k)} & 0 & \dots & 0 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \dots & 0 & \alpha_{n}^{(k)} & 0 & \dots & 0 \end{pmatrix}.$$

De plus, l'étape qui mène du système $A^{(k)}\, x = b^{(k)}$ au nouveau système $A^{(k+1)}\, x = b^{(k+1)}$ nécessite

- $\star$ $(n-k)$ divisions,
- $\star$ $(n-k+1)\,(n-k) = (n-k)^2 + (n-k)$ multiplications,
- $\star$ $(n-k+1)\,(n-k) = (n-k)^2 + (n-k)$ additions.

En itérant ce processus jusqu'à la $(n-1)$-ème étape, nous obtenons finalement un système de la forme $A^{(n)}\, x = b^{(n)}$ c'est-à-dire

$$\begin{cases} a_{1,1}^{(1)}\, x_1 + & a_{1,2}^{(1)}\, x_2 + & \dots & + a_{1,n}^{(1)}\, x_n & = & b_1^{(1)}, \\ & a_{2,2}^{(2)}\, x_2 + & \dots & + a_{2,n}^{(2)}\, x_n & = & b_2^{(2)}, \\ & & \ddots & \vdots & \vdots & \vdots \\ & & & a_{n,n}^{(n)}\, x_n & = & b_n^{(n)}. \end{cases}$$

Nous remarquons que $A^{(n)}$ est une matrice triangulaire supérieure. Pour conclure la méthode d'élimination de Gauss, il reste à résoudre le système $A^{(n)}\, x = b^{(n)}$: nous commençons par la dernière ligne

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

Ensuite, connaissant x_n, nous calculons x_{n-1} en utilisant la $(n-1)$-ème ligne

$$x_{n-1} = \frac{1}{a_{n-1,n-1}^{(n-1)}} \left(b_{n-1}^{(n-1)} - a_{n-1,n}^{(n-1)} x_n \right).$$

Puis, nous continuons

$$x_k = \frac{1}{a_{k,k}^{(k)}} \left(b_k^{(k)} - \sum_{i=k}^{n} a_{k,i}^{(k)} x_i \right), \quad k = n-2, .., 1.$$

Pour résoudre le système triangulaire, nous avons recours à n divisions, $\sum_{k=1}^{n} k = n(n+1)/2$ multiplications et additions.

Au total,

$\star$ le nombre de divisions est de

$$n.\,d. \;=\; (n-1) + (n-2) + \ldots + 1 \;=\; \frac{n(n-1)}{2},$$

$\star$ le nombre de multiplications est de

$$\begin{aligned} n.\,m. &= (n-1)^2 + (n-2)^2 + \ldots + 1^2 + (n-1) + (n-2) + \ldots + 1 \;+\; \frac{n(n+1)}{2} \\ &= \frac{n(n-1)(2n-1)}{6} + \frac{n(n-1)}{2} = \frac{n^3 - n}{3}, \end{aligned}$$

$\star$ le nombre d'additions est de

$$\begin{aligned} n.\,a. &= (n-1)^2 + (n-2)^2 + \ldots + 1^2 + (n-1) + (n-2) + \ldots + 1 \;+\; \frac{n(n+1)}{2} \\ &= \frac{n(n-1)(2n-1)}{6} + \frac{n(n-1)}{2} = \frac{n^3 - n}{3}. \end{aligned}$$

Nous concluons que la méthode est en $O\left(2n^3/3\right)$.

Pour mieux comprendre et analyser la méthode d'élimination de Gauss, nous pouvons réécrire l'algorithme de manière plus abstraite en utilisant seulement des produits de matrices. En effet, la méthode d'élimination de Gauss présentée plus haut consiste en fait à décomposer la matrice A en le produit d'une matrice triangulaire inférieure et une matrice triangulaire supérieure. C'est la décomposition LU, où L est une matrice triangulaire inférieure (comme "Low", bas) et U une matrice triangulaire supérieure (comme "Up", haut). En effet, nous avons vu que

$$A^{(n)} = L^{(n-1)}\, A^{(n-1)} = L^{(n-1)} \,\ldots\, L^{(1)}\, A$$

et donc en appliquant le résultat du Lemme 3.1[(ii)]

$$\begin{aligned} A &= \left(L^{(n-1)} \,\ldots\, L^{(1)}\right)^{-1} A^{(n)} \\ &= \left(L^{(1)}\right)^{-1} \ldots \left(L^{(n-1)}\right)^{-1} A^{(n)} \\ &= \left(I_n + B^{(1)}\right) \ldots \left(I_n + B^{(n-1)}\right) A^{(n)}. \end{aligned}$$

Puis, en appliquant une propriété similaire à celle du Lemme 3.1[(iii)] (en changeant $B^{(k)}$ en $-B^{(k)}$), nous avons

$$\left(I_n + B^{(1)}\right) \ldots \left(I_n + B^{(n-1)}\right) \;=\; \left(I_n + B^{(1)} \;+\; \ldots \;+\; B^{(n-1)}\right)$$

et obtenons finalement

$$A = \left(I_n + B^{(1)} \;+\; \ldots \;+\; B^{(n-1)}\right)\, A^{(n)},$$

où la matrice $L = \left(I_n + B^{(1)} + \ldots + B^{(n-1)}\right)$ est une matrice triangulaire inférieure tandis que $U = A^{(n)}$ est une matrice triangulaire supérieure. Cette décomposition est utilisée en analyse numérique pour résoudre des systèmes d'équations linéaires. Nous démontrons le résultat suivant qui détermine un domaine d'application de la méthode d'élimination de Gauss sans pivot.

Théorème 3.1 *Soit $A \in \mathcal{M}_{n,n}(\mathbb{K})$ une matrice inversible. Nous supposons que toutes les sous-matrices principales d'ordre 1 à $n-1$ de A sont inversibles. Alors, il existe une unique matrice L triangulaire inférieure à diagonale unité et une matrice U triangulaire supérieure inversible telles que*

$$A \;=\; L\,U.$$

Démonstration. Établissons d'abord le résultat d'unicité. Soient (L_α, U_α) et (L_β, U_β) telles que

$$A = L_\alpha\,U_\alpha, \quad \text{et} \quad A = L_\beta\,U_\beta,$$

où L_α et L_β sont des matrices triangulaires inférieures à diagonale unité et donc inversibles. En revanche, U_α et U_β sont des matrices triangulaires supérieures inversibles. Nous pouvons alors écrire

$$L_\beta^{-1}\,L_\alpha = U_\beta\,U_\alpha^{-1}.$$

Or, $L_\beta^{-1}\,L_\alpha$ est une matrice triangulaire inférieure à diagonale unité et $U_\beta\,U_\alpha^{-1}$ est une matrice triangulaire supérieure ; elle est donc diagonale et $L_\beta^{-1}\,L_\alpha = I_n$ et $U_\beta\,U_\alpha^{-1} = I_n$. Ainsi,

$$L_\alpha \;=\; L_\beta, \quad U_\alpha \;=\; U_\beta,$$

ce qui prouve que la décomposition est unique.

Passons maintenant au résultat d'existence. Nous avons vu que pour pouvoir appliquer la méthode d'élimination de Gauss sans pivot, il suffit qu'à chaque étape le coefficient $a_{i,i}^{(i)}$ soit non nul. Prouvons par récurrence que l'hypothèse que A a toutes ses sous-matrices principales inversibles conduit au fait que $a_{i,i}^{(i)}$ est différent de zéro.

Tout d'abord, puisque la première sous-matrice est de déterminant non nul, nous avons $a_{1,1} \neq 0$ et donc le premier pivot $a_{1,1}^{(1)} = a_{1,1}$ est bien non nul ; nous pouvons donc effectuer la première étape de l'élimination de Gauss.

Nous supposons alors que pour tout $k \in \{1, ..., i-1\}$ le coefficient $a_{k,k}^{(k)}$ est différent de zéro, dans ce cas en suivant la méthode d'élimination de Gauss, nous avons

$$A \;=\; A^{(1)} \;=\; \left(L^{(i-1)} \ldots L^{(1)}\right)^{-1}\,A^{(i)}.$$

Développons alors le produit par blocs, nous obtenons

$$\begin{pmatrix} A_i & \times \\ \\ \times & \times \end{pmatrix} \;=\; \begin{pmatrix} L_i & 0 \\ \\ \times & \times \end{pmatrix} \;\begin{pmatrix} A_i^{(i)} & \times \\ \\ \times & \times \end{pmatrix},$$

où A_i (respectivement $A_i^{(i)}$) est la sous-matrice principale de A (respectivement $A^{(i)}$) tandis que $L_i \in \mathcal{M}_{i,i}(\mathbb{K})$ est une matrice tridiagonale inférieure avec uniquement des 1 sur sa diagonale. Puisque $A_i^{(i)}$ est une matrice triangulaire supérieure, nous avons

$$det\,(A_i) \;=\; det\left(L_i\,A_i^{(i)}\right) \;=\; det\,(L_i) \;\times\; det\left(A_i^{(i)}\right) \;=\; 1 \;\times\; a_{1,1}^{(1)} \;\ldots\; a_{i,i}^{(i)}.$$

Or, puisque toutes les sous-matrices principales sont inversibles $det(A_i)$ est différent de zéro et donc le produit $\prod_{k=1}^{i} a_{k,k}^{(k)}$ est aussi non nul et donc en particulier $a_{i,i}^{(i)}$. Nous pouvons donc effectuer une nouvelle étape de la méthode d'élimination de Gauss.

Au final, nous obtenons la décomposition

$$A = (L^{(1)})^{-1}\,L^{(1)}A^{(1)} = (L^{(1)})^{-1}\,A^{(2)} = ... = (I_n \;+\; B^{(1)} \;...\; +\; B^{(n-1)})\,A^{(n-1)}.$$

Notons U la matrice triangulaire supérieure $A^{(n-1)}$ et $L = (L^{(1)})^{-1} \,...\, (L^{(n-1)})^{-1}$. Nous obtenons bien $A = LU$. □

À l'aide de ce dernier résultat, nous pouvons démontrer le corollaire suivant qui est particulièrement important pour les applications.

Corollaire 3.1 *Soit $A \in \mathcal{M}_{n,n}(\mathbb{K})$ une matrice hermitienne définie positive. Alors, A admet une décomposition $L\,U$.*

Démonstration. Il suffit de vérifier que les sous-matrices principales sont inversibles. En effet, les sous-matrices principales sont symétriques définies positives et donc inversibles. Par application du Théorème 3.1, A admet une décomposition $L\,U$. □

La décomposition LU fournit donc un algorithme exact pour résoudre un système linéaire.

Remarque 3.1 *Étant donnée l'équation algébrique*

$$A\,x \;=\; b;$$

nous avons vu que sous certaines hypothèses, nous pouvons écrire ce système sous la forme :

$$\begin{cases} \textit{Trouver } y \in \mathbb{K}^n \textit{ tel que } L\,y \;=\; b. \\ \\ \textit{Trouver } x \in \mathbb{K}^n \textit{ tel que } U\,x \;=\; y. \end{cases}$$

Lorsque nous voulons inverser A pour un b donné. Les matrices triangulaires peuvent être inversées aisément en utilisant l'élimination de Gauss-Jordan. C'est pourquoi, si nous voulons résoudre ce système pour divers b, il est plus intéressant de réaliser la décomposition LU une fois pour toute et d'inverser les matrices triangulaires pour les différents b plutôt que d'utiliser l'élimination de Gauss-Jordan à de multiples reprises. Les matrices L et U peuvent être utilisées pour déterminer l'inverse d'une matrice. Les programmes informatiques qui implémentent ce type de calcul, utilisent généralement cette méthode.

L'algorithme correspondant peut être décrit sous une forme compacte comme suit :

Algorithme 1. Élimination de Gauss sans pivot.

Pour $k = 1, ..., n-1$ et pour $i = k+1, ..., n$:

-nous calculons

$$\alpha_i^{(k)} = \frac{a_{i,k}^{(k)}}{a_{k,k}^{(k)}}.$$

Pour $j = k+1, ..., n$

$$a_{i,j}^{(k+1)} = a_{i,j}^{(k)} - \alpha_i^{(k)} a_{k,j}^{(k)}.$$

Fin de la boucle sur j,

-puis le second membre

$$b_i^{(k+1)} = b_i^{(k)} - \alpha_i^{(k)} b_k^{(k)}.$$

Fin de la boucle sur i et fin de la boucle sur k.

Nous observons finalement qu'à partir de cet algorithme numérique, nous pouvons effectuer un décomposition $L\,U$ de la matrice A. En effet, $L = I_n + B^{(1)} + \ldots + B^{(n-1)}$ et $U = A^{(n)}$.

Remarque 3.2 *Cet algorithme peut être utilisé sur un ordinateur pour des systèmes avec des milliers d'inconnues et d'équations. Il est cependant numériquement instable, les erreurs d'arrondis effectuées pendant le calcul sont accumulées et le résultat trouvé peut être loin de la solution surtout lorsque le système est mal conditionné.*

Méthode de Gauss avec pivot.

Observons bien qu'ici nous avons supposé à chaque étape que $a_{k,k}^{(k)} \neq 0$ mais nous ne sommes pas en mesure d'assurer que cela se produit vraiment pour une matrice inversible quelconque. Dans le cas général, nous avons plutôt recours à une étape supplémentaire de permutations de lignes pour remplacer un pivot éventuellement nul par un autre pivot qui lui sera non nul à coup sûr. Nous commençons par présenter un exemple puis nous énonçons un résultat qui assure la validité de la méthode d'élimination de Gauss avec pivot pour une matrice inversible quelconque.

Exemple : application de la méthode de Gauss avec pivot

Nous reprenons le système d'équations déjà rencontré lors du premier exemple sur la méthode de Gauss sans pivot mais nous l'écrivons de manière différente. Le système d'équations considéré est le suivant :

$$\begin{cases} \varepsilon x_1 + 2\varepsilon x_2 + 2\varepsilon x_3 = 2\varepsilon, \\ 3x_1 + 7x_2 + 8x_3 = 8. \\ x_1 + 3x_2 + 3x_3 = 2, \end{cases}$$

où $\varepsilon > 0$ est un petit paramètre proche de zéro.

Nous établissons la matrice correspondante

$$A^{(1)} = \begin{pmatrix} \varepsilon & 2\varepsilon & 2\varepsilon \\ 3 & 7 & 8 \\ 1 & 3 & 3 \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, \quad b^{(1)} = \begin{pmatrix} 2\varepsilon \\ 8 \\ 2 \end{pmatrix}$$

Tout d'abord, nous recherchons sur la première colonne le coefficient le plus grand en valeur absolue ; celui-ci jouera le rôle du pivot. Nous comprenons bien qu'ici il vaut mieux éviter de choisir ε, qui est proche de zéro, comme pivot puisque nous allons diviser toute une ligne par ce coefficient (lors de l'étape de l'élimination de Gauss), ce qui pourrait générer des valeurs de plus en plus grande.

Ici, le coefficient 3 de la deuxième ligne jouera le rôle du pivot, nous permutons alors la première ligne et la dernière. Le système s'écrit alors

$$(PA)^{(1)} = \begin{pmatrix} 3 & 7 & 8 \\ \varepsilon & 2\varepsilon & 2\varepsilon \\ 1 & 3 & 3 \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, \quad (Pb)^{(1)} = \begin{pmatrix} 8 \\ 2\varepsilon \\ 2 \end{pmatrix},$$

où le symbole P signifie que nous avons effectué une permutation de lignes de $A^{(1)}$ et $b^{(1)}$.

Ensuite, nous ajoutons un multiple de la première ligne aux deux autres lignes pour obtenir des zéros, c'est l'étape classique de l'élimination de Gauss-Jordan que nous avons déjà vu

$$A^{(2)} = \begin{pmatrix} 3 & 7 & 8 \\ 0 & -\varepsilon/3 & -2\varepsilon/3 \\ 0 & 2/3 & 1/3 \end{pmatrix}, \quad b^{(2)} = \begin{pmatrix} 8 \\ -2\varepsilon/3 \\ -2/3 \end{pmatrix}.$$

Encore une fois nous recherchons le coefficient qui jouera le rôle du pivot. Puisque $-2\varepsilon/3$ peut prendre des valeurs proches de zéro, nous tenons à éviter qu'il joue ce rôle. Nous devons donc effectuer une nouvelle étape de permutation durant laquelle nous échangeons la deuxième et troisième ligne ; il vient

$$(PA)^{(2)} = \begin{pmatrix} 3 & 7 & 8 \\ 0 & 2/3 & 1/3 \\ 0 & -\varepsilon/3 & -2\varepsilon/3 \end{pmatrix}, \quad (Pb)^{(2)} = \begin{pmatrix} 8 \\ -2/3 \\ -2\varepsilon/3 \end{pmatrix}.$$

Ensuite, nous appliquons l'étape de l'élimination de Gauss classique

$$A^{(3)} = \begin{pmatrix} 3 & 5 & 8 \\ 0 & 2/3 & 1/3 \\ 0 & 0 & -\varepsilon/2 \end{pmatrix}, \quad b = \begin{pmatrix} 8 \\ -2/3 \\ -\varepsilon \end{pmatrix}.$$

Nous retrouvons bien la solution du système :

$$x = \begin{pmatrix} 2 \\ -2 \\ 2 \end{pmatrix}.$$

La méthode du pivot de Gauss est une méthode directe de résolution de système linéaire qui permet de transformer un système en un autre système équivalent composé d'une matrice triangulaire. L'algorithme est le même que celui décrit pour la méthode d'élimination de Gauss sans pivot mais avec une étape supplémentaire de permutations. Après avoir obtenu un système triangulaire, nous résolvons le système à l'aide d'un algorithme de remontée.

Algorithme 2. Élimination de Gauss avec pivot.

Pour $k = 1, ..., n-1$ et pour $i = k+1, ..., n$:

-nous cherchons $k_0 \in \{k, .., n\}$ tel que

$$|a_{k_0,k}^{(k)}| = \max\{|a_{l,k}^{(k)}|, l \in \{k, .., n\}\}$$

et permute les lignes k et k_0,

-nous calculons

$$\alpha_i^{(k)} = \frac{a_{i,k}^{(k)}}{a_{k,k}^{(k)}}.$$

Pour $j = k+1, ..., n$

$$a_{i,j}^{(k+1)} = a_{i,j}^{(k)} - \alpha_i^{(k)} a_{k,j}^{(k)}.$$

Fin de la boucle sur j,

-puis le second membre

$$b_i^{(k+1)} = b_i^{(k)} - \alpha_i^{(k)} b_k^{(k)}.$$

Fin de la boucle sur i et fin de la boucle sur k.

Nous avons donc le résultat suivant

Théorème 3.2 *Soit $A \in \mathcal{M}_{n,n}(\mathbb{K})$ une matrice inversible. Alors, la matrice A admet une décomposition $L\,U$ à quelques permutations près*

$$P^{-1}\,A = L\,U, \quad \textit{ou} \quad A = P\,L\,U,$$

où P est une matrice de permutation (de même pour P^{-1}), tandis que L est une matrice triangulaire inférieure et U une matrice triangulaire supérieure. Parfois, la matrice de permutation peut être choisie afin d'être une matrice identité. Dans ce cas, la décomposition devient $A = L\,U$.

Démonstration. Nous ne présentons pas le détail de la preuve qui est essentiellement technique, nous nous reporterons par exemple au livre [3] pour plus de détails. L'idée de la preuve consiste à démontrer que la méthode d'élimination de Gauss avec pivot est comme la méthode

d'élimination de Gauss sans pivot une décomposition $L\,U$ mais avec en plus à chaque étape une permutation des lignes. Cette permutation permet d'assurer à chaque étape que le pivot $a_{k,k}^{(k)}$ est bien non nul. En effet, la méthode d'élimination de Gauss avec pivot peut s'écrire comme le produit suivant

$$A^{(k)} \;=\; L^{(k-1)}\,P^{(k-1)}\ldots L^{(1)}\,P^{(1)}\,A^{(1)},$$

où $L^{(i)}$ est la matrice déjà construite lors de la méthode d'élimination de Gauss sans pivot et $P^{(i)}$ est une matrice de permutation. Ainsi,

$$\begin{aligned}
|det(A)| \;&=\; |det(A^{(k)}| = \left|\begin{pmatrix}
a_{1,1}^{(1)} & a_{1,2}^{(1)} & \ldots & \ldots & a_{1,n}^{(1)} \\
 & a_{2,2}^{(2)} & \ldots & \ldots & a_{2,n}^{(2)} \\
 & & \ddots & & \vdots \\
 & & a_{k,k}^{(k)} & \ldots & a_{n,n}^{(k)} \\
 & & a_{n,k}^{(k)} & \ldots & a_{n,n}^{(k)}
\end{pmatrix}\right| \\
&=\; a_{1,1}^{(1)}\ldots a_{k-1,k-1}^{(k-1)}\left|\begin{pmatrix}
a_{k,k}^{(k)} & \ldots & a_{k,n}^{(k)} \\
a_{n,k}^{(k)} & \ldots & a_{n,n}^{(k)}
\end{pmatrix}\right|.
\end{aligned}$$

Puisque $det(A)$ est non nul, nous avons le déterminant de la sous-matrice qui est aussi non nul et donc il existe forcément un coefficient $a_{i,k}^{(k)}$ qui est non nul. Nous pouvons donc appliquer la stratégie de l'élimination de Gauss en ayant au préalable effectué si nécessaire une permutation des lignes. □

3.2 Factorisation de Choleski

La factorisation de Choleski, nommée d'après André-Louis Choleski, consiste pour une matrice symétrique définie positive A à déterminer une matrice triangulaire inférieure L telle que : $A \;=\; L\,L^T$.

La matrice L est en quelque sorte une "racine carrée" de A. Cette décomposition permet notamment de calculer la matrice inverse A^{-1}, de calculer le déterminant de A (égal au carré du produit des éléments diagonaux de L).

Pour débuter, voyons sur un exemple et considérons la matrice symétrique A :

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 5 & 5 & 5 \\ 1 & 5 & 14 & 14 \\ 1 & 5 & 14 & 15 \end{pmatrix}$$

Nous recherchons une matrice L telle que L soit triangulaire inférieure et $A = L\,L^T$. Pour cela, nous écrivons

$$L = \begin{pmatrix} l_{11} & 0 & 0 & 0 \\ l_{21} & l_{22} & 0 & 0 \\ l_{31} & l_{32} & l_{33} & 0 \\ l_{41} & l_{42} & l_{43} & l_{44} \end{pmatrix}$$

et calculons le produit $L\,L^T$.

En effectuant le produit scalaire de la première ligne de L et de la première colonne de L^T, il vient $l_{11} = \sqrt{1} = 1$. Ensuite, le produit scalaire de la première ligne de L et de la deuxième colonne de L^T donne $l_{21} = 1/1 = 1$. En poursuivant, la calcul pour la deuxième ligne de L et les suivantes, nous trouvons finalement

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

Dans la suite, nous démontrons le résultat plus général qui justifie l'existence et l'unicité d'une telle décomposition pour une matrice symétrique et fournit un algorithme de calcul dont ce dernier exemple est inspiré.

Théorème 3.3 (Factorisation de Choleski d'une matrice) *Si $A \in \mathcal{M}_{n,n}(\mathbb{R})$ est une matrice symétrique définie positive, il existe au moins une matrice réelle triangulaire inférieure L telle que :*

$$A = L\,L^T.$$

Nous pouvons également imposer que les éléments diagonaux de la matrice L soient tous positifs, et la factorisation correspondante est alors unique.

Démonstration. Nous allons prouver l'existence, l'unicité de la factorisation de Choleski puis construire l'algorithme permettant de calculer les coefficients de la matrice L. Commençons par l'existence d'une telle décomposition. Puisque A est une matrice définie positive, nous vérifions facilement que toutes les sous-matrices principales sont également définies positives, ce qui signifie que les sous-matrices principales sont toutes inversibles. D'après le Théorème 3.1, la matrice A admet une décomposition $L\,U$. En développant, le produit par bloc, nous obtenons

$$det(A_k) \;=\; det((LU)_k) \;=\; 1 \times u_{1,1} \times \ldots \times u_{k,k}.$$

D'une part, $det(A_k)$ correspond au produit des valeurs propres de A_k qui sont strictement positives, donc $det(A_k) > 0$. Par récurrence, nous obtenons que $u_{k,k} > 0$ pour tout $1 \leq k \leq n$. En posant,

$$\Lambda \;=\; diag(\sqrt{u_{1,1}}, \ldots, \sqrt{u_{n,n}}), \quad \tilde{L} \;=\; L\,\Lambda, \quad \tilde{U} \;=\; \Lambda^{-1}\,U,$$

nous avons

$$A \;=\; (L\,\Lambda)\,(\Lambda^{-1}\,U) \;=\; \tilde{L}\,\tilde{U}.$$

De plus, comme A est symétrique, nous avons

$$\tilde{L}\,\tilde{U} = \tilde{U}^T\,\tilde{L}^T,$$

ou encore puisque $\tilde{U}$ est inversible

$$(\tilde{U}^T)^{-1}\,\tilde{L} \;=\; \tilde{L}^T\,\tilde{U}^{-1}.$$

Maintenant, à gauche nous avons le produit de deux matrices triangulaires inférieures $\tilde{L}$ et $(\tilde{U}^T)^{-1}$, puisque $\tilde{U}^T$ est une matrice triangulaire inférieure et l'inverse d'une matrice triangulaire inférieure est elle aussi triangulaire inférieure, la matrice $(\tilde{U}^T)^{-1}\,\tilde{L}$ est donc une matrice triangulaire inférieure. A droite, nous montrons de la même manière que $\tilde{L}^T\,\tilde{U}^{-1}$ est une matrice triangulaire supérieure. Ainsi, nous avons prouvé que la matrice $\tilde{L}^T\,\tilde{U}^{-1}$ est à la fois triangulaire inférieure et supérieure, elle est donc diagonale et

$$(\tilde{L}^T\,\tilde{U}^{-1})_{i,i} \;=\; \sqrt{u_{i,i}}\,\frac{1}{\sqrt{u_{i,i}}} = 1.$$

Il vient alors

$$\tilde{L}^T\,\tilde{U}^{-1} = I_n,$$

ou encore

$$\tilde{L}^T \;=\; \tilde{U}.$$

Ensuite, la démonstration de l'unicité découle directement de l'unicité de la décomposition $L\,U$ que nous avons déjà prouvé.

Enfin, en pratique pour construire la matrice L, nous cherchons la matrice :

$$L = \begin{pmatrix} l_{1,1} & & & \\ l_{2,1} & l_{2,2} & & \\ \vdots & \vdots & \ddots & \\ l_{n,1} & l_{n,2} & \dots & l_{n,n} \end{pmatrix}.$$

De l'égalité $A = L\,L^T$ nous déduisons :

$$a_{i,j} = (L\,L^T)_{i,j} = \sum_{k=1}^{n} l_{i,k}\,l_{j,k} = \sum_{k=1}^{\min(i,j)} l_{i,k}\,l_{j,k}, \quad 1 \leq i,j \leq n,$$

puisque $l_{p,q} = 0$ si $1 \leq p < q \leq n$.

La matrice A étant symétrique, il suffit que les relations ci-dessus soient vérifiées pour $i \leq j$, c'est-à-dire que les éléments $l_{i,j}$ de la matrice L doivent satisfaire :

$$a_{i,j} = \sum_{k=1}^{i} l_{i,k}\,l_{j,k}, \quad 1 \leq i,j \leq n.$$

Pour $i = 1$, nous déterminons la première colonne de L :

$$\begin{aligned} &[j=1] \quad a_{1,1} = l_{1,1}\,l_{1,1} \quad \Rightarrow \quad l_{1,1} = \sqrt{a_{1,1}}, \\ &[j=2] \quad a_{1,2} = l_{1,1}\,l_{2,1} \quad \Rightarrow \quad l_{2,1} = \frac{a_{1,2}}{l_{1,1}}, \\ &\ldots\ldots \\ &[j=n] \quad a_{1,n} = l_{1,1}\,l_{n,1} \quad \Rightarrow \quad l_{n,1} = \frac{a_{1,n}}{l_{1,1}}. \end{aligned}$$

Pour $i \geq 1$ fixé, nous déterminons la j-ème colonne de L, après avoir calculé les $(j-1)$

premières colonnes :

$$
\begin{aligned}
&[j=i] \qquad a_{i,i} = l_{i,1}\, l_{i,1} \,+\, ... \,+\, l_{i,i}\, l_{i,i} \qquad\qquad \Rightarrow \quad l_{i,i} = \sqrt{a_{i,i} - \sum_{k=1}^{i-1} l_{i,k}^2},\\
&[j=i+1] \quad a_{i,i+1} = l_{i,1}\, l_{i+1,1} \,+\, ... \,+\, l_{i,i}\, l_{i+1,i} \quad \Rightarrow \quad l_{i+1,i} = \frac{a_{i,i+1} - \sum_{k=1}^{i-1} l_{i,k}\, l_{i+1,k}}{l_{i,i}},\\
&\vdots\\
&[j=n] \qquad a_{i,n} = l_{i,1}\, l_{n,1} \,+\, ... \,+\, l_{i,i}\, l_{n,i} \qquad\qquad \Rightarrow \quad l_{n,i} = \frac{a_{i,n} - \sum_{k=1}^{i-1} l_{i,k}\, l_{n,k}}{l_{i,i}}.
\end{aligned}
$$

Puisque nous avons montré l'existence d'une décomposition $L\,L^T$ de la matrice A, où tous les coefficients $l_{i,i} >$ sont strictement positifs, cela permet d'assurer que toutes les quantités

$$a_{i,i} - \sum_{k=1}^{i-1} l_{i,k}^2, ...$$

sont positives. □

L'algorithme correspondant peut être décrit sous une forme compacte comme suit :

Algorithme 3. Factorisation de Choleski.

Pour $i = 1, ..., n$

-nous calculons

$$l_{i,i} \;=\; \sqrt{a_{i,i} - \sum_{k=1}^{i-1} l_{i,k}}.$$

Pour $j = i+1, ..., n$

$$l_{i,j} \;=\; \frac{a_{i,j} \,-\, \sum_{k=1}^{i-1}\, l_{i,k}\, l_{j,k}}{l_{i,i}}.$$

Fin de la boucle sur j,

Fin de la boucle sur i.

Remarque 3.3 *Nous pouvons également énoncer un résultat analogue pour des matrices à coefficients complexes $A \in \mathcal{M}_{n,n}(\mathbb{C})$ hermitienne et définie positive dans $\mathbb{C}$. Nous notons alors la décomposition*

$$A \,=\, R^{\star}\, R,$$

avec R triangulaire supérieure et à coefficients diagonaux réels strictement positifs. Dans le cas $\mathbb{K} = \mathbb{R}$, nous retrouvons le résultat du Théorème 3.3 en posant $L = R^{\star} = R^T$.

4 Méthodes itératives

Nous allons étudier une méthode itérative générale pour la résolution ou disons plutôt l'approximation de la solution d'un système linéaire. Ici, l'objectif est de construire une suite vectorielle convergente vers la solution du système linéaire.

Nous verrons dans cette partie deux méthodes : la méthode de Jacobi puis celle de Gauss-Seidel.

4.1 Méthodologie générale

Nous nous intéressons à la résolution de système linéaire de la forme

$$A\, x \,=\, b.$$

Pour cela, nous construisons une suite $(x^{(k)})_{k\in\mathbb{N}}$ qui converge vers un point fixe x, solution du système d'équations linéaires. Nous cherchons à construire l'algorithme pour $x^{(0)}$ donné, la suite de terme général $x^{(k+1)} = F(x^{(k)})$ avec $k \in \mathbb{N}$. Nous posons alors

$$A \,=\, M \,-\, N,$$

où M est une matrice inversible et surtout facile à inverser (par exemple une matrice diagonale, triangulaire ou orthogonale). Ainsi, la résolution du système $A\, x \,=\, b$ sera équivalente à résoudre le système $M\, x \,=\, N\, x \,+\, b$ ou encore dès que M est facile à inverser

$$x \,=\, M^{-1}\, N\, x \,+\, M^{-1}\, b = F(x),$$

où F est une fonction affine.

Remarque 4.1 *Nous verrons dans le chapitre suivant que certaines méthodes itératives proposées peuvent également être utilisées pour des problèmes non linéaires où F est une fonction quelconque. Pour l'instant, nous nous consacrons uniquement à la résolution d'un problème linéaire.*

À partir de cette dernière équation, nous proposons la méthode itérative suivante : pour $x^{(0)}$ donné, et $k \geq 0$

$$x^{(k+1)} = F(x^{(k)}) = M^{-1}\, N\, x^{(k)} \,+\, M^{-1}\, b. \tag{4.4}$$

Il se pose alors plusieurs problèmes :

- Pour quelle décomposition de la matrice A obtenons-nous la convergence de la méthode ? Une décomposition possible est $M = I_n$ et $N = I_n - A$ mais nous rencontrons généralement des problèmes de convergence de la méthode itérative.
- Y-a-t-il un moyen d'accélérer la convergence de la méthode ? Nous voyons bien que le choix de $M = A$ et $N = 0$ permet la convergence de la méthode en une itération mais cette fois ce choix n'est en général pas intéressant puisque nous ne connaissons pas $M^{-1} = A^{-1}$.
- Comment déterminons-nous l'arrêt de l'itération ? Nous ne pouvons pas calculer l'écart exact entre $x^{(k)}$ et la solution x puisque nous ne connaissons pas la solution.

Dans la suite nous tentons d'apporter des réponses à ces questions. Nous commençons par la notion de convergence de la méthode itérative. Ensuite, nous proposons plusieurs décompositions possibles (méthode de Jacobi, Gauss-Seidel) et tentons d'étudier la vitesse de convergence des différentes méthodes.

Définition 4.1 *Soit $A = M - N \in \mathcal{M}_{n,n}(\mathbb{K})$ avec M une matrice inversible. Nous dirons que l'algorithme itératif (4.4) converge globalement si pour tout $b \in \mathbb{K}^n$ et tout $x^{(0)} \in \mathbb{K}^n$, la suite $(x^{(k)})_{k\geq 0}$ converge vers la solution $x = A^{-1}\,b$ dans $\mathbb{K}^n$:*

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

en rappelant que toutes les normes sont équivalentes sur $\mathbb{K}^n$.

D'autre part, nous dirons que l'algorithme itératif (4.4) converge localement si pour tout $b \in \mathbb{K}^n$, il existe un voisinage V de $x = A^{-1}\,b$, tel que pour tout $x^{(0)} \in V$, la suite $(x^{(k)})_{k\geq 0}$ converge vers la solution $x = A^{-1}\,b$ dans $\mathbb{K}^n$:

Bien sûr, la plupart du temps nous essaierons de construire un algorithme qui converge globalement de manière à pouvoir choisir $x^{(0)}$ au hasard sans qu'il soit forcément proche de la solution.

Essayons d'établir maintenant un critère sur les matrices M et N pour assurer la convergence de la méthode. Pour cela, nous introduisons l'erreur $e^{(k)}$ entre la solution approchée $x^{(k)}$ et la solution exacte x qui est donnée par

$$e^{(k+1)} \;=\; x^{(k+1)} - x.$$

Bien sûr, nous rappelons qu'a priori nous ne connaissons pas la solution exacte x, en pratique nous ne pouvons donc pas calculer l'erreur $e^{(k)}$. Notre objectif est plutôt de trouver un critère sur la décomposition assurant que cette erreur $e^{(k)}$ converge vers zéro lorsque k tend vers l'infini.

En utilisant l'algorithme itératif (4.4) et le fait que la solution x est une solution stationnaire de l'algorithme $x = A^{-1}\,b = M^{-1}\,(N\,x - b)$, nous sommes en mesure de calculer l'erreur $e^{(k+1)}$ à l'étape $k+1$ en fonction de l'erreur $e^{(k)}$ à l'étape k

$$e^{(k+1)} \;=\; M^{-1}\,N\,(x^{(k)} - x) \;=\; M^{-1}\,N\,e^{(k)}.$$

Nous posons alors $B = M^{-1}\,N$, ce qui donne

$$e^{(k+1)} \;=\; B\,e^{(k)} \;=\; B^{k+1}\,e^{(0)}.$$

D'après la définition de convergence, l'algorithme itératif va converger dès que

$$\lim_{k\to\infty} \|B^k\| = 0. \tag{4.5}$$

Le théorème suivant établit des critères sur la matrice B pour que la propriété (4.5) soit vérifiée ; ce qui par la même occasion assurera la convergence de l'algorithme itératif.

Théorème 4.1 *Soit $B \in \mathcal{M}_{n,n}(\mathbb{K})$ une matrice carrée d'ordre n alors les quatre propositions suivantes sont équivalentes*

(i) La limite $\lim_{k\to\infty} \|B^k\| = 0$, pour une norme quelconque.

(ii) La limite $\lim_{k\to\infty} B^k\,v = 0_{\mathbb{K}^n}$, pour tout vecteur $v \in \mathbb{K}^n$.

(iii) Le rayon spectral de B vérifie

$$\rho(B) < 1.$$

(iv) Il existe une norme matricielle subordonnée (dont le choix dépend de B) telle que $\|B\| < 1$.

Démonstration. Nous allons montrer que $(i) \Rightarrow (ii) \ldots \Rightarrow (iv)$ et enfin $(iv) \Rightarrow (i)$.
Montrons d'abord que $(i) \Rightarrow (ii)$. Nous supposons que

$$\lim_{k\to\infty} \|B^k\| = 0,$$

pour $\|.\|$ une norme quelconque. Vu qu'en dimension finie toutes les normes sont équivalentes, ceci est en particulier vraie pour une norme subordonnée. Soit $u \in \mathbb{R}^n$, nous avons pour une norme subordonnée

$$0 \leq \|B^k\,u\| \leq \|B^k\|\,\|u\|.$$

Or, d'après (i), nous savons que $\|B^k\|$ converge vers zéro lorsque k tend vers l'infini et donc pour tout vecteur $u \in \mathbb{R}^n$

$$\lim_{k\to\infty} \|B^k\,u\| = 0;$$

ce qui signifie bien que $B^k\,u$ tend vers zéro lorsque k tend vers l'infini.

Nous supposons maintenant que (ii) est vraie et montrons (iii). Pour cela, nous considérons $\lambda \in Sp(B)$ et w un vecteur propre associé à la valeur propre λ : $B\,w = \lambda\,w$; nous avons alors

$$B^k\,w \;=\; B^{k-1}\,(B\,w) \;=\; B^{k-1}\,(\lambda\,w) \;=\; \ldots \;=\; \lambda^k w.$$

Or, nous savons que $\lim_{k\to\infty} \|B^k\,w\| = 0$, ce qui signifie que

$$\lim_{k\to\infty} |\lambda^k|\,\|w\| \;=\; 0,$$

où $w \in I\!R^n$ ne dépend pas de k et il existe $i \in \{1, .., n\}$ tel que $w_i \neq 0$, nous avons donc

$$\lim_{k\to\infty} |\lambda|^k = 0,$$

ce qui implique que $|\lambda| < 1$. Ainsi, pour tout $\lambda \in Sp(B)$, nous avons $|\lambda| < 1$, c'est-à-dire $\rho(B) < 1$.

Montrons ensuite que $(iii) \Rightarrow (iv)$. Nous supposons que pour tout $\lambda \in Sp(B)$, nous avons $|\lambda| < 1$. En appliquant le théorème de Shur, il existe une matrice unitaire telle que

$$T = U^T B U = \begin{pmatrix} \lambda_1 & t_{1,2} & \dots & t_{1,n} \\ & \ddots & \ddots & \vdots \\ & & \lambda_{n-1} & t_{n-1,n} \\ 0 & & & \lambda_n \end{pmatrix}$$

et nous avons $|\lambda_i| < 1$ pour tout $i \in \{1, ..., n\}$. En appliquant la Proposition 2.4, nous pouvons alors calculer $\|T\|_\infty$ ou $\|T\|_1$ qui est donnée par

$$\|T\|_1 = \max\left(|\lambda_1|, \dots, |\lambda_n| + \sum_{i=1}^{n-1} |t_{i,n}|\right).$$

Nous ne pouvons pas conclure directement, nous introduisons alors pour $\delta > 0$ (un petit paramètre) la matrice diagonale

$$D = diag(1, \delta, \dots, \delta^{n-1})$$

et $D^{-1} = diag(1, \delta^{-1}, \dots, \delta^{1-n})$. Alors,

$$D^{-1} T D = \begin{pmatrix} \lambda_1 & \delta t_{1,2} & \dots & \delta^n t_{1,n} \\ & \ddots & \ddots & \vdots \\ & & \lambda_{n-1} & \delta t_{n-1,n} \\ 0 & & & \lambda_n \end{pmatrix}.$$

Ainsi, pour δ assez petit, nous avons $\|D^{-1}TD\|_1 < 1$ ou alors

$$\|D^{-1} U^T B U D\|_1 < 1.$$

Nous choisissons alors l'application qui pour $A \in \mathcal{M}_{n,n}(I\!R)$ associe

$$\|A\|_B = \|D^{-1} U^T A U D\|_1 < 1.$$

Nous vérifions aisément que $\|.\|_B$, dépend de B par l'intermédiaire de U et D et est une norme matricielle subordonnée à la norme vectorielle

$$v \longrightarrow \|v\|_B = \|D^{-1}\, U^T\, v\|_1.$$

Finalement, nous montrons que $(iv) \Rightarrow (i)$. Nous supposons que pour une norme subordonnée donnée $\|.\|_\alpha$, nous avons

$$\|B\|_\alpha < 1.$$

En utilisant que toutes les normes sont équivalentes, il existe C_1 et C_2 telles que pour une norme subordonnée quelconque $\|.\|$ et pour tout $k \geq 0$,

$$C_1\, \|B^k\| \;\leq\; \|B^k\|_\alpha \;\leq\; C_2\, \|B^k\|.$$

Or, nous savons que $\|B\|_\alpha < 1$, et donc

$$C_1\, \|B^k\| \;\leq\; \|B^k\|_\alpha \;\leq\; \|B\|_\alpha^k.$$

Ce qui montre que $\|B^k\|$ converge vers zéro lorsque k tend vers l'infini. □

4.2 Méthode de Jacobi

Nous choisissons la décomposition suivante :

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

avec

- ⋆ D la diagonale
- ⋆ $-E$ la partie en dessous de la diagonale
- ⋆ $-F$ la partie au dessus.

Dans la méthode de Jacobi, nous choisissons pour M une matrice facile à inverser, c'est-à-dire $M = D$ et $N = E + F$. La suite $(x^{(k)})_{k\geq 0}$ est alors donné par :

$$\begin{cases} x^{(0)} \in I\!K^n \\ \\ D\, x^{(k+1)} = (E+F)\, x^{(k)} + b, \quad k \geq 0. \end{cases}$$

Nous pouvons aussi écrire la méthode de Jacobi sous la forme

$$\begin{cases} x^{(0)} \in I\!K^n, \\ \\ a_{i,i}\, x_i^{(k+1)} \;=\; -\sum_{j<i} a_{i,j}\, x_j^{(k)} \;-\; \sum_{j>i} a_{i,j}\, x_j^{(k)} \;+\; b_i,\; i = 1, \ldots, n; \quad k \geq 0. \end{cases} \tag{4.6}$$

L'algorithme correspondant peut être décrit sous une forme compacte comme suit :

Algorithme 4. Méthode de Jacobi.

Pour $x_0 \in I\!K^n$ donné, nous posons $x^{(0)} = x_0$ et $\varepsilon = 1$.

Tant que $\varepsilon \geq 1.\,10^{-8}$

-pour $k \geq 0$, nous calculons

$$x^{(k+1)} = D^{-1}\left((E+F)\,x^{(k)} + b\right);$$

-nous calculons un "résidu" qui doit être proche de zéro

lorsque $x^{(k+1)}$ approche la solution

$$\varepsilon = \|A\,x^{(k+1)} - b\|.$$

- nous itérons $k = k+1$

Fin de tant que.

Maintenant, nous présentons un résultat de convergence de la méthode de Jacobi lorsque la matrice A est une matrice à diagonale strictement dominante.

Théorème 4.2 *Soit $A \in \mathcal{M}_{n,n}(I\!K)$ une matrice à diagonale strictement dominante, c'est-à-dire telle que*

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

Alors, pour tout $x^{(0)}$, la suite $(x^{(k)})_{k \in I\!N}$, donnée par la méthode de Jacobi (4.6), converge vers la solution x du système $A\,x = b$.

Démonstration. D'après le Théorème 4.1, il suffit de prouver qu'il existe une norme matricielle subordonnée $\|.\|_*$ telle que que la matrice $B = D^{-1}(E+F)$ vérifie

$$\|B\|_* < 1.$$

Considérons par exemple la norme $\|B\|_\infty$, nous savons d'après la Proposition 2.4 que cette norme peut être calculée exactement à parti des coefficients de B

$$\|B\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^{n} |b_{i,j}|.$$

Or, par construction de B, nous avons aussi

$$b_{i,j} = \begin{cases} 0, & \text{si } i = j, \\ -\dfrac{a_{i,j}}{a_{i,i}}, & \text{si } i \neq j. \end{cases}$$

D'une part, puisque A est à diagonale strictement dominante, nous vérifions facilement que pour tout $i \in \{1, ..., n\}$

$$\sum_{j=1}^{n} |b_{i,j}| = \frac{\sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{i,j}|}{|a_{i,i}|} < 1,$$

et donc

$$\|B\|_\infty < 1.$$

Ainsi, par application du Théorème 4.1, la méthode de Jacobi est convergente. □

Pour terminer, nous donnons un exemple de matrice symétrique définie positive pour laquelle la méthode de Jacobi ne converge pas.

Exemple 4.1 *Soit $A \in \mathcal{M}_{3,3}(\mathbb{R})$ donnée par*

$$A = \begin{pmatrix} 1 & a & a \\ a & 1 & a \\ a & a & 1 \end{pmatrix},$$

où a est un réel. Nous voulons résoudre le système $A\,x = b$ par une méthode itérative.

Nous vérifions d'abord que A est symétrique définie positive si et seulement si $a \in (-1/2, 1)$. En effet, en calculant les valeurs propres de A, nous obtenons le polynôme caractéristique

$$P(\lambda) = (1-\lambda)^3 - 3\,a^2\,(1-\lambda) + 2a^3 = -(\lambda - (1-a))^2\,(\lambda - (1+2\,a))$$

et les valeurs propres sont données par

$$\lambda_1 = \lambda_2 = 1 \,-\, a,\ \lambda_3 = 1 \,+\, 2\,a.$$

Les valeurs propres sont strictement positives dès que $-1/2 < a < 1$ et dans ce cas la matrice est bien définie positive.

Ensuite, nous vérifions que la méthode Jacobi est convergente si et seulement si $a \in [-1/2, 1/2]$. En effet, la méthode de Jacobi s'écrit

$$x^{(k+1)} \,=\, D^{-1}(D - A)\,x^{(k)} + D^{-1}\,b,$$

avec ici $D = I_n$ et donc la méthode converge si et seulement si $\rho(D - A) < 1$. Or, les valeurs propres de $D - A$ sont de la forme $\mu = 1 - \lambda$ où λ est une valeur propre de A et donc $\mu_1 = \mu_2 = -a$ et $\mu_3 = 2\,a$. Nous en concluons que la méthode de Jacobi converge si et seulement si $-1/2 < a < 1/2$.

Cet exemple nous montre bien que la méthode de Jacobi n'est pas toujours la plus adaptée pour traiter des applications concrètes, où les systèmes linéaires à résoudre font apparaître des matrices définies positives. Une alternative possible est de recourir à la méthode de Gauss-Seidel.

4.3 Méthode de Gauss-Seidel

Nous décomposons également la matrice A de la façon suivante :

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

avec le même choix pour D, E et F que dans la méthode de Jacobi. Cependant, pour la méthode de Gauss-Seidel nous choisissons $M = D - E$ et $N = F$

$$(D - E)\, x^{(k+1)} = F\, x^{(k)} + b, \quad k \geq 0.$$

Dans ce cas, $D - E$ est une matrice triangulaire inférieure, elle est donc facilement inversible en utilisant un algorithme de descente (nous calculons d'abord x_1, puis x_2...). Nous avons cette fois-ci

$$\begin{cases} x^{(0)} \in I\!K^n, \\ a_{i,i}\, x_i^{(k+1)} = -\sum_{j<i} a_{i,j}\, x_j^{(k+1)} - \sum_{j>i} a_{i,j}\, x_j^{(k)} + b_i,\ i = 1, \ldots, n; \quad k \geq 0. \end{cases}$$

L'algorithme correspondant peut être décrit sous une forme compacte comme suit :

Algorithme 5. Méthode de Gauss-Seidel.

Pour $x_0 \in I\!K^n$ donné, nous posons $x^{(0)} = x_0$ et $\varepsilon = 1$.

Tant que $\varepsilon \geq 1.\,10^{-8}$

-pour $k \geq 0$, nous résolvons par un algorithme de remontée

$$x^{(k+1)} = (D - E)^{-1} \left(F\, x^{(k)} + b\right);$$

-nous calculons un "résidu" qui doit être proche de zéro

lorsque $x^{(k+1)}$ approche la solution

$$\varepsilon = \|A\, x^{(k+1)} - b\|.$$

- nous itérons $k = k + 1$

Fin de tant que.

Nous démontrons alors un théorème qui établit la convergence de la méthode de Gauss-Seidel pour les matrices symétriques définies positives. Avant cela, nous énonçons un lemme qui permettra de faire la démonstration de ce résultat.

Lemme 4.1 *Soit $A \in \mathcal{M}_{n,n}(I\!R)$ une matrice symétrique définie positive. Nous décomposons $A = M - N$ alors si $M^T + N$ est symétrique définie positive, nous avons $\rho(M^{-1}\, N) < 1$.*

Démonstration. Soit $A = M - N$ telle que $M + N$ est symétrique définie positive. D'après le Théorème 4.1, la condition $\rho(M^{-1}\, N) < 1$ est équivalente à démontrer qu'il existe une norme matricielle subordonnée $\|.\|_*$ telle que

$$\|M^{-1}\, N\|_* < 1.$$

Puisque la matrice A est définie positive, nous pouvons considérer la norme matricielle subordonnée à la norme vectorielle donnée par $\|x\|_* = \sqrt{(A\, x, x)}$. Alors, il vient

$$\begin{aligned} \|M^{-1}\, N\|_*^2 &= \sup_{x \in I\!R^n} \frac{\|M^{-1}\, N\, x\|_*^2}{\|x\|_*^2}, \\ &= \sup_{x \in I\!R^n} \frac{(A\, M^{-1}\, N\, x, M^{-1}\, N\, x)}{(A\, x, x)}. \end{aligned}$$

Or, en écrivant $N = M - A$ puis $y = M^{-1}\, A\, x$

$$\begin{aligned} (A\, M^{-1}\, N\, x, M^{-1}\, N\, x) &= (A\, (I_n - M^{-1}\, A)\, x, (I_n - M^{-1}\, A)\, x), \\ &= (A\, x, x) - (A\, x, y) - (A\, y, x) + (A\, y, y) \end{aligned}$$

et puisque A est symétrique

$$\|M^{-1}\,N\,x\|_*^2 \;=\; (A\,M^{-1}\,N\,x, M^{-1}\,N\,x) \;=\; (A\,x,x) \;-2\,(A\,x,y) \;+\; (A\,y,y).$$

Ce dernier terme est positif et donc pour démontrer que

$$\frac{\|M^{-1}\,N\,x\|_*^2}{\|x\|_*^2} < 1,$$

il nous suffit de prouver que $-2\,(A\,y,x) + (A\,y,y)$ est négatif. Nous vérifions puisque $A\,x = M\,y$

$$-2\,(A\,x,y) + (A\,y,y) = -2(M\,y,y) \;+\; (A\,y,y).$$

Or, nous voyons que

$$(M\,y,y) = \sum_{i=1}^{n}\left(\sum_{j=1}^{n} M_{i,j}\,y_j\right)\,y_i \;=\; \sum_{j=1}^{n}\left(\sum_{i=1}^{n} M_{i,j}\,y_i\right)\,y_j \;=\; (M^T\,y,y)$$

et donc

$$-2\,(A\,x,y) + (A\,y,y) \;=\; -\left(M^T\,y,y\right) \;-\; ((M-A)\,y,y) \;=\; -\left((M^T+N)\,y,y\right).$$

Comme $M^T + N$ est définie positive, nous obtenons bien que

$$-2\,(A\,x,y) + (A\,y,y) < 0$$

et donc $\|M^{-1}\,N\|_* < 1$ ou encore $\rho(M^{-1}\,N) < 1$. □

Nous pouvons alors le résultat suivant.

Théorème 4.3 *Si $A \in \mathcal{M}_{n,n}(I\!R)$ est une matrice symétrique définie positive. Alors, pour tout $x^{(0)}$ la méthode de Gauss-Seidel est bien définie et converge vers la solution x du système $A\,x = b$.*

Démonstration. D'abord, nous vérifions que la méthode de Gauss-Seidel est bien définie. Nous posons $A = M - N$, avec $M = D - E$ et $N = F$, D représente la diagonale de A, $-E$ la partie inférieure de A et $-F$ la partie supérieure. Alors, pour que la méthode de Gauss-Seidel soit bien définie, il suffit de vérifier que M est inversible. Puisque M est une matrice triangulaire, nous avons

$$det(M) \;=\; det(D) \;=\; \prod_{i=1}^{n} a_{i,i}$$

et puisque A est définie positive tous les terme diagonaux sont positifs (pour le vérifier il suffit de prendre le vecteur e_i de la base cannonique, dont seule la i-ème composante est non nulle et vaut un ; nous obtenons $0 \;<\; e_i^T\,A\,e_i \;=\; a_{i,i}$ pour tout $i \in \{1,...,n\}$) ; nous avons alors

$$det(M) > 0.$$

Maintenant, pour prouver que la méthode de Gauss-Seidel converge, nous voulons appliquer le Lemme 4.1, il nous suffit donc de vérifier que la matrice $M^T + N$ est symétrique définie positive

$$M^T + N \;=\; (D - E)^T + F \;=\; D - E^T + F$$

et puisque A est symétrique $E^T = F$ et donc

$$M^T + N \;=\; D,$$

qui est symétrique définie positive. Par application du Lemme 4.1, nous obtenons $\rho(M^{-1}\,N) < 1$ et donc la méthode de Gauss-Seidel est convergente. □

4.4 Test d'arrêt et nombre d'itérations

Finalement, comme test d'arrêt, nous utilisons le vecteur résidu $r^{(k)} = b - A\,x^{(k)}$, ce qui donne, pour une précision donnée ε :

$$\frac{\|r^{(k)}\|}{\|b\|} \;=\; \frac{\|b \;-\; A\,x^{(k)}\|}{\|b\|} \;\leq\; \varepsilon.$$

Remarquons aussi que pour une méthode itérative qui converge vers la solution, il est parfois possible de calculer le nombre d'itération maximal en fonction de l'erreur souhaitée. En effet, supposons que nous sommes en mesure de calculer pour une norme donnée $\|.\|$ telle que $\|B\| = \|M^{-1}\,N\| < 1$, alors nous avons déjà démontré que

$$\|e^{(k)}\| \;\leq\; \|B\|^k\,\|e^{(0)}\|.$$

D'autre part, nous avons aussi

$$\|e^{(0)}\| \;\leq\; \|x^{(0)} - x^{(1)}\| \;+\; \|e^{(1)}\| \;\leq \|x^{(0)} - x^{(1)}\| \;+\; \|B\|\,\|e^{(0)}\|$$

et donc puisque $\|B\| < 1$, nous obtenons

$$\|e^{(0)}\| \;\leq\; \frac{1}{1 - \|B\|}\,\|x^{(0)} - x^{(1)}\|,$$

d'où en regroupant les deux résultats

$$\|e^{(k)}\| \;\leq\; \frac{\|B\|^k}{1 - \|B\|}\;\|x^{(1)} - x^{(0)}\|.$$

Ainsi, si nous souhaitons que l'erreur $\|e^{(k)}\|$ soit inférieure à 10^{-N}, il suffit de calculer un nombre d'itérations $k_0 \in I\!N^*$ tel que

$$\frac{\|B\|^{k_0}}{1 - \|B\|}\;\|x^{(1)} - x^{(0)}\| \leq 10^{-N},$$

c'est-à-dire

$$k_0 \geq \frac{\log\left(\frac{1-\|B\|}{\|x^{(1)}-x^{(0)}\|}\right) - N\log(10)}{\log(\|B\|)}.$$

5 Complément du Chapitre 1

Dans cette partie, nous présentons une méthode de factorisation permettant de résoudre facilement un système linéaire, c'est la factorisation $Q\,R$. Ensuite, nous présentons deux méthodes itératives (méthodes de relaxation et de Richardson) directement inspirées de la méthode de Gauss-Seidel.

5.1 La factorisation $Q\,R$

la décomposition $Q\,R$ d'une matrice A est une décomposition de la forme

$$A \;=\; Q\,R,$$

où Q est une matrice orthogonale ($Q\,Q^T \;=\; I_n$), et R est une matrice triangulaire supérieure.

Nous remarquons que lorsqu'une telle décomposition est connue, nous pouvons résoudre facilement un système linéaire de la forme $A\,x = b$. En effet, nous recherchons $x \in I\!R^n$ tel que

$$Q\,R\,x \;=\; b.$$

En multipliant alors par Q^T et puisque Q est orthogonale, il vient

$$R\,x = Q^T\,b,$$

lequel peut être résolu facilement puisque R est triangulaire supérieure.

Il existe plusieurs méthodes pour réaliser cette décomposition :

- la méthode de Householder où Q est obtenue par produits successifs de matrices orthogonales élémentaires ;
- la méthode de Givens où Q est obtenue par produits successifs de matrices de rotation plane
- la méthode de Schmidt, basée sur le procéde de Graham-Schmidt.

Nous présentons ici la méthode de Householder qui mène à la factorisation $Q\,R$ d'une matrice A inversible.

Définition 5.1 *Nous appelons matrices de Householder, les matrices $H_u \in \mathcal{M}_{n,n}(I\!K)$ qui vérifient*

$$H_u = I_n - 2\,u\,u^*, \quad u \in I\!K^n, \quad \|u\| = 1. \tag{5.7}$$

Cette matrice de Householder satisfait les propriétés suivantes

Proposition 5.1 *Soit $u \in I\!K^n$, la matrice de Householder H_u vérifie alors*

(a) $H_u = H_u^*$,

(b) $H_u^{-1} = H_u$

(c) $H_u\,v \;=\; -v$, *pour tout* $v \;\in\; I\!K^n$ *colinéaire à* u *et* $H_u\,v \;=\; v$ *pour tout* $v \;\in\; I\!K^n$ *orthogonal à* u.

(d) $|det(H_u)| \,=\, 1$.

Démonstration. D'abord, d'après la définition de la matrice de Householder, nous voyons facilement que la matrice H_u est hermitienne (ou symétrique dans le cas $\mathbb{K} = \mathbb{R}$).

Nous vérifions ensuite la propriété (b). Pour cela, nous calculons

$$H_u\, H_u \,=\, I_n - 4u\, u^* + 4(u\, u^*)\, (u\, u^*).$$

Or, comme $\|u\|^2 = u^*\, u = 1$, nous avons

$$H_u\, H_u \,=\, I_n - 4u\, u^* + 4u\, u^* = I_n$$

et donc $(H_u)^{-1} \,=\, H_u$.

Ensuite pour démontrer (c), nous prenons d'une part $v \in \mathbb{K}^n$ tel que $(v, u) \,=\, u^*\, v \,=\, 0$ alors

$$H_u\, v \,=\, v - 2\, u\, (u^*\, v) = v.$$

D'autre part, pour $v \in \mathbb{K}^n$ colinéaire à u, c'est-à-dire $v = \alpha\, u$ avec $\alpha \in \mathbb{K}$, nous avons

$$H_u\, v = \alpha\, u - 2\, \alpha (u\, u^*)\, u = \alpha\, u - 2\, \alpha\, u \,=\, -v.$$

Enfin, nous souhaitons démontrer (d). D'abord, lorsque $u = 0_{\mathbb{K}^n}$, nous avons $H_u = I_n$ et donc $det(H_u) \,=\, 1$. En revanche lorsque $u \in \mathbb{K}^n$ et $u \neq 0$ nous posons $u_1 \,=\, u$, puis nous formons une matrice à partir du vecteur u_1 et des vecteurs $u_2, \ldots, u_n$ tels que $(u_i, u_1) = 0$ pour tout $i \geq 2$ et $(u_j)_{1 \leq j \leq n}$ forme une base orthonormée de $\mathbb{K}^n$. Nous notons alors U la matrice composée des colonnes $(u_j)_{1 \leq j \leq n}$

Alors, nous avons en utilisant (c)

$$H_u\, U \,=\, H_u\, (u_1, u_2, \ldots, u_n) \,=\, (u_1, u_2, \ldots, u_n) \begin{pmatrix} -1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & 1 \end{pmatrix}$$

et donc $det(H_u\, U) \,=\, det(H_u)\, det(u_1, \ldots, u_n) = -1 \times det(u_1, \ldots, u_n)$, c'est-à-dire $det(H_u) = -1$. □

À partir de la Proposition 5.1, nous vérifions que pour tout $v \in \mathbb{K}^n$, qui peut être décomposé comme $v = x + y$ avec x colinéaire à u et y orthogonal à u, nous avons

$$H_u\, v \,=\, H_u\, (x + y) \,=\, y - x.$$

Nous disons que H_u est une réflection par rapport à l'hyperplan $u^\perp$ de $\mathbb{R}^n$ et nous vérifions la proposition suivante

Proposition 5.2 *Pour tout $v \in \mathbb{R}^n$, il existe $u \in \mathbb{R}^n$ tel que*

$$H_u\, v \,=\, \|v\|\, e_1, \quad e_1 \,=\, (1, 0, \ldots, 0)^T \in \mathbb{R}^n.$$

Démonstration. Si $v \,=\, \|v\|\, e_1$; nous posons $u \,=\, 0$ et $H_u \,=\, I_n$. Nous supposons que $v \neq \|v\|\, e_1$ et posons alors

$$u = \frac{v - \|v\|\, e_1}{\|v \,-\, \|v\|\, e_1\|}.$$

Alors, un calcul direct nous montre que

$$H_u\, v \,=\, v \,-\, 2\, \frac{(v \,-\, \|v\|\, e_1)(v \,-\, \|v\|\, e_1)^T}{\|v \,-\, \|v\|\, e_1\|^2}\, v \,=\, v \,-\, 2\,(v \,-\, \|v\|\, e_1)\, \frac{(v \,-\, \|v\|\, e_1, v)}{\|v \,-\, \|v\|\, e_1\|^2}.$$

De plus,

$$\begin{aligned} \|v \,-\, \|v\|\, e_1\|^2 \;&=\; (v \,-\, \|v\|\, e_1, v) \,-\, (v \,-\, \|v\|\, e_1, \|v\|\, e_1) \\ &=\; (v \,-\, \|v\|\, e_1, v) \,+\, (v, v) \,-\, (v, \|v\|\, e_1) \\ &=\; 2\,(v - \|v\|\, e_1, v). \end{aligned}$$

En combinant, ces deux derniers résultats, nous obtenons $H_u\, v = \|v\|\, e_1$. □

Maintenant que nous avons étudié les propriétés des matrices de Householder, nous pouvons effectuer la factorisation $Q\, R$.

Avant de présenter la méthode générale, nous étudions un exemple

Exemple 5.1 *Considérons la matrice inversible $A^{(1)} \in \mathcal{M}_{3,3}(\mathbb{R})$*

$$A^{(1)} = \begin{pmatrix} 0 & 3 & 2 \\ 1 & 0 & 2 \\ 0 & 4 & 1 \end{pmatrix}.$$

Dans un premier temps, nous voulons construire une matrice $A^{(2)}$ telle que la première colonne soit colinéaire au vecteur de la base cannonique e_1 de manière à faire apparaître des zéros sur la première colonne de $A^{(2)}$ (excepté sur la première ligne bien sûr !).

Puisque la première colonne de $A^{(1)}$ n'est pas colinéaire au vecteur de la base cannonique e_1, nous formons la matrice H_1 à partir du vecteur u_1 construit dans la démonstration de la Proposition 5.2, nous obtenons le vecteur $u_1 \in \mathbb{R}^3$ donné par

$$u_1 = \frac{1}{\sqrt{2}}\, (-1, 1, 0)^T$$

et la matrice de Householder correspondante $H_1 = H_{u_1}$, qui est donnée par

$$H_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Nous construisons alors $A^{(2)}$ comme suit

$$A^{(2)} = H_1 A^{(1)} = \begin{pmatrix} 1 & 0 & 2 \\ 0 & 3 & 2 \\ 0 & 4 & 1 \end{pmatrix}.$$

Dans un deuxième temps, nous voulons faire apparaître une matrice $A^{(3)}$ triangulaire supérieure. Il reste donc à modifier la deuxième colonne. Nous considérons seulement la matrice extraite de $A^{(2)}$ (à partir de la deuxième ligne et de la deuxième colonne de $A^{(2)}$)

$$\tilde{A}^{(2)} = \begin{pmatrix} 3 & 2 \\ 4 & 1 \end{pmatrix}.$$

Pour faire apparaître des zéros sur la première colonne (excepté sur la première ligne), nous formons le vecteur $u_2 \in \mathbb{R}^2$ à partir de la première colonne de $\tilde{A}^{(2)}$, il vient alors

$$u_2 = \frac{1}{\sqrt{5}}(-1, 2)^T$$

et nous posons $H_2 \in \mathcal{M}_{3,3}(\mathbb{R})$ formée à partir de I_3 et de $H_{u_2} \in \mathcal{M}_{2,2}(\mathbb{R})$ complétée par des zéros

$$H_2 == \frac{1}{5} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 3 & 4 \\ 0 & 4 & -3 \end{pmatrix}.$$

Puis, finalement

$$A^{(3)} = H_2 A^{(2)} = \begin{pmatrix} 1 & 0 & 2 \\ 0 & 5 & 2 \\ 0 & 0 & 1 \end{pmatrix}.$$

En posant $R = A^{(3)}$ et $Q = H_1\, H_2$, nous obtenons une factorisation QR de la matrice A.

Passons maintenant au résultat général.

Théorème 5.1 (Décomposition QR) *Soit $A \in \mathcal{M}_{n,n}(\mathbb{K})$ une matrice inversible. Alors, il existe une unique matrice Q hermitienne, c'est-à-dire $Q^*\, Q = I_n$ et une matrice triangulaire supérieure R telles que*

$$A \;=\; Q\, R.$$

Démonstration. Nous considérons $A \in \mathcal{M}_{n,n}(\mathbb{K})$ une matrice inversible. Nous posons $A^{(1)} = A$ et pour $i \in \{1, \ldots, n\}$, notons $a_i^{(1)}$ la i-ème colonne de la matrice $A^{(1)}$, c'est-à-dire

$$A^{(1)} = \left(a_1^{(1)}, \ldots, a_n^{(1)}\right).$$

Supposons que $a_1^{(1)}$ n'est pas colinéaire au vecteur e_1 et nous posons u_1, le vecteur donné par

$$u_1 = \frac{a_1^{(1)} \;-\; \|a_1^{(1)}\|\, e_1}{\|a_1^{(1)} \;-\; \|a_1^{(1)}\|\, e_1\|}.$$

Nous notons alors $H_1 \;=\; H_{u_1}$. D'après la Proposition 5.2, nous avons $H_1\, a_1^{(1)} \;=\; \|a_1^{(1)}\|\, e_1$. Ainsi, nous obtenons

$$A^{(2)} = H_1\, A^{(1)}$$

avec $a_{1,1}^{(2)} = \|a_1^{(1)}\|$ et $a_{1,j}^{(2)} = 0$ pour tout $j \geq 2$.

En réitérant ce procédé et en s'inspirant de ce qui précède, nous mettons en place un algorithme (dit d'élimination de Householder) permettant de construire une matrice $A^{(k)}$ à l'issue de l'étape $k-1$, donnée par

$$A^{(k)} = \begin{pmatrix} a_{1,1}^{(k)} & a_{1,2}^{(k)} & \ldots & a_{1,k}^{(k)} & \ldots & a_{1,n}^{(k)} \\ 0 & a_{2,2}^{(k)} & \ldots & a_{2,k}^{(k)} & \ldots & a_{2,n}^{(k)} \\ \vdots & 0 & \ddots & \vdots & & \vdots \\ \vdots & & 0 & a_{k,k}^{(k)} & \ldots & a_{k,n}^{(k)} \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \ldots & 0 & a_{n,k}^{(k)} & \ldots & a_{n,n}^{(k)} \end{pmatrix},$$

telle que

$$A^{(k)} \;=\; H_{k-1}\, A^{(k-1)} \;=\; H_{k-1} \ldots H_1\, A^{(1)}$$

Nous notons

$$a_k^{(k)} = (0, \ldots, 0, a_{k,k}^{(k)}, \ldots, a_{n,k}^{(k)}), \quad e_k = (0, \ldots, 0, 1, 0, \ldots, 0),$$

où seule la k-ème composante du vecteur e_k est non nulle. Supposons que $a_k^{(k)} \neq \|a_k^{(k)}\| \, e_k$ et en posant

$$u_k = \frac{a_k^{(k)} - \|a_k^{(k)}\| \, e_k}{\|a_k^{(k)} - \|a_k^{(k)}\| \, e_k\|}, \quad H_k = H_{u_k} = I_n - 2 \, u_k \, u_k^T.$$

D'après la Proposition 5.2, nous avons $H_k \, a_k^{(k)} = \|a_k^{(k)}\| \, e_k$ et obtenons que

$$A^{(k+1)} = H_k \, A^{(k)},$$

avec

$$A^{(k+1)} = \begin{pmatrix} a_{1,1}^{(k+1)} & a_{1,2}^{(k+1)} & \ldots & a_{1,k+1}^{(k+1)} & \ldots & a_{1,n}^{(k+1)} \\ 0 & a_{2,2}^{(k+1)} & \ldots & a_{2,k+1}^{(k+1)} & \ldots & a_{2,n}^{(k+1)} \\ \vdots & 0 & \ddots & \vdots & & \vdots \\ \vdots & & 0 & a_{k+1,k+1}^{(k+1)} & \ldots & a_{k+1,n}^{(k+1)} \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \ldots & 0 & a_{n,k+1}^{(k+1)} & \ldots & a_{n,n}^{(k+1)} \end{pmatrix}.$$

Ainsi, en $(n-1)$ itérations, nous réduisons la matrice A en une matrice $A^{(n)}$ qui est triangulaire supérieure et telle que

$$H_{n-1} \ldots H_1 \, A = A^{(n)}.$$

Puisque chaque matrice H_k est orthogonale, nous posons

$$Q = H_{n-1} \ldots H_1, \quad \text{et} \quad R = A^{(n)}$$

et nous aboutissons ainsi à une factorisation $A = Q \, R$ avec Q une matrice orthogonale et R une matrice triangulaire supérieure. Nous démontrons facilement que cette décomposition est unique.

□

Remarque 5.1 *Une autre méthode pour démontrer qu'une matrice inversible A admet une factorisation $Q \, R$, consiste à considérer la matrice $A^T \, A$. Cette matrice est bien sûr symétrique mais aussi définie positive. En effet pour tout $x \in \mathbb{R}^n$ non nul, nous introduisons $y = A \, x$*

$$x^T \, A^T \, A \, x = y^T \, y > 0,$$

car A étant inversible, $y = A\,x$ ne peut pas être nul. Ainsi, la matrice $A^T\,A$ admet une factorisation de Choleski, c'est-à-dire il existe une unique matrice triangulaire inférieure L avec des éléments diagonaux strictement positifs telle que

$$A^T\,A = L\,L^T$$

Nous posons alors $R = L^T$ qui est une matrice triangulaire supérieure avec des éléments diagonaux strictement positifs. Nous posons alors Q telle que

$$Q^T = L^{-1}\,A^T$$

Il reste à vérifier que Q est orthogonale

$$\begin{aligned} Q^T\,Q &= \left(L^{-1}\,A^T\right)\,\left(L^{-1}\,A^T\right)^T \\ &= L^{-1}\,A^T\,A\,(L^{-1})^T \end{aligned}$$

Or, nous savons

$$A^T\,A \,=\, L\,L^T,$$

d'où

$$Q^T\,Q \,=\, L^{-1}\,L\,L^T\,(L^{-1})^T \,=\, L^T\,(L^T)^{-1} \,=\, I_n.$$

et donc ($L^T = R$)

$$\begin{aligned} A &= (A^T)^{-1}\,L\,L^T \\ &= (A^T)^{-1}\,(L^{-1})^{-1}\,R \\ &= (L^{-1}\,A^T)^{-1}\,R \\ &= (Q^T)^{-1}\,R \end{aligned}$$

et en utilisant que Q est orthogonale, nous obtenons le résulat

$$A = Q\,R.$$

Nous observons ainsi qu'il est facile de construire un algorithme pour effectuer une factorisation $Q\,R$ en utilisant la factorisation de Choleski de $A^T\,A$.

5.2 Méthode de relaxation

La méthode de relaxations successives est définie comme suit. Tout d'abord, nous décomposons la matrice A comme nous l'avons fait précédement

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

où D est diagonale, E est strictement triangulaire inférieure, F est strictement triangulaire supérieure.
Dans la méthode de Gauss-Seidel, nous résolvons

$$(D \,-\, E)\,\tilde{x}^{(k+1)} \,=\, F\,\tilde{x}^{(k)} \,+\, b.$$

Nous avons donc $M = D - E$ et $N = F$. L'idée de la méthode de relaxations successives est de dire que $x^{(k+1)}$ est une combinaison linéaire de $x^{(k)}$ et de $\tilde{x}^{(k+1)}$ calculée par Gauss-Seidel.

$$\begin{cases} x^{(k+1)} = \omega\, \tilde{x}^{(k+1)} + (1-\omega)\, x^{(k)}, \\ D\, \tilde{x}^{(k+1)} - E\, x^{(k+1)} = F\, x^{(k)} + b. \end{cases}$$

Nous remarquons d'abord que pour $\omega = 1$ nous retrouvons la méthode de Gauss-Seidel. En éliminant $\tilde{x}^{(k+1)}$ du système précédent, il vient

$$\tilde{x}^{(k+1)} = \frac{1}{\omega}\, x^{(k+1)} - \frac{1-\omega}{\omega}\, x^{(k)}$$

soit encore

$$D\left(\frac{1}{\omega}\, x^{(k+1)} - \frac{1-\omega}{\omega}\, x^{(k)}\right) - E\, x^{(k+1)} = F\, x^{(k)} + b,$$

c'est-à-dire

$$\left(\frac{D}{\omega} - E\right) x^{(k+1)} = \left(\frac{1-\omega}{\omega}\, D + F\right) x^{(k)} + b.$$

Nous posons alors

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

Nous voulons maintenant montrer que lorsque $0 < \omega < 2$, la méthode de relaxations successives converge ce qui revient à prouver que $\rho(M^{-1}\, N) < 1$. Pour cela nous appliquons le Lemme 4.1 : la matrice A est symétrique définie positive, il suffit alors de vérifier que $M^T + N$ est symétrique définie positive. En effet,

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

Et donc $M^T + N$ est symétrique définie positive si et seulement si $(2-\omega)/\omega > 0$, c'est-à-dire $0 < \omega < 2$. Par conséquent la méthode converge lorsque $0 < \omega < 2$.

5.3 Méthode itérative de Richardson

Nous considérons la décomposition $A = I_n - (I_n - A)$. Ceci conduit à l'algorithme suivant

$$x^{(k+1)} = (I_n - A)\, x^{(k)} - b = x^{(k)} - \left(A\, x^{(k)} + b\right).$$

Pour étudier la convergence, nous posons $B = M^{-1}\, N = I_n - A$ et voulons montrer que $\rho(B) < 1$. Or, les valeurs propres $\lambda(B)$ de $B = I_n - A$ sont données par

$$\lambda(B) = 1 - \lambda(A),$$

où $\lambda(A)$ est une valeur propre de A. En effet, il suffit de considérer $(\lambda(A), v_{\lambda(A)})$ un élément propre de A, alors $(1 - \lambda(A), v_{\lambda(A)})$ est un élément propre de $I_n - A_n$. Par conséquent, la méthode converge si et seulement si $|1 - \lambda(A)| < 1$ pour tout $\lambda(A) \in Sp(A)$.

Cependant, cette dernière condition à peu de chance d'être vérifiée en général. Pour remédier à ce problème, nous introduisons la décomposition suivante

$$A = \frac{1}{\gamma} I_n - \left(\frac{1}{\gamma} I_n - A \right),$$

avec γ un réel strictement positif. Ceci conduit à l'algorithme suivant

$$x^{(k+1)} \;=\; (I_n - \gamma\, A)\; x^{(k)} \;-\; \gamma\, b \;=\; x^{(k)} \;-\; \gamma\, \left(A\, x^{(k)} \;-\; b \right).$$

Nous utilisons toujours le même critère pour l'étude de la convergence. Ainsi, la matrice d'itération $B = M^{-1}\, N$ est donnée par

$$B \;=\; I_n - \gamma\, A$$

et donc les valeurs propres de B sont $\lambda(B) \;=\; 1 - \gamma\, \lambda(A)$, où $\lambda(A)$ est valeur propre de A.

Cet algorithme définit une méthode convergente si et seulement si

$$|1 - \gamma\, \lambda(A)| < 1, \quad \forall\, \lambda(A) \;\in\; Sp(A).$$

Nous notons par $\lambda_{max} := \rho(A)$ le rayon spectral correspondant au module de la plus grande valeur propre en module. Alors, en choisissant $0 < \gamma < 2/\rho(A)$, la méthode est bien convergente.

Nous nous intéressons ensuite à l'évolution de l'erreur au cours des itérations successives. Nous notons $r^{(k)} := -A\, x^{(k)} + b$, le résidu à l'étape k. La méthode de Richardson s'écrit alors

$$x^{(k+1)} \;=\; x^{(k)} \;+\; \gamma\, r^{(k)}.$$

Soit $e^{(k)} = x^{(k)} - x$ l'erreur à l'étape k, nous avons alors

$$\begin{aligned} r^{(k+1)} \;&=\; b \;-\; A\, x^{(k+1)}, \\ &=\; b \;-\; A\, (x^{(k)} \;+\; \gamma\, r^{(k)}), \\ &=\; b \;-\; A\, x^{(k)} \;-\; \gamma\, A\, r^{(k)}, \\ &=\; (I_n \;-\; \gamma\, A)\, r^{(k)}. \end{aligned}$$

Par récurrence, nous avons donc

$$r^{(k)} \;=\; (I_n \;-\; \gamma\, A)^k\, r^{(0)}.$$

Ainsi, pour le terme d'erreur, nous remarquons que

$$r^{(k)} \;=\; b \;-\; A\, x^{(k)} \;=\; A\, x \;-\; A\, x^{(k)} \;=\; -A e^{(k)}$$

ou encore

$$A\,e^{(k)} = (I_n \,-\, \gamma\,A)^k\,A\,e^{(0)}.$$

En remarquant que nous pouvons commuter la multiplication des matrices A^{-1} et $(I_n \,-\, \gamma\,A)$, nous avons finalement

$$e^{(k)} = (I_n \,-\, \gamma\,A)^k\,e^{(0)}$$

et observons que ce calcul permet d'estimer la vitesse de convergence de la méthode.

Pour terminer, nous choisissons $x^{(0)} = 0$ et donc $r^{(0)} = b$. La calcul de $x^{(k)}$ par itérations successives donne alors

$$\begin{aligned} x^{(k)} &= x^{(k-1)} \,-\, \gamma\,r^{(k-1)}, \\ &= x^{(k-2)} \,-\, \gamma\,r^{(k-2)} \,-\, \gamma\,r^{(k-1)}, \\ &= x^{(0)} \,-\, \gamma \sum_{l=0}^{k-1} r^{(l)}. \end{aligned}$$

Comme $x^{(0)} = 0$, nous obtenons donc

$$x^{(k)} \,=\, -\gamma \sum_{l=0}^{k-1} r^{(l)} \,=\, -\gamma\,\left(I_n + (I_n - \gamma A) + \ldots + (I_n - \gamma A)^{k-1}\right)\,b.$$

Par conséquent $x^{(k)}$ est une combinaison linéaire de b, $A\,b$, …, $A^{k-1}\,b$, nous disons que $x^{(k)}$ appartient à l'espace de Krylov $\mathcal{K}_k(A,b)$ donné par

$$\mathcal{K}_k(A,b) \,=\, Vect\left(b, A\,b, \ldots, A^{k-1}\,b\right).$$

Chapitre 2

Le calcul de valeurs propres

1 Mouvement de ressorts

Considérons un système de deux billes de masse unité reliées par trois ressorts de raideur unité. Notons $x_1(t)$ et $x_2(t)$ les positions des deux billes au temps t, par rapport à leur position d'équilibre. Soit $F_1(t)$, $F_2(t)$ et $F_3(t)$ les forces appliquées sur les billes dues aux forces de rappel des trois ressorts. Nous allons voir que l'étude des positions des billes au cours du temps

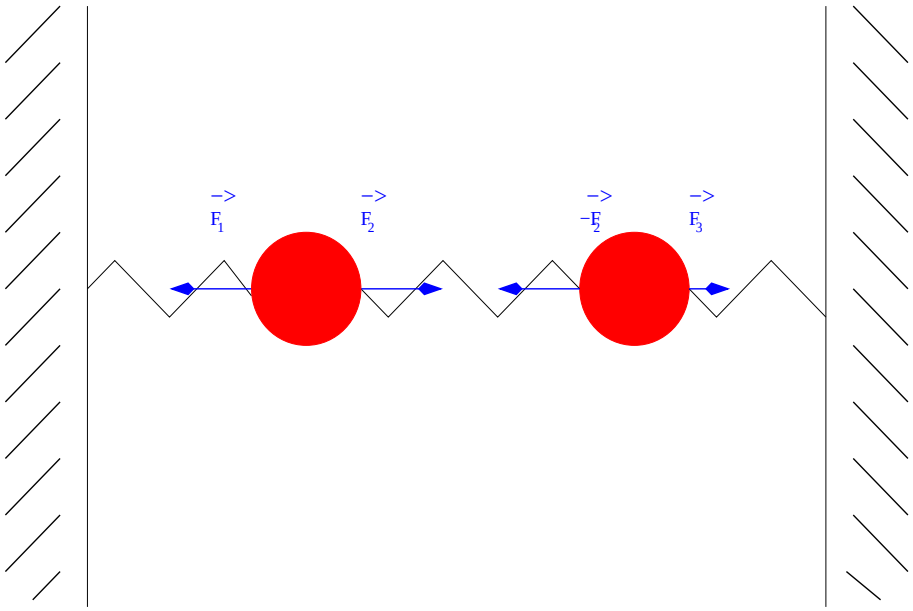

FIG. 2.1 – Étude du mouvements de deux billes maintenues par trois ressorts.

$x_1(t)$ et $x_2(t)$ se ramène à un calcul de valeurs propres. Pour cela, nous écrivons d'abord les équations de Newton pour les deux billes (*masse* $\times$ *accélération* = *forces extérieures*), nous

obtenons

$$\begin{cases} x_1''(t) \;=\; +F_1(t) \;+\; F_2(t), \\ \\ x_2''(t) \;=\; -F_2(t) \;+\; F_3(t). \end{cases}$$

En considérant le cas simplifié où les forces sont proportionnelles à l'allongement du ressort, nous avons

$$F_1(t) = -x_1(t), \quad F_2(t) = x_2(t) - x_1(t), \quad F_3(t) = -x_2(t),$$

et donc

$$\begin{cases} x_1''(t) = -2\,x_1(t) \;+\; x_2(t), \\ \\ x_2''(t) = x_1(t) \;-\; 2\,x_2(t). \end{cases}$$

Cette relation peut s'écrire sous forme matricielle en posant

$$A = \begin{pmatrix} +2 & -1 \\ \\ -1 & +2 \end{pmatrix}, \quad x(t) = \begin{pmatrix} x_1(t) \\ \\ x_2(t) \end{pmatrix},$$

les équations du mouvement deviennent alors

$$x''(t) \;+\; A\,x(t) \;=\; 0.$$

Le vecteur $x(t)$ est alors solution d'un système d'équations aux dérivées ordinaires linéaire qui peut être résolu de manière exacte. Les positions des billes $x_1(t)$ et $x_2(t)$ s'écrivent sous la forme

$$x_1(t) = \alpha_1\,\cos(\omega\,t), \quad x_2(t) = \alpha_2\,\cos(\omega\,t),$$

où les grandeurs α_1, α_2 et ω restent à déterminer. En substituant ces relations dans les équations du mouvement, il vient

$$\begin{cases} x_1''(t) = -\omega^2\,\alpha_1\,\cos(\omega\,t), \\ \\ x_2''(t) = -\omega^2\,\alpha_2\,\cos(\omega\,t), \end{cases}$$

nous obtenons après avoir simplifié par $\cos(\omega\,t)$

$$A \begin{pmatrix} \alpha_1 \\ \\ \alpha_2 \end{pmatrix} \;=\; \omega^2 \begin{pmatrix} \alpha_1 \\ \\ \alpha_2 \end{pmatrix}. \tag{1.1}$$

Le problème consiste donc à chercher α_1, et α_2 tels que les deux équations ci-dessus soient satisfaites.

Ainsi, la connaissance des deux valeurs propres et vecteurs propres de la matrice A, c'est-à-dire λ_1, $\lambda_2 \in \mathbb{R}$ et v_1, $v_2 \in \mathbb{R}^2$ tels que

$$A\,v_1 \;=\; \lambda_1\,v_1, \quad A\,v_2 \;=\; \lambda_2\,v_2.$$

permet de résoudre complètement le système différentiel. En effet, puisque A est symétrique définie positive, les valeurs propres λ_1 et λ_2 sont réelles strictement positives ; les grandeurs α_1, α_2 et ω solutions de (1.1) sont alors données par

$$\begin{pmatrix} \alpha_1 \\ \\ \alpha_2 \end{pmatrix} = v_1, \qquad \omega = \sqrt{\lambda_1}$$

ou

$$\begin{pmatrix} \alpha_1 \\ \\ \alpha_2 \end{pmatrix} = v_2, \qquad \omega = \sqrt{\lambda_2}.$$

Dans le cas présent $\lambda_1 = 3$ et $v_1 = (1,-1)^T$ ou $\lambda_2 = 1$ et $v_2 = (1,1)^T$.

Nous avons considéré un système de deux billes et trois ressorts et nous nous sommes ramenés à la résolution d'un problème de valeurs propres pour une matrice 2×2. Cependant, nous pouvons généraliser au cas du système de n billes et $n+1$ ressorts (n grand) que nous pourrons résoudre en recherchant les valeurs propres d'une matrice $n \times n$.

Souvent dans la pratique, étant donné une matrice, il n'est pas nécessaire de chercher toutes les valeurs propres et vecteurs propres de cette matrice. Par exemple, si nous admettons que la matrice est issue de l'étude de la réponse dynamique d'un pont, nous chercherons à déterminer les valeurs propres correspondant aux fréquences induites par des piétons marchant sur ce pont.

Dans la suite, nous mettons au point des algorithmes permettant le calcul des valeurs propres et des vecteurs propres associés.

Définition 1.1 *Soit $A \in \mathcal{M}_{n,n}(\mathbb{K})$. Nous appelons élément propre le couple formé d'une valeur propre et d'un vecteur propre associé : $(\lambda, v) \in \mathbb{K} \times \mathbb{K}^n$ tel que*

$$A\,v \;=\; \lambda\,v.$$

Dans la suite nous présentons quelques résultats intermédiaires sur la localisation des valeurs propres. Nous poursuivons avec la méthode de la puissance permettant d'approcher la plus grande des valeurs propres. Enfin, la méthode de Jacobi pour le calcul des valeurs propres et des vecteurs propres.

2 Localisation des valeurs propres

2.1 Approximation des valeurs propres

Commençons par enoncer un résultat facile à démontrer et qui pourra se réverler utile par la suite pour le calcul d'élément propre.

Théorème 2.1 (Gershgorine) *Soit $A \in \mathcal{M}_{n,n}(\mathbb{K})$. Si $\lambda \in Sp(A)$ est une valeur propre de A, alors il existe un indice i tel que*

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

c'est-à-dire que toutes les valeurs propres de $\lambda \in Sp(A)$ se trouvent dans l'union des disques $\mathcal{D}_i$

$$D = \cup_{i=1}^{n} \{\lambda \in \mathbb{K}; \quad |\lambda - a_{i,i}| \leq \sum_{j \neq i} |a_{i,j}|\}.$$

Démonstration. Soit $v \in \mathbb{K}^n$ un vecteur propre ($v \neq 0$), nous choisissons l'indice $i \in \{1, \ldots, n\}$ tel que $|v_i| \geq |v_j|$ pour tout $j \in \{1, \ldots, n\}$. La ligne i de l'équation $A\,v = \lambda\,v$ donne

$$\sum_{j=1}^{n} a_{i,j}\, v_j = \lambda\, v_i$$

ou encore

$$\sum_{\substack{j=1 \\ j \neq i}}^{n} a_{i,j}\, v_j = (\lambda - a_{i,i})\, v_i.$$

En divisant par $|v_i|$ et en utilisant l'inégalité triangulaire, nous obtenons

$$\sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{i,j}| \geq \left| \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{i,j} \frac{v_j}{|v_i|} \right| = |\lambda - a_{i,i}|.$$

Nous avons donc pour tout $\lambda \in Sp(A)$, il existe $i \in \{1, \ldots, n\}$ tel que

$$\lambda \in \mathcal{D}_i := \left\{ r \in \mathbb{K}, \quad |r - a_{i,i}| \leq \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{i,j}| \right\};$$

d'où

$$Sp(A) \subset \cup_{i=1}^{n} \mathcal{D}_i.$$

□

Ce théorème donne une première estimation de la localisation des valeurs propres de A. Cependant, il ne fournit pas un algorithme pour la calcul des valeurs propres. Avant de construire de tels algorithmes, étudions la sensibilité des valeurs propres par rapport aux coefficients de la matrice. Autrement dit, une petite erreur sur les coefficients de la matrice A implique-t-elle une forte variation des valeurs propres de A ?

Il arrive que suite à des erreurs d'arrondi, les coefficients d'une matrice A ne sont pas connus de manière exacte. Nous supposons que ces coefficients vérifient

$$\hat{a}_{i,j} = a_{i,j}\,(1 + \varepsilon_{i,j}),$$

avec $|\varepsilon_{i,j}| \leq \varepsilon$ (ε est de l'ordre de la précision de l'ordinateur, et est supposée être très petite). Lorsque nous voulons calculer une approximation des valeurs propres de la matrice A, il est très important d'étudier l'influence qu'aurons ces perturbations à chaque étape du calcul. Nous avons alors le résultat suivant

Théorème 2.2 *Soit $A \in \mathcal{M}_{n,n}(\mathbb{K})$ une matrice diagonalisable, c'est-à-dire il existe U avec*

$$U^{-1}\,A\,U \;=\; diag(\lambda_i),$$

où λ_i est valeur propre de A et soit $A(\varepsilon) = A + \varepsilon C$ donnée par

$$a_{i,j}(\varepsilon) = a_{i,j} + \varepsilon\, c_{i,j}.$$

Alors, pour chaque valeur propre $\lambda(\varepsilon)$ de $A(\varepsilon)$, il existe un λ_i avec

$$|\lambda(\varepsilon) \;-\; \lambda_i| \;\leq\; \varepsilon\, cond_\infty(U)\,\|C\|_\infty.$$

Démonstration. Soit $\lambda(\varepsilon)$ une valeur propre de la matrice approchée $A(\varepsilon)$. D'abord, nous transformons la matrice $A(\varepsilon)$ de manière à rendre la matrice A sous sa forme diagonale :

$$U^{-1}\,A(\varepsilon)\,U = U^{-1}\,A\,U \;+\; \varepsilon\,U^{-1}\,C\,U.$$

Nous notons par $e_{i,j}$ les éléments de $U^{-1}\,C\,U$ et appliquons le Théorème 2.1 à la matrice $U^{-1}\,A(\varepsilon)\,U$, ce qui implique l'existence d'un indice i tel que

$$|\lambda(\varepsilon) - (U^{-1}\,A(\varepsilon)\,U)_{i,i}| \;\leq\; \sum_{\substack{j=1\\ j\neq i}}^{n} |(U^{-1}\,A(\varepsilon)\,U)_{i,j}|.$$

Or, nous savons que $(U^{-1}\,A(\varepsilon)\,U)_{i,i} = \lambda_i + \varepsilon\,e_{ii}$, où λ_i est une valeur propre de A. L'inégalité triangulaire donne alors

$$|\lambda(\varepsilon) - \lambda_i| \leq |\lambda(\varepsilon) - (\lambda_i \;+\; \varepsilon\,e_{ii})| \;+\; \varepsilon\,|e_{ii}|.$$

En rassemblant les inégalités

$$\begin{aligned} |\lambda(\varepsilon) - \lambda_i| &\leq \varepsilon \sum_{\substack{j=1 \\ j\neq i}}^{n} |e_{i,j}| + \varepsilon\,|e_{ii}| \\ &= \varepsilon \sum_{j=1}^{n} |e_{i,j}| \\ &\leq \varepsilon \max_{1\leq i\leq n} \sum_{j=1}^{n} |e_{i,j}| \\ &= \varepsilon \|U^{-1}\,C\,U\|_\infty \leq \varepsilon\, cond_\infty(U)\,\|C\|_\infty, \end{aligned}$$

ce qui démontre l'affirmation du théorème. □

2.2 Ce qu'il ne faut pas faire

La première méthode (déjà utilisée par Lagrange) pour calculer les valeurs propres d'une matrice est la suivante :

- calculer d'abord les coefficients du polynôme caractéristique $P(\lambda)$
- déterminer ensuite les racines de ce polynôme

$$P(\lambda) = (-1)^n\lambda^n + a_{n-1}\lambda^{n-1} + \ldots + a_1\lambda + a_0.$$

Si la dimension de A est très petite (disons $n \leq 3$) ou si nous faisons le calcul en arithmétique exacte, cet algorithme peut être très utile. En revanche, si nous faisons le calcul avec des erreurs d'arrondi, cet algorithme peut donner des mauvaises surprises.

Considérons, par exemple, le problème de calculer les valeurs propres de la matrice diagonale

$$A = diag(1, \ldots, n),$$

dont le polynôme caractéristique est

$$\begin{aligned} P(\lambda) &= (1-\lambda)\,\ldots\,(n-\lambda), \\ &= (-1)^n\lambda^n + a_{n-1}\lambda^{n-1} + \ldots + a_1\lambda + a_0. \end{aligned}$$

Les coefficients calculés satisfont $\hat{a}_i = a_i(1+\varepsilon_i)$ avec $\varepsilon_i \leq \varepsilon \simeq 10^{-8}$. Cette perturbation $\hat{P}(\lambda)$ dans les coefficients provoque une grande erreur dans les racines du polynôme $P(\lambda)$. Les résultats numériques pour $n = 13$ sont dessinés dans la Figure 2.2.

Conclusion : Il faut éviter le calcul des coefficients du polynôme caractéristique. Un tel algorithme est numériquement instable.

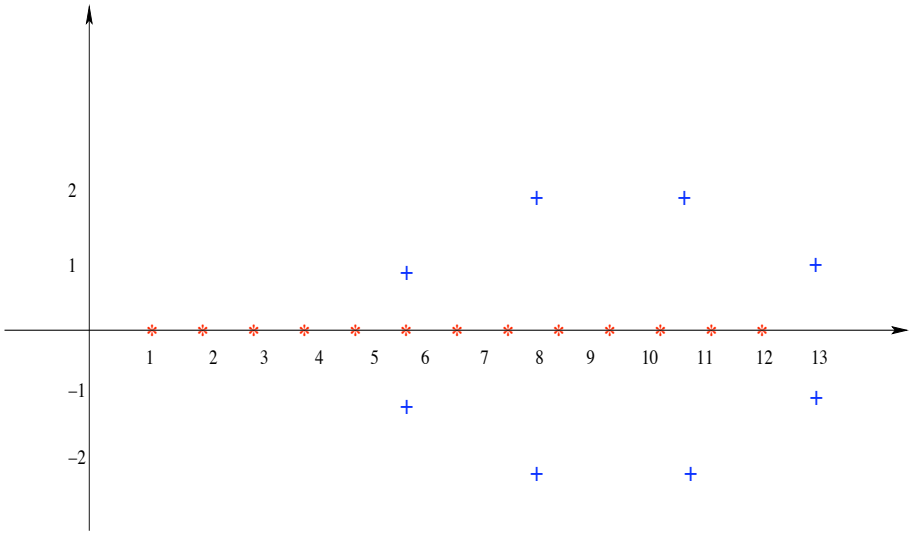

FIG. 2.2 – Résultats numériques des solutions de $P(\lambda) = (1-\lambda)\dots(n-\lambda) = 0$ et pour une petite perturbation des coefficients du polynôme pour $n = 13$.

3 Méthode de la puissance

3.1 L'algorithme

La méthode de la puissance est une méthode numérique qui permet de déterminer la valeur propre λ_1 de module maximal d'une matrice $A \in \mathcal{M}_{n,n}(\mathbb{R})$. Nous supposons que λ_1 est de multiplicité p, c'est-à-dire,

$$|\lambda_1| = \dots |\lambda_p| > |\lambda_{p+1}| \geq \dots \geq |\lambda_n|.$$

Nous choisissons un vecteur colonne $x_0 \in \mathbb{K}^n$ au hasard (en espérant que x_0 ne soit pas orthogonal à l'espace vectoriel $Ker(A - \lambda_1 I_n)$ et pour $k \geq 0$ calculons la suite récurrente $(x^{(k)})_{k\in\mathbb{N}}$ de la manière suivante

$$\begin{cases} x^{(0)} = x_0 \in \mathbb{K}^n, \\ x^{(k+1)} = A\,x^{(k)}, \quad k \in \mathbb{N}. \end{cases} \tag{3.2}$$

Notons bien que d'un point de vue pratique la méthode de la puissance ne peut pas être implémentée en l'état car l'algorithme pourrait conduire à un "overflow", c'est-à-dire que la suite prend des valeurs trop grande. Pour remédier à cela, il suffit de rajouter une étape de renormalisation du vecteur : on choisit $x^{(k+1)}$ tel que

$$\|x^{(k+1)}\| = 1. \tag{3.3}$$

Plus précisément, l'algorithme s'écrit sous la forme suivante

Algorithme 1. Méthode de la puissance.

Nous choisissons $x^{(0)} = x_0 \in I\!R^n$ et posons $\varepsilon = 1$, $z^{(0)} = x^{(0)}$, $k = 0$.

Tant que $\varepsilon \geq 1.10^{-8}$

-nous calculons d'abord $z^{(k+1)}$ tel que $z^{(k+1)} = A\, z^{(k)}$;

-nous normalisons le résultat $x^{(k+1)} = z^{(k+1)}/\|z^{(k+1)}\|$.

-nous calculons une approximation de la valeur propre

$$\beta^{(k+1)} = (x^{(k+1)})^T\, A\, x^{(k+1)}.$$

-nous effectuons un test d'arrêt : $\varepsilon = |\beta^{(k+1)} - \beta^{(k)}|$.

-nous itérons $k = k + 1$.

Fin de tant que

3.2 Un résultat de convergence

Nous avons le résultat de convergence suivant pour les matrices symétriques définies positives

Théorème 3.1 *Soit $A \in \mathcal{M}_{n,n}(I\!R)$ une matrice symétrique définie positive c'est-à-dire qu'elle vérifie $x^T\, A\, x > 0$ pour tout $x \neq 0_{I\!R^n}$. Si nous choisissons $x_0 \in I\!R^n$ tel que*

$$x_0 \notin Ker(A - \lambda_1\, I_n)^T.$$

Alors, la suite récurrente $(x^{(k)})_{k\in I\!N}$, fournie par l'algorithme (3.2)-(3.3), converge vers un vecteur propre $x \in I\!R^n$ associé à la valeur propre λ_1 :

$$A\, x = \lambda_1\, x.$$

De plus, $\beta^{(k)} = A\, x^{(k)} \cdot x^{(k)}$ converge vers $\lambda_1 = \rho(A)$.

Démonstration. Comme A est symétrique définie positive, A est diagonalisable dans $I\!R$ et il existe une base orthonormée de vecteurs propres $(v_i)_{1\leq i\leq n}$. Pour $1 \leq i \leq n$, nous notons par λ_i la valeur propre associée au vecteur propre v_i, c'est-à-dire

$$A\, v_i = \lambda_i\, v_i.$$

De plus, nous ordonnons les valeurs propres $\lambda_1 \geq \lambda_2 ... \geq \lambda_n > 0$ (les valeurs propres sont réelles et strictement positives puisque A est définie positive).

Nous écrivons d'abord le vecteur $x_0 \in I\!R^n$ dans la base $(v_i)_{1 \leq i \leq n}$

$$x_0 = \sum_{i=1}^{n} \alpha_i \, v_i$$

et donc

$$A \, x_0 = \sum_{i=1}^{n} \alpha_i \, \lambda_i \, v_i.$$

En appliquant une première fois l'algorithme de la puissance (3.2),

$$z^{(1)} \; = \; A \, x_0,$$

puis l'étape de normalisation donne encore

$$x^{(1)} \; = \; \frac{z^{(1)}}{\|z^{(1)}\|}.$$

En utilisant la décomposition de x_0 sur la base orthonormée formée des vecteurs propres $(v_i)_{1 \leq i \leq n}$ nous obtenons

$$z^{(1)} \; = \; A \, x_0 \; = \; \sum_{i=1}^{n} \alpha_i \, \lambda_i \, v_i.$$

Ainsi en appliquant successivement l'algorithme (3.2), nous avons à l'étape k

$$\left\{ \begin{array}{l} z^{(k)} \; = \; A \, z^{(k-1)} \; = \; A^k \, x_0 \; = \; \displaystyle\sum_{i=1}^{n} \alpha_i \, \lambda_i^k \, v_i, \\ \\ x^{(k)} \; = \; \displaystyle\frac{z^{(k)}}{\|z^{(k)}\|}. \end{array} \right.$$

D'une part en divisant par λ_1^k, la première ligne donne

$$\frac{z^{(k)}}{\lambda_1^k} \; = \; \sum_{i=1}^{n} \alpha_i \, \left(\frac{\lambda_i}{\lambda_1} \right)^k \, v_i.$$

D'autre part, puisque A est définie positive, nous savons que $-\lambda_1$ n'est pas valeur propre, nous déduisons que

$$\lim_{k \to \infty} \left(\frac{\lambda_i}{\lambda_1} \right)^k = 0, \quad \text{pour} \quad \lambda_i \neq \lambda_1.$$

Nous notons alors par $p \in \{1, .., n\}$ l'entier tel que $\lambda_1 = .. = \lambda_p$ et $\lambda_{p+1} < \lambda_1$; nous avons donc

$$\lim_{k \to \infty} \frac{z^{(k)}}{\lambda_1^k} = \sum_{i=1}^{p} \alpha_i \, v_i \; =: \; z$$

mais aussi puisque la norme est une application continue

$$\lim_{k\to\infty} \frac{\|z^{(k)}\|}{\lambda_1^k} \;=\; \|z\|.$$

Nous posons alors $x \;=\; z/\|z\|$ et obtenons que la limite du produit est bien le produit des limites

$$\lim_{k\to\infty} x^{(k)} = \lim_{k\to\infty} \frac{z^{(k)}}{\lambda_1^k}\,\frac{\lambda_1^k}{\|z^{(k)}\|} \;=\; \frac{z}{\|z\|} \;=\; x.$$

Étant donné que les vecteurs $(v_i)_{1\leq i\leq p}$ sont des vecteurs propres associés à la valeur propre λ_1, nous vérifions que

$$A\,x = \lambda_1\,x.$$

Enfin, nous vérifions bien que $x \neq 0_{I\!R^n}$ puisque $x_0 \notin Ker(A-\lambda_1 I_n)^{\perp} = vect\{v_{p+1},..,v_n\}$ et donc il existe $1 \leq i_0 \leq p$ tel que $\alpha_{i_0} \neq 0$.

Enfin, nous posons $\beta^{(k)} = Ax^{(k)} \cdot x^{(k)}$ et montrons que la suite $(\beta^{(k)})_{k\in I\!N}$ converge bien vers λ_1. En effet, puisque $(x^{(k)})$ converge vers le vecteur propre x associé à la valeur propre λ_1, nous avons

$$A\,x^{(k)} \longrightarrow A\,x \;=\; \lambda_1\,x, \quad \text{lorsque}\, k\to\infty,$$

et donc $\beta^{(k)} = (x^{(k)})^T\,A\,x^{(k)}$ converge bien vers $\lambda_1\|x\|^2 = \lambda_1$ lorsque k tend vers l'infini. □

Nous pouvons rencontrer des problèmes de convergence en utilisant la méthode de la puissance lorsque la valeur propre de module maximal n'est pas unique.

Exemple 3.1 *Lorsque* $|\lambda_1| = |\lambda_2|$ *avec* $\lambda_2 = -\lambda_1$ *ou* $\lambda_2 = \overline{\lambda}_1$*, la méthode ne converge pas. Nous pouvons constater certaines de ces difficultés en observant, par exemple sous* `Matlab`*, le comportement de la méthode pour les exemples suivants*

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & -1 \end{pmatrix}.$$

Pour le cas de la première matrice, tout se passe bien l'ensemble des valeurs propres est donné par $\{1,1\}$*. En revanche, dans le deuxième cas, les valeurs propres sont* $\{-1,-1,1,1\}$ *et la*

méthode de la puissance donne

k	$x^{(k)}$	$\|x^{(k)}\|$
0	$(1,\ 1,\ 1,\ 1)^T$	2.00
1	$(2,\ 1,-1,-1)^T$	2.65
2	$(3,\ 1,\ 1,\ 1)^T$	3.46

k	$x^{(k)}$	$\|x^{(k)}\|$
3	$(4,\ 1,-1,-1)^T$	4.36
4	$(5,\ 1,\ 1,\ 1)^T$	5.29
5	$(6,\ 1,-1,-1)^T$	6.24

Ici $x^{(k)}$ ne converge pas vers un vecteur propre de A même si le rapport des normes $\|x^{(k+1)}\|/\|x^{(k)}\|$ converge vers 1, le module de la plus grande valeur propre, lorsque k tend vers l'infini.

3.3 Méthode de la puissance inverse

Nous présentons une méthode pour calculer une approximation de la plus petite des valeurs propres en valeur absolue, c'est la méthode de la puissance inverse. Puis, nous fournissons l'algorithme pour calculer une approximation des autres valeurs propres.

Si A est inversible, les valeurs propres de A^{-1} sont les inverses des valeurs propres de A. La méthode de la puissance peut donc aussi servir à calculer la valeur propre de plus petit module en l'appliquant à A^{-1}. Cela se fait sans calcul d'inverse mais à partir d'une factorisation $L\,U$ faite une fois pour tout en début d'algorithme. L'algorithme est alors le suivant

Algorithme 2. Méthode de la puissance inverse.

Nous effectuons une décomposition $L\,U$ de la matrice A

(ou $L^T\,L$ si A est symétrique).

Nous choisissons $x^{(0)} = x_0 \in I\!R^n$ et posons $\varepsilon = 1$, $z^{(0)} = x^{(0)}$, $k = 0$

Tant que $\varepsilon \geq 1.\,10^{-8}$

-nous calculons d'abord $z^{(k+1)}$ tel que $A\,z^{(k+1)} = z^{(k)}$;

-nous normalisons le résultat $x^{(k+1)} = z^{(k+1)}/\|z^{(k+1)}\|$.

-nous calculons une approximation de la valeur propre

$$\beta^{(k+1)} = (x^{(k+1)})^T\,A\,x^{(k+1)}.$$

-nous effectuons un test d'arrêt : $\varepsilon = |\beta^{(k+1)} - \beta^{(k)}|$.

-nous itérons $k = k + 1$.

Fin de tant que

Cet algorithme converge sous les mêmes conditions que le précédent ; la valeur limite est cette fois la valeur propre de plus petit module $|\lambda_n|$. Nous avons en particulier le résultat suivant

Théorème 3.2 *Soit $A \in \mathcal{M}_{n,n}(I\!R)$ une matrice symétrique définie positive c'est-à-dire qu'elle vérifie $x^T\,A\,x > 0$ pour tout $x \neq 0_{I\!R^n}$. Si nous choisissons $x_0 \in I\!R^n$ tel que*

$$x_0 \notin Ker(A - \lambda_n\,I_n)^T.$$

Alors, la suite récurrente $(x^{(k)})_{k\in I\!N}$, définie par la méthode de la puissance inverse, converge vers un vecteur propre $x \in I\!R^n$ associé à la valeur propre λ_n : $A\,x = \lambda_n\,x$. De plus, $\beta^{(k)} = A\,x^{(k)} \cdot x^{(k)}$ converge vers λ_n =.

Démonstration. Le preuve s'inspire directement de celle du Théorème 3.1.

□

4 Méthode de Jacobi

Nous cherchons à déterminer numériquement les valeurs propres et vecteurs propres d'une matrice A symétrique réelle ($A^T = A$). Nous savons qu'une telle matrice est diagonalisable, c'est-à-dire qu'il existe une matrice réelle U telle que $D = U^{-1}\,A\,U$ est diagonale, la diagonale étant composée des valeurs propres de A. Comme A est symétrique, U est orthogonale, c'est-à-dire ($U^{-1} = U^T$ la transposée de U), et $D = U^T\,A\,U$. La diagonalisation consiste donc à trouver la matrice U, c'est-à-dire trouver une base dans laquelle la représentation de A est diagonale.

4.1 Cas de la dimension deux

Soit $A \in \mathcal{M}_{2,2}(\mathbb{K})$ une matrice symétrique et P une matrice de rotation

$$A = \begin{pmatrix} \alpha & \beta \\ \beta & \gamma \end{pmatrix}, \quad P = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix}.$$

Nous décidons de choisir θ pour que la matrice $P\,AP^T$ soit diagonale.

Pour cela nous notons $A^{(0)} = A$ et calculons

$$A^{(1)} = P\,A^{(0)}\,P^T = \begin{pmatrix} \alpha' & \beta' \\ \beta' & \gamma' \end{pmatrix},$$

il suffit donc de chercher pour quel angle $\theta \in [0, 2\pi]$, le coefficient β' devient nul.

D'une part, puisque une rotation est une matrice orthogonale, cette transformation conserve la norme (nous vérifions facilement que le produit d'une matrice orthogonale par un vecteur préserve la norme du vecteur) et en particulier la norme de Froebenus (définie Chapitre 1)

$$\alpha'^2 + 2\,\beta'^2 + \gamma'^2 = \alpha^2 + 2\,\beta^2 + \gamma^2.$$

Ainsi pour un coefficient $\beta' = 0$, nous avons

$$\alpha'^2 + \gamma'^2 \geq \alpha^2 + \gamma^2.$$

La méthode de Jacobi consiste donc à accroître la somme des carrés des termes diagonaux et à diminuer la somme des carrés des termes extra-diagonaux. En dimension deux, une seule étape de la méthode de Jacobi suffit à diagonaliser la matrice A, la matrice A' contient les valeurs propres de A et les vecteurs colonnes de P fournissent les vecteurs propres de la matrice A. Il suffit de choisir $\theta \in [0, 2\pi]$ tel que

$$\coth(2\,\theta) = \frac{\cos^2\theta - \sin^2\theta}{2\,\cos\theta\,\sin\theta} = \frac{\alpha - \gamma}{2\,\beta}.$$

Voyons maintenant comment nous pouvons étendre cette méthode en dimension supérieure.

4.2 Cas général

La méthode de Jacobi consiste à écrire la matrice U sous forme d'un produit de matrices de rotation, chaque rotation étant choisie de façon à annuler des éléments non diagonaux de la représentation de A. Une rotation, d'un angle dans le plan défini par les vecteurs d'indices p et q, est donnée par la matrice orthogonale (*rotation de Givens*)

$$(P_{pq})_{i,j} \;=\; \begin{cases} 1 & \text{si } i=j, \text{ avec } j\neq p, j\neq q, \\ c & \text{si } i=j=p, \text{ ou }, i=j=q, \\ s & \text{si } i=p, \text{ et }, j=q, \\ -s & \text{si } i=q, \text{ et }, j=p, \\ 0 & \text{sinon,} \end{cases}$$

où $c=\cos\theta$ et $s=\sin\theta$. Nous avons aussi

$$P_{pq} = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & \ddots & \ddots & & & & & \vdots \\ \vdots & \ddots & c & \ddots & & s & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & & & \vdots \\ \vdots & & & \ddots & 1 & \ddots & & \vdots \\ \vdots & & -s & & \ddots & c & \ddots & \vdots \\ \vdots & & & & & \ddots & \ddots & 0 \\ 0 & \dots & \dots & \dots & \dots & \dots & 0 & 1 \end{pmatrix}.$$

La matrice P_{pq} ainsi construite satisfait les propriétés suivantes.

Lemme 4.1 *La matrice P_{pq} vérifie*

(i) pour la norme de Froebenus donnée par $\|A\|_F=\sqrt{tr(A^*A)}$, nous avons

$$\|P_{pq}\|_F^2 \;=\; \sum_{i,j=1}^{n} |(P_{pq})_{i,j}|^2 \;=\; n.$$

(ii) la matrice P_{pq} est orthogonale, c'est-à-dire $P_{pq}\,P_{pq}^T = I_n$

Démonstration. Le premier résultat est évident

$$\|P_{pq}\|_F^2 = (n-2)\times 1^2 + 2\times(c^2+s^2) = n.$$

D'autre part le produit $P_{pq}\,P_{pq}^T$ s'écrit

$$\begin{pmatrix} 1 & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & \ddots & \ddots & & & & & \vdots \\ \vdots & \ddots & c & \ddots & & s & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & & & \vdots \\ \vdots & & & \ddots & 1 & \ddots & & \vdots \\ \vdots & & -s & & \ddots & c & \ddots & \vdots \\ \vdots & & & & & \ddots & \ddots & 0 \\ 0 & \dots & \dots & \dots & \dots & \dots & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & \ddots & \ddots & & & & & \vdots \\ \vdots & \ddots & c & \ddots & & -s & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & & & \vdots \\ \vdots & & & \ddots & 1 & \ddots & & \vdots \\ \vdots & & s & & \ddots & c & \ddots & \vdots \\ \vdots & & & & & \ddots & \ddots & 0 \\ 0 & \dots & \dots & \dots & \dots & \dots & 0 & 1 \end{pmatrix},$$

ce qui donne bien $P_{pq}\,P_{pq}^T = I_n$. □

Nous posons alors $A^{(0)} = A$ puis pour $k \geq 0$ nous calculons par récurrence $A^{(k+1)}$ comme la rotation de $A^{(k)}$ par la matrice de rotation P_{pq}

$$A^{(k+1)} = P_{pq}\,A^{(k)}\,P_{pq}^T.$$

La représentation de la matrice $A^{(k+1)} = (a_{i,j}^{(k+1)})_{1\leq i,j\leq n}$ (qui reste symétrique) dans la nouvelle base obtenue après rotation s'écrit :

$$a_{i,j}^{(k+1)} = \begin{cases} a_{i,j}^{(k)} & \text{si } j \neq p,q \text{ et } i \neq p,q, \\ c\,a_{i,p}^{(k)} - s\,a_{i,q}^{(k)} & \text{si } j = p \text{ avec } i \neq p,q, \\ c\,a_{i,q}^{(k)} + s\,a_{i,p}^{(k)} & \text{si } j = q \text{ avec } i \neq p,q, \\ c^2\,a_{p,p}^{(k)} + s^2\,a_{q,q}^{(k)} - 2s\,c\,a_{p,q}^{(k)} & \text{si } i = p \text{ et } j = p, \\ s^2\,a_{p,p}^{(k)} + c^2\,a_{q,q}^{(k)} + 2s\,c\,a_{p,q}^{(k)} & \text{si } i = q \text{ et } j = q, \\ (c^2 - s^2)\,a_{p,q}^{(k)} + c\,s\,(a_{p,p}^{(k)} - a_{q,q}^{(k)}) & \text{si } (i,j) = (p,q) \text{ ou } (i,j) = (q,p). \end{cases}$$

Nous observons que seuls les éléments sur les lignes et colonnes p et q sont modifiés. L'idée consiste alors à choisir l'angle de la rotation de façon à annuler le terme $a_{p,q}^{(k+1)}$, ce qui conduit à

$$\coth(2\,\theta) \;=\; \frac{c^2 \;-\; s^2}{2\,c\,s} = \frac{a_{q,q}^{(k)} \;-\; a_{p,p}^{(k)}}{2\,a_{p,q}^{(k)}},$$

mais ce choix modifie les autres éléments non diagonaux.

Nous résumons les modifications induites par la rotation dans la proposition suivante

Proposition 4.1 *Supposons que la matrice $A^{(0)} \;=\; A$ est symétrique. Alors, pour tout $k \geq 0$ nous construisons la matrice $A^{(k+1)} = P_{pq}^T\, A^{(k)}\, P_{pq}$, laquelle est symétrique et vérifie*

$$\|A^{(k+1)}\|_F \;=\; \|A^{(k+1)}\|_F$$

D'autre part

$$S^{(k+1)} \;=\; \sum_{i=1}^{n} [a_{i,i}^{(k+1)}]^2 \;=\; S^{(k)} + 2\,[a_{p,q}^{(k)}]^2 \geq S^{(k)}.$$

Démonstration. D'une part, nous rappelons que

$$\|A\|_F^2 = tr(A\,A^*)$$

et donc dans le cas d'une matrice A à coefficient réel et symétrique et pour U une matrice orthogonale nous avons

$$\|AU\|_F^2 = tr(AU\,(AU)^T) = tr(AU\,U^T\,A^T) = tr(A\,A^T) = \|A\|_F^2,$$

d'où le résultat pour $A^{(k+1)}$ et $A^{(k)}$

$$\|A^{(k)}\|_F \;=\; \|A^{(k+1)}\|_F.$$

Ensuite, en calculant la somme des éléments extra-diagonaux au carré, il vient

$$\begin{aligned}
\sum_{\substack{i,j=1\\ i\neq j}}^{n} |a_{i,j}^{(k+1)}|^2 \;&=\; \sum_{\substack{i=1\\ i\neq p,q}}^{n} \left(\sum_{\substack{j=1\\ j\neq i,p,q}}^{n} |a_{i,j}^{(k)}|^2 \;+\; (c\,a_{i,p}^{(k)} \;-\; s\,a_{i,q}^{(k)})^2 + (c\,a_{i,q}^{(k)} \;+\; s\,a_{i,p}^{(k)})^2 \right) \\
&=\; \sum_{\substack{i,j=1\\ i\neq j}}^{n} |a_{i,j}^{(k)}|^2 \;-\; 2\,[a_{p,q}^{(k)}]^2.
\end{aligned}$$

En combinant ce dernier résultat avec la conservation de la norme de Froebenus, nous avons donc

$$S^{(k+1)} \;=\; S^{(k)} \;+\; 2\,[a_{p,q}^{(k)}]^2.$$

□

Nous voyons donc que la somme $S^{(k)}$ croît lorsque l'itération k augmente et est bornée par la somme des carrés de tout les terme c'est-à-dire la norme de Froebenus de $\|A\|_F^2$.

Algorithme 3. Méthode de Jacobi pour le calcul de valeurs propres.

$A^{(0)} = A, \quad k = 0$

Tant que $\max_{p \neq q} |a^{(k)}_{p,q}| \geq \varepsilon$

-nous cherchons $p_0, q_0 \in \{1, .., n\}$ tel que $p_0 \neq q_0$

$$|a^{(k)}_{p_0,q_0}| = \max\{|a^{(k)}_{p,q}|, p, q \in \{k, .., n\}, q \neq q\}$$

-nous construisons la matrice orthogonale $P_{p_0 q_0}$

de rotation d'angle θ tel que

$$\coth(2\,\theta) \;=\; \frac{c^2 \;-\; s^2}{2\,c\,s} = \frac{a^{(k)}_{q_0,q_0} \;-\; a^{(k)}_{p_0,p_0}}{2\,a^{(k)}_{p_0,q_0}}.$$

-nous construisons la matrice $A^{(k+1)} = P^T_{p_0 q_0}\, A^{(k)}\, P_{p_0 q_0}$.

-nous itérons $k = k + 1$.

Fin de Tant que.

Ainsi, en itérant plusieurs fois ce procédé tout en effectuant le bon choix (p, q), cela nous permet de faire tendre tous les éléments non diagonaux vers 0, et de rendre la matrice diagonale.

Il reste à déterminer les vecteurs propres. Les valeurs propres (réelles) et vecteurs propres x de la matrice A sont tels que $A\,x = \lambda\,x$. Pour une matrice diagonale, les vecteurs propres sont les vecteurs de la base. Il faut trouver leur expression dans la base initiale de la matrice A. Pour une matrice de rotation P, posons

$$y \;=\; P^T_{pq}\,x$$

alors

$$A^{(k+1)}\,y \;=\; P^T_{pq}\,A^{(k)}\,P_{pq}\,P^T_{pq}\,x \;=\; P^T_{pq}A^{(k)}\,x \;=\; \lambda\,P^T_{pq}\,x \;=\; \lambda\,y$$

et y est vecteur propre de $A^{(k)}$ associé à la valeur propre λ. Les vecteurs propres de A s'obtiennent donc par itérations successives.

5 Complément du Chapitre 2 : les valeurs propres du Laplacien

Nous nous intéressons au problème de vibration d'une membrane en cherchant des solutions (u, λ) telles que u et λ vérifient l'équation de Laplace suivante en dimension une sur l'intervalle

$[0,1]$. Nous résolvons alors : trouver une valeur $\lambda \in I\!R^+$ et une fonction u telles que

$$-u''(x) = \lambda\, u(x),$$

avec les conditions aux limites $u(0) = u(1) = 0$.

Nous pouvons résoudre explicitement ce problème linéaire en utilisons les séries de Fourier et les fonctions de bases $x \rightarrow \sin(k\,\pi)$ et trouvons

$$\lambda_k = k^2\,\pi^2, \qquad \varphi_k(x) = C_k\,\sin(k\,\pi),$$

la constante C_k vient du fait que si u est solution alors $\alpha\, u$ est également solution du problème au valeur propre. Il faut rajouter une condition de normalisation pour fixer cette constante, comme par exemple

$$\|u\|_\infty = 1.$$

Remarque 5.1 *Dans le cas où Ω est un carré $[0,1] \times [0,1]$. Nous vérifions en utilisant la technique des variables séparées que les fonctions*

$$\varphi_{n,m}(x,y) = \sin(n\pi\, x)\,\sin(m\pi\, y)\,, \quad (x,y) \in [0,1]\times[0,1]$$

sont identiquement nulles sur le bords et qu'elles sont vecteurs propres du Laplacien.

Nous admettons ici que la famille $(u_k, \lambda_k)_{k\in I\!N}$ donne tous les modes propres du Laplacien sur l'intervalle $[0,1]$. Nous allons nous servir de cette solution exacte pour **tester** les algorithmes numériques. En effet, cette étape est primordiale lorsque nous nous intéressons au développement d'un algorithme numérique, il faut se servir des solutions exactes pour tester la taille de l'erreur en comparant la solution numérique et la solution exacte lorsque nous pouvons le faire.

Voyons d'abord comment résoudre ce problème d'un point de vue numérique. Nous utilisons une méthode de différences finies basée sur un développement de Taylor de $u(x)$ comme dans le Chapitre 1. Nous verrons plus tard au Chapitre 7, comment nous effectuons l'analyse de la convergence des méthodes aux differences finies. Pour l'instant, interessons nous simplement à la mise au point de l'algorithme. Nous notons $x_i = i\,h$ avec $h = 1/(n+1)$ et n est le nombre total de points de discrétisation

$$u(x_{i+1}) \;=\; u(x_i) \;+\; h\,u'(x_i) \;+\; \frac{h^2}{2}\,u''(x_i) \;+\; \frac{h^3}{6}\,u'''(x_i) \;+\; O(h^3)$$

et

$$u(x_{i-1}) \;=\; u(x_i) \;-\; h\,u'(x_i) \;+\; \frac{h^2}{2}\,u''(x_i) \;-\; \frac{h^3}{6}\,u'''(x_i) \;+\; O(h^3).$$

En sommant ces deux égalités, en divisant par h^2 et en remplaçant $u(x_i)$ par son approximation u_i, nous obtenons

$$u''(x_i) \;\simeq\; \frac{u_{i+1} \;-\; 2\,u_i \;+\; u_{i-1}}{h^2} \;+\; O(h^2).$$

Nous posons alors $u_h = (u_1, \ldots, u_n) \in I\!R^n$ et A la matrice symétrique réelle de $\mathcal{M}_{n,n}(I\!R)$ définie par

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

Nous négligeons alors les termes en $O(h^2)$ et considérons le problème discret suivant

$$\begin{cases} \frac{1}{h^2} A\, u_h \;=\; \lambda_h\, u_h & u \neq 0, \\ u_h \in I\!R^n & \lambda_h \in I\!R^+, \end{cases} \tag{5.4}$$

avec la convention $u_0 = u_{n+1} = 0$ pour approcher les conditions de bords.

Puisque la matrice A est une matrice symétrique définie positive, nous pouvons utiliser la méthode de la puissance pour calculer la plus grande valeur propre. D'autre part, nous savons aussi qu'il existe une base orthonormée de vecteur propre de la matrice A.

Voyons maintenant comment nous pouvons calculer une approximation de l'ensemble des valeurs propres de la matrice A. Nous supposons que l'élément propre (λ_1, v_1) est connu et vérifie

$$\frac{1}{h^2} A\, v_1 \;=\; \lambda_1\, v_1,$$

où λ_1 est la plus grande valeur propre en module et nous choisissons v_1 tel que $\|v_1\| \;=\; 1$.

La méthode va consister à construire une suite $(\lambda_2^{(k)}, v_2^{(k)})_{k \in I\!N} \subset I\!R^+ \times I\!R^n$, telle que pour tout $k \in I\!N$, le vecteur $v_2^{(k)}$ soit orthogonal au vecteur propre approché v_1. Ainsi, nous prenons

$$x^{(0)} \;=\; x_0 \;-\; (x_0, v_1)\, v_1.$$

Puis, nous construisons un vecteur orthogonale à v_1

$$z^{(1)} \;=\; A\, x^{(0)} - (A\, x^{(0)}, v_1)\, v_1$$

et l'étape de normalisation

$$v_2^{(1)} \;=\; \frac{z^{(1)}}{\|z^{(1)}\|}.$$

En réitérant ce procédé, nous construisons une suite $(v_2^{(k)})_{k \in I\!N}$ qui converge vers un vecteur propre de la matrice $\frac{1}{h^2} A$, ce vecteur propre est associé à la deuxième valeur propre la plus grande en module.

Nous proposons alors l'algorithme suivant qui permet le calcul de λ_i, la i-ème plus grande valeur propre de $\frac{1}{h^2}\,A$ en module.

Algorithme 4. Calcul de l'ensemble des valeurs propres

-Nous choisissons $x_0 \in I\!R^n$

Pour $i = 1, \ldots, n$

nous calculons

$$v_i^{(0)} \;=\; x_0 - \sum_{l=1}^{i-1} (x_0, v_l)\, v_l.$$

et posons $k = 0$, $z^{(0)} = v_i^{(0)}$, $\varepsilon^{(0)} = 1$.

Tant que $\varepsilon^{(k)} \geq \varepsilon$

-nous calculons $\tilde{z}^{(k+1)} \;=\; A\, z^{(k)}$,

-nous obtenons $z^{(k+1)} \;=\; \tilde{z}^{(k+1)} \;-\; \sum_{l=1}^{i-1} (\tilde{z}^{(k+1)}, v_l)\, v_l$,

-nous avons alors

$$v_i^{(k+1)} \;=\; \frac{z^{(k+1)}}{\|z^{(k+1)}\|}$$

et $\lambda_i^{(k+1)} \;=\; (A\, v_i^{(k+1)}, v_i^{(k+1)})$,

-nous effectuons un test d'arrêt

$$\varepsilon^{(k+1)} := |\lambda_i^{(k+1)} \;-\; \lambda_i^{(k)}| \;\leq\; \varepsilon\, |\lambda_i^{(k)}|,$$

où ε est fixé et petit 10^{-6}.

-nous itérons $k = k + 1$.

Fin de pour k

Nous posons alors $v_i = v_i^{(k)}$ et $\lambda_i = \lambda_i^{(k)}$.

Fin de pour i

Nous avons montrer que cet algorithme converge et que $\lambda^{(k)}$ tend vers la plus grande valeur propre de A tandis que le vecteur $x^{(k)}$ tend vers le vecteur propre associé.

Nous pouvons vérifier que la matrice A symétrique et définie positive elle est donc inversible. De plus, nous calculons exactement ses éléments propres donnés par

$$\lambda^{(k)} = \frac{4}{h^2} \sin^2\left(\frac{k\,\pi}{2(n+1)}\right)$$

et

$$v_j^{(k)} = \sin\left(\frac{j\,k\,\pi}{(n+1)}\right), \quad j = 1, ..., n.$$

Ainsi, nous pouvons estimer la précision de l'algorithme en comparant les valeurs propres de A avec les valeurs données par l'algorithme de la puissance.

Chapitre 3

Les systèmes non linéaires

1 Introduction aux problèmes non linéaires

Nous commençons par un exemple simple de problème non linéaire.

1.1 Motivation : le remplissage d'un réservoir

Nous souhaitons construire un réservoir sphérique contenant la quantité d'eau potable nécessaire pour "alimenter" une population de 30 personnes pendant une semaine. Nous voulons donc calculer la hauteur d'eau possible sachant que le rayon R de la sphère est donné et le réservoir doit pouvoir contenir un volume d'eau de $500\,l \;=\; 0.5\,m^3$.

Le volume d'eau à l'intérieur du réservoir est alors donné par

$$V = \int_{R-h}^{R} \int_{\{x^2+y^2 \leq R^2-z^2\}} dx\,dy\,dz \;=\; \pi\,\frac{h^2(3\,R-h)}{3}.$$

La hauteur d'eau h est alors la solution strictement positive de l'équation

$$h^3 \;-\; 3\,R\,h \;+\; \frac{3\,V}{\pi} \;=\; 0.$$

Hélàs, nous ne savons pas résoudre de manière explicite une équation aussi simple. Nous recherchons alors une valeur approchée et avons recours pour cela à des algorithmes numériques pour résoudre de telles équations non linéaires.

Plus généralement, pour une fonction $f : x \in K \subset I\!R^n \rightarrow I\!R^n$, nous cherchons à déterminer les solutions $\bar{x} \;\in\; K \subset I\!R^n$ telles que

$$f(\bar{x}) = 0.$$

Ce problème est non linéaire dans le sens ou la fonction f n'est plus seulement une fonction du type $f(x) = A\,x - b$. Nous ne pouvons donc pas utiliser les algorithmes développés dans la partie précédente mais nous verrons que la stratégie est proche.

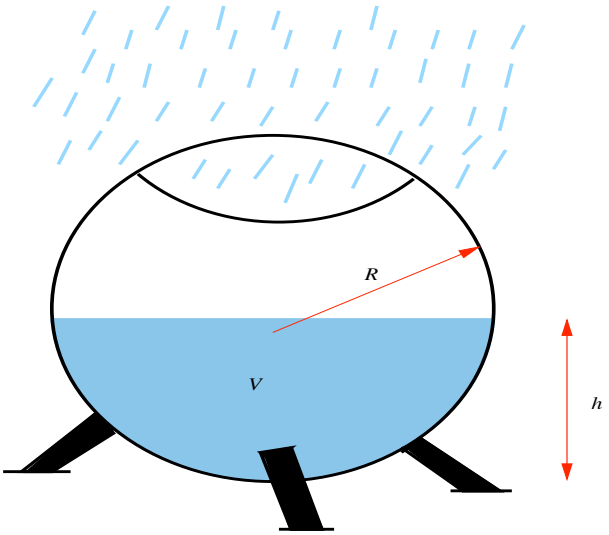

FIG. 3.1 – Calcul de la hauteur d'eau h permettant de contenir un volume d'eau V donné.

Le plan du chapitre est le suivant. Dans une première partie, nous présentons quelques résultats généraux sur l'existence de solutions pour une équation non linéaire posée dans $\mathbb{R}$ ou un intervalle $[a, b]$ de $\mathbb{R}$. Puis, nous proposons des algorithmes pour la résolution numérique de problèmes non linéaires dans le cas de fonctions à valeurs dans $\mathbb{R}$. Ensuite, dans le cas $\mathbb{R}^n$, nous aurons besoin de fonctions régulières, c'est-à-dire différentiables, nous rappelons donc quelques résultats classiques de calcul différentiel pour des fonctions $f : \mathbb{R}^n \to \mathbb{R}^n$. Nous présentons enfin la méthode de Newton pour approcher les solutions $\bar{x} \in \mathbb{R}^n$ de $f(\bar{x}) = 0_{\mathbb{R}^n}$.

1.2 Résultats généraux et définitions

Soit $f : K \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ une fonction continue. Nous nous intéressons à la résolution de l'équation suivante

$$\begin{cases} \text{Trouver } \bar{x} \in K, \\ f(\bar{x}) = 0. \end{cases} \tag{1.1}$$

Avant de procéder à la présentation de méthodes numériques pour approcher la solution de (1.1), nous rappelons le théorème des valeurs intermédaires qui assure l'existence d'une solution lorsque $n = 1$

Théorème 1.1 (Théorème des valeurs intermédiares) *Supposons que $f : [a, b] \to \mathbb{R}$ est une fonction continue et k un réel entre $f(a)$ et $f(b)$. Alors, il existe un réel $\xi \in [a, b]$ tel que $k = f(\xi)$. En particulier, si $f(a)\, f(b) \leq 0$, alors l'équation $f(x) = 0$ admet au moins une solution.*

Nous verrons plus tard que l'application successive de ce théorème est à la base de la méthode de dichotomie. Pour l'instant, indiquons simplement que les algorithmes que nous mettons au point sont basés sur des méthodes itératives comme nous l'avons fait pour les systèmes linéaires, c'est-à-dire nous construisons une suite d'approximation $(x^{(k)})_{k\in\mathbb{N}}$ telle que la limite $\bar{x} = \lim_{k\to\infty} x^{(k)}$ soit solution de (1.1). Avant tout, nous précisons ce que signifie la convergence de la suite $(x^{(k)})_{k\in\mathbb{N}}$, puis nous introduisons la notion d'ordre de convergence.

Définition 1.1 *Soit $(x^{(k)})_{k\in\mathbb{N}}$ une suite d'approximation de la solution $\bar{x}$ de (1.1). Nous disons que la suite $(x^{(k)})_{k\in\mathbb{N}}$ converge vers $\bar{x}$ si*

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

De plus, nous disons que la méthode itérative fournissant les valeurs de $(x^{(k)})_{k\in\mathbb{N}}$ est d'ordre p, s'il existe deux constantes C_1 et $C_2 > 0$ telles que

$$C_1 \leq \lim_{p\to\infty} \frac{\|x^{(k+1)} - \bar{x}\|}{\|x^{(k)} - \bar{x}\|^p} \leq C_2.$$

La notion de vitesse de convergence est évidemment essentielle en analyse numérique. Ainsi, afin de diminuer le temps de calcul, nous choisissons souvent d'utiliser l'algorithme qui converge le plus rapidement, c'est-à-dire d'ordre le plus élevé.

Une autre notion importante est la convergence *locale* et *globale*. En effet, nous avons les définitions suivantes.

Définition 1.2 *Soit $(x^{(k)})_{k\in\mathbb{N}}$ une suite d'approximation de la solution $\bar{x}$ de (1.1) telle que $x^{(0)} = x_0$ est le point de départ.*

- *Nous disons que la suite $(x^{(k)})_{k\in\mathbb{N}}$ converge globalement vers $\bar{x}$ si pour tout $x_0 \in K$, la suite $(x^{(k)})_{k\in\mathbb{N}}$ converge vers $\bar{x}$.*
- *Nous disons que la suite $(x^{(k)})_{k\in\mathbb{N}}$ converge localement vers $\bar{x}$ s'il existe un voisinage V de $\bar{x}$, tel que pour tout $x_0 \in V$, la suite $(x^{(k)})_{k\in\mathbb{N}}$ converge vers $\bar{x}$.*

En pratique, nous préférons souvent des méthodes qui donnent la convergence globale de la solution à moins bien sûr d'avoir une idéee *a priori* de la localisation de la solution.

2 Méthode de point fixe

La méthode de point fixe consiste à d'abord remplacer l'équation (1.1) par un nouveau problème : trouver $x \in K$, où K est le domaine de définition de la fonction f tel que

$$\Phi(x) = x,$$

où la fonction Φ dépend de la fonction f, par exemple $\Phi(x) = f(x) + x$.

Nous présentons d'abord la méthode de Héron qui permet d'approcher des racines carrées et qui est basée sur ce principe de point fixe. Nous proposons ensuite une méthode générale de point fixe et indiquons les critères que doivent satisfaire la fonction Φ pour que la méthode soit effectivement convergente.

2.1 La méthode de Héron

Héron d'Alexandrie vécut vers I^{er} siècle après J. C. et fût un des grands mécaniciens de l'antiquité ; L'algorithme de calcul des racines carrées lui est attribué bien qu'il semble que cet algorithme fût déjà connu des Babyloniens.

Pour a un nombre réel positif, nous voulons calculer sa racine carrée $\alpha = \sqrt{a}$. Obervons que α n'est rien d'autre que la longueur du côté d'un carré dont l'aire est a. L'idée de base de la méthode va consister à construire un tel carré en partant d'un rectangle que nous allons peu à peu transformer au carré recherché.

Si nous considérons une valeur $x^{(0)}$ donnée comme une estimation grossière de α, nous construisons d'abord un rectangle d'aire a ayant un côté de longueur $L = x^{(0)}$ et un second côté égal à $l = a/L = a/x^{(0)}$. Pour construire un nouveau rectangle qui ressemble plus à un carré, il est raisonnable de choisir pour longueur d'un côté la moyenne des longueurs des côtés du rectangle précédent. Cette longueur sera

$$L = x^{(1)} = \frac{1}{2}\left(x^{(0)} + \frac{a}{x^{(0)}}\right)$$

tandis que le second côté aura pour longueur $l = a/L = a/x^{(1)}$. Nous continuons ce procédé jusqu'à ce que nous ne soyons plus en mesure de distinguer les longueurs des deux côtés

Dans l'exemple qui suit, nous cherchons à calculer $\alpha = \sqrt{8000}$ à l'aide de cet algorithme : la valeur initiale est $x^{(0)} = 160$ et le rectangle initial a pour côtés 160 et 50. Pour construire un rectangle qui se rapproche d'un carré, nous le remplaçons par le rectangle dont un côté est la moyenne des deux précédents, c'est-à-dire

$$x^{(1)} = \frac{1}{2}\left(x^{(0)} + \frac{a}{x^{(0)}}\right) = 105$$

de sorte que l'autre côté a pour longueur $l = 8000/105 == 76.1905$.

Nous construisons une suite $(x^{(k)})_{k\geq 0}$ en appliquant la formule

$$x^{(k+1)} = \frac{1}{2}\left(x^{(k)} + \frac{a}{x^{(k)}}\right)$$

Nous obtenons alors les valeurs suivantes

k	$x^{(k)}$	$\frac{a}{x^{(k)}}$
0	160.00000000000000	50.00000000000000
1	105.00000000000000	76.19047619047619
2	90.59523809523810	88.30486202365309
3	89.45005005944560	89.43538874135297
4	89.44271940039928	89.44271879958389
5	89.44271909999159	89.44271909999158

Nous avons obtenu la racine cherchée à la précision machine

Voyons maintenant comment généraliser cette approche pour la résolution d'une équation du type (1.1).

2.2 Méthode de point fixe

Comme nous l'avons déjà précisé, la méthode de point fixe consiste à d'abord remplacer l'équation (1.1) par

$$\Phi(x) = x. \tag{2.2}$$

Nous sommes ainsi ramenés à la recherche de points fixes de l'application Φ. Le remplacement de (1.1) par (2.2) est toujours possible en posant $\Phi(x) = f(x) + x$. Cependant, ce n'est pas le seul choix possible.

Par exemple la méthode de Héron, qui approche la racine carrée de $a \in \mathbb{R}^+$, revient à résoudre l'équations $f(x) = x^2 - a = 0$ et nous avons pris pour Φ, la fonction

$$\Phi(x) = \frac{1}{2}\left(x + \frac{a}{x}\right)$$

laquelle vérifie bien $\Phi(\sqrt{a}) = \sqrt{a}$. Pourtant, nous aurions pu prendre

- $\star$ $\Phi(x) = x^2 - a + x$,
- $\star$ $\Phi(x) = a/x$.

L'algorithme de point fixe général pour le problème (1.1) est alors donné pour $x_0 \in K \subset \mathbb{R}^n$

$$\begin{cases} x^{(0)} = x_0, \\ \text{Trouver } x^{(k+1)} = \Phi(x^{(k)}). \end{cases} \tag{2.3}$$

Il nous faut donc déterminer des critères sur la fonction Φ pour que la méthode de point fixe soit effectivement convergente.

Nous construisons une suite d'approximation $(x^{(k)})_{k\in I\!N}$ fournie par l'algorithme (2.3) et cherchons à établir la convergence de la suite $(x^{(k)})_{k\in I\!N}$ vers une solution $\bar{x}$ de l'équation $f(\bar{x}) = 0$. Pour cela, nous utiliserons le théorème suivant.

Théorème 2.1 *Soit E un espace métrique complet c'est-à-dire que toutes les suites de Cauhy sont convergentes) muni de la distance d et Φ une contraction stricte dans E, c'est-à-dire il existe $0 < L < 1$*

$$d(\Phi(x), \Phi(y)) \leq L\, d(x, y), \quad (x, y) \in E \times E. \tag{2.4}$$

Alors, pour toute donnée initiale $x_0 \in E$, la suite d'approximation définie par (2.3) converge vers le point fixe de Φ et vérifie

$$d(x^{(k+1)}, \bar{x}) \leq L\, d(x^{(k)}, \bar{x}).$$

Nous disons alors que la convergence est linéaire.

Démonstration. Soit $(x^{(k)})_{k\in I\!N}$ la suite définie par

$$x^{(k+1)} = \Phi(x^{(k)}), \quad k = 0, 1, \ldots$$

Nous voulons démontrer que c'est une suite de Cauchy.

D'après l'hypothèse (2.4), nous avons

$$d(x^{(k+1)}, x^{(k)}) \leq L\, d(x^{(k)}, x^{(k-1)}) \leq \ldots \leq L^k\, d(x^{(1)}, x^{(0)}).$$

Soient m et n deux entiers, nous avons alors

$$\begin{aligned} d(x^{(k+m)}, x^{(k)}) &\leq d(x^{(k+m)}, x^{(k+m-1)}) + d(x^{(k+m-1)}, x^{(k+m-2)}) + \ldots \\ &+ d(x^{(k+1)}, x^{(k)}) \\ &\leq \left(L^{k+m-1} + \ldots + L^k\right) d(x^{(1)}, x^{(0)}) \leq \frac{L^k}{1-L}\, d(x^{(1)}, x^{(0)}). \end{aligned}$$

Comme L^k converge vers zéro, lorsque k tend vers l'infini ; nous avons prouvé que $(x^{(k)})_{k\in I\!N}$ est une suite de Cauchy et donc puisque E est complet, la suite $(x^{(k)})_{k\in I\!N}$ est convergente vers un point $\bar{x}$. En passant à la limite dans le schéma itératif nous obtenons alors

$$\bar{x} = \Phi(\bar{x}).$$

Puis, enfin

$$d(x^{(k+1)}, \bar{x}) \leq L\, d(x^{(k)}, \bar{x}).$$

Au passage, nous vérifions que si Φ est une contraction, alors elle possède un seul point fixe. En effet, soient $\bar{x}$ et $\bar{y}$ deux points fixes différents alors

$$0 < d(\bar{x}, \bar{y}) = d(\Phi(\bar{x}), \Phi(\bar{y})) \leq L\, d(\bar{x}, \bar{y}) < d(\bar{x}, \bar{y}),$$

d'où

$$0 < d(\bar{x}, \bar{y}) < d(\bar{x}, \bar{y}),$$

ce qui est absurde, donc forcément $\bar{x} = \bar{y}$. □

Corollaire 2.1 *Soit Φ une fonction dérivable à dérivée continue sur l'intervalle $[a, b]$. Supposons que $\bar{x} \in K \subset \mathbb{R}^n$ est un point fixe de Φ tel que $\|\Phi'(\bar{x})\| < 1$. Alors, il existe δ tel que pour tout $x^{(0)} = x_0 \in B(\bar{x}, \delta)$, boule de centre $\bar{x}$ et de rayon $\delta > 0$, la suite $(x^{(k)})_{k \in \mathbb{N}}$ converge vers le point fixe $\bar{x}$.*

L'avantage de cette méthode est qu'elle se généralise sans difficulté à des fonctions $f : K \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ puisque qu'aucune dérivée n'apparaît dans l'algorithme. Cependant, cette méthode est généralement moins efficace que la méthode de Newton donnée par la suite.

3 Vers la méthode de Newton

Dans cette partie, nous étudions plusieurs méthodes itératives qui ont pour but d'approcher la solution $\bar{x} \in [a, b]$ de l'équation $f(\bar{x}) = 0$. Nous commençons par la méthode de dichotomie. C'est la méthode plus naturelle mais pas forcément la plus efficace en terme d'ordre de convergence. Nous présentons ensuite l'algorithme de la sécante et nous poursuivons par la méthode de Newton.

3.1 Méthode de dichotomie

Supposons par exemple que $f(a)\, f(b) < 0$. Dans ce cas, puisque f est continue il existe forcément $\bar{x} \in (a, b)$ tel que $f(\bar{x})$.

La méthode de dichotomie est la plus simple et sûrement la plus intuitive. Elle consiste à encadrer la solution $\bar{x}$ par un intervalle de plus en plus petit. Nous posons $m = \frac{a+b}{2}$ et calculons $f(m)$, puis testons le signe de la quantité $f(a)\, f(m)$.

Ainsi, si $f(a)\, f(m) < 0$, ou autrement dit $f(a)$ et $f(m)$ sont de signes différents, cela signifie qu'il existe au moins un $\bar{x} \in [a, m]$ solution de $f(\bar{x}) = 0$.

En revanche si $f(a)\, f(m) > 0$, ou autrement dit $f(a)$ et $f(m)$ sont de même signe, alors $f(m)\, f(b) < 0$ et par continuité de f, il existe $\bar{x} \in [m, b]$ tel que $f(\bar{x}) = 0$.

En itérant ce procédé, nous obtenons alors l'algorithme suivant

Algorithme 1. Méthode de dichotomie.

Nous posons $a^{(0)} = a$ et $b^{(0)} = b$ et soit $f : [a^{(0)}, b^{(0)}] \longrightarrow I\!R$

continue et telle que $f(a^{(0)})\, f(b^{(0)}) < 0$.

Pour $k = 0,\ 1,\ \ldots$

-nous posons

$$m := \frac{a^{(k)} + b^{(k)}}{2},$$

si $f(a^{(k)})\, f(m) < 0$, alors

$$a^{(k+1)} = a^{(k)} \quad \text{et} \quad b^{(k+1)} = m$$

sinon

$$a^{(k+1)} = m \quad \text{et} \quad b^{(k+1)} = b^{(k)}$$

Fin de la boucle sur k

Nous voyons bien qu'à chaque itération l'intervalle encadrant la solution $\bar{x}$ telle que $f(\bar{x}) = 0$ est divisé par deux, soit par récurrence

$$(b^{(k)} - a^{(k)}) = \frac{b^{(0)} - a^{(0)}}{2^k}.$$

Il en résulte que la méthode de dichotomie converge puisque $b^{(k)} - a^{(k)}$ tend vers zéro lorsque k tend vers l'infini. Nous pouvons donc choisir le temps d'arrêt N tel que

$$\frac{1}{2^N}\,(b^{(0)} - a^{(0)}) \leq \varepsilon,$$

où ε est la précision choisie.

Cet algorithme paraît très intéressant mais nous verrons qu'en fait cette méthode converge plutôt lentement. Néanmoins elle présente l'avantage d'être très simple et nécessite seulement que la fonction f soit continue.

3.2 Méthode de la sécante

Une autre méthode consiste simplement à remplacer f par son interpolée linéaire. En effet, nous prenons $a^{(0)}$ et $b^{(0)} \in [a, b]$ et nous nous plaçons sur l'intervalle $[a^{(0)}, b^{(0)}]$, puis remplaçons

f par une polynôme de degré égal à un

$$y(x) := f(a^{(0)}) + (x - a^{(0)}) \frac{f(b^{(0)}) - f(a^{(0)})}{b^{(0)} - a^{(0)}},$$

$y(x)$ est la seule fonction affine telle que $y(a^{(0)}) = f(a^{(0)})$ et $y(b^{(0)}) = f(b^{(0)})$. L'avantage ici est que nous pouvons résoudre exactement l'équation

$$y(x) = 0, \quad x \in [a^{(0)}, b^{(0)}].$$

En effet, nous vérifions que $y(x^{(1)}) = 0$ avec

$$x^{(1)} = a^{(0)} - f(a^{(0)}) \frac{b^{(0)} - a^{(0)}}{f(b^{(0)}) - f(a^{(0)})}.$$

Géométriquement, ceci revient à remplacer la courbe $y = f(x)$ par la droite sécante (AB)

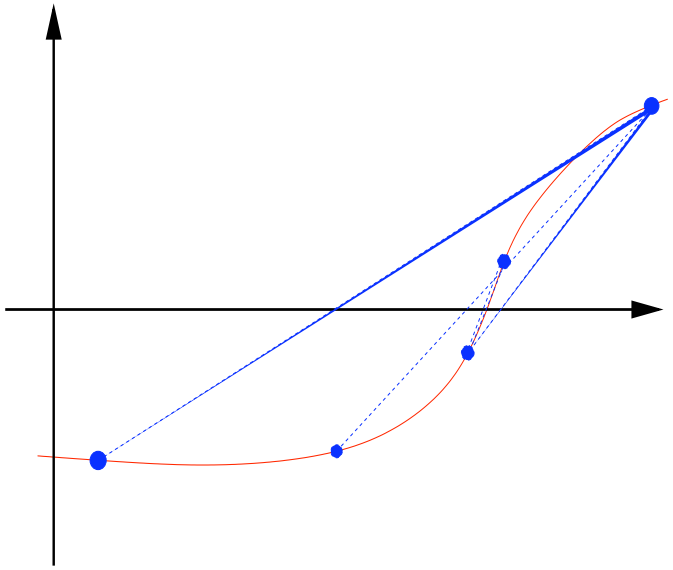

FIG. 3.2 – Illustration de la méthode de la sécante et de la convergence vers le point $\bar{x}$ tel que $f(\bar{x}) = 0$.

passant par les points $(a^{(0)}, f(a^{(0)}))$ et $(b^{(0)}, f(b^{(0)}))$ tandis que $x^{(1)}$ est simplement l'intersection de l'axe des abscisses avec la droite (AB). Pour trouver une meilleure approximation que $x^{(1)}$, il suffit de répéter le procédé soit sur l'intervalle $[a^{(0)}, x^{(1)}]$ ou $[x^{(1)}, b^{(0)}]$ selon le signe de $f(a^{(0)})\, f(x^{(1)})$ comme pour la méthode de dichotomie. Cette méthode peut se révéler plus efficace puisqu'à chaque étape nous réduisons la taille de l'intervalle bien plus vite que pour la méthode de dichotomie.

Algorithme 2. Méthode de la sécante.

Nous choisissons $a^{(0)}$ et $b^{(0)} \in [a, b]$.

Soit $f : [a^{(0)}, b^{(0)}] \longrightarrow I\!R$

continue et telle que $f(a^{(0)})\, f(b^{(0)}) < 0$.

Pour $k = 0,\ 1,\ \ldots$

-nous calculons $x^{(k)} := a^{(k)} - f(a^{(k)}) \dfrac{a^{(k)} - b^{(k)}}{f(a^{(k)}) - f(b^{(k)})}$

si $f(a^{(k)})\, f(x^{(k)}) < 0$, alors

$$a^{(k+1)} = a^{(k)} \quad \text{et} \quad b^{(k+1)} = x^{(k)},$$

sinon

$$a^{(k+1)} = x^{(k)} \quad \text{et} \quad b^{(k+1)} = b^{(k)}.$$

Fin de la boucle sur k

Nous montrons alors le résultat de convergence suivant qui donne seulement la convergence locale de la suite d'approximations

Théorème 3.1 *Soit $f : [a, b] \longrightarrow I\!R$ une fonction deux fois dérivable et dont la deuxième dérivée est continue (f est de classe $\mathcal{C}^2([a, b], I\!R)$). Supposons qu'il existe $\bar{x} \in [a, b]$ tel que $f(\bar{x}) = 0$ et $f'(\bar{x})$ est non nul. Alors, il existe $\delta > 0$ tel que pour tout $a^{(0)}, b^{(0)} \in [\bar{x} - \delta, \bar{x} + \delta]$ la suite $(x^{(k)})_{k \in I\!N}$ donnée par la méthode de la sécante est bien définie et converge vers la solution $\bar{x} \in [a, b]$.*

Démonstration. La preuve de ce théorème est analogue à celle présentée pour la méthode de Newton. Nous choisissons plutôt de détailler la preuve de convergence de la méthode de Newton. □

Cette méthode donne la convergence locale de la suite $(x^{(k)})_{k \in I\!N}$ vers la solution $\bar{x}$; il faut donc partir d'un intervalle $[a^{(0)}, b^{(0)}]$ suffisament petit et tel que $\bar{x} \in [a^{(0)}, b^{(0)}]$, il faut donc avoir une idée *a priori* de la localiation de $\bar{x}$.

3.3 Méthode de Newton

Soit $f : [a,b] \longrightarrow \mathbb{R}$ une fonction dérivable et à dérivée continue. Nous cherchons à calculer numériquement les solutions de l'équation non linéaire $f(x) = 0$. Nous considérons $x_0 \in [a,b]$, une valeur approchée de la solution du problème. Nous posons $x^{(0)} = x_0$, puis par un développement de Taylor, nous obtenons

$$f(x) \;=\; f(x^{(0)}) \;+\; f'(x^{(0)})\,(x - x^{(0)}) \;+\; \frac{1}{2} f''(\eta)\,(x - x^{(0)})^2,$$

avec $\eta \in [a,b]$. En négligeant, le terme d'ordre deux qui sera petit dès que f'' est bornée et $x^{(0)}$ suffisamment proche de la solution x, nous avons une nouvelle approximation $x^{(1)}$ de x donnée par

$$x^{(1)} \;=\; x^{(0)} \;-\; \frac{f(x^{(0)})}{f'(x^{(0)})}.$$

Ainsi, en réitérant le procédé, nous obtenons un schéma itératif appelé méthode de Newton

$$\left\{ \begin{array}{l} x^{(0)} \;=\; x_0, \\ \\ x^{(k+1)} \;=\; x^{(k)} \;-\; \dfrac{f(x^{(k)})}{f'(x^{(k)})}, \quad k \geq 0. \end{array} \right. \tag{3.5}$$

Pour cette méthode nous démontrons le théorème suivant

Théorème 3.2 *Soit $f : [a,b] \longrightarrow \mathbb{R}$ une fonction deux fois dérivable et dont la deuxième dérivée est continue (f est de classe $\mathcal{C}^2([a,b],\mathbb{R})$). Supposons qu'il existe $\bar{x} \in [a,b]$ tel que $f(\bar{x}) = 0$ et $f'(\bar{x})$ est non nul. Alors, il existe $\delta > 0$ tel que pour tout $x_0 \in [\bar{x} - \delta, \bar{x} + \delta]$ la suite $(x^{(k)})_{k \in \mathbb{N}}$ donnée par (3.5) est bien définie et converge vers la solution $\bar{x} \in [a,b]$. De plus il existe une constante $C > 0$ telle que*

$$|x^{(k+1)} - \bar{x}| \;\leq\; C\,|x^{(k)} - \bar{x}|^2.$$

Nous disons que la méthode de Newton est au pire d'ordre deux.

Démonstration. Comme f est de classe $\mathcal{C}^2([a,b],\mathbb{R})$ et $f'(\bar{x})$ est différent de zéro, il existe $\delta > 0$, $L > 0$ et $M > 0$ tels que pour tout $x \in [\bar{x} - \delta, \bar{x} + \delta]$, les valeurs $f'(x)$ sont différentes de zéro et

$$\frac{1}{|f'(x)|} \;\leq\; \frac{1}{L}, \qquad |f''(x)| \;\leq\; M.$$

En effet, en utilisant la continuité de f' au point $\bar{x}$, pour tout $\varepsilon > 0$, il existe $\delta > 0$, tel que pour tout $x \in [\bar{x} - \delta, \bar{x} + \delta]$,

$$|f'(x) - f'(\bar{x})| \leq \varepsilon,$$

alors en notant $\alpha = f'(\bar{x})$ que nous supposons par exemple strictement positif et en prenant $\varepsilon = \alpha/2$, nous obtenons

$$\frac{\alpha}{2} \leq f'(x) \leq \frac{3\,\alpha}{2}.$$

Donc, en prenant $L = \alpha/2$, nous avons pour tout $x \in\in [\bar{x} - \delta, \bar{x} + \delta]$

$$|f'(x)| \geq L.$$

De plus, f étant de classe $\mathcal{C}^2$ sur l'intervalle $[\bar{x} - \delta, \bar{x} + \delta]$, il existe $M > 0$ tel que

$$|f''(x)| \leq M.$$

Ainsi, en prenant au besoin une valeur de δ suffisamment petite, nous pouvons prendre $\delta > 0$ qui vérifie

$$\frac{M\,L\,\delta}{2} \,<\, 1.$$

En choisissant $x^{(0)} \,\in\, (\bar{x} - \delta, \bar{x} + \delta)$ et en effectuant un développement de Taylor au voisinage du point $x^{(0)}$, nous obtenons

$$f(\bar{x}) \,=\, f(x^{(0)}) \,+\, f'(x^{(0)})\,(\bar{x} - x^{(0)}) \,+\, \frac{1}{2}\,f''(\eta^{(0)})\,(\bar{x} - x^{(0)})^2.$$

Puis, en écrivant

$$\begin{aligned} x^{(1)} \;&=\; x^{(0)} \,-\, [f'(x^{(0)})]^{-1}\,\left(f(x^{(0)}) \,-\, f(\bar{x})\right), \\ &=\; x^{(0)} \,+\, (\bar{x} - x^{(0)}) \,-\, \frac{1}{2}\,\frac{f''(\eta^{(0)})}{f'(x^{(0)})}\,(\bar{x} - x^{(0)})^2, \\ &=\; \bar{x} \,-\, \frac{1}{2}\,\frac{f''(\eta^{(0)})}{f'(x^{(0)})}\,(\bar{x} - x^{(0)})^2. \end{aligned}$$

Ainsi, en utilisant une majoration de $|f''|$ et une minoration de $|f'|$, nous avons finalement

$$|x^{(1)} - \bar{x}| \,\leq\, \frac{M}{2\,L}|x^{(0)} - \bar{x}|^2 \,\leq\, \frac{M}{2\,L}\delta^2 \,\leq\, \delta.$$

Nous procédons ensuite par récurrence : nous supposons que $|x^{(k)} - \bar{x}| \leq \delta$. Alors, par un calcul identique nous obtenons

$$|x^{(k+1)} - \bar{x}| \,\leq\, \frac{M}{2\,L}|x^{(k)} - \bar{x}|^2 \,\leq\, \delta. \tag{3.6}$$

Ainsi, en posant

$$e^{(k)} \,=\, \frac{M}{2\,L}|x^{(k)} - \bar{x}|,$$

nous obtenons en utilisant (3.6)

$$e^{(k)} \,\leq\, [e^{(k-1)}]^2 \,\leq\, [e^{(k-2)}]^{2^2} \,\leq \ldots \leq\, [e^{(0)}]^{2^k}.$$

Or, en ayant choisi au préalable $x^{(0)}$ tel que

$$e^{(0)} = |x^{(0)} - \bar{x}| \leq \frac{M}{2\,L}\,\delta < 1,$$

nous montrons que la méthode de Newton est bien convergente. □

Cet algorithme n'est évidemment pas complet tant que nous ne précisons pas un critère d'arrêt. Nous venons de voir que, généralement, $x^{(k)}$ converge vers la solution $\bar{x}$. Un critère d'arrêt souvent utilisé consiste à terminer l'algorithme lorsque $|x^{(k+1)} - x^{(k)}| < \varepsilon$. Pour autant, cela n'assure pas la convergence de la solution vers une solution $\bar{x}$ tel que $f(\bar{x}) = 0$. En effet, en prenant $x^{(k)} = \sum_{l=1}^{k} 1/l$ nous avons

$$\lim_{k \longrightarrow \infty} |x^{(k+1)} - x^{(k)}| = 0$$

et pour autant $x^{(k)}$ tend vers l'infini. Cependant, il n'est pas toujours possible de faire autrement. Un autre choix possible est $|f(x^{(k)})| < \varepsilon$.

Exemple 3.1 *Prenons $f(x) = x\, e^{-x^2}$, nous avons alors $f'(x) = (1 - 2\,x^2)\, e^{-x^2}$. L'algorithme de la méthode de Newton s'écrit alors*

$$x^{(k+1)} = x^{(k)} - -\frac{x^{(k)}}{1 - 2\,(x^{(k)})^2} = \frac{(x^{(k)})^3}{(x^{(k)})^2 - 1/2}.$$

D'une part, en prenant $x^{(0)} = x_0 = 0.3$, nous obtenons

k	$x^{(k)}$	$f(x^{(k)})$
0	0.3	0.274
1	$-6.58\,10^{-2}$	$-6.56\,10^{-2}$
2	$5.76\,10^{-4}$	$5.76\,10^{-4}$

k	$x^{(k)}$	$f(x^{(k)})$
3	$-3.82\,10^{-10}$	$-3.82\,10^{-10}$
4	$1.12\,10^{-28}$	$1.12\,10^{-28}$
5	$-2.81\,10^{-84}$	X

D'autre part, en prenant $x^{(0)} = x_0 = 0.5$, nous avons

k	$x^{(k)}$	$f(x^{(k)})$
0	0.5	0.3894
1	-0.5	-0.3894
2	0.5	0.3894

Ici la méthode ne converge pas.

Enfin en prenant $x^{(0)} = x_0 = 1.$*, nous avons*

k	$x^{(k)}$	$f(x^{(k)})$
0	1	0.3678
1	2	$3.66\,10^{-2}$
2	2.2857	$1.23\,10^{-2}$
3	2.5275	$4.24\,10^{-3}$

la suite $(x^{(k)})_{k\in\mathbb{N}}$ *croît vers l'infini tandis que* $f(x^{(k)})$ *tend vers zéro mais la méthode ne converge pas.*

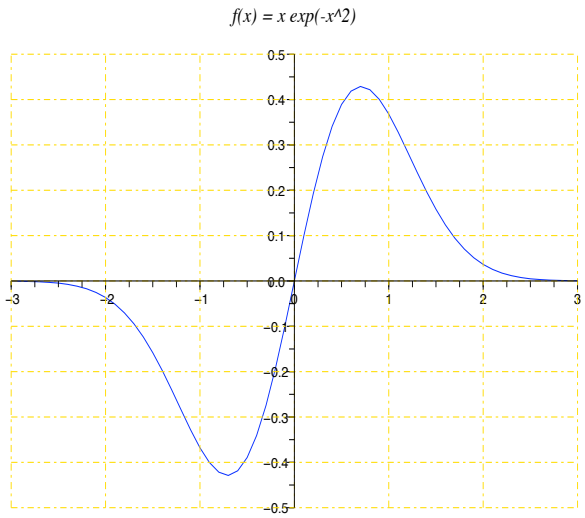

FIG. 3.3 – Tracé de la fonction $f(x) = x\ \exp(-x^2)$.

3.4 Combinaison de méthodes

La méthode de Newton est incontestablement la meilleure parmi celles présentées dans ce chapitre mais à la seule condition d'être assez proche de la solution recherchée (où "assez proche" dépend de la fonction f). Plus précisément, lorsque nous connaissons bien les dérivées premières et secondes de f au voisinage de la solution nous pouvons combiner deux

méthodes de façon pragmatique. À l'initialisation, nous disposons d'un intervalle $[a^{(0)}, b^{(0)}]$ tel que $f(a^{(0)})\, f(b^{(0)}) < 0$ et choisissons une des bornes $a^{(0)}$ ou $b^{(0)}$ pour $x^{(0)}$. Nous calculons $x^{(1)}$ à l'aide de la méthode de Newton.

- Si $x^{(1)} \notin [a^{(0)}, b^{(0)}]$, nous le rejetons et effectuons une itération de dichotomie pour trouver $[a^{(1)}, b^{(1)}]$.
- Si $x^{(1)} \in [a^{(0)}, b^{(0)}]$, alors :
  - Si $f(x^{(1)})\, f(a^{(0)}) < 0$, nous posons $a^{(1)} = a^{(0)}$ et $b^{(1)} = x^{(1)}$.
  - Si $f(x^{(1)})\, f(b^{(0)}) < 0$, nous posons $a^{(1)} = x^{(1)}$ et $b^{(1)} = b^{(0)}$.
- Nous continuons en partant de l'intervalle $[a^{(1)}, b^{(1)}]$.

4 Méthode de Newton dans $\mathbb{R}^n$

4.1 Quelques rappels de calcul différentiel

Nous considérons deux $\mathbb{R}$-espaces vectoriels normés E et F, que nous supposons complets, nous disons aussi que ce sont des espaces de Banach.

Nous notons $\mathcal{L}(E; F)$ l'espace des applications linéaires continues de E dans F, muni de la norme

$$\|l\| = \sup_{h \in E} \frac{\|l\, h\|_F}{\|h\|_E},$$

laquelle correspond à la norme matricielle lorsque $E = F = \mathbb{R}^n$. C'est un espace de Banach, c'est-à-dire un espace vectoriel pour lequel toutes les suites de Cauchy convergent. Nous supposons que U est un ouvert (non vide !) de E, et nous considérons une fonction $f : U \to F$.

Définition 4.1 *Nous disons que la fonction f est différentiable en un point $x \in U$ si elle est continue au point x et s'il existe $l(x) \in \mathcal{L}(E; F)$ tel que*

$$\lim_{h \to 0_E} \frac{\|f(x+h) - f(x) - l(x)\, h\|_F}{\|h\|_E} = 0. \tag{4.7}$$

Cette définition est (volontairement) redondante : en fait, l'existence d'une application linéaire continue $l(x)$ telle que nous ayons (4.7) impose à f d'être continue en x. Car, en notant

$$\varepsilon(h) = \frac{\|f(x+h) - f(x) - l(x)\, h\|_F}{\|h\|_E}$$

si $h \neq 0_E$ et $\varepsilon(0) = 0$, nous avons par l'inégalité triangulaire

$$\|f(x+h) - f(x)\|_F \leq \varepsilon(h)\|h\|_E + \|l(x)\, h\|_F,$$

où le membre de droite tend vers 0 lorsque h tend vers 0_E grâce à (4.7) et à la continuité de $l(x)$ en 0_E. Inversement, si nous supposons f continue en x et s'il existe $l(x)$ linéaire telle que nous ayons (4.7), alors par l'inégalité triangulaire

$$\|l(x)\, h\|_F \leq \varepsilon(h)\|h\|_E + \|f(x+h) - f(x)\|_F$$

tend vers 0 lorsque h tend vers 0_E.

L'application $l(x)$ dépend du point x, nous la noterons désormais $l(x) = \nabla f(x)$, de sorte que (4.7) se réécrit

$$\lim_{h \to 0_E} \frac{\|f(x+h) - f(x) - \nabla f(x)\, h\|_F}{\|h\|_E} = 0.$$

L'application $\nabla f(x)$ est appelée différentielle de f au point x.

Définition 4.2 *Nous disons que la fonction f est différentiable sur U si elle est différentiable en tout point $x \in U$. Dans ce cas, nous appelons différentielle de f la fonction*

$$\nabla f : \begin{array}{l} U \to \mathcal{L}(E;F) \\ \\ x \to \nabla f(x). \end{array}$$

Si de plus ∇f est continue (par rapport à x), nous disons que f est continûment différentiable, ou de façon équivalente que f est de classe C^1.

Dérivées partielles. Pour tout $x \in U$ de composantes $(x_1, \ldots, x_p)$ et pour tout $i \in \{1, \ldots, p\}$, l'ensemble

$$V_i(x) := \{t \in I\!R; \quad (x_1, \ldots, x_{i-1}, t, x_{i+1}, \ldots, x_p) \in U\}$$

est un voisinage ouvert de x_i.

Supposons $f : U \subset I\!R^p \to F$ différentiable. Alors l'application partielle $g_i : V_i(x) \to F$ telle que

$$t \to f(x_1, \ldots, x_{i-1}, t, x_{i+1}, \ldots, x_p) = f(x + (t - x_i)e_i)$$

est dérivable en x_i et

$$g_i'(x_i) = \nabla f(x)\, e_i,$$

(dérivée de f dans la direction e_i au point x). Il est d'usage de noter cette dérivée $\frac{\partial f}{\partial x_i}(x)$. Nous appelons dérivées partielles de f les fonctions

$$\frac{\partial f}{\partial x_i} : \begin{array}{l} U \longrightarrow F \\ \\ x \longrightarrow \frac{\partial f}{\partial x_i}(x) \end{array}$$

pour $i \in \{1, \ldots, p\}$. Par linéarité de $\nabla f(x)$, nous voyons que pour tout $h \in I\!R^p$, de composantes $(h_1, \ldots, h_p)$,

$$\nabla f(x)(h) = \sum_{i=1}^{p} h_i \frac{\partial f}{\partial x_i}(x).$$

Quel que soit $i \in \{1, \ldots, p\}$, l'application $h \in \mathbb{R}^p \to h_i \in \mathbb{R}$ est une forme linéaire continue (c'est-à-dire un élément de $\mathcal{L}(\mathbb{R}^p; \mathbb{R})$). Nous la notons dx_i. Ainsi, la différentielle de f au point x s'écrit

$$\nabla f(x) = \sum_{i=1}^{p} \frac{\partial f}{\partial x_i}(x) dx_i.$$

L'existence de dérivées partielles n'est pas suffisante en général pour qu'une fonction soit différentiable.

Théorème 4.1 *Une application $f : U \subset \mathbb{R}^p \to F = \mathbb{R}^q$ est continûment différentiable si et seulement si ses p dérivées partielles existent et sont continues sur U.*

Matrice jacobienne. Si une fonction $f : U \subset \mathbb{R}^p \to \mathbb{R}^q$, de composantes $(f_1, \ldots, f_q)$, est différentiable au point x, on définit sa matrice jacobienne au point x comme la matrice de l'application linéaire $\nabla f(x)$ dans les bases canoniques de $\mathbb{R}^p$ et $\mathbb{R}^q$. Elle est donnée par

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(x) & \ldots & \frac{\partial f_1}{\partial x_n}(x) \\ \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1}(x) & \ldots & \frac{\partial f_n}{\partial x_n}(x) \end{pmatrix} \in \mathcal{M}_{n,n}(\mathbb{R}).$$

Autrement dit, le coefficient de la matrice jacobienne de f d'indice $i \in \{1, \ldots, q\}$ en ligne et $j \in \{1, \ldots, p\}$ en colonne est $(\nabla f(x))_{i,j} = \frac{\partial f_i}{\partial x_j}(x)$.

En particulier, si $q = 1$, $\nabla f(x)$ est une matrice ligne. De façon générale, les lignes des $\nabla f(x)$ sont les $\nabla f_i(x)$.

Théorème 4.2 *Soit $f : U \subset E \to F$ une fonction différentiable sur un ouvert convexe U. Nous supposons qu'il existe $k > 0$ tel que*

$$\|\nabla f(u)\| \leq k; \quad u \in U.$$

Alors

$$\|f(x) - f(y)\|_F \leq k\|x - y\|_E; \quad (x, y) \in U \times U.$$

Remarque 4.1 *La démonstration donne en fait l'inégalité plus fine*

$$\|f(y) - f(x)\|_F \leq \sup_{t \in [0,1]} \|\nabla f(x + t(y - x))\| \, \|y - x\|_E.$$

Théorèmes d'inversion locale et des fonctions implicites. Soient U et V des ouverts (non vides) d'espaces de Banach E et F respectivement.

Définition 4.3 *Nous disons qu'une application $f : U \to V$ est un difféomorphisme (de U sur V) si et seulement si*
(i) f est une bijection,
(ii) f est de classe C^1, c'est-à-dire continûment différentiable sur U,
(iii) f^{-1} est de classe C^1 sur V.

Le théorème d'inversion locale est le suivant

Théorème 4.3 *Si $f : U \to V$ est de classe C^1, si $a \in U$ est tel que que $\nabla f(a)$ soit un isomorphisme de E sur F, il existe un voisinage ouvert U_a de a dans U et un voisinage ouvert V_b de $b = f(a)$ dans V tel que la restriction de f à U_a soit un difféomorphisme de U_a sur V_b.*

Parmi les conséquences fondamentales du théorème d'inversion locale, nous trouvons un résultat tout aussi important, connu sous le nom de théorème des fonctions implicites. Il concerne la résolution d'équations non-linéaires de la forme

$$f(x, y) = 0,$$

et doit son nom au fait que, sous les hypothèses que nous allons préciser, nous pouvons en tirer y comme fonction de x : nous disons alors que $f(x, y) = 0$ définit implicitement y, ou encore y comme fonction implicite de x.

Dans l'énoncé qui suit, E, F et G sont trois espaces de Banach.

Théorème 4.4 *Soit U un ouvert de $E \times F$ et $f : U \to G$ une fonction de classe C^1. Nous supposons qu'il existe $(a, b) \in U$ tel que $f(a, b) = 0_G$ et la différentielle partielle de f par rapport à y, $\nabla^2 f$ est telle que $\nabla^2 f(a, b)$ soit un isomorphisme de F sur G. Alors il existe un voisinage ouvert $U(a, b)$ de (a, b) dans U, un voisinage ouvert W_a de a dans E et une fonction de classe C^1*

$$\varphi : W_a \to F$$

telle que

$$(x, y) \in U(a, b), \quad \text{et} \quad f(x, y) = 0_G \ \Leftrightarrow \ y = \varphi(x).$$

Dérivées d'ordre supérieur et développements de Taylor. Nous commençons par la définition

Définition 4.4 *Une fontion f définie sur un ouvert (non vide) U d'un $\mathbb{R}$-espace de Banach E et à valeurs dans un $\mathbb{R}$-espace de Banach F est dite deux fois différentiable en $x \in U$ si elle est différentiable dans un voisinage ouvert U_x de x et si sa différentielle $\nabla f : U_x \to \mathcal{L}(E; F)$ est différentiable en x. Nous disons que f est deux fois différentiable dans U si elle est différentiable en tout point de U.*

Par définition, la différentielle de ∇f en x, $\nabla^2 f := \nabla(\nabla f)(x)$ est une application linéaire continue de E dans $\mathcal{L}(E;F)$.

Théorème 4.5 *Si U est un ouvert d'un espace de Banach E, si F est un espace de Banach et $f : U \to F$ est une fonction de classe C^{n+1}, alors, pour tout $(x,h) \in U \times E$ tel que le segment $[x, x+h]$ soit inclus dans U,*

$$f(x+h) = f(x) + \sum_{p=1}^{n} \frac{1}{p!} \nabla^p f(x)(h^{[p]}) + \int_0^1 \frac{(1-t)^n}{n!} \nabla^{n+1} f(x+th)(h^{[n+1]}) dt. \qquad (4.8)$$

Voyons une dernière version de la formule de Taylor, valable sous des hypothèses encore moins fortes, et qui pour cette raison donne un résultat local seulement.

Théorème 4.6 *Si U est un ouvert d'un espace de Banach E, si F est un espace de Banach et $f : U \to F$ est une fonction n fois différentiable en $x \in U$ alors*

$$\left\| f(x+h) - f(x) - \sum_{p=1}^{n} \frac{1}{p!} \nabla^p f(x)(h^{[p]}) \right\| = o(\|h\|^n).$$

4.2 Méthode de Newton

Dans la partie précédente nous avons construit la méthode de Newton à partir d'un développement de Taylor sur $\mathbb{R}$. Pour une fonction de $K \subset \mathbb{R}^n$ à valeur dans $\mathbb{R}^n$. Nous considérons donc $f : K \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ et $\bar{x} \in K$ tel que

$$f(\bar{x}) = 0.$$

Comme dans le cas réel, nous choisissons une première approximation x_0 de $\bar{x}$ et remplaçons f par son développement de Taylor au point x_0 à l'ordre un, c'est-à-dire

$$\begin{cases} x^{(0)} = x_0 \in \mathbb{R}^n, \\ f(x^{(k)}) + \nabla f(x^{(k)})\,(x^{(k+1)} - x^{(k)}) = 0_{\mathbb{R}^n}, \quad k \geq 0. \end{cases} \qquad (4.9)$$

Pour chaque $k \in \mathbb{N}$, nous devons donc résoudre le problème (4.9), il nous faut donc

- calculer la matrice $A = \nabla f(x^{(k)})$,
- s'assurer que cette matrice $A = \in \mathcal{M}_{n,n}(\mathbb{R})$ est bien inversible,
- résoudre le système

$$A\,x^{(k+1)} = -f(x^{(k)}) + A\,x^{(k)} \in \mathbb{R}^n.$$

Lorsque $n = 1$, nous avons facilement montré que la méthode de Newton est convergente et d'ordre deux. Nous allons voir que ce résultat est toujours vrai en dimension n.

Théorème 4.7 *Soit $f : K \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ une fonction deux fois dérivable et dont la deuxième dérivée est continue (f est de classe $\mathcal{C}^2(K, \mathbb{R}^n)$). Supposons qu'il existe $\bar{x} \in K$ tel que $f(\bar{x}) = 0$ et $f'(\bar{x})$ est inversible. Alors,*

(i) il existe $\delta > 0$ tel que pour tout $x_0 \in B(\bar{x}, \delta)$, la suite $(x^{(k)})_{k\in\mathbb{N}}$ donnée par (4.9) est bien définie et $x^{(k)} \in B(\bar{x}, \delta)$;
(ii) la suite $(x^{(k)})_{k\in\mathbb{N}}$ donnée par (4.9) est convergente vers la solution $\bar{x} \in K$;
(iii) il existe une constante $C > 0$ telle que

$$\|x^{(k+1)} - \bar{x}\| \leq C \|x^{(k)} - \bar{x}\|^2.$$

Ici $B(\bar{x}, \delta)$ représente la boule de $\mathbb{R}^n$ de centre $\bar{x}$ et de rayon δ.

Avant de présenter la preuve de ce théorème, nous proposons trois lemmes qui nous seront utiles. D'abord, nous proposons un premier rappel sur les matrices.

Lemme 4.1 *Soit $B \in \mathcal{M}_{n,n}(\mathbb{K})$ une matrice dans un corps $\mathbb{K}$ telle que pour une norme donnée $\|.\|$*

$$\|B\| < 1.$$

Alors, la matrice $(I_n - B)$ est inversible et

$$(I_n - B)^{-1} = \sum_{k\leq 0} B^k, \qquad \|(I_n - B)^{-1}\| \leq \frac{1}{1 - \|B\|}.$$

Démonstration. Soit $p \in \mathbb{N}$, nous calculons alors

$$(I_n - B) \sum_{k=0}^{p} B^k = \sum_{k=0}^{p} (I_n - B) B^k = I_n - B^{p+1}.$$

En passant à la limite, nous avons

$$(I_n - B) \sum_{k\geq 0} B^k = I_n - \lim_{p\to\infty} B^{p+1}$$

et puisque $\|B\| < 1$, nous avons

$$0 \leq \|(I_n - B) \sum_{k\geq 0} B^k - I_n\| \leq \lim_{p\to\infty} \|B\|^{p+1} = 0.$$

ce qui signifie que

$$(I_n - B)^{-1} = \sum_{k\geq 0} B^k.$$

Ensuite, nous avons donc pour $p \in \mathbb{N}$

$$\|(I_n - B)^{-1}\| = \|\sum_{k\geq 0} B^k\| \leq \sum_{k\geq 0} \|B\|^k \leq \frac{1}{1 - \|B\|}.$$

□

Ensuite, nous proposons un résultat sur les propriétés de la fonction f du problème $f(\bar{x}) = 0$.

Lemme 4.2 *Soit $f : K \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ une fonction de classe $\mathcal{C}^2(K, \mathbb{R}^n)$). Supposons qu'il existe $\bar{x} \in K$ tel que $f(\bar{x}) = 0$ et $\nabla f(\bar{x})$ est inversible. Alors, il existe $\delta > 0$, $L > 0$ et $M > 0$ tels que pour tout $x \in B(\bar{x}, \delta)$*

$$\|\nabla f(x)^{-1}\| \leq \frac{1}{L},$$

et de plus

$$\|\nabla f(x) - \nabla f(y)\| \leq M \|x - y\|.$$

Démonstration. Puisque $f \in \mathcal{C}^2(K, \mathbb{R}^n)$ donc pour $r > 0$ fixé quelconque, il existe $M(r) > 0$ tel que pour tout $x, y \in B(\bar{x}, r)$

$$\|\nabla f(x) - \nabla f(y)\| \leq M(r) \|x - y\|.$$

Remarquons bien que r est fixé et quelconque, il reste donc à montrer qu'il existe $\delta > 0$ et $L > 0$ tels que pour tout $x \in B(\bar{x}, \delta)$, la matrice $\nabla f(x)$ est inversible et vérifie

$$\|\nabla f(x)^{-1}\| \leq \frac{1}{L}.$$

Pour cela, nous écrivons

$$\nabla f(x) = \nabla f(\bar{x}) - \nabla f(\bar{x}) + \nabla f(x) = \nabla f(\bar{x}) \left[I_n - \nabla f^{-1}(\bar{x}) (\nabla f(\bar{x}) - \nabla f(x))\right]$$

et posons $B = f'^{-1}(\bar{x}) (\nabla f(\bar{x}) - \nabla f(x))$ et donc

$$\nabla f(x) = \nabla f(\bar{x}) [I_n - B].$$

La matrice $\nabla f(x)$ est inversible dès que la matrice $[I_n - B]$ est elle même inversible. En vue d'appliquer le Lemme 4.1, il nous suffit de montrer qu'il existe $\delta > 0$ tel que pour tout $x \in B(\bar{x}, \delta)$, nous avons $\|B\| > 1$. Ainsi,

$$\|B\| \leq \|\nabla f(\bar{x}))^{-1}\| \|\nabla f(x) - \nabla f(\bar{x})\| \leq \|\nabla f(\bar{x}))^{-1}\| M(r) \|x - \bar{x}\|$$

et en prenant $\delta < 1/(M(r) \|\nabla f(\bar{x}))^{-1}\|)$, nous avons

$$\|B\| < 1,$$

ce qui signifie que pour tout $x \in B(\bar{x}, \delta)$, la matrice $\nabla f(x)$ est inversible et

$$\|\nabla f(x))^{-1}\| \leq \|\nabla f(\bar{x}))^{-1}\| \|(I_n - B)^{-1}\| \leq \|\nabla f(\bar{x}))^{-1}\| \frac{1}{1 - \|B\|} =: \frac{1}{L},$$

ce qui conclut la démonstration. □

Puis, le troisième lemme.

Lemme 4.3 *Soit ω un ouvert de $\mathbb{R}^n$. Nous considérons $f : \omega \to \mathbb{R}^n$ une fonction telle que $f \in \mathcal{C}^1(\omega, \mathbb{R}^n)$, $\nabla f(x)$ est inversible pour tout $x \in \omega$ et il existe $L > 0$ et $M > 0$*

$$\|\nabla f(x)^{-1}\| \leq \frac{1}{L}, \quad \|\nabla f(x) - \nabla f(y)\| \leq M \|x - y\|. \tag{4.10}$$

et pour $k \geq 1$, $x^{(k-1)} \in \omega$.

Alors, le terme $x^{(k)}$ donné par la méthode de Newton est bien défini et tel que

$$\|x^{(k)} - x^{(k-1)}\| \leq \frac{1}{L} \|f(x^{(k-1)})\|. \tag{4.11}$$

et

$$\|f(x^{(k)})\| \leq M \|x^{(k)} - x^{(k-1)}\|^2. \tag{4.12}$$

Démonstration. D'une part, puisque $x^{(k-1)} \in \omega$, la matrice Jacobienne $\nabla f(x^{(k-1)})$ est inversible, nous pouvons donc construire

$$x^{(k)} = x^{(k-1)} - [\nabla f(x^{(k-1)})]^{-1} f(x^{(k-1)}),$$

d'où nous déduisons que grâce à l'hypothèse (4.10)

$$\|x^{(k)} - x^{(k-1)}\| = \|[\nabla f(x^{(k-1)})]^{-1} f(x^{(k-1)})\| \leq C \|f(x^{(k-1)})\|.$$

D'autre part, à l'aide de la formule de Taylor-Young avec reste intégral, nous avons aussi

$$f(x^{(k)}) = f(x^{(k-1)}) + \int_0^1 \nabla f(x^{(k-1)} + t\,(x^{(k)} - x^{(k-1)}))\,(x^{(k)} - x^{(k-1)})\,dt.$$

Puis, en utilisant la méthode de Newton à l'étape $k-1$

$$f(x^{(k-1)}) + \nabla f(x^{(k-1)})\,(x^{(k)} - x^{(k-1)}) = 0,$$

nous avons

$$f(x^{(k)}) = \int_0^1 \left[\nabla f(x^{(k-1)} + t\,(x^{(k)} - x^{(k-1)})) - \nabla f(x^{(k-1)})\right]\,(x^{(k)} - x^{(k-1)})\,dt.$$

D'où

$$\begin{aligned}
\|f(x^{(k)})\| &\leq \sup_{t\in(0,1)} \left(\|\nabla f(x^{(k-1)} + t\,(x^{(k)} - x^{(k-1)})) - \nabla f(x^{(k-1)})\|\right)\,\|x^{(k)} - x^{(k-1)}\|, \\
&\leq \sup_{t\in(0,1)} \left(M\,t\|x^{(k)} - x^{(k-1)}\|\right)\,\|x^{(k)} - x^{(k-1)}\|,
\end{aligned}$$

c'est-à-dire

$$\|f(x^{(k)})\| \leq M \|x^{(k)} - x^{(k-1)}\|^2.$$

□

Nous sommes maintenant en mesure de démontrer le Théorème 4.7.

Preuve du Théorème 4.7. D'abord, en appliquant le Lemme 4.2, puisque la matrice $\nabla f(\bar{x})$ est inversible et $f \in \mathcal{C}^2(K, \mathbb{R}^n)$, il existe $\delta_0 > 0$, $L_0 > 0$ et M_0 tels que pour tout $x \in B(\bar{x}, \delta_0)$, la matrice $\nabla f(x)$ est inversible et

$$\|\nabla f(x)^{-1}\| \leq \frac{1}{L_0}.$$

De plus, pour tout $x, y \in B(\bar{x}, \delta_0)$,

$$\|\nabla f(x) - \nabla f(y)\| \leq M_0 \|x - y\|.$$

Nous posons alors $\delta = \min(\delta_0\,,\, L_0/(4\,M_0))$ de sorte que

$$\frac{M_0}{2\,L_0}\,\delta < 1.$$

Nous voulons démontrer que $x^{(k)}$ est bien défini pour tout $k \in \mathbb{N}$ et ensuite que la suite $(f(x^{(k)}))_{k\geq 0}$ tend vers zéro lorsque k tend vers l'infini. Pour cela, nous raisonnons par récurrence.

Soit $x_0 \in B(\bar{x}, \delta)$. Dans un premier temps, nous montrons que si $x^{(k)} \in B(\bar{x}, \delta)$, alors $x^{(k+1)} \in B(\bar{x}, \delta)$.

D'une part, en appliquant le Lemme 4.3, la valeur $x^{(k+1)}$ est bien définie. D'autre part, il nous faut vérifier que $x^{(k+1)} \in B(\bar{x}, \delta)$. En effet, à l'aide d'un développement de Taylor-Young à l'ordre deux de la fonction f et puisque $f(\bar{x}) = 0$, il vient

$$\|f(x^{(k)}) + \nabla f(x^{(k)})\,(\bar{x} - x^{(k)})\| \leq M_0\,\|\bar{x} - x^{(n)}\|^2.$$

Aussi, pa définition de $x^{(k+1)}$

$$0 = f(x^{(k)}) + \nabla f(x^{(k)})\,(x^{(k+1)} - x^{(k)}),$$

nous avons donc

$$\|\nabla f(x^{(k)})\,(x^{(k+1)} - \bar{x})\| \leq M_0\,\|\bar{x} - x^{(k)}\|^2.$$

Ensuite,

$$\begin{aligned} \|x^{(k+1)} - \bar{x}\| &= \|\nabla f(x^{(k)})^{-1}\,\nabla f(x^{(k)})\,(x^{(k+1)} - \bar{x})\| \\ &\leq \|\nabla f(x^{(k)})^{-1}\|\,\|\nabla f(x^{(k)})\,(x^{(k+1)} - \bar{x})\| \\ &\leq \frac{M_0}{L_0}\,\|x^{(k)} - \bar{x}\|^2 \end{aligned} \tag{4.13}$$

et donc par définition de δ, nous avons

$$\|x^{(k+1)} - \bar{x}\| \leq \frac{M_0\,\delta^2}{L_0} \leq \delta,$$

ce qui montre que $x^{(k+1)} \in B(\bar{x}, \delta)$ et donc le point (i) du Théorème 4.7. Aussi, nous avons démontré l'existence d'une constante $C = M_0/L_0 > 0$ telle que

$$\|x^{(k+1)} - \bar{x}\| \leq C \|x^{(k)} - \bar{x}\|^2,$$

ce qui démontre le point (iii) du Théorème 4.7.

Intéressons nous alors au point (ii), c'est-à-dire à la convergence de la suite $(x^{(k)})_{k\geq 0}$. Nous posons

$$e^{(k)} = \frac{M_0}{L_0} \|x^{(k)} - \bar{x}\|$$

il vient alors en multipliant (4.13) par M_0/L_0,

$$e^{(k+1)} \leq [e^{(k)}]^2 \leq [e^{(k-1)}]^{2^2} \leq \ldots \leq [e^{(0)}]^{2^{k+1}}.$$

Or, en ayant choisi au préalable $x^{(0)}$ tel que

$$e^{(0)} = |x^{(0)} - \bar{x}| \leq \frac{M_0}{2\,L_0}\,\delta < 1,$$

nous montrons que la méthode de Newton est bien localement convergente. □

Ce théorème montre la convergence "locale" de la méthode de Newton, c'est-à-dire que si x_0 est suffisamment proche d'une solution $f(\bar{x}) = 0$, alors la suite $(x^{(k)})_{k\in I\!N}$ converge vers $\bar{x}$. Concernant la convergence globale, nous en savons très peu et nous pouvons analyser seulement que quelques cas de fonctions simples. L'exemple le plus connu est

$$f(z) = z^3 - 1,$$

ou encore

$$f(x, y) = \begin{pmatrix} x^3 - 3x\,y^2 - 13x^2\,y - y^3 \end{pmatrix},$$

avec $z = x + iy$ et pour lequel l'itération devient

$$z^{(k+1)} = z^{(k)} - \frac{(z^{(k)})^3 - 1}{3(z^{(k)})^2} = \frac{1}{3}\left(2\,z^{(k)} - \frac{1}{(z^{(k)})^2}\right).$$

Nous posons alors

$$A(a) = \{z_0 \in \mathbb{C};\quad (z^{(k)})_k \text{ converge vers } a\},$$

avec $a = 1, (-1 \pm i\sqrt{3})/2$ de $f(z) = 0$. Dans ce cas, nous vérifions à l'aide d'un ordinateur que pour toutes valeurs de z_0, l'algorithme de Newton converge vers une solution.

En pratique, il arrive souvent que la forme analytique de la matrice $\nabla f(x)$ est inconnue. Dans cette situation nous approchons les éléments $\partial f_i/\partial x_j$ de la matrice Jacobienne par

$$\frac{\partial f_i}{\partial x_j}(x_1, \ldots, x_n) \simeq \frac{f_i(x_1, \ldots, x_j + \delta, \ldots, x_n) - f_i(x_1, \ldots, x_j, \ldots, x_n)}{\delta},$$

mais dans ce cas il faut faire très attention aux erreurs d'arrondi.

4.3 Calcul d'éléments propres

Nous considérons $A \in \mathcal{M}_{n,n}(\mathbb{R})$ une matrice symétrique et $\lambda \in \mathbb{R}$ une valeur propre *simple* de A. Il existe donc un vecteur $v \in \mathbb{R}^n$ un vecteur propre associé à la valeur propre λ tel que

$$\|v\|_2 = 1.$$

Pour calculer l'élément propre (v, λ) nous appliquons la méthode de Newton au système non linéaire d'inconnues $(u, r) \in \mathbb{R}^n \times \mathbb{R}$ donné par

$$\begin{cases} A\,u - r\,u = 0, \\ \|u\|_2^2 = \sum_{i=1}^n u_i^2 = 1. \end{cases} \tag{4.14}$$

Nous posons alors $x = (u, r) \in \mathbb{R}^{n+1}$ et $f(x) = (f_1(x), \ldots, f_{n+1}(x))$ donnée par

$$f_i(x) = \begin{cases} \displaystyle\sum_{j=1}^{n} a_{i,j}\, x_j - x_{n+1}\, x_i = 0, & i = 1, \ldots, n; \\ \displaystyle\sum_{j=1}^{n} x_j^2 - 1 = 0 & i = n+1 \end{cases}$$

et appliquons la méthode de Newton. En utilisant les résultats de convergence de la partie précédente, nous pouvons démontrer que la méthode est localement convergente.

5 Complément du Chapitre 3

5.1 Recherche de racines de polynômes

Dans cette partie, nous mettons en place des algorithmes itératifs effectuant la recherche de racines de polynômes.

Soit $p(x)$ un polynôme de degré n à coefficients réels et admettant n racines distinctes

$$x_n < x_{n-1} < \ldots < x_1.$$

Nous souhaitons appliquer la méthode de Newton pour approcher les racines $(x_i)_{1 \le i \le n}$, l'algorithme itératif est alors fourni par

$$\begin{cases} x^{(0)} = x_0, \\ x^{(k+1)} = x^{(k)} - \dfrac{p(x^{(k)})}{p'(x^{(k)})}. \end{cases}$$

Nous démontrons alors le théorème de convergence suivant

Théorème 5.1 *Soit p un polynôme de degré $n \geq 2$, à coefficients réels et admettant n racines réelles*

$$x_n \, < \, x_{n-1} \, < \, \ldots \, < \, x_1.$$

Alors, pour toute valeur $x_0 > x_1$, la suite $(x^{(k)})_{k\geq 0}$ donnée par la méthode de Newton est strictement décroissante et converge vers x_1 la plus grande racine.

Démonstration. D'une part, en supposant que le coefficient du polynôme $p(x)$ devant le terme x^n est strictement positif, alors pour tout $x > x_1$, la polynôme $p(x)$ est strictement positif.

D'autre part, en appliquant le théorème de Rolle, nous savons que p' admet $(n-1)$ racines notées $(\eta_i)_{1\leq i\leq n-1}$ telles que

$$x_n \, < \, \eta_{n-1} \, < \, x_{n-1} \, < \, \ldots \, < \, \eta_1 \, < \, x_1.$$

Alors, en utilisant le même argument que précédemment nous démontrons que pour tout $x > x_1$, $p'(x)$ est strictement positif. En appliquant encore le théorème de Rolle, nous prouvons que $p''(x)$ est strictement positif pour tout $x > x_1$ et finalement $p'''(x)$ est positif ou nul pour $x > x_1$.

Ensuite, à l'aide d'un développement de Taylor-Young, nous obtenons

$$0 \, = \, p(x_1) \, = \, p(x_0) \, + \, (x_1 - x_0)\, p'(x_0) \, + \, \frac{1}{2}(x_1 - x_0)^2\, p''(\theta_0),$$

avec $\theta_0 \in (x_1, x_0)$. Puis, en appliquant l'algorithme de Newton et ce dernier résultat, nous avons

$$x^{(1)} \, = \, x_0 \, - \, \frac{p(x_0) - p(x_1)}{p'(x_0)} \, = \, x_1 + \frac{p''(\theta_0)}{2\, p'(x_0)}\, (x_1 - x_0)^2.$$

Ainsi, en utilisant la stricte positivité de $p'(x)$ et $p''(x)$ pour $x > x_1$, nous démontrons que

$$x_1 < x^{(1)} < x^{(0)} = x_0.$$

En utilisant un raisonnement identique, nous établissons que la suite $(x^{(k)}_{k\geq 0}$ est strictement décroissante et minorée

$$x_1 \, < \, x^{(k+1)} \, < \, x^{(k)}.$$

La suite $(x^{(k)})_{k\geq 0}$ est donc convergente, nous notons cette limite par $\bar{x}$. Il nous reste à démontrer que la $\bar{x}$ est bien la racine x_1. Par passage à la limite dans l'algorithme de Newton, il vient

$$\bar{x} = \bar{x} \, - \, \frac{p(\bar{x})}{p'(\bar{x})},$$

ou encore

$$\frac{p(\bar{x})}{p'(\bar{x})} = 0,$$

comme $\bar{x} \geq x_1$, nous savons que $p'(\bar{x}) > 0$ et $\bar{x}$ est la plus grande racine de p donc $\bar{x} = x_1$.

D'autre part, puisque $p'''(x)$ est positif ou nul nous avons

$$p''(x) \leq p''(x^{(0)}) = p''(x_0)$$

et

$$p'(x^{(k)}) > p'(x_1)$$

et donc

$$(x^{(k+1)} - x_1) \leq \frac{p''(x_0)}{2\,p'(x_1)} \left(x^{(k)} - x_1\right)^2,$$

la méthode est bien quadratique. □

Chapitre 4

Optimisation

1 Motivation

L'objectif de ce chapitre est de rechercher des minima ou des maxima d'une fonction $f \in C(\mathbb{R}^d, \mathbb{R})$ avec ou sans contrainte. Le problème d'optimisation sans contrainte s'écrit :

$$\begin{cases} \text{Trouver } x \in \mathbb{R}^d \text{ tel que :} \\ f(x) \leq f(y), \forall\, y \in \mathbb{R}^d. \end{cases}$$

Le problème d'optimisation sous contraintes s'écrit :

$$\begin{cases} \text{Trouver } x \in K \text{ tel que :} \\ f(x) \leq f(y), \forall\, y \in K. \end{cases}$$

où $K \subset \mathbb{R}^d$ et $K \neq \mathbb{R}^d$.

Si x est solution du premier problème, nous disons que $x \in \text{argmin}_{\mathbb{R}^d} f$, et si x est solution du deuxième problème, nous disons que $x \in \text{argmin}_K f$.

2 Optimisation sans contrainte

Considérons $f \in C(E, \mathbb{R})$ et E un espace vectoriel normé. Nous recherchons un minimum global de f, c'est-à-dire

$$\bar{x} \in E, \quad \text{tel que} \quad f(\bar{x}) \leq f(y), \quad \forall\, y \in E, \tag{2.1}$$

ou un minimum local, c'est-à-dire :

$$\bar{x} \in E \quad \text{tel qu'il existe } \alpha > 0, f(\bar{x}) \leq f(y), \quad \forall y \in B(\bar{x}, \alpha). \tag{2.2}$$

Proposition 2.1 *(Condition nécessaire d'optimalité) Soient E un espace vectoriel normé, $f \in C(E, \mathbb{R})$ et $\bar{x} \in E$ tels que f soit différentiable en $\bar{x}$. Si $\bar{x}$ est solution de (2.1) ou (2.2) alors $\nabla f(\bar{x}) = 0$.*

Démonstration. Supposons qu'il existe $\alpha > 0$ tel que $f(\bar{x}) \leq f(y)$ pour tout $y \in B(\bar{x}, \alpha)$.

Nous choisissons alors $z \in E$ tel que $z \neq 0$, alors en prenant $|t| < \alpha/\|z\|$, nous avons $x + t\,z \in B(\bar{x}, \alpha)$ (où $B(\bar{x}, \alpha)$ désigne la boule ouverte de centre $\bar{x}$ et de rayon α).

Nous avons donc

$$f(\bar{x}) \leq f(\bar{x} + t\,z).$$

Comme la fonction f est différentiable en $\bar{x}$, nous avons :

$$f(\bar{x} + t\,z) = f(\bar{x}) + \nabla f(\bar{x}) \cdot t\,z + |t|\epsilon_z(t),$$

où $\epsilon_z(t) \to 0$ lorsque $t \to 0$. Nous avons donc

$$f(\bar{x}) + t\,\nabla f(\bar{x}) \cdot z + |t|\,\epsilon_z(t) \geq f(\bar{x}).$$

Ensuite, pour $\alpha/\|z\| > t > 0$, nous avons $\nabla f(\bar{x}) \cdot z + \epsilon_z(t) \geq 0$. En faisant tendre t vers 0, nous obtenons que

$$\nabla f(\bar{x}) \cdot z \geq 0, \quad \forall\, z \in E.$$

Nous avons aussi $-\nabla f(\bar{x}) \cdot z \geq 0$ pour tout $z \in E$, et donc $-\nabla f(\bar{x}) \cdot z \geq 0$ pour tout $z \in E$. Nous en concluons que

$$\nabla f(\bar{x}) = 0.$$

□

La proposition précédente donne une condition nécessaire mais non suffisante. En effet, $\nabla f(\bar{x}) = 0$ n'entraîne pas forcément que f atteint un minimum (ou un maximum) même local, en x. En effet, pour le vérifier prenons $E = \mathbb{R}$, $x = 0$ et la fonction f définie par : $f(x) = x^3$, nous avons bien $f'(0) = 0$ mais 0 n'est pas un extremum.

Proposition 2.2 *(Existence) Soient $E = \mathbb{R}^d$ 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 $\bar{x} \in E$ tel que $f(\bar{x}) \leq f(y)$ pour tout $y \in E$.

Démonstration. La condition (ii) peut encore s'écrire pour tout $A \in \mathbb{R}$, il existe $R \in \mathbb{R}$; tel que pour tout $x \in E$ vérifiant $\|x\| \geq R$ alors $f(x) \geq A$.

Nous prenons $A = f(0)$ et obtenons alors qu'il existe $R_0 \in \mathbb{R}$ tel que pour tout $x \in E$ vérifiant $\|x\| \geq R_0$, nous avons $f(x) \geq f(0)$.

Nous en déduisons que

$$\inf_{x \in \mathbb{R}^d} f(x) = \inf_{x \in \overline{B}_{R_0}} f(x),$$

où $\overline{B}_{R_0} = \{x \in I\!R^d; \|x\| \leq R_0\}$. Or, $\overline{B}_{R_0}$ est un ensemble compact de $I\!R^d$ et f est continue donc il existe $\bar{x} \in \overline{B}_{R_0}$ tel que $f(\bar{x}) = \inf_{y \in \overline{B}_{R_0}} f(y)$ et donc $f(\bar{x}) = \inf_{y \in I\!R^d} f(y)$. □

Remarquons tout de suite que ce théorème peut être généralisé sous certaines condition sur l'espace E.

Remarque 2.1 *La proposition précédente ne s'applique que lorsque l'espace vectoriel normé E est de dimension finie ou plus généralement lorsque E est un espace de Banach. En effet, lorsque E est simplement de dimension infinie, la boule fermée $\overline{B}_R$ n'est pas compacte.*

D'autre part, l'hypothèse (ii) du théorème peut être remplacée par l'hypothèse

$$\exists\, b \in I\!R^d, \quad \exists\, R > 0 \text{ tel que } \|x\| \geq R \Rightarrow f(x) \geq f(b).$$

Enfin sous les hypothèses de la proposition précédente il n'y a pas toujours unicité de $\bar{x}$ même dans le cas $n = 1$. Pour s'en convaincre la fonction f définie de $I\!R$ dans $I\!R$ par $f(x) = x^2(x-1)(x+1)$. Pour obtenir l'unicité de la solution $\bar{x}$, nous avons besoin d'une notion supplémentaire : la convexité.

Définition 2.1 *Soient E un espace vectoriel et $f : E \to I\!R$.*
- *Nous disons que f est une fonction convexe lorsqu'elle vérifie*

$$f(t\,x + (1-t)\,y) \leq t\,f(x) + (1-t)\,f(y),$$

  pour tout $(x, y) \in E^2$ tel que $x \neq y$ et $t \in [0, 1]$.
- *Nous disons que f est une fonction strictement convexe si*

$$f(t\,x + (1-t)\,y) < t\,f(x) + (1-t)\,f(y)$$

  pour tout $(x, y) \in E^2$ tel que $x \neq y$ et $t \in]0, 1[$.

Proposition 2.3 *(Condition suffisante d'unicité) Soient E un espace vectoriel normé et f une fonction de $E \to I\!R$ strictement convexe alors il existe au plus un $\bar{x} \in E$ tel que $f(\bar{x}) \leq f(y)$, pour tout $y \in E$.*

Démonstration. Prenons f une fonction strictement convexe et supposons qu'il existe $\bar{x}$ et $\bar{y} \in E$ tels que $f(\bar{x}) = f(\bar{y}) = \inf_{z \in E} f(z)$. Puisque f est strictement convexe, si $\bar{x} \neq \bar{x}$ alors nous avons

$$f(\frac{1}{2}\bar{x} + \frac{1}{2}\bar{y}) < \frac{1}{2}\,f(\bar{x}) + \frac{1}{2}\,f(\bar{y}) = \inf_{z \in E} f(z),$$

ce qui est impossible ; donc $\bar{x} = \bar{y}$. □

Cette dernière proposition n'assure pas l'existence d'une solution au problème de minimisation (2.1) ou (2.2). Par exemple dans le cas $n = 1$ la fonction f définie par $f(x) = e^x$

n'atteint pas son minimum, car $\inf_{x\in\mathbb{R}} f(x) = 0$ et $f(x) \neq 0$ pour tout $x \in \mathbb{R}$, et pourtant f est strictement convexe.

En revanche, en réunissant les hypothèses des propositions précédentes, nous obtenons le résultat d'existence et unicité suivant :

Théorème 2.1 *(Existence et unicité) Soient $E = \mathbb{R}^d$ et f une fonction de $E \to \mathbb{R}$, nous supposons que :*
(i) *la fonction f est continue,*
(ii) *la focntion f vérifie $f(x) \to +\infty$ quand $\|x\| \to +\infty$,*
(iii) *la fonction f est strictement convexe.*
Alors il existe un unique $\bar{x} \in E$ tel que

$$f(\bar{x}) = \inf_{y\in E} f(y).$$

Le théorème reste vrai lorsque E est un espace de Banach ; nous avons besoin dans ce cas pour la partie existence des hypothèses (i), (ii) et de la convexité de f.

Nous présentons dans la suite des propriétés permettant justement de caractériser la convexité d'une fonction.

Proposition 2.4 *(Caractérisation de la convexité) Soit E un espace vectoriel normé (sur $\mathbb{R}$) et $f \in C^1(E, \mathbb{R})$ alors :*
- *f est convexe si et seulement si $f(y) \geq f(x) + \nabla f(x)(y - x)$, pour tout couple $(x, y) \in E^2$,*
- *f est strictement convexe si et seulement si $f(y) > f(x)+\nabla f(x)(y-x)$ pour tout couple $(x, y) \in E^2$ tel que $x \neq y$.*

Démonstration. Supposons que f est convexe : soit $(x, y) \in E^2$; nous voulons montrer que

$$f(y) \geq f(x) + \nabla f(x)(y - x).$$

Soit $t \in [0, 1]$, alors

$$f(ty + (1 - t)x) \leq tf(y) + (1 - t)f(x)$$

Nous avons donc

$$f(x + t(y - x)) - f(x) \leq t(f(y) - f(x)).$$

Comme f est différentiable,

$$f(x + t(y - x)) = f(x) + \nabla f(x)(t(y - x)) + t\varepsilon(t)$$

où $\varepsilon(t)$ tend vers 0 lorsque t tend vers 0. Donc en reportant dans l'inégalité précédente, il vient

$$\varepsilon(t) + \nabla f(x)(y - x) \leq f(y) - f(x), \quad \forall t \in]0, 1[.$$

En faisant tendre t vers 0, nous obtenons alors

$$f(y) \geq \nabla f(x)(y - x) + f(x).$$

Montrons maintenant la réciproque : Soit $(x, y) \in E^2$, et $t \in]0, 1[$ (pour $t = 0$ ou $t = 1$ nous n'avons rien à démontrer). Nous voulons montrer que

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y).$$

Nous posons $z = tx + (1-t)y$ et avons alors par hypothèse

$$f(y) \geq f(z) + \nabla f(z)(y - z),$$

et

$$f(x) \geq f(z) + \nabla f(z)(x - z).$$

En multipliant la première inégalité par $1 - t$, la deuxième par t et en les additionnant, nous obtenons

$$(1-t)f(y) + tf(x) \geq f(z) + (1-t)\nabla f(z)(y - z) + t\nabla f(z)(x - z)$$

et

$$(1-t)f(y) + tf(x) \geq f(z) + \nabla f(z)((1-t)(y - z) + t(x - z)).$$

Et comme $(1-t)(y - z) + t(x - z) = 0$, nous avons donc

$$(1-t)f(y) + tf(x) \geq f(z) = f(tx + (1-t)y).$$

□

Proposition 2.5 *(Caractérisation de la convexité) Soient $E = \mathbb{R}^d$ et f une fonction appartenant à $C^2(E, \mathbb{R})$. Nous notons $H_f(x)$ la hessienne de f au point x, c'est-à-dire $H_f(x))_{i,j} = \partial_{i,j} f(x)$. Alors, nous avons*

(i) la fonction f est convexe si et seulement si la hessienne $H_f(x)$ est symétrique et positive pour tout $x \in E$, c'est-à-dire

$$H_f(x)^T = H_f(x), \quad \text{et} \quad H_f(x)\, y \cdot y \geq 0$$

pour tout $y \in E$.

(ii) f est strictement convexe si $H_f(x)$ est symétrique définie positive pour tout $x \in E$ mais la réciproque est généralement fausse.

Démonstration. Commençons par démontrer le point (i). Nous voulons prouver que lorsque la fonction f est convexe, alors la matrice hessienne $H_f(x)$ est symétrique et positive.

D'une part, il est clair que la hesseinne $H_f(x)$ est symétrique car $\partial_{i,j} f = \partial_{j,i} f$ puisque la fonction f est de classe C^2.

D'autre part, en utilisant la définition de $H_f(x) = \nabla^2 f(x)$ et $\nabla f \in C^1(E, E)$, nous considérons pour un couple $(x, y) \in E^2$, comme f est convexe et de classe C^1, nous avons grâce à la Proposition 2.4 :

$$f(y) \geq f(x) + \nabla f(x) \cdot (y - x).$$

Prenons $\varphi \in C^2(\mathbb{R}, \mathbb{R})$ et 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

$$\varphi''(t) = D(\nabla f(x + t(y - x))(y - x) \cdot (y - x) = H_f(x + t(y - x))(y - x) \cdot (y - x).$$

Nous avons donc :

$$f(y) - f(x) = \nabla f(x)(y - x) + \int_0^1 H_f(x + t(y - x))(y - x) \cdot (y - x)(1 - t)dt.$$

Ces inégalités entraînent :

$$\int_0^1 H_f(x + t(y - x))(y - x) \cdot (y - x)(1 - t)dt \geq 0$$

pour tout $x, y \in E$, nous avons donc

$$\int_0^1 H_f(x + tz)z \cdot z(1 - t)dt \geq 0$$

pour tout $x, z \in E$. En fixant $x \in E$, nous écrivons cette inégalité avec $z = \epsilon y$, $\epsilon > 0$ et $y \in \mathbb{R}^d$, nous obtenons

$$\epsilon^2 \int_0^1 H_f(x + t\epsilon y)y \cdot y(1 - t)dt \geq 0$$

pour tout $x, y \in E$, et $\epsilon > 0$, et donc :

$$\int_0^1 H_f(x + t\epsilon y)y \cdot y(1 - t)dt \geq 0, \quad \epsilon > 0.$$

Pour $(x, y) \in E^2$ fixé, $H_f(x + t\epsilon y)$ tend vers $H_f(x)$ uniformément lorsque ϵ tend vers 0, pour $t \in [0, 1]$. Nous avons donc :

$$\int_0^1 H_f(x)y \cdot y(1 - t)dt \geq 0,$$

c'est-à-dire

$$\frac{1}{2} H_f(x)y \cdot y \geq 0.$$

Donc pour tout $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$, $H_f(x)y \cdot y \geq 0$ donc $H_f(x)$ est positive.

Montrons maintenant la réciproque. Nous supposons que $H_f(x)$ est positive pour tout $x \in E$. Nous voulons démontrer que f est convexe ; nous allons pour cela utiliser la Proposition 2.4 et montrer que : $f(y) \geq f(x) + \nabla f(x) \cdot (y - x)$ pour tout $(x, y) \in E^2$. Grâce à (? ?), nous avons :

$$f(y) - f(x) = \nabla f(x) \cdot (y - x) + \int_0^1 H_f(x + t(y - x))\,(y - x) \cdot (y - x)(1 - t)dt.$$

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 $[0, 1]$. Nous avons 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.

Démontrons ensuite le point (ii). Nous supposons d'abord que $H_f(x)$ est strictement positive pour tout $x \in E$, et nous voulons montrer que f est strictement convexe. Nous allons encore utiliser la caractérisation de la Proposition 2.4. Pour un couple $(x, y) \in E^2$ tel que $y \neq x$. Alors :

$$f(y) = f(x) + \nabla f(x) \cdot (y - x) + \int_0^1 H_f(x + t(y - x))(y - x) \cdot (y - x)(1 - t)dt > 0.$$

Donc $f(y) > f(x) + \nabla f(x)(y - x)$ si $x \neq y$, ce qui prouve que f est strictement convexe. □

Dans le cas d'une fonctionnelle quadratique, nous considérons $A \in \mathcal{M}_{d,d}(\mathbb{R})$, $b \in \mathbb{R}^d$ et f la fonction de $\mathbb{R}^d$ dans $\mathbb{R}^d$ définie par

$$f(x) = \frac{1}{2} Ax \cdot x \; - \; b \cdot x.$$

Alors $f \in C^\infty(\mathbb{R}^d, \mathbb{R})$. Le calcul du gradient de f et de sa hessienne donne

$$\nabla f(x) = \frac{1}{2}(Ax + A^T x) - b.$$

Donc si A est symétrique $\nabla f(x) = Ax - b$ tandis que le calcul de la hessienne de f donne

$$H_f(x) = D(\nabla f(x)) = \frac{1}{2}(A + A^T).$$

Nous en déduisons que si A est symétrique, $H_f(x) = A$ et pouvons montrer en particulier que si A est symétrique définie positive alors il existe un unique $x \in \mathbb{R}^d$ tel que $f(x) \leq f(y)$ pour tout $y \in \mathbb{R}^d$, et que ce x est aussi l'unique solution du système linéaire $Ax = b$.

2.1 Algorithmes d'optimisation sans contrainte

Soient $E = \mathbb{R}^d$ et $f \in C(E, \mathbb{R})$. Nous supposons qu'il existe $\bar{x} \in E$ tel que

$$f(\bar{x}) \; = \; \inf_{x \in E} f(x).$$

Nous recherchons à calculer $\bar{x}$ (si f est de classe C^1, nous avons nécessairement $\nabla f(\bar{x}) = 0$) et nous allons donc maintenant développer des algorithmes permettant d'approcher le point $\bar{x}$ qui réalise le minimum de f.

Méthodes de descente

Définition 2.2 *Soient $f \in C(E, I\!R)$ et $E = I\!R^d$.*

- *Soit $x \in E$, nous disons que $w \in E$ avec $w \neq 0$ est une direction de descente en x s'il existe $\rho_0 > 0$ tel que*
$$f(x + \rho\, w) \leq f(x); \quad \forall \rho \in [0, \rho_0]$$
- *Soit $x \in E$, nous disons que $w \in E \setminus \{0\}$ est une direction de descente stricte en x s'il existe $\rho_0 > 0$ tel que*
$$f(x + \rho\, w) < f(x); \quad \forall \rho \in]0, \rho_0]$$

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

$$f(\bar{x}) = \inf_{x \in E} f(x)$$

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

(a) Initialisation $x^{(0)} = x_0 \in E$;

(b) Pour $k \geq 0$, nous recherchons w_k une direction de descente stricte de $x^{(k)}$.
Puis, nous prenons
$$x^{(k+1)} = x^{(k)} + \rho_k\, w_k; \quad \text{avec} \quad \rho_k > 0$$
bien choisi.

Pour construire, un tel algorithme nous avons besoin d'avoir une indication pour bien choisir la direction de descente stricte w_k au point $x^{(k)}$ et le scalaire $\rho_k > 0$. Pour cela, nous démontrons le résultat suivant

Proposition 2.6 *Soient $E = I\!R^d$, $f \in C^1(E, I\!R)$, $x \in E$ et $w \in E$ avec $w \neq 0$; nous avons alors*

- *si w direction de descente en x alors $w \cdot \nabla f(x) \leq 0$;*
- *si $\nabla f(x) \neq 0$ alors $w = -\nabla f(x)$ est une direction de descente stricte en x.*

Démonstration. Nous définissons la fonction φ de $I\!R$ dans $I\!R$ définie par : $\varphi(\rho) = f(x + \rho\, w)$. Nous avons alors $\varphi \in C^1(I\!R, I\!R)$ et $\varphi'(\rho) = \nabla f(x + \rho\, w) \cdot w$.

Commençons par le premier résultat et considérons $w \in E \setminus \{0\}$ une direction de descente au point x alors par définition, nous avons l'existence de $\rho_0 > 0$ tel que

$$f(x + \rho\, w) < f(w), \quad \forall \rho \in]0, \rho_0].$$

Comme w est une direction de descente, nous pouvons écrire pour tout $\rho \in [0, \rho_0]$

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

et donc pour tout $\rho \in]0, \rho_0]$

$$\frac{\varphi(\rho) - \varphi(0)}{\rho - 0} \leq 0,$$

en passant à la limite lorsque ρ tend vers 0, nous déduisons que $\varphi'(0) \leq 0$, ce qui signifie que

$$\nabla f(x) \cdot w \leq 0.$$

Montrons maintenant le deuxième résultat. Pour cela nous prenons $w = -\nabla f(x) \neq 0$ et souhaitons démontrer 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 plus tôt. Nous avons

$$\varphi'(0) = \nabla f(x) \cdot w = -|\nabla f(x)|^2 \ < \ 0.$$

Comme φ' est continue, il existe $\rho_0 > 0$ tel que si $\rho \in [0, \rho_0]$ alors $\varphi'(\rho) < 0$.

Lorsque $\rho \in]0, \rho_0]$ alors $\varphi(\rho) - \varphi(0) = \int_0^\rho \varphi'(s)ds \ < \ 0$, et donc nous avons bien pour tout $\rho \in]0, \rho_0]$

$$\varphi(\rho) \ < \ \varphi(0),$$

ce qui prouve que w est une direction de descente stricte en x.

Pour la réciproque, nous supposons que pour $w \in E$ tel que

$$\nabla f(x) \cdot w \leq 0,$$

ce qui signifie que $\varphi'(0) \leq 0$. Par continuité de la fonction φ', il existe $\rho_0 > 0$

$$\varphi'(\rho) \leq 0$$

□

Algorithme du gradient à pas fixe

Soient $f \in C^1(E, I\!R)$ et $E = I\!R^d$, l'algorithme du gradient à pas fixe s'écrit : pour $\rho > 0$ fixé

Algorithme 1. Algorithme du gradient à pas fixe

- Initialisation : $x^{(0)} \ = \ x_0 \ \in \ E$.

- Pour $k \geq 0$

  Nous calculons $w_k = -\nabla f(x^{(k)})$.

  - Nous posons $x^{(k+1)} \ = \ x^{(k)} \ + \ \rho \, w_k$.

Fin de la boucle sur k.

Théorème 2.2 (Convergence du gradient à pas fixe) *Soient $E = \mathbb{R}^d$ et $f \in C^1(E, \mathbb{R})$, nous supposons que :*

- *il existe $\alpha > 0$ tel que $(\nabla f(x) - \nabla f(y)) \cdot (x - y) \geq \alpha \|x - y\|^2$, pour tout $(x, y) \in E^2$,*
- *il existe $M > 0$ tel que $\|\nabla f(x) - \nabla f(y)\|^2 \leq M \|x - y\|^2$, pour tout $(x, y) \in E^2$,*

alors :

- *f est strictement convexe,*
- *$f(x) \to \infty$ quand $\|x\|$ tend vers l'infini,*
- *il existe un et un seul $\bar{x} \in E$ tel que*

$$f(\bar{x}) = \inf_{x \in E} f(x)$$

- *si $0 < \rho < 2\alpha/M^2$ alors la suite $(x^{(k)})_{k \in \mathbb{N}}$ construite par l'algorithme du gradient à pas fixe converge vers $\bar{x}$ lorsque k tend vers l'infini.*

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 $E = \mathbb{R}^d$ cet algorithme s'écrit :

Algorithme 2. Algorithme du gradient à pas optimal

-Initialisation : $x^{(0)} = x_0 \in E$.

-Pour $k \geq 0$

Nous calculons $w_k = -\nabla f(x^{(k)})$.

Nous choisissons $\rho_k \geq 0$ tel que

$$f(x^{(k)} + \rho_k w_k) \leq f(x^{(k)} + \rho\, w_k), \quad \rho \geq 0.$$

Nous posons $x^{(k+1)} = x^{(k)} + \rho_k\, w_k$.

Fin de la boucle sur k.

Les questions auxquelles on doit répondre pour s'assurer du bien fondé de ce nouvel algorithme sont les suivantes :

- Existe-t-il ρ_k tel que $f(x^{(k)} + \rho_k w_k) \leq f(x^{(k)} + \rho w^{(k)})$, pour tout $\rho \geq 0$.
- Comment calcule-t-on ρ_k ?
- La suite $(x^{(k)})_{k \in \mathbb{N}}$ construite par l'algorithme converge-t-elle ?

2.2 La méthode du gradient conjugué

Nous notons par $(.,.)$ le produit scalaire dans $\mathbb{R}^d$, pour tout x et $y \in \mathbb{R}^d$

$$(x\,,\,y) = x^T\,y.$$

La méthode du gradient conjugué consiste à minimiser la fonctionnelle quadratique $f(x)$ de la forme

$$f(x) = \frac{1}{2}\,(x\,,\,A\,x)\,-\,(b\,,\,x)\,,$$

où $A \in \mathcal{M}_{d,d}(\mathbb{R})$ est une matrice symétrique définie positive et $b \in \mathbb{R}^d$. D'une part, la fonction f est minorée

$$f(x)\,\geq\,\frac{\lambda_d}{2}\,\|x\|^2\,-\,\|b\|\,\|x\|\,\geq\,-\frac{\|b\|}{2\,\lambda_d}.$$

où λ_d est la plus petite valeur propre de A. La fonction f étant minorée, il existe $\alpha \in \mathbb{R}$ tel que

$$\alpha\;=\;\inf_{y\in\mathbb{R}^d} f(y).$$

De plus, en utilisant le résultat précédent, nous pouvons trouver un $R > 0$ tel que pour tout $\|x\| \geq R$, $f(x) \geq \alpha + 1$, cela signifie que

$$\inf_{y\in\mathbb{R}^d} f(y)\;=\;\inf_{\substack{y\in\mathbb{R}^d\\ \|y\|<R}} f(y).$$

Or, $\overline{\mathcal{B}}(0,R)$, la boule fermée de centre $O \in \mathbb{R}^d$ et de rayon R, est un ensemble fermé borné de $\mathbb{R}^d$, c'est donc un compact et comme f est continue, il existe $\bar{x} \in \mathcal{B}(0,R)$ tel que

$$f(\bar{x})\;=\;\inf_{\substack{y\in\mathbb{R}^d\\ \|y\|<R}} f(y).$$

De plus, $\nabla f(\bar{x}) = A\,\bar{x} - b = 0$. Comme ce système admet une unique solution, le minimiseur est unique.

Avant de décrire la méthode du gradient conjugué, nous introduisons la définition de *vecteurs conjugués*.

Définition 2.3 *Soit $A \in \mathcal{M}_{d,d}(\mathbb{R})$ une matrice symétrique définie positive. Nous dirons que deux vecteurs non nuls v et w de $\mathbb{R}^d$ sont A-conjugués lorsque*

$$w^T\,A\,v\;=\;v^T\,A\,w\;=\;0.$$

De plus, la famille $\{w^{(1)},\ldots,w^{(p)}\}$ de $\mathbb{R}^d$, composée de vecteurs tous non nuls, est dite A-conjuguée si $\left(w^{(i)}\,,\,A\,w^{(j)}\right)\,=\,0$ pour tout couple $(i,j) \in \{1,...,p\}^2$ tel que $i \neq j$.

Proposition 2.7 *Soient $A \in \mathcal{M}_{d,d}(\mathbb{R})$ une matrice symétrique définie positive et $\{w^{(1)},\ldots,w^{(p)}\}$ une famille de $\mathbb{R}^d$, alors*

– *si la famille* $\{w^{(1)}, \ldots, w^{(p)}\}$ *est* A*-conjuguée alors elle est libre ;*
– *dans le cas où* $p = d$*, si la famille* $\{w^{(1)}, \ldots, w^{(d)}\}$ *est* A*-conjuguée alors elle forme une base de* $\mathbb{R}^d$.

Démonstration. La deuxième partie de la Proposition est immédiate dès que nous avons démontré la première partie, nous détaillons donc seulement la première partie. Supposons que $\{w^{(1)}, ..., w^{(p)}\}$ est une famille A-conjuguée, c'est-à-dire le vecteur $w^{(i)}$ est non nul et vérifie

$$\left(w^{(i)}\,,\, A\,w^{(j)}\right) \,=\, 0, \quad i \neq j.$$

Pour tout $i = 1, ..., p$, nous considérons $\alpha_i \in \mathbb{R}$ tels que

$$\sum_{i=1}^{p} \alpha_i\, w^{(i)} \,=\, 0,$$

nous voulons démontrer que cela implique que tous les coefficients α_i, pour $i = 1, \ldots, p$ sont nuls. En effet, fixons $i = 1, \ldots, p$ et multiplions cette dernière égalité par $A\,w^{(i)}$ nous avons donc

$$\sum_{k=1}^{p} \alpha_k \,\left(w^{(k)}\,,\, A\,w^{(i)}\right) \,=\, 0$$

et donc en utilisant la définition de vecteurs A-conjugués, nous déduisons que

$$\alpha_i \,\left(w^{(i)}\,,\, A\,w^{(i)}\right) \,=\, 0.$$

Or, $\left(w^{(i)}\,,\, A\,w^{(i)}\right) \,\neq\, 0$ car $w^{(i)}$ est non nul et A est symétrique définie positive. Nous en déduisons que $\alpha_i = 0$ pour tout $i = 1, \ldots, p$. La famille $(w^{(1)}, \ldots, w^{(p)})$ est donc libre. □

La méthode du gradient conjugué va consister à construire une suite $(w^{(k)})_k$ de direction A-conjuguées et une suite de solutions approchées $(x^{(k)})_k$ telles que les gradients successifs $\nabla f(x^{(k)})$ soient orthogonaux deux à deux et aux directions $(w^{(l)})_{0 \leq l < k}$ précédentes. Ainsi, le vecteur

$$\nabla f(x^{(k)}) = A\,x^{(k)} \,-\, b$$

sera orthogonal à l'ensemble des vecteurs libres $\{w^{(0)}, \ldots, w^{(k-1)}\}$. Nous construisons donc une suite $x^{(0)}, \ldots, x^{(k)}$ et $w^{(0)}, \ldots, w^{(k-1)}$, puis à l'étape suivante nous recherchons une direction $w^{(k)}$ telle que :

– $w^{(k)}$ soit A-conjuguée avec $w^{(p)}$ pour tout $p < k$.
– $w^{(k)}$, non nul, soit une direction orthogonale au gradient $\nabla f(x^{(k-1)})$,

Ainsi, lorsque nous trouvons $w^{(k)}$, nous posons

$$x^{(k+1)} \,=\, x^{(k)} \,+\, \rho_k\, w^{(k)},$$

avec ρ_k tel que $\nabla f(x^{(k+1)})$ soit orthogonal à la direction $w^{(k)}$.

Nous proposons alors l'algorithme suivant pour une matrice $A \in \mathcal{M}_{d,d}(\mathbb{R})$ symétrique définie positive et $b \in \mathbb{R}^d$.

Algorithme 4. Méthode du gradient conjugué.

Nous posons

$$f(x) = \frac{1}{2}\,(x\,,\,A\,x)\;-\;(b\,,\,x)\,.$$

Pour $x_0 \in I\!R^d$, nous prenons $x^{(0)} = x_0$ et $w^{(-1)} = 0 \in I\!R^d$.

Pour $k \geq 0$ nous construisons

$$r^{(k)} \;=\; -\nabla f(x^{(k)}) \;=\; b \;-\; A\,x^{(k)}.$$

-Si $r^{(k)} \;=\; 0$ nous avons $A\,x^{(k)} = b$ donc $x^{(k)} = \bar{x}$

auquel cas l'algorithme s'arrête.

-Si $r^{(k)} \neq 0$, alors nous posons $w^{(k)} = r^{(k)} + \alpha_{k-1}\,w^{(k-1)}$,

avec α_{k-1} tel que $\left(w^{(k)}\,,\,A\,w^{(k-1)}\right) \;=\; 0$.

Puis, nous construisons $x^{(k+1)} \;=\; x^{(k)} \;+\; \rho_k\,w^{(k)}$

et choisissons ρ_k tel que

$$\left(\nabla f(x^{(k+1)})\,,\,w^{(k)}\right);$$

c'est-à-dire

$$\rho_k \;=\; \frac{\left(r^{(k)}\,,\,w^{(k)}\right)}{\left(A\,w^{(k)}\,,\,w^{(k)}\right)}.$$

Théorème 2.3 *Soient $A \in \mathcal{M}_{d,d}(I\!R)$ une symétrique définie positive et $b \in I\!R^d$. Nous posons*

$$f(x) = \frac{1}{2}\,(x\,,\,A\,x)\;-\;(b\,,\,x)\,.$$

Alors, l'algorithme du gradient conjugué définit une suite $(x^{(k)})_{k=0,\ldots,p}$ avec $p \leq n$ telle que $x^{(p)} = \bar{x}$ avec $A\,\bar{x} \;=\; b$.

Démonstration. L'objectif est d'écrire la solution dans une base formée par les vecteurs $(w^{(k)})_{0\leq k\leq d-1}$ et donc que la méthode converge en au plus d itérations.

D'abord si $r^{(0)} = 0$, alors $A\,x^{(0)} \;=\; b$ et donc $x^{(0)} = \bar{x}$ auquel cas $p = 0$.

Ensuite, lorsque $r^{(0)} \neq 0$, comme par construction

$$w^{(0)} = r^{(0)} = b - A\,x^{(0)} = -\nabla f(x^{(0)}),$$

Nous calculons la valeur de ρ_0 obtenue par

$$\rho_0 \;=\; \frac{\big(r^{(0)}\,,\,w^{(0)}\big)}{\big(w^{(0)}\,,\,A\,w^{(0)}\big)}.$$

L'élément $x^{(1)}$ est donc bien défini et donné par

$$x^{(1)} \;=\; x^{(0)} \;+\; \rho_0\,w^{(0)}.$$

Notons que

$$r^{(1)} \;=\; b \;-\; A\,x^{(1)} \;=\; r^{(0)} \;-\; \rho_0\,A\,w^{(0)}$$

et donc $(r^{(1)},\,w^{(0)}) \;=\; 0$.

À l'itération k, nous supposons que nous avons construit $x^{(0)},\ldots,x^{(k)}$ et $w^{(0)},\ldots,w^{(k-1)}$. Alors, nous posons

$$r^{(k)} \;=\; b \;-\; A\,x^{(k)}.$$

et supposons

$$\begin{cases} (r^{(k)}\,,\,r^{(j)}) \;=\; 0, & \text{pour } 0 \leq j < k \\ (r^{(k)}\,,\,w^{(j)}) \;=\; 0, & \text{pour } 0 \leq j < k \\ (w^{(k-1)}\,,\,A\,w^{(j)}) \;=\; 0, & \text{pour } 0 \leq j < k-1 \end{cases} \tag{2.3}$$

D'une part, si $r^{(k)} = 0$ alors $A\,x^{(k)} = b$ et donc $x^{(k)} = \bar{x}$ auquel cas l'algorithme s'arrête et $p = k$.

D'autre part, lorsque $r^{(k)}$ est non nul, nous posons

$$w^{(k)} = r^{(k)} + \alpha_{k-1}\,w^{(k-1)}.$$

Comme $w^{(k-1)}$ est non nul nous pouvons choisir α_{k-1} tel que

$$\big(w^{(k)}\,,\,A\,w^{(k-1)}\big) = 0,$$

c'est-à-dire

$$\big(r^{(k)} + \alpha_{k-1}\,w^{(k-1)}\,,\,A\,w^{(k-1)}\big) \;=\; 0,$$

pour cela, il suffit de prendre

$$\alpha_{k-1} \;=\; \frac{\big(r^{(k)}\,,\,A\,w^{(k-1)}\big)}{\big(w^{(k-1)}\,,\,A\,w^{(k-1)}\big)}.$$

De plus, pour ce choix de α_{k-1} et lorsque $j < k$ nous avons

$$\left(w^{(k)}\,,\,A\,w^{(j)}\right) \;=\; \left(r^{(k)} + \alpha_{k-1}\,w^{(k-1)}\,,\,A\,w^{(j)}\right) \;=\; 0,$$

ce qui prouve la troisième égalité de (2.3) à l'étape $(k+1)$, c'est-à-dire que la famille $\{w^{(0)}, \ldots, w^{(k)}\}$ est A-conjuguée.

Ensuite, à partir de $w^{(k)}$ nous construisons ρ_k et posons

$$\rho_k \;=\; \frac{\left(r^{(k)}\,,\,w^{(k)}\right)}{\left(A\,w^{(k)}\,,\,w^{(k)}\right)}.$$

Nous pouvons alors bien définir $x^{(k+1)} = x^{(k)} + \rho_k w^{(k)}$ et

$$r^{(k+1)} = -\nabla f(x^{(k+1)}) \;=\; b \;-\; A\,x^{(k+1)} = r^{(k)} \;-\; \rho_k\,A\,w^{(k)}.$$

Remarquons que le choix de ρ_k entraîne que

$$\left(r^{(k+1)}\,,\,w^{(k)}\right) = 0$$

et en utilisant les deux premières hypothèses de recurrence (2.3), il vient également pour tout $j < k$

$$\left(r^{(k+1)}\,,\,w^{(j)}\right) = \left(r^{(k)} \;-\; \rho_k\,A\,w^{(k)}\,,\,w^{(j)}\right) \;=\; 0$$

ce qui démontre la deuxième égalité de (2.3) à l'étape $(k+1)$.

Enfin, par construction pour tout $j < k$, nous avons $r^{(j)} = w^{(j)} - \alpha_{j-1}\,w^{(j-1)}$

$$\left(r^{(k+1)}\,,\,r^{(j)}\right) \;=\; \left(r^{(k+1)}\,,\,w^{(j)} - \alpha_{j-1}\,w^{(j-1)}\right)$$

et en utilisant ce que nous avons démontré précédemment

$$\left(r^{(k+1)}\,,\,r^{(j)}\right) \;=\; \left(r^{(k+1)}\,,\,w^{(j)}\right) \;-\; \alpha_{j-1}\,\left(r^{(k+1)}\,,\,w^{(j-1)}\right) \;=\; 0,$$

ce qui conclut la récurrence.

Grâce à ce résultat, nous obtenons que lorsque $r^{(k)}$ est non nul pour $k = 0, ..., d-1$, la famille $\{w^{(0)},\ldots,w^{(d-1)}\}$ est donc A-conjuguée, elle forme donc une base de $\mathbb{R}^d$ et $w^{(k)}$ est orthogonal à $r^{(d)} = b - A\,x^{(d)}$ pour tout $k \leq d-1$. Nous en déduisons que $r^{(d)}$ est forcément nul et donc $x^{(d)} = \bar{x}$. □

3 Optimisation sous contraintes

Soient $E = \mathbb{R}^d$, $f \in C(E, \mathbb{R})$, et K un sous ensemble de E.Nous nous intéressons à la recherche de $\bar{x} \in K$ tel que

$$\begin{aligned} &\bar{x} \in K \\ &f(\bar{x}) = \inf_{x\in K} f(x) \end{aligned} \tag{3.4}$$

Ce problème est un problème de minimisation avec contrainte (ou "sous contrainte") au sens où nous recherchons $\bar{x}$ qui minimise f tout en imposant à $\bar{x}$ d'appartenir à K. Voyons quelques exemples de ces contraintes (définies par l'ensemble K), que nous allons expliciter à l'aide des p fonctions continues, $g_i \in C(E, \mathbb{R})$ pour tout $i = 1, \ldots, p$.

Contraintes d'égalités. Nous posons

$$K = \{x \in E, \quad g_i(x) = 0, \quad i = 1, \ldots, p\}.$$

Nous verrons plus loin que le problème de minimisation de f peut alors être résolu grâce au théorème des multiplicateurs de Lagrange (voir Théorème 3.5).

Contraintes d'inégalités. Nous posons

$$K = \{x \in E, \quad g_i(x) \leq 0, \quad i = 1, \ldots, p\}.$$

Ce problème de minimisation de f peut alors être résolu grâce au théorème de Kuhn-Tucker (nous ne traiterons pas ce cas ici [3]).

Programmation linéaire. Avec un tel ensemble de contraintes K, si de plus f est linéaire, c'est-à-dire qu'il existe $b \in \mathbb{R}^d$ tel que $f(x) = b \cdot x$, et les fonctions g_i sont affines, c'est-à-dire qu'il existe $b_i \in \mathbb{R}^d$ et $c_i \in \mathbb{R}$ tels que $g_i(x) = b_i \cdots x + c_i$, alors nous devons résoudre un problème de programmation linéaire. Ces problèmes sont souvent résolus numériquement à l'aide de l'algorithme de Dantzig, inventé dans les années 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

$$f(x) = \frac{1}{2} A\, x \cdots x - b \cdots x,$$

et les fonctions g_i sont affines, alors nous devons résoudre un problème de programmation quadratique.

Programmation convexe. Dans le cas où f est convexe et K est convexe, nous devons résoudre un problème de programmation convexe.

3.1 Existence et unicité, conditions d'optimalité simple

Théorème 3.1 (Existence) *Soient $E = \mathbb{R}^d$ et $f \in C(E, \mathbb{R})$.*

Si K est un sous-ensemble fermé borné de E, alors il existe $\bar{x} \in K$ tel que

$$f(\bar{x}) = \inf_{x \in K} f(x).$$

Si K est un sous-ensemble fermé de E, et si $f(x)$ tend vers l'infini lorsque $\|x\|$ tend vers l'infini, alors il existe $\bar{x} \in K$ tel que $f(\bar{x}) = \inf_{x \in K} f(x)$

Démonstration. 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 $\bar{x}$.

Si f tend vers l'infini lorsque $\|x\|$ tend vers l'infini, alors il existe $R > 0$ tel que si $\|x\| > R$ alors $f(x) > f(0)$; donc

$$\inf_{x \in K} f(x) = \inf_{x \in K \cap B(0,R)} f(x),$$

où $B(0, R)$ désigne la boule de centre 0 et de rayon R. L'ensemble $K \cap B(0, R)$ est compact, comme l'intersection d'un ensemble fermé et d'un ensemble compact. D'où, d'après ce qui précède, il existe $\bar{x} \in K$ tel que

$$f(\bar{x}) = \inf_{x \in K \cap B(0,R)} f(x) = \inf_{x \in B(0,R)} f(x).$$

□

Théorème 3.2 *Soient $E = I\!R^d$ et $f \in C(E, I\!R)$. Nous supposons que f est strictement convexe et que K est convexe. Alors il existe au plus un élément $\bar{x}$ de K tel que $f(\bar{x}) = \inf_{x \in K}(x)$.*

Démonstration. Supposons que $\bar{x}$ et $\tilde{x}$ soient deux solutions du problème (3.4), avec $\bar{x} \neq \tilde{x}$. Alors

$$f(\frac{1}{2}\bar{x} + \frac{1}{2}\tilde{x}) < \frac{1}{2}f(\bar{x}) + \frac{1}{2}f(\tilde{x}) = \inf_{x \in K} f(x).$$

Nous aboutissons alors à une contradiction. □

Des deux théorèmes précédent nous déduisons immédiatement le théorème d'existence et d'unicité suivant :

Théorème 3.3 (Existence et unicité) *Soient $E = I\!R^d$, $f \in C(E, I\!R)$ une fonction strictement convexe et K un sous ensemble convexe fermé de E. Si K est borné ou si f est croissante à l'infini, c'est-à-dire si $f(x) \to +\infty$ quand $\|x\|$ tend vers l'infini, alors il existe un unique élément $\bar{x}$ de K solution du problème de minimisation (3.4).*

Remarque 3.1 *Nous pouvons remplacer $E = I\!R^d$ par E espace de Hilbert de dimension infinie dans le dernier théorème, mais nous avons besoin dans ce cas de l'hypothèse de convexité de f pour assurer l'existence de la solution.*

Proposition 3.1 (Condition simple d'optimalité) *Soient $E = I\!R^d$, $f \in C(E, I\!R)$ et $\bar{x} \in K$ tel que*

$$f(\bar{x}) = \inf_{x \in K} f(x).$$

Nous supposons que f est différentiable en $\bar{x}$.

- *Si $\bar{x}$ appartient à l'intérieur de K alors $\nabla f(\bar{x}) = 0$.*
- *Si K est convexe, alors $\nabla f(\bar{x}) \cdot (x - \bar{x}) \geq 0$ pour tout $x \in K$.*

Démonstration. Supposons d'abord que $\bar{x}$ appartient à l'intérieur de K, alors il existe $r > 0$ tel que $B(\bar{x}, r) \subset K$ et $f(\bar{x}) \leq f(x)$ pour tout $x \in B(\bar{x}, r)$. Alors nous avons déjà vu que ceci implique $\nabla f(\bar{x}) = 0$.

Supposons maintenant $x \in K$. Comme $\bar{x}$ réalise le minimum de f sur K, nous avons :

$$f(\bar{x} + t(x - \bar{x})) = f(tx + (1-t)\bar{x}) \geq f(\bar{x})$$

pour tout $t \in]0, 1]$, par convexité de K, nous en déduisons que

$$\frac{f(\bar{x} + t(x - \bar{x})) - f(\bar{x})}{t} \geq 0$$

pour tout $t \in]0, 1]$.

En passant à la limite lorsque t tend vers 0 dans cette dernière inégalité, nous obtenons : $\nabla f(\bar{x}) \cdot (x - \bar{x}) \geq 0$. □

3.2 Conditions d'optimalité dans le cas de contraintes d'égalité

Commençons par rappeler le Théorème des fonctions implicites

Théorème 3.4 *Soient p et q deux entiers naturels et $h \in C^1(\mathbb{R}^q \times \mathbb{R}^p, \mathbb{R}^p)$, et soient $(\bar{x}, \bar{y}) \in \mathbb{R}^q \times \mathbb{R}^p$ et $c \in \mathbb{R}^p$ tels que $h(\bar{x}, \bar{y}) = c$. Supposons que la matrice de la différentielle $\nabla_y h(\bar{x}, \bar{y}) \in \mathcal{M}_{p,p}(\mathbb{R})$ soit 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 $h(x, y) = c$. Nous pouvons ainsi définir une application ϕ de $B(\bar{x}, \varepsilon)$ dans $B(\bar{y}, \nu)$ par $\phi(x) = y$. Nous avons $\phi(\bar{x}) = \bar{y}$, $\phi \in C^1(\mathbb{R}^p, \mathbb{R}p)$ et*

$$\nabla\phi(x) = -[\nabla_y h(x, \phi(x))]^{-1} \nabla_x h(x, \phi(x)).$$

Dans tout ce paragraphe, nous considérerons les hypothèses et notations suivantes :

– nous supposons $f \in C(\mathbb{R}^d, \mathbb{R})$, $g_i \in C^1(\mathbb{R}^d, \mathbb{R})$, $i = 1, \ldots, p$;

– le domaine K est défini par

$$K = \{u \in R^d, g_i(x) = 0 \quad i = 1, \ldots, p\};$$

avec $g = (g_1, \ldots, g_p)^T \in C^1(\mathbb{R}^d, \mathbb{R}^p)$

Nous avons alors le théorème suivant

Théorème 3.5 (Multiplicateurs de Lagrange) *Soit $\bar{x} \in K$ tel que*

$$f(\bar{x}) = \inf_{x \in K} f(x).$$

Nous supposons que f est différentiable en $\bar{x}$ et $\dim(Im(\nabla g(\bar{x}))) = p$ ou autrement dit $rang(\nabla g(\bar{x})) = p$), alors il existe $(\lambda_1, \ldots, \lambda_p)^T \in \mathbb{R}^p$ tels que

$$\nabla f(\bar{x}) + \sum_{i=1}^{p} \lambda_i \nabla g_i(\bar{x}) = 0.$$

Démonstration. Par hypothèse, nous avons $\nabla g(\bar{x}) \in \mathcal{L}(I\!R^d, I\!R^p)$ et $Im(\nabla g(\bar{x})) = I\!R^p$, il existe donc un sous-espace vectoriel F de $I\!R^d$ de dimension p, tel que $\nabla g(\bar{x})$ soit bijective de F dans $I\!R^p$. En effet, soit $(e_1, \ldots, e_p)$ la base canonique de $I\!R^p$, alors pour tout $i \in \{1, \ldots, p\}$, il existe $y_i \in I\!R^d$ tel que

$$\nabla g(\bar{x})\, y_i = e_i.$$

Soit F le sous espace engendré par la famille $\{y_1, \ldots, y_p\}$. Montrons d'abord que cette famille est libre : soit $(\lambda_i)_{1 \leq i \leq p}$ tels que

$$\sum_{i=1}^{p} \lambda_i\, y_i = 0.$$

Montros que cela implique que $\lambda_i = 0$ pour tout $i = 1, ..., p$. En effet, en multipliant par $\nabla g(\bar{x})$, nous obtenons

$$\sum_{i=1}^{p} \lambda_i\, e_i = 0,$$

et donc $\lambda_i = 0$ pour tout $i = 1, \ldots, p$, ce qui démontre bien que la famille est libre. La famille $\{y_1, ..., y_p\}$ est libre et génératrice de F, elle forme donc une base de F

Nous avons ainsi montré l'existence d'un sous-espace F de dimension p telle que $\nabla g(\bar{x})$ soit bijective, puisqu'elle est surjective, de F dans $I\!R^p$.

Dans la suite, nous allons faire une séparation des variables pour appliquer le théorème des fonctions implicites.

Nous construisons alors un sous espace-vectoriel G de $I\!R^d$, tel que

$$I\!R^d = F \oplus G.$$

Puis, pour $y \in F$ et $z \in G$, nous définissons les fonctions $\hat{g}$ et $\hat{f}$ à partir des fonctions g et f :

– $\hat{g}(y, z) = g(y + z)$;

– $\hat{f}(y, z) = f(y + z)$;

La régularité des fonctions g et f assure que $\hat{f} \in C(F \times G, I\!R)$ et $\hat{g} \in C^1(F \times G, I\!R)$.

D'autre part, nous pouvons définir les dérivées par rapport à y et z de $\hat{g}$ et $\hat{f}$, nous avons

$$\nabla_y \hat{g}(y, z) \in \mathcal{L}(F, I\!R^p),$$

et pour tout $w \in F \subset I\!R^d$, nous avons

$$\nabla_y \hat{g}(y, z)\, w = \nabla g(y + z)\, w.$$

Soit $(\hat{y}, \bar{z}) \in F \times G$ tel que $\bar{x} = \bar{y} + \bar{z}$, nous avons pour $\hat{g}$

$$\nabla_y \hat{g}(\bar{y}, \bar{z})\, w = \nabla g(\bar{x})\, w$$

pour tout $w \in F \subset I\!R^d$. Par définition de F, $\nabla g(\bar{x})$ est bijective de F sur $I\!R^p$, ce qui signifie que l'application $\nabla_y \hat{g}(\bar{y}, \bar{z})$ est aussi une bijection de F sur $I\!R^p$.

Nous rappelons que l'ensemble K a été défini par $K = \{x \in I\!R^d; g(x) = 0\}$ et nous construisons alors l'ensemble $\hat{K}$ par

$$\hat{K} = \{(y,z) \in F \times G, \quad \hat{g}(y,z) = 0\}.$$

Par définition de $\hat{f}$ et de $\hat{g}$, nous avons

$$\begin{cases} (\bar{y},\bar{z}) \in \hat{K} \\ \\ \hat{f}(\bar{y},\bar{z}) \leq \hat{f}(y,z), \quad \forall\,(y,z) \ \in \ \hat{K}. \end{cases} \tag{3.5}$$

Par construction de la fonction $\hat{g}$ et de l'ensemble des contraintes $\hat{K}$, nous pouvons appliquer le théorème des fonctions implicites à la fonction $\hat{g}$. Ceci entraîne l'existence de $\varepsilon > 0$ et $\nu > 0$ tels que pour tout $z \in B_G(\bar{z},\varepsilon)$ il existe un unique $y \in B_F(\bar{y},\nu)$ tel que $\hat{g}(y,z) = 0$. Nous pouvons alors définir une fonction ϕ

$$\begin{array}{lll} B_F(\bar{y},\nu) & \longrightarrow & B_G(\bar{z},\varepsilon) \\ \\ y & \longrightarrow & z = \phi(y), \end{array}$$

où cette application $\phi \in C^1(B_G(\bar{z},\varepsilon), B_F(\bar{y},\eta))$.

Nous déduisons alors de (3.5) que :

$$\hat{f}(\phi(\bar{z}),\bar{z}) \leq \hat{f}(\phi(z),z), \quad \forall\, z \ \in B_G(\bar{z},\eta),$$

ou encore

$$f(\bar{x}) = f(\bar{z}+\phi(\bar{z})) \leq f(z+\phi(z)), \quad \forall\, z \ \in B_G(\bar{z},\varepsilon).$$

En posant $\psi(z) \ = \ \hat{f}(\phi(z)+z)$, nous pouvons donc écrire

$$\psi(\bar{z}) \ = \ \hat{f}(\phi(z),z) \ \leq \ \psi(z), \quad \forall\, z \ \in B_G(\bar{z},\varepsilon).$$

Nous avons donc, grâce à la Proposition 3.1,

$$\nabla\psi(\bar{z}) \ = \ 0.$$

En utilisant la définition de ψ, nous pouvons calculer le gradient de ψ en fonction de $\hat{f}$ et ϕ au point $\bar{z} \in G$

$$\nabla\psi(\bar{z}) \ = \ \Big(\nabla_z\hat{f} \ + \ \nabla_y\hat{f}\,\nabla\phi\Big)\,(\phi(\bar{z}),\bar{z}).$$

D'après le théorème des fonctions implicites, nous pouvons exprimer $\nabla\phi$

$$\nabla\phi(\bar{z}) \ = \ [\nabla_y\hat{g}]^{-1} \ \nabla_z\hat{g}(\phi(\bar{z}),\bar{z})$$

Nous en déduisons donc que

$$\left(\nabla_z \hat{f} \,+\, \nabla_y \hat{f}\, [\nabla_y \hat{g}]^{-1}\, \nabla_z \hat{g}\right)(\phi(\bar{z}), \bar{z})\, z \,=\, 0, \quad \forall\, z \in G. \tag{3.6}$$

De plus, comme

$$[\nabla_y \hat{g}(\phi(\bar{z}), \bar{z})]^{-1}\, \nabla_y \hat{g}(\phi(\bar{z}), \bar{z}) \,=\, I_p$$

nous avons donc

$$\left(\nabla_y \hat{f} \,-\, \nabla_y \hat{f}\, [\nabla_y \hat{g}]^{-1}\, \nabla_z \hat{g}\right)(\phi(\bar{z}), \bar{z})\; y = \,0, \quad \forall\, y \,\in\, F. \tag{3.7}$$

Soient $x \in I\!R^d$, et $(y, z) \in F \times G$ tel que $x = y + z$; en additionant (3.6) et (3.7), et en notant $\Lambda = \nabla_y \hat{f}\, [\nabla_y \hat{g}]^{-1}$; nous obtenons finalement

$$\left(\nabla \hat{f} \,-\, \Lambda\, \nabla g\right)(\phi(\bar{z}), \bar{z})\, x \,=\, 0.$$

ou encore

$$\nabla f(\bar{x}) + \sum_{i=1}^{p} \lambda_i\, \nabla g(\bar{x})\, w = 0,$$

où $\lambda_i = \Lambda_i$ pour $i = 1, ..., p$ □

Chapitre 5

Les polynômes

1 Motivation : l'interpolation de fonctions

1.1 Un exemple en Analyse

Commençons par un exemple où l'analyse numérique vient au secours de l'analyse. Nous avons déjà vu le Théorème de Weierstrass qui permet d'approcher une fonction continue

Théorème 1.1 (Weierstrass) *Pour toute fonction f continue à support compact sur $\mathbb{R}$, il existe une suite de fonctions entières qui tend vers f uniformément sur $\mathbb{R}$; en particulier, il existe une suite de fonctions polynomiales qui tend vers f uniformément sur tout compact.*

Pour démontrer le théorème d'approximation polynomiale de Weierstrass sur la droite réelle, nous considérons la fonction réelle définie sur $\mathbb{R}$ par

$$\gamma(x) = e^{-x^2/2}/\sqrt{2\pi}.$$

Il s'agit bien entendu de la densité gaussienne standard ; cette fonction est réelle, positive et son intégrale sur $\mathbb{R}$ vaut un. Par ailleurs, elle est développable en série entière de rayon de convergence infini. On peut dès lors imaginer qu'une convolution $f \star \gamma$, pour f continue à support compact, sera encore développable en série entière de rayon infini, donc nous pouvons l'approcher sur tout compact par des fonctions polynomiales (tout simplement les sommes partielles de la série entière) ; d'après les résultats classiques d'approximation par convolution, on sait qu'on peut approcher f uniformément par des convolées qui sont en gros du type $f \star \gamma$ (et qui par conséquent sont aussi des fonctions entières) : c'est la stratégie conceptuellement très simple et originelle qui est employée par Weierstrass en 1885 pour prouver ce théorème.

Cependant, cette démonstration n'est en rien constructive et nous n'avons aucune idée de la forme que prennent de tels polynômes. En fait, on peut démontrer que les polynômes de Bernstein

$$B_i^m(x) \;=\; \left(\begin{array}{c} m \\ \\ i \end{array} \right) x^i \, (1-x)^{m-i},$$

où

$$\left(\begin{array}{c} m \\ \\ i \end{array} \right) = \frac{m!}{i!\,(m-i)!}$$

sont les coefficients binomiaux, conviennent parfaitement et peuvent être calculés exactement en utilisant les opérations de bases comme l'addition, la division et la multiplication. Ceci signifie que nous pouvons calculer des approximations aussi précise que nous le souhaitons de fonctions seulement continues ($\cos$, $\sin$, etc).

1.2 Courbes de Bézier

Les courbes de Bézier sont des courbes polynomiales paramétriques décrites pour la première fois en 1962 par l'ingénieur français Pierre Bézier qui les utilisa pour concevoir des pièces d'automobiles à l'aide d'ordinateurs. Elles ont de nombreuses applications dans la synthèse d'images et le rendu de fontes. Elles ont donné naissance à de nombreux autres objets mathématiques.

Pour $n+1$ points de contrôle $(P_0, \ldots, P_n)$, on définit une courbe de Bézier par l'ensemble des points

$$\sum_{i=0}^{n} B_i^n(x)\, P_i, \quad x \in [0,1],$$

où les B_i^n sont les polynômes de Bernstein. Le polygone $P_0, \ldots, P_n$ est appelé "polygone convexe de Bézier".

Quatre points P_0, P_1, P_2 et P_3 définissent une courbe de Bézier cubique. La courbe se trace en partant du point P_0, en se dirigeant vers P_1 et en arrivant au point P_3 selon la direction $P_2 - P_3$. En général, la courbe ne passe ni par P_1 ni par P_2 : ces points sont simplement là pour donner une information de direction. La distance entre P_0 et P_1 détermine la "longueur" du déplacement dans la direction de P_1 avant de tourner vers P_3.

La forme paramétrique de la courbe s'écrit :

$$P(t) \;=\; P_0\,(1-t)^3 \;+\; 3\,P_1\,t\,(1-t)^2 \;+\; 3\,P_2\,t^2\,(1-t) \;+\; P_3\,t^3$$

pour $t \in [0,1]$.

Les courbes de Bézier de degré trois sont intéressantes pour le traitement des images pour deux raisons principales :

- les points peuvent être calculés rapidement (méthode de Horner) ;
- les courbes de Bézier cubiques permettent d'assurer la continuité en tangence et en courbure de deux courbes raccordées. Par exemple, pour une courbe définie par les points A, B, C, D, E, F et G, nous utilisons les courbes cubiques définies par A, B, C, et D, et par D, E, F, et G et la continuité est ainsi assurée. Pour avoir une courbe C^1 en D, il faut que $[C, D] = [D, E]$, et si en plus on veut qu'elle soit C^2 en D, alors $[B, D] = [D, F]$, et de même pour les dérivées successives.

Les courbes de Bézier cubiques, les plus utilisées, se retrouvent en graphisme et dans de multiples systèmes de synthèse d'images, tels que PostScript, Metafont et The GIMP, pour dessiner des courbes "lisses" joignant des points ou des polygones de Bézier.

Dans ce chapitre, nous commencerons par donner une première approche basée sur l'interpolation de Lagrange, nous verrons ensuite d'autres techniques plus efficaces (polynômes orthogonaux, approximation au sens des moindres carrés).

2 Polynômes de Lagrange

Soit f une fonction de l'intervalle $[a, b]$ à valeur dans $\mathbb{R}$ connue en $n + 1$ points distincts x_0,...,x_n de l'intervalle $[a, b]$. Il s'agit de construire un polynôme P_n de degré inférieur ou égal à n tel que

$$P_n(x_i) \;=\; f(x_i), \qquad i = 0, ..., n. \tag{2.1}$$

2.1 Construction et convergence de l'interpolation de Lagrange

Nous avons alors le résultat suivant

Théorème 2.1 *Il existe un unique polynôme de degré inférieur ou égal à n solution de (2.1). Ce polynôme s'écrit alors*

$$P_n(x) := \sum_{i=0}^{n} f(x_i)\, L_i(x), \tag{2.2}$$

avec

$$L_i(x) \;:=\; \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{x - x_j}{x_i - x_j}. \tag{2.3}$$

Démonstration. Nous vérifions d'abord que le polynôme (2.2)-(2.3) vérifie bien la propriété (2.1) ; pour $i = 0, ..., n$

$$P_n(x_i) = \sum_{j=0}^{n} f(x_j)\, L_j(x_i).$$

Or, nous remarquons que $L_j(x_i) = \delta_{i,j}$ où

$$\delta_{i,j} = \begin{cases} 1 & \text{si } i = j, \\ 0 & \text{si } i \neq j. \end{cases}$$

Ainsi,

$$P_n(x_i) \;=\; f(x_i) \times 1 \;+\; \sum_{j\neq i} f(x_j) \times 0 \;=\; f(x_i).$$

De plus, le polynôme P_n est le produit de n polynômes de degré un, c'est donc un polynôme de degré n.

Enfin, nous vérifions que ce polynôme est unique. Soit Q_n un autre polynôme solution de (2.1). Alors, nous avons pour $i = 0, ..., n$

$$P_n(x_i) - Q_n(x_i) \;=\; f(x_i) - f(x_i) \;=\; 0.$$

Ainsi, $P_n - Q_n$ est un polynôme de degré inférieur ou égal à n s'annulant en $n+1$ points. Il est donc identiquement nul. □

L'écriture (2.2)-(2.3) est certainement intéressante d'un point de vue théorique mais peu du point de vue pratique. D'une part elle a un caractère relativement peu algorithmique et d'autre part son évaluation requiert trop d'opérations élémentaires. Nous préférons donc la formule de Newton qui consiste à écrire la polynôme P_n aux points $x_0, ..., x_n$ sous la forme

$$P_n(x) = a_0 + a_1\,(x - x_0) + ... + a_n\,(x - x_0)\ldots(x - x_{n-1}).$$

D'abord nous observons que tout polynôme de degré inférieur ou égal à n peut s'écrire sous cette forme du moment que les points x_i sont tous distincts. Ensuite, cette formule présente un intérêt puisque elle donne une récurrence : la partie tronquée

$$a_0 + a_1\,(x - x_0) + ... + a_{n-1}\,(x - x_0)\,...(x - x_{n-2})$$

n'est pas autre chose que le polynôme P_{n-1} écrit aux points $x_0, ..., x_{n-1}$. En effet, P_{n-1} est un polynôme de degré inférieur ou égal à $n-1$ et tel que pour tout $i = 0, ..., n-1$

$$P_{n-1}(x_i) = f(x_i).$$

Donc, connaissant P_{n-1}, il suffit de calculer a_n pour connaître P_n. Les coefficients $(a_i)_{0\le n\le n}$ sont donnés par la formule de Newton

Théorème 2.2 *Le polynôme d'interpolation de Lagrange de la fonction f aux points distincts $(x_i)_{1\le i\le n}$ est donné par*

$$P_n(x) = \sum_{i=0}^{n} f[x_0, ..., x_i] \prod_{k=0}^{i-1} (x - x_k),$$

où $f[\,]$ désigne les différences divisées de f définies par

$$f[x_i] = f(x_i), \quad i = 0, ..., n$$

et

$$f[x_0, ..., x_k] = \frac{f[x_1, ..., x_k] - f[x_0, ..., x_{k-1}]}{x_k - x_0}.$$

Démonstration. Nous faisons la démonstration par récurrence sur le nombre de points x_i. Nous vérifions d'abord que la formule est vraie à l'ordre $n = 1$

$$P_1(x) = f(x_0) + (x - x_0)\,\frac{f(x_1) - f(x_0)}{x_1 - x_0} = f[x_0] + (x - x_0)\,\frac{f[x_1] - f[x_0]}{x_1 - x_0}.$$

Nous supposons ensuite que la formule est vraie à l'ordre $n - 1$; nous avons alors

$$P_{n-1}(x) = \sum_{i=0}^{n-1} f[x_0, .., x_i] \prod_{k=0}^{i-1}(x - x_k)$$

et pour tout $0 \leq k < l \leq n - 1$

$$f[x_k, .., x_l] = \frac{f[x_{k+1}, \ldots, x_l] - f[x_k, \ldots, x_{l-1}]}{x_l - x_k}.$$

Nous construisons alors un polynôme de Lagrange P_n construit à partir de P_{n-1} et tel que $P_n(x_n) = f(x_n)$, nous avons alors

$$\begin{aligned} P_n(x) &= P_{n-1}(x) + a_n \prod_{k=0}^{n}(x - x_k) \\ &= \sum_{i=0}^{n-1} f[x_0, .., x_i] \prod_{k=0}^{i-1}(x - x_k) \;+\; a_n \prod_{k=0}^{n-1}(x - x_k). \end{aligned}$$

Le coefficient a_n est donc tel que $P_n(x_n) = f(x_n)$ et il nous reste à calculer sa valeur. Considérons donc le polynôme $Q(x)$ écrit sous la forme

$$Q(x) = \frac{x - x_0}{x_n - x_0}\, Q_{n-1}(x) \;+\; \frac{x_n - x}{x_n - x_0} P_{n-1}(x),$$

où

$$Q_{n-1}(x) = \sum_{i=0}^{n} f[x_1, ..., x_i] \prod_{k=1}^{i-1}(x - x_k)$$

est le polynôme d'interpolation de f aux points $x_1, ..., x_n$. Nous vérifions immédiatement que pour tout $i = 0, ..., n$

$$Q(x_i) = f(x_i).$$

Ainsi, puisque Q est un polynôme de degré inférieur ou égal à n, nous avons $P_n \equiv Q$. Or, cette nouvelle expression de P_n permet de déterminer le coefficient de x^n de P_n

$$a_n = \frac{1}{x_n - x_0} f[x_1, ..., x_n] - \frac{1}{x_n - x_0} f[x_0, ..., x_{n-1}] = f[x_0, ..., x_n].$$

□

À partir de cette formulation, nous pouvons établir le théorème suivant

Théorème 2.3 *Soit f une fonction définie de l'intervalle $[a,b]$ à valeur dans $\mathbb{R}$. Nous supposons que f est $n+1$ fois continûment différentiable et P_n est le polynôme de Lagrange défini aux points distincts $(x_i)_{0\leq i\leq n}$ de $[a,b]$. Alors*

$$|P_n(x) - f(x)| \leq \frac{M_{n+1}}{(n+1)!}|\pi_n(x)|, \tag{2.4}$$

où $M_{n+1} = \max_{a\leq x\leq b} |f^{(n+1)}(x)|$ et

$$\pi_n(x) = \prod_{i=0}^{n}(x - x_i).$$

Avant de procéder à la preuve de ce théorème nous prouvons le lemme suivant

Lemme 2.1 *Soit f une fonction définie de l'intervalle $[a,b]$ à valeur dans $\mathbb{R}$. Nous supposons que f est p fois continûment différentiable. Alors, il existe $\xi \in [a,b]$ tel que*

$$f[x_0, ..., x_p] = \frac{f^{(p)}(\xi)}{p!}.$$

Démonstration. Nous introduisons la fonction $E_p(x) = f(x) - P_p(x)$, qui s'annule en $p+1$ points distincts. D'après le théorème de Rolle, puisque E_p est continue et dérivable, sa dérivée E_p' s'annule en au moins p points,..., $E_p^{(p)}$ est continue et s'annule en un point $\xi \in [a,b]$, c'est-à-dire

$$f^{(p)}(\xi) = P_p^{(p)}(\xi).$$

Comme P_p est de degré p, nous savons que $P_p^{(p)}(\xi) = a_p\, p!$, où a_p est le coefficient de x^p, soit ici $a_p = f[x_0, ..., x_p]$. Par transitivité, nous obtenons le résultat

$$f[x_0, ..., x_p] = a_p = \frac{f^{(p)}(\xi)}{p!}.$$

□

Nous sommes maintenant en mesure de prouver le Théorème 2.3

Démonstration. Soit $x \in [a,b]$, nous introduisons Q le polynôme d'interpolation aux points x_0,...,x_n et x. D'après la formule de Newton, ce polynôme est donné au point x par

$$Q(x) = P_n(x) + f[x_0, ..., x_n, x]\pi_n(x).$$

Or, puisque $Q(x) = f(x)$ nous avons

$$f(x) = P_n(x) + f[x_0, ..., x_n, x]\pi_n(x).$$

Finalement, en appliquant le Lemme 2.1, il existe $\xi \in [a,b]$ tel que

$$f(x) = P_n(x) + \frac{f^{(n+1)}(\xi)}{(n+1)!}\,\pi_n(x).$$

D'où le résultat. □

2.2 Phénomène de Runge

Ce résultat ne signifie pas que l'interpolé de Lagrange va converger vers la fonction f du moment que la fonction f est régulière (disons de classe $\mathcal{C}^\infty([a,b],\mathbb{R})$). En effet, l'erreur dépend de la taille de l'intervalle $[a,b]$ et de la répartition des points $(x_i)_{0\leq i\leq n}$ sur l'intervalle $[a,b]$. Par exemple, si nous prenons des points x_i équidistants sur l'intervalle $[-5,5]$ et la fonction f donnée par

$$f(x)=\frac{1}{25\,x^2+1},\quad x\in[-5,5].$$

Nous traçons ci-dessous, les polynômes de Lagrange correspondants de degré $n=5$, 7 et 9 ; nous constatons graphiquement (et nous pourrions même le démontrer mais c'est compliqué) que lorsque n tend vers l'infini, le polynôme de Lagrange défini à partir d'un ensemble de points équidistants ne va pas converger vers la fonction f. Pour remédier à ce problème, nous

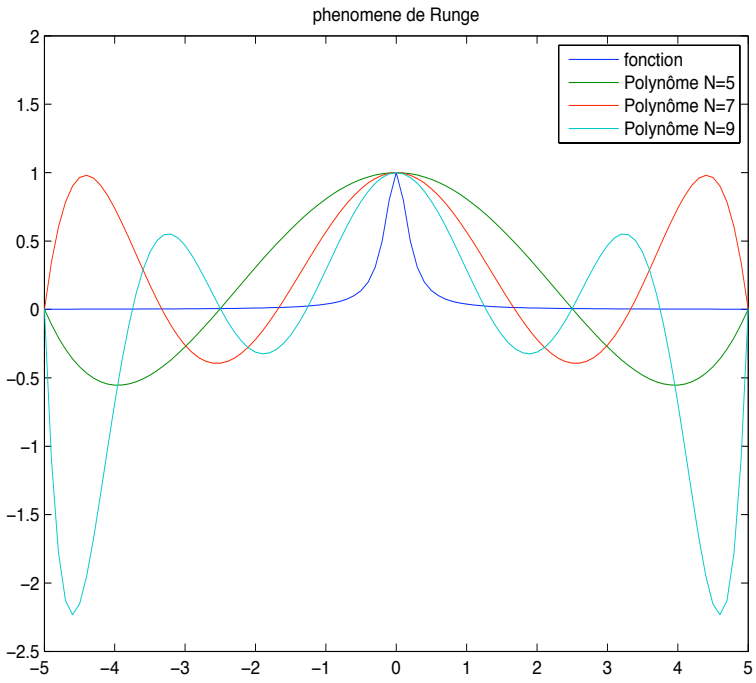

FIG. 5.1 – La fonction f et les polynômes de Lagrange correspondants

proposons une autre stratégie qui est le plus souvent utilisée en pratique.

2.3 Interpolation composée

Pour éviter les instabilité numériques du type "phénomène de Runge" , nous n'utilisons l'interpolation polynomiale de degré peu élevé et sur des intervalles de petites tailles. En pratique,

nous décomposons une première fois l'intervalle $[a, b]$ en $n + 1$ points

$$a = a_0 < a_1 < \ldots < a_{n-1} < a_n = b.$$

Puis sur chaque intervalle $[a_i, a_{i+1}]$, nous construisons un polynôme d'interpolation de Lagrange $P_{i,m}$ de la fonction f à l'aide de $m + 1$ points, notés $(x_j^i)_{0 \leq j \leq m}$ de $[a_i, a_{i+1}]$

$$a_i = x_0^i < x_1^i < \ldots < x_{m-1}^i < x_m^i = a_{i+1}$$

Pour assurer que la reconstruction sur tout l'intervalle $[a, b]$ est continue, nous prenons toujours les extrémités des intervalles.

Nous démontrons alors très facilement en utilisant les résultats précédents le théorème suivant

Théorème 2.4 *Soient n et m deux entiers positifs et un ensemble de points $(a_i)_{0 \leq i \leq n}$ de l'intervalle $[a, b]$, nous choisissons sur chaque intervalle $[a_i, a_{i+1}]$ un ensemble de points $(x_j^i)_{0 \leq j \leq m}$ tels que*

$$a_i = x_0^i < x_1^i < \ldots < x_{m-1}^i < x_m^i = a_{i+1}.$$

Alors il existe une seule fonction continue $f_{m,n}$ telle que

$$\begin{cases} f_{m,n|[a_i,a_{i+1}]} \text{ est un polynôme de degré } m \\ \text{Sur l'intervalle } [a_i, a_{i+1}], \quad f_{m,n}(x_j^i) = f(x_j^i), \quad 0 \leq j \leq m \text{ et } 0 \leq i \leq n. \end{cases}$$

De plus, si la fonction $f \in \mathcal{C}^{m+1}([a, b], \mathbb{R})$, nous avons l'estimation d'erreur

$$\|f - f_{m,n}\|_\infty \leq \frac{h^{m+1}}{(m+1)!} \|f^{(m+1)}\|_\infty,$$

avec $h = \max_{0 \leq i \leq n} |a_{i+1} - a_i|$.

3 Polynômes d'Hermite

Soient $x_0, \ldots, x_n$, $(n + 1)$ points distincts de l'intervalle $[a, b]$ et $f \in \mathcal{C}^1([a, b], \mathbb{R})$ une fonction donnée. Nous supposons que les valeurs de la fonction f et de sa dérivées sont connues aux points $(x_i)_{0 \leq i \leq n}$; nous cherchons alors un polynôme H_n de degré minimal tel que

$$H_n(x_i) = f(x_i), \quad H_n'(x_i) = f'(x_i), \quad i = 0, \ldots, n. \tag{3.5}$$

Nous rappelons que

$$l_i(x) = \prod_{\substack{j=0 \\ j \neq i}} \frac{x - x_j}{x_i - x_j}.$$

Nous démontrons le résultat suivant

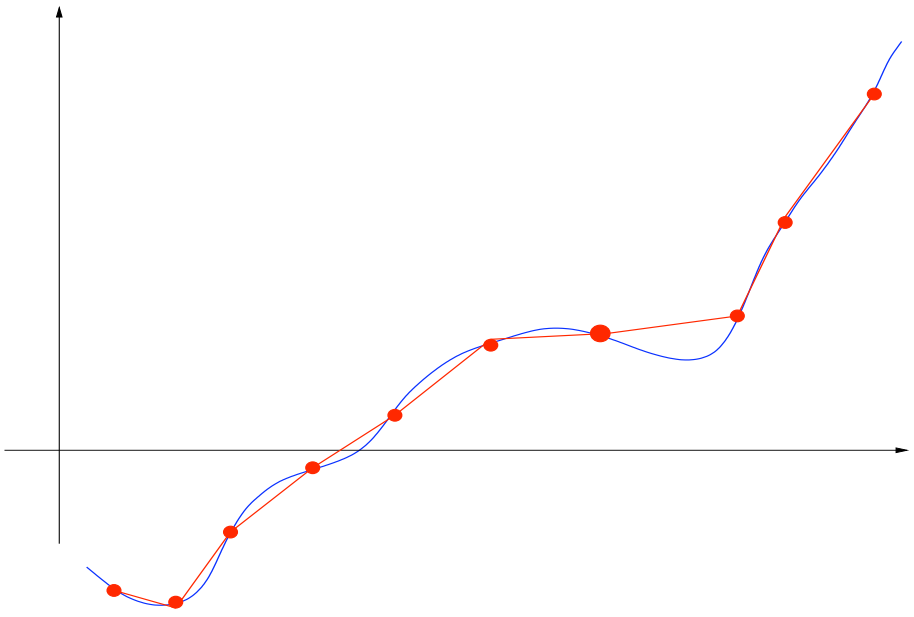

FIG. 5.2 – Illustration de l'interpolation composée pour l'approximation d'une fonction f.

Théorème 3.1 *Le polynôme H_n s'écrit*

$$H_n(x) = \sum_{i=0}^{n} f(x_i)\, h_i(x) \;+\; \sum_{i=0}^{n} f'(x_i)\, \tilde{h}_i(x)$$

avec

$$h_i(x) = (1 - 2\, L_i'(x_i)(x - x_i))L_i^2(x); \quad \tilde{h}_i(x) = (x - x_i)\, L_i^2(x).$$

De plus, si $f \in \mathcal{C}^{2(n+1)}([a,b], I\!R)$

$$|f(x) - H_n(x)| \;\leq\; \frac{\|f^{(2(n+1))}\|_\infty}{(2n+2)!} \prod_{i=0}^{n} (x - x_i)^2. \tag{3.6}$$

Démonstration. Nous procédons en plusieurs étapes et vérifions d'abord que

$$h_i(x_j) = \delta_{i,j}, \quad h_i'(x_j) = 0$$

et

$$\tilde{h}_i(x_j) = 0, \quad \tilde{h}_i'(x_j) = \delta_{i,j}.$$

Ainsi en prenant H_n donné par

$$H_n(x) = \sum_{i=0}^{n} f(x_i)\, h_i(x) \;+\; \sum_{i=0}^{n} f'(x_i)\, \tilde{h}_i(x)$$

nous prouvons qu'il existe bien un polynôme de degré $2n+1$ vérifiant les conditions requises (3.5). Par l'absurde, nous vérifions que ce polynôme est forcément unique.

Pour démontrer l'estimation de l'erreur, nous la vérifions d'abord en chaque point x_i, pour $i = 0, \ldots, n$. Puis, nous pouvons alors prendre x dans l'intervalle $[a, b]$ tel que x est différent des points $(x_i)_{\leq i \leq n}^{\circ}$. Nous posons

$$\pi_n^2(x) = \prod_{i=0}^{n} (x - x_i)^2$$

et

$$F(y) = f(y) - H_n(y) - \frac{f(x) - H_n(x)}{\pi_n^2(x)}\, \pi_n^2(y)$$

et montrons que x est une racine de F, et x_i est racine double de F, c'est-à-dire la fonction F admet au moins $2n+3$ racines. En appliquant le théorème de Rolle d'abord sur F, puis sur F' de manière succéssive, nous prouvons de manière analogue à ce que nous avons effectué pour l'interplation de Lagrange qu'il existe η_x tel que

$$F^{(2n+2)}(\eta_x) = 0.$$

En dérivant $2n+2$ fois la fonction $F(y)$ et en nous plaçant au point η_x, nous obtenons directement

$$(2n+2)!\, \frac{f(x) - H_n(x)}{\pi_n^2(x)} \;=\; f^{(2(n+1))}(\eta_x),$$

ce qui conduit facilement au résultat final. □

4 Méthode des moindres carrés discrète

Jusqu'ici nous avons toujours considéré l'approximation d'une fonction par un procédé d'interpolation à l'aide des valeurs $f(x)$ de f en des points $(x_i)_{0 \leq i \leq n}$. Ceci suppose que ces valeurs soient connues exactement. Or, il y a beaucoup de situations où ce n'est pas le cas en particulier lorsque les valeurs $f(x)$ proviennent de mesures physiques. Le résultat de mesures physiques tel que celui présenté ci-dessus conduit à penser que $f(x)$ doit être une fonction affine $f(x) = a_1\, x + a_0$. Il ne semble pas très raisonnable de remplacer $f(x)$ par un polynôme d'interpolation en les points x_i dont le calcul dépendrait manifestement des valeurs erronées $f(x_i)$. Une analyse succincte du phénomène ci-dessus conduit à penser que ces valeurs $f(x_i)$ contiennent une information juste (variant lentement) mais aussi un certain "bruit" (variant rapidement mais de faible amplitude). L'ajustement de données consiste à éliminer le "bruit". Dans cet exemple, le principe de la méthode va consister à chercher la fonction $f(x)$ sous la forme $f(x) = a_1\, x + a_0$ où a_0 et a_1 sont calculées de manière à rendre le bruit le plus petit possible. Il reste à définir correctement la notion de plus petit possible, nous pourrions par exemple

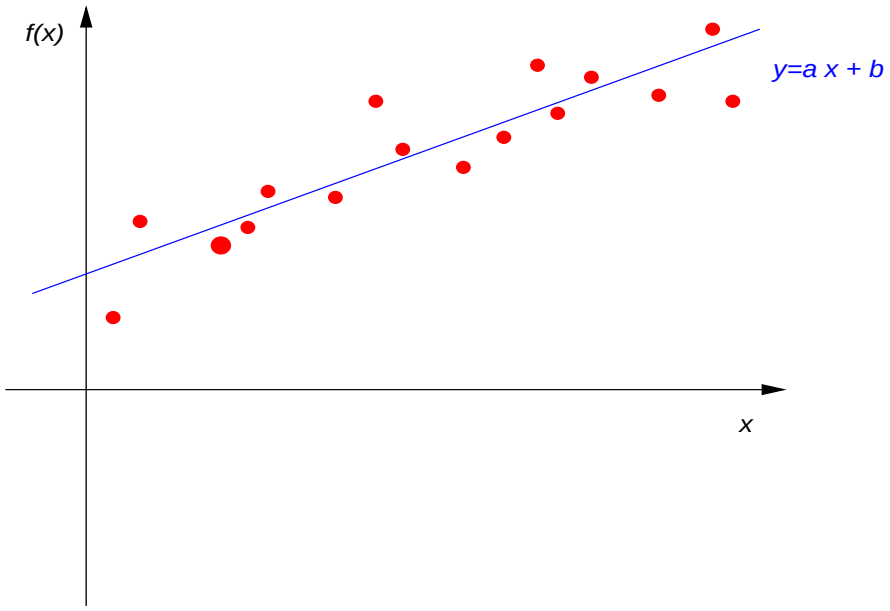

FIG. 5.3 – Approximation linéaire (en bleu) au sens des moindres carrés à partir de données physiques (en rouge)

chercher à minimiser la quantité (approximation uniforme) a_0 et a_1 sont tels que

$$\max_{0\leq i\leq n} |f(x_i) - (a_0 + a_1\, x_i)|$$

soit minimal. Cependant, le plus souvent nous choisissons dans la pratique à minimiser la quantité (méthode des moindres carrés)

$$\sum_{i=0}^{n} |f(x_i) - (a_0 + a_1\, x_i)|^2 \,,$$

car ce problème peut être résolu plus facilement.

4.1 Rappel du théorème de projection

Avant de passer à la description de la méthode des moindes carrés, nous donnons le théorème de la projection qui sera très utile pour la suite.

Théorème 4.1 (Théorème de la projection en dimension finie) *Soient E espace vectoriel complet muni d'un produit scalaire, F un sous-espace vectoriel de dimension finie de E et $f \in E$. Alors, il existe un unique $p^* \in F$ tel que*

$$\|p^* - q\| = \min_{q\in F} \|q - f\|.$$

De plus, p^* est la projection orthogonale de f sur l'espace F : $p^* = Proj_F\, f$ qui est telle que

$$(p^* - f, q) = 0, \quad q \in F.$$

Nous disons que p^* est la projection de $f \in E$ sur le sous-espace .

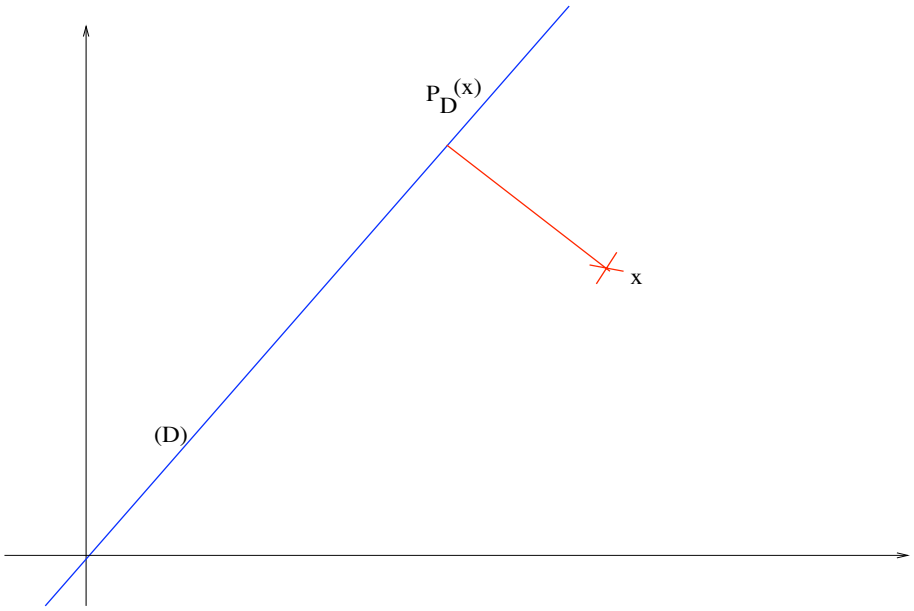

FIG. 5.4 – Illustration du Théorème de la projection dans $\mathbb{R}^2$: nous projetons un point $x \in \mathbb{R}^2$ sur une droite $\mathcal{D}$ du plan ; c'est la projection orthgonale de x sur $\mathcal{D}$.

Démonstration. Nous notons d la distance induite par le produit scalire de E : pour tout f et g de E

$$d(f, g) = \|f - g\| = (f - g, f - g)^{1/2}.$$

Soient $f \in E$ et F un sous-espace vectoriel de E, nous définissons alors

$$d(f, F) = \inf_{q \in F} \|f - q\|.$$

Par définition de l'infimum, il existe une suite $(q_k)_{k \in \mathbb{N}} \subset F$ telle que

$$\lim_{k \to \infty} \|f - q_k\| = d(f, F).$$

Nous voulons démontrer que la suite $(q_k)_{k \in \mathbb{N}}$ est une suite de Cauchy. D'une part par définition de la limite, pour tout $\varepsilon > 0$, il existe $n_\varepsilon \in \mathbb{N}$ tel que pour tout $k \geq n_\varepsilon$

$$\|f - q_k\| \leq \varepsilon/2$$

et donc pour tout k et $l \in I\!N$, tels que $k > n_\varepsilon$, $l > n_\varepsilon$

$$\|q_k - q_l\| \leq \|f - q_k\| + \|f - q_l\| \leq \varepsilon.$$

Ainsi, $(q_k)_{k\in I\!N}$ est une suite de Cauchy, et puisque P est un sous-espace vectoriel de E de dimension finie, P est fermé et de plus $F \subset E$ est complet, il existe dont $q^* \in F$ tel que $q_k \to v^*$ et donc

$$\|f - p^*\| = d(f, F) = \inf_{q\in F} \|f - q\|.$$

De plus, sa limite est unique. En effet, soient p^* et q^* deux éléments réalisant le minimum ; alors

$$\begin{aligned} 0 \leq \|p^* - q^*\|^2 &= 2\|p^* - f\|^2 + 2\|f - q^*\|^2 - 4\|f - \frac{p^* + q^*}{2}\|^2, \\ &= 4\,[d(f, F)]^2 - 4\|f - \frac{p^* + q^*}{2}\|^2. \end{aligned}$$

Or, puisque F est un espace vectoriel, $(p^* + q^*)/2 \in F$ et

$$d(f, F) \leq \|f - \frac{p^* + q^*}{2}\|,$$

donc

$$0 \leq \|p^* - q^*\|^2 \leq 4\,[d(f, F)]^2 - 4\,[d(f, F)]^2 = 0.$$

Ainsi, la limite est unique.

Nous posons alors $p^* = Proj_F(f)$, soient $q \in F$ et $\lambda \in I\!R$ alors $p^* + \lambda\, q \in F$

$$\|f - (p^* + \lambda\, q\|^2 = \|f - p^*\|^2 - 2\,\lambda\,(f - p^*, q) + \lambda^2\|q\|^2,$$

il vient donc pour tout $\lambda \in I\!R$

$$-2\,\lambda\,(f - p^*, q) + \lambda^2\|q\|^2 \leq 0.$$

En prenant d'une part $\lambda > 0$, nous avons aussi

$$-2\,(f - p^*, q) + \lambda\|q\|^2 \leq 0$$

et ensuite en faisant tendre λ vers zéro, nous obtenons le résulat : pour tout $q \in F$

$$(f - p^*, q) \leq 0.$$

Finalement, puisque F est un espace vectoriel $-q$ appartient ausi à F et donc : pour tout $q \in F$

$$(f - p^*, q) = 0.$$

□

4.2 Résolution du problème des moindres carrés discrets

Nous voulons démontrer le résultat suivant

Théorème 4.2 *Considérons $n+1$ points (x_i, y_i) pour $0 \leq i \leq n$ et $N < n+1$ fonctions ϕ_1,..., ϕ_N. Nous notons $\mathcal{U}$ l'espace vectoriel engendré par les fonctions ϕ_1,...,ϕ_N*

$$\mathcal{U} = vect\left\{\phi_1, \ldots, \phi_N\right\}.$$

Alors si la famille $(\phi)_{1\leq i\leq N}$ est libre, il existe une unique solution $\phi^* \in \mathcal{U}$ tel que

$$\sum_{i=0}^{n} |y_i - \phi^*(x_i)|^2 = \min_{\phi\in\mathcal{U}} \sum_{i=0}^{n} |\phi(x_i) - y_i|^2.$$

Ainsi, nous souhaitons démontrer que
- ce problème de minimisation admet une solution ;
- cette solution est unique ;
- nous calculons cette solution en résolvant un système linéaire.

La première étape consiste à mettre le problème sous forme matricielle. Nous posons $y \in \mathbb{R}^{n+1}$ avec $y = (y_0, \ldots, y_n)^T \in \mathbb{R}^{n+1}$. Ensuite, nous remarquons que l'application

$$x \in \mathbb{R}^{n+1} \;\rightarrow\; \|x\| = \left(\sum_{i=0}^{n} |x_i|^2\right)^{1/2} \;\in\; \mathbb{R}^+$$

définit la norme euclidienne de $\mathbb{R}^{n+1}$. Alors, en écrivant pour $\phi \in \mathcal{U}$, nous obtenons

$$\phi(x_i) \;=\; \sum_{j=1}^{N} u_j\, \phi_j(x_i) \;\equiv\; B\, u,$$

avec

$$B = \begin{pmatrix} \phi_1(x_0) & \ldots & \phi_N(x_0) \\ \vdots & & \vdots \\ \phi_1(x_n) & \ldots & \phi_N(x_n) \end{pmatrix}$$

et $u = (u_1, \ldots, u_N)^T \in \mathbb{R}^N$. Par conséquent, nous avons

$$\sum_{i=0}^{n} |\phi(x_i) - y_i|^2 \;=\; \|B\, u \;-\; y\|^2.$$

En notant $v \;=\; B\, u$ qui appartient à $\mathbb{R}^{n+1}$, le problème revient alors à trouver $v^* \in F := \{v \in \mathbb{R}^n; \quad v = B\, u\} = Im(B)$ réalisant le minimum de $\|v - y\|$ pour tout $v \in F$, c'est-à-dire $v^* \in F$ tel que

$$\|v^* - y\| = \min_{v\in F} \|v - y\|.$$

Démonstration. Tout d'abord en appliquant le Théorème 4.1, nous démontrons qu'il existe un unique $v^* \in F$ réalisant le minimum de $\|y - v\|$. Cette solution est caractérisée par $v^* \in F$

$$(y - v^*, v) = 0, \quad \forall\, v \in F.$$

Comme la solution v^* appartient à F, il existe $u^* \in I\!R^N$ tel que $v^* = B\,u^*$; nous avons donc

$$(B\,u^* - y, B\,u) = 0, \quad \forall\, u \in I\!R^N$$

ou de manière équivalente

$$(B^T\,B\,u^* - B^T\,y, u) = 0, \quad \forall\, u \in I\!R^N$$

et donc

$$B^T\,B\,u^* = B^T\,y.$$

Nous avons donc existence d'une solution u^*. Pour autant, ceci n'assure pas l'unicité. En effet, nous savons que v^* est unique d'après le théorème de la projection mais il peut exister plusieurs u^* tel que $B\,u^* = v^*$. L'unicité est liée au caractère injectif de B. En effet, supposons que B est de rang N, c'est-à-dire

$$dim(Im(B)) = dim F = N.$$

Or, nous savons que

$$dim(Im(B)) + dim(Ker(B)) = dim I\!R^N = N$$

Donc $Ker(B)$ est de dimension 0 et donc B est injective. Nous en déduisons que $B^T\,B$ est définie positive et donc est inversible. Il existe donc un unique $u^* \in I\!R^N$. □

5 Vers la méthode des moindres carrés continue

Avant de présenter la méthode des moindres carrés en détail, nous introduisons la notion de polynôme orthogonaux qui sert de base théorique à la méthode.

5.1 Quelques rappels théoriques

Soit $\omega(x)$ une fonction positives sur $[a, b]$; nous notons par

$$L^2_\omega(a, b) = \left\{ f \text{ mesurable } : \int_a^b |f(x)|^2\,\omega(x)\,dx < \infty \right\}.$$

Dans la suite nous utiliserons les notations suivantes

– Soit $\omega(x)$ une fonction positive sur $[a,b]$, nous notons par $(.,.)_\omega$ le produit scalaire pondéré par le poids ω : pour toute fonction $f, g \in L^2_\omega(a,b)$

$$(f,g)_\omega = \int_a^b f(x)\, g(x)\, \omega(x)\, dx.$$

– L'espace $L^2_\omega(a,b)$ est l'espace de Hilbert des fonctions de carré sommable sur $[a,b]$ par rapport au poids ω. C'est un espace vectoriel normé (donc chaque élément est une fonction) complet, muni d'un produit scalaire.
– La norme associée au produit scalaire est donnée pour toute fonction $f \in L^2(a,b)$ par l'application $f \in L^2_\omega(a,b) :\to \|f\|_\omega \in I\!R^+$ telle que

$$\|f\|^2_{L^2_\omega} = (f,f)_\omega = \int_a^b |f(x)|^2\, \omega(x)\, dx.$$

– Nous introduisons également la norme de la convergence uniforme sur $[a,b]$ définie par

$$\|f\|_{L^\infty} = \sup_{x\in[a,b]} |f(x)|,$$

celle-ci est correctement définie si f est par exemple continue.
– Un polynôme dont le coefficient du terme de plus haut degré est égal à 1 est dit *unitaire*.

Définition 5.1 *Nous disons que la famille $(e_n)_{n\in I}$ est orthogonale par rapport au produit scalaire $(.\,,.)_\omega$ lorsque*

$$(e_n\,,\,e_p)_\omega = 0, \quad p \neq n, \qquad (e_n\,,\,e_n)_\omega > 0.$$

Nous disons que la famille $(e_n)_{n\in I}$ est orthonormale par rapport au produit scalaire $(.\,,.)_\omega$ lorsque

$$(e_n\,,\,e_n)_\omega = 0, \quad p \neq n; \quad \text{et} \quad (e_n\,,\,e_n)_\omega = 1, \quad n \in I.$$

Toute famille orthogonale constituée de vecteurs non nuls est libre, il suffit d'effectuer le produit scalaire pour le vérifier. Nous avons alors les propriétés suivantes

Proposition 5.1 *Soit $(e_n)_{n\in I}$ une famille orthogonale de l'espace vectoriel H muni d'un produit scalaire (appelé espace préhilbertien). Alors pour tout élément $x \in H$, nous avons l'inégalité de Bessel*

$$\sum_{i\in I} |(x\,,\,e_n)|^2 \;\leq\; \|x\|^2_{L^2_\omega}.$$

De plus, si $(e_n)_{n\in I}$ forme une base de l'espace H, nous avons l'égalité

$$\sum_{i\in I} |(x\,,\,e_n)|^2 \;=\; \|x\|^2_{L^2_\omega}.$$

connue sous le nom de formule de Parseval ; de plus la série

$$\sum_{i\in I} (x\,,\,e_n)\, e_n$$

converge.

Démonstration. voir [8][Chapitre II]. □

5.2 Polynômes orthogonaux

Concentrons nous maintenant sur le cas particulier de fonctions polynômiales, nous avons d'abord la définition suivante.

Définition 5.2 *Nous appelons polynômes orthogonaux par rapport au produit scalaire $(.,.)_\omega$et de degré n, la suite de polynômes $P_n(x)$ tels que :*
- *le degré de P_n est exactement n et le coefficient de x^n est positif*
- *les P_n sont orthonormés dans $L^2_\omega(a,b)$, c'est-à-dire*

$$\int_a^b P_n(x)\,P_m(x)\,\omega(x)\,dx = C_m\,\delta_{m,n} = \begin{cases} 1, & \textit{si } m = n, \\ 0 & \textit{sinon.} \end{cases}$$

où C est une constante strictement positive.

Dans la suite, nous supposerons que

$$\int_a^b |x|^n\,\omega(x)\,dx < \infty, \qquad \forall\, n \geq 0$$

ce qui assure que tous les polynômes sont contenus dans l'espace $L^2_\omega(a,b)$.

Théorème 5.1 *Soit ω une fonction positive sur l'intervalle $[a,b]$. Alors, il existe une unique suite de polynômes, orthogonaux par rapport au produit scalaire $(.,.)_\omega$ et de degré n et dont le coefficient devant le terme x^n vaut un. Plus précisément, nous avons la relation de récurrence suivante*

$$\begin{cases} P_0(x) = 1, \\ P_1(x) = x - \alpha_0\,P_0(x), \\ P_{n+1}(x) = (x-\alpha_n)\,P_n(x) - \lambda_n\,P_{n-1}(x), \quad n \geq 1, \end{cases} \tag{5.7}$$

avec

$$\alpha_n = \frac{1}{\|P_n\|^2_{L^2_\omega}} \int_a^b x\,P_n(x)\,P_n\,\omega(x)\,dx, \quad \lambda_n = \frac{\|P_n\|^2_{L^2_\omega}}{\|P_{n-1}\|^2_{L^2_\omega}}. \tag{5.8}$$

Démonstration. Nous allons construire les polynômes orthogonaux P_n par le procédé de Graham-Schmidt. Nous posons d'abord $P_0(x) = 1$ et $P_1(x) = x - \alpha_0 P_0(x)$. La condition d'orthogonalité entre P_0 et P_1 donne

$$0 = \int_a^b P_1(x)\, P_0(x)\, \omega(x)\, dx \;=\; \int_a^b x\, \omega(x)\, dx - \alpha_0 \int_a^b \omega(x)\, dx,$$

nous en déduisons la valeur de α_0,

$$\alpha_0 := \frac{\int_a^b x\, \omega(x)\, dx}{\int_a^b \omega(x)\, dx} = \frac{1}{\|P_0\|^2_{L^2_\omega}} \int_a^b x\, P_0(x)\, P_0(x)\, \omega(x)\, dx.$$

Nous calculons pour la suite

$$\int_a^b P_1^2(x)\, \omega(x)\, dx \;=\; \int_a^b P_1(x)\, (x - \alpha_0\, P_0(x))\, \omega(x)\, dx$$

ce qui donne en utilisant la condition d'orthogonlité et puisque $P_0(x) = 1$

$$\int_a^b P_1^2(x)\, \omega(x)\, dx \;=\; \int_a^b x\, P_1(x)\, P_0(x)\, \omega(x)\, dx. \tag{5.9}$$

Nous posons ensuite

$$P_2(x) = (x - \alpha_1)\, P_1(x) - \lambda_1\, P_0(x).$$

Des conditions $(P_2, P_1)_\omega = (P_2, P_0)_\omega = 0$, nous déduisons que

$$\alpha_1 = \frac{1}{\|P_1\|^2_{L^2_\omega}} \int_a^b x\, P_1(x)\, P_1(x)\, \omega(x)\, dx,$$

et en utilisant (5.9), nous obtenons

$$\lambda_1 = \frac{1}{\|P_0\|^2_{L^2_\omega}} \int_a^b x\, P_1(x)\, P_0(x)\, \omega(x)\, dx.$$

Ce qui démontre le résultat pour $n = 1$.

Nous procédons ensuite par récurrence en supposons que (5.8) est vrai au rang n, c'est-à-dire que nous avons construit une famille de polynômes orthogonaux $\{P_0, \ldots, P_{n+1}\}$.

Nous posons alors

$$P_{n+2}(x) = (x - \alpha_{n+1})\, P_{n+1}(x) - \lambda_{n+1}\, P_n(x).$$

D'une part, nous voulons démontrer que P_{n+1} est orthogonal à tout polynôme P_i, construit précédemment, de degré $i \leq n - 1$. Pour cela, pour $i \in \{0, ..., n - 1\}$, nous écrivons

$$x\, P_i(x) = \sum_{k=0}^{i+1} c_k\, P_k(x).$$

Il vient alors par hypothèse de récurrence $(P_{n+1}, P_k)_\omega = 0$ pour tout $0 \leq k \leq n$, $(P_{n+1}, P_i)_\omega = (P_n, P_i)_\omega = 0$, pour tout $0 \leq i \leq n-1$ et donc

$$(P_{n+2}, P_i)_\omega = \sum_{k=0}^{i+1} c_k \, (P_{n+1}, P_k)_\omega - \alpha_{n+1} \, (P_{n+1}, P_i)_\omega - \lambda_{n+1} \, (P_n, P_i)_\omega = 0.$$

D'autre part, en imposant les conditions $(P_{n+2}, P_{n+1})_\omega = (P_{n+2}, P_n)_\omega = 0$, nous en déduisons que

$$\alpha_{n+1} = \frac{1}{\|P_{n+1}\|^2_{L^2_\omega}} \int_a^b x \, P_{n+1}(x) \, P_{n+1}(x) \, \omega(x) \, dx$$

et

$$\lambda_{n+1} = \frac{1}{\|P_n\|^2_{L^2_\omega}} \int_a^b x \, P_n(x) \, P_{n+1}(x) \, \omega(x) \, dx.$$

La valeur de λ_{n+1} n'est pas exactement celle recherchée. Montrons alors que

$$\int_a^b x \, P_n(x) \, P_{n+1}(x) \, \omega(x) \, dx = \|P_{n+1}\|^2_{L^2_\omega}.$$

En effet, puisque les polynômes $P_0, \ldots, P_{n+1}$ sont orthogonaux deux à deux ; ils forment une base de l'espace vectoriel formé par les polynômes de degré $n+1$. En développant $x \, P_n(x)$ dans la base $\{P_0, \ldots, P_{n+1}\}$ et en utilisant le fait que le coefficient devant de terme de degré $n+1$ du polynôme $x \, P_n$ vaut un, nous avons

$$x \, P_n(x) = P_{n+1}(x) + \sum_{i=0}^{n} c_i P_i(x)$$

et déduisons que

$$\int_a^b x \, P_n(x) \, P_{n+1}(x) \, \omega(x) \, dx = (P_{n+1}, P_{n+1})_\omega + \sum_{i=0}^{n} c_i (P_i, P_{n+1})_\omega = \|P_{n+1}\|^2_{L^2_\omega},$$

ce qui donne

$$\alpha_{n+1} = \frac{1}{\|P_{n+1}\|^2_{L^2_\omega}} \int_a^b x \, P_{n+1}(x) \, P_{n+1}(x) \, \omega(x) \, dx$$

et

$$\lambda_{n+1} = \frac{\|P_{n+1}\|^2_{L^2_\omega}}{\|P_n\|^2_{L^2_\omega}}.$$

□

Nous énonçons ici deux résultats classiques sur les racines des polynômes orthogonaux.

Proposition 5.2 *Pour tout $n \geq 1$, nous avons*

(i) le polynôme P_n possède exactement n racines simples contenues dans l'intervalle (a, b),

(ii) les racines de P_n séparent les racines de P_{n+1}, c'est-à-dire

$$a < x_1^{n+1} < x_1^n < x_2^{n+2} < ... < x_n^{n+1} < x_n^n < x_{n+1}^{n+1} < b,$$

où x_1^n,..,x_n^n sont les n racines de P_n.

Pour y voir plus clair, donnons un exemple de polynômes orthogonaux :

Exemple 5.1 *Les polynômes d'Hermite $H_n(x)$ définis par*

$$H_n(x) = (-1)^n e^{x^2/2} \frac{d^n}{dx^n} \left(e^{-x^2/2} \right).$$

sont orthogonaux pour le produit scalaire associé au poids ω

$$\omega(x) = e^{-x^2/2}, \quad x \in I\!R$$

Nous pouvons démontrer que les polynômes d'Hermite sont donnés par la récurrence suivante pour $x \in I\!R$

$$H_0(x) = 1, \quad H_1(x) = x$$

et

$$H_{n+1}(x) = x\, H_n(x) - nH_{n-1}(x).$$

Exemple 5.2 *Les polynômes de Legendre $P_n(x)$ sont définis pour $x \in [-1, 1]$ par la récurrence suivante pour*

$$P_0(x) = 1, \quad P_1(x) = x$$

et

$$P_{n+1}(x) = \frac{(2n+1)\, x\, P_n(x) - n\, P_{n-1}(x)}{n+1}.$$

Cette suite de polynômes forme une famille orthogonale pour le produit scalaire associé au poids ω

$$\omega(x) = 1, \quad x \in [-1, 1].$$

Un autre exemple important est l'ensemble des polynômes de Chebychev (voir Complément du chapitre).

5.3 Méthode des moindres carrés

Dans la partie précédente, nous avons présenté la méthode des moindres carrés discrète qui consiste à trouver par exemple un polynôme p^*, qui approche une fonction connue seulement en certains points $(x_i, y_i)_{0\leq i\leq n}$; nous travaillons alors dans l'espace $I\!R^{n+1}$ et le produit scalaire associé. La méthode des moindres carrés continue consiste à approcher une fonction f, connue en tout point, dans l'espace formé par des polynômes et selon le produit scalaire induit par cet espace.

Tout d'abord, commençons par démontrer un théorème de projection qui permettra de généraliser la méthode des moindres carrés à l'espace formé par des polynômes quelconques.

Théorème 5.2 (Théorème de la projection en dimension infinie) *Soit $(E, \|.\|_E)$ un espace vectoriel normé fonctionnel, c'est-à-dire chaque élément de E est une fonction et $\mathcal{P}$ un sous-espace vectoriel, de dimension finie de E, composé de polynômes.*

Alors, pour toute fonction $f \in E$, il existe (au moins) un polynôme $p \in \mathcal{P}$ meilleure approximation de f dans E, c'est-à-dire

$$\|f - p^*\|_E = \min_{p\in\mathcal{P}} \|f - p\|_E.$$

Démonstration. Nous décomposons la démonstration en deux étapes : dans la première, nous ramenons le problème en nous situant sur un compact. Dans la deuxième partie, nous utilisons un théorème de compacité pour d'émontrer l'existence d'une solution.

Tout d'abord, nous voulons reformuler le problème en nous situant sur un ensemble compact. D'une part, nous avons en réduisant le domaine de minimisation

$$\inf_{p\in\mathcal{P}} \|f\ -\ p\|_E\ \leq \inf_{\substack{p\in\mathcal{P} \\ \|p\|_E\leq 2\,\|f\|_E}} \|f\ -\ p\|_E.$$

D'autre part, pour montrer qu'il y a en fait égalité, nous raisonnons par l'absurde. En effet, supposons que l'inégalité est stricte, c'est-à-dire

$$\inf_{p\in\mathcal{P}} \|f\ -\ p\|_E\ > \inf_{\substack{p\in\mathcal{P} \\ \|p\|_E\leq 2\,\|f\|_E}} \|f\ -\ p\|_E$$

ce qui signifie donc que nous pouvons réduire le domaine de recherche de l'infimum à l'ensemble

$$p \in \mathcal{P}, \quad \text{tel que} \quad \|p\|_E > 2\,\|f\|_E.$$

Or, puisque $\mathcal{P}$ est un sous-espace vectoriel de E, nous avons $0 \in \mathcal{P}$ et donc

$$\inf_{p\in\mathcal{P}} \|f\ -\ p\|_E\ \leq\ \|f - 0\|_E\ =\ \|f\|_E.$$

De plus, en considérant $p \in \mathcal{P}$ tel que $\|p\|_E\ >\ 2\,\|f\|_E$, nous obtenons

$$\|f\ -\ p\|_E\ \geq\ \|p\|_E\ -\ \|f\|_E\ >\ \|f\|_E$$

Ainsi,

$$\inf_{q\in\mathcal{P}} \|f - q\|_E \leq \|f\|_E < \inf_{\substack{p\in\mathcal{P} \\ \|p\|_E > 2\|f\|_E}} \|f - p\|_E,$$

ce qui est faux par hypothèse, nous concluons donc que

$$\inf_{p\in\mathcal{P}} \|f - p\|_E = \inf_{\substack{p\in\mathcal{P} \\ \|p\|_E \leq 2\|f\|_E}} \|f - p\|_E.$$

Ainsi, le problème revient à trouver $p^* \in \mathcal{P}$ tel que $\|p\|_E \leq 2\|f\|_E$ et

$$\|f - p^*\|_E = \min_{\substack{p\in\mathcal{P} \\ \|p\|_E \leq 2\|f\|_E}} \|f - p\|_E,$$

ce qui permet de conclure la première étape.

Dans un deuxième temps, nous considérons alors l'ensemble

$$\mathcal{A} = \{p \in \mathcal{P} \subset E; \quad \|p\|_E \leq 2\|f\|_E\},$$

qui est fermé, borné dans E, l'ensemble $\mathcal{A}$ est donc compact. D'une part, par définition de l'infimum, il existe un suite $(p_n)_{n\in\mathbb{N}}$ telle que

$$\lim_{n\to\infty} \|f - p_n\|_E = \inf_{p\in\mathcal{A}} \|f - p\|_E.$$

D'autre part, d'après le Théorème de Bolzano-Weirstrass, puisque $\mathcal{A}$ est compact, nous pouvons extraire une sous-suite qui converge vers $p^* \in \mathcal{P}$. Nous avons alors

$$\|f - p^*\|_E = \inf_{p\in\mathcal{A}} \|f - p\|_E.$$

□

Ce théorème permet de démontrer qu'il existe au moins une solution à notre problème mais il n'est en aucun cas constructif.

C'est pourquoi pour la construction d'une approximation au sens des moindres carrés nous utilisons le plus souvent des polynômes orthogonaux plutôt que la base canonique usuelle 1,..,x^n. Nous avons alors le résultat suivant

Théorème 5.3 *Pour tout $n \geq 1$, nous choisissons P_0,.., P_n une suite de polynômes orthogonaux par rapport au produit scalaire $(.,.)_\omega$ et*

$$\mathcal{P} = vect\{P_0, \ldots, P_n\}.$$

Considérons une fonction f définie sur l'intervalle $[a, b] \subset \mathbb{R}$ et à valeur dans $\mathbb{K}$ ($\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$) telle que pour un poids ω

$$\int_a^b |f(x)|^2 \omega(x)\, dx < \infty.$$

Alors il existe un unique polynôme $Q \in \mathcal{P}$ de la forme

$$Q(x) = \sum_{i=0}^{n} c_i \, P_i(x)$$

solution du problème de minimisation

$$\|f \, - \, Q\|_{L^2_\omega}^2 \, = \, \inf_{P \, \in \mathcal{P}} \|f \, - \, P\|_{L^2_\omega}.$$

De plus, les coefficients $(c_i)_{0 \leq i \leq n}$ sont donnés par

$$c_i = \frac{(P_i, f)_\omega}{(P_i, P_i)_\omega}, \quad i = 0, .., n.$$

Démonstration. La démonstration de l'existence de $Q \in \mathcal{P}$ découle directement du Théorème 5.2. Il reste à démontrer que le polynôme solution du problème de minimisation s'écrit

$$Q(x) = \sum_{i=0}^{n} c_i \, P_i(x)$$

où les coefficients c_i, $0 \leq i \leq n$ sont donnés par

$$c_i = \frac{(P_i, f)_\omega}{(P_i, P_i)_\omega}, \quad i = 0, .., n.$$

Pour cela il suffit d'appliquer le Théorème 4.1, ce qui signifie que la solution vérifie

$$\sum_{i=0}^{n} c_i \, (P_i \, , \, P)_\omega = (f \, , \, P)_\omega, \quad \forall \, P \, \in \mathcal{P}$$

et donc en choisissant $P = P_j$, pour $j = 0, \ldots, n$ et en utilisant l'orthogonalité des polynômes, nous obtenons le résultat

$$c_i \, (P_i \, , \, P_i)_\omega = (f \, , \, P_i)_\omega, \quad i = 0, \ldots, n.$$

□

Ce dernier théorème nous aide bien à comprendre pourquoi en pratique nous utilisons un espace fonctionnel muni d'un produit sclaire ; ce cadre permet de donner une expression simple de la solution du problème de minimisation.

Exemple 5.3 *Trouver le polynôme P de degré inférieur ou égal à trois qui minimise*

$$\int_{-1}^{1} |e^x - P(x)|^2 \, dx.$$

Nous utilisons pour cela la base de polynômes de Legendre

$$L_0(x) \,=\, 1, \quad L_1(x) \,=\, x,$$

et

$$L_2(x) \,=\, \frac{3}{2}\left(x^2 - \frac{1}{3}\right),$$

puis

$$L_3(x) \,=\, \frac{5}{2}\left(x^3 - \frac{3}{5}x\right).$$

Nous calculons alors

$$\begin{aligned}
(f, P_0)_1 &= \int_{-1}^{1} e^x\, dx = e^{-1/e}, \\
(f, P_1)_1 &= \int_{-1}^{1} x\, e^x\, dx = 2/e, \\
(f, P_2)_1 &= \frac{3}{2}\int_{-1}^{1}\left(x^2 - \frac{1}{3}\right) e^x\, dx = e - 7/e, \\
(f, P_3)_1 &= \frac{5}{2}\int_{-1}^{1}\left(x^3 - \frac{3}{5}\,x\right) e^x\, dx = -5\,e + 37/e.
\end{aligned}$$

D'autre part, nous montrons que $(P_k, P_k)_1 = 2/(2\,k+1)$. *Nous déduisons alors l'expression du polynôme de degré inférieur ou égal à trois qui est la meilleure approximation au sens des moindres carrés de* e^x.

6 Transformation de Fourier rapide

Pour une fonction f *périodique* de période T, on peut se contenter de considérer l'ensemble discret de fréquences T/n où n est un entier relatif (le cas $n = 0$ correspondant à une fréquence infinie c'est-à-dire à une période nulle) et associer à f un coefficient c_n pour chacune de ces fréquences. Pour simplifier la présentation, on suppose désormais $T = 1$.

Si f est une fonction 1-périodique et intégrable[1] sur tout intervalle de longueur 1, on définit, pour tout $n \in \mathbb{Z}$,

$$c_n \,:=\, \int_0^1 f(x)\, \mathrm{e}^{-2\,i\,\pi\,n\,x}\, \mathrm{d}x\,.$$

Les nombres complexes c_n sont appelés *coefficients de Fourier* de f.

[1] au sens de Lebesgue, qu'il n'est pas indispensable de connaître pour la suite ; il nous suffira de supposer la fonction continue par morceaux.

Exemple. Si f est précisément une fonction sinusoïdale $f_k(x) = \mathrm{e}^{2\,i\,\pi\,k\,x}$, on voit immédiatement que $c_k = 1$, et pour tout $n \neq k$, $c_n = 0$. Plus généralement, si f est une superposition de fonctions sinusoïdales[2]

$$f(x) = \sum_{k=k_1}^{k_2} a_k \, \mathrm{e}^{2\,i\,\pi\,k\,x} \,,$$

ses coefficients de Fourier sont nuls pour $n \notin \{k_1, \ldots, k_2\}$, et $c_n = a_n$ pour $n \in \{k_1, \ldots, k_2\}$.

Une fois définis les coefficients de Fourier d'une fonction 1-périodique, on lui associe la *série de Fourier*

$$\sum_n c_n \, \mathrm{e}^{2\,i\,\pi\,n\,x} \,.$$

Attention, cette série n'est pas nécessairement convergente. Il faut faire quelques hypothèses sur f pour cela (voir plus loin le théorème de Dirichlet). C'est pourquoi il plus prudent de considérer les sommes partielles de la série de Fourier de f :

$$S_N(f) = \sum_{n=-N}^{N} c_n \, \mathrm{e}^{2\,i\,\pi\,n\,x} \,.$$

Remarque. D'après l'exemple vu ci-dessus, si f est un *polynôme trigonométrique*, il existe N_0 tel que pour tout $N \geq N_0$, $S_N(f) = f$.

6.1 Théorie Hilbertienne des séries de Fourier.

Pour $p \in \mathbb{N}^*$, on note $L^p(\mathbb{T})$ l'espace des fonctions 1-périodiques (au sens où $f(x+1) = f(x)$ pour *presque tout*[3] $x \in \mathbb{R}$) et de puissance p-ième intégrable sur $[0,1]$, que l'on munit de la *norme*[4] naturelle

$$\|f\|_{L^p(\mathbb{T})} = \Big(\int_0^1 |f(x)|^p \, \mathrm{d}x \Big)^{1/p} .$$

Remarque. Si $f \in L^2(\mathbb{T})$, alors $f \in L^1(\mathbb{T})$ car

$$\|f\|_{L^1(\mathbb{T})} \leq \|f\|_{L^2(\mathbb{T})} \,.$$

En effet, d'après l'inégalité de Cauchy-Schwarz :

$$\|f\|_{L^1(\mathbb{T})} = \int_0^1 |f(x)| \, \mathrm{d}x \leq \Big(\int_0^1 |f(x)|^2 \, \mathrm{d}x \Big)^{1/2} \Big(\int_0^1 1 \, \mathrm{d}x \Big)^{1/2} = \|f\|_{L^2(\mathbb{T})} \,.$$

[2] on dit alors que f est un *polynôme trigonométrique*.

[3] l'expression *presque tout* veut dire à l'exception de points contenus dans un ensemble de mesure nulle pour la mesure de Lebesgue ; pour une fonction continue par morceaux, cela signifie que "sa valeur" aux points de discontinuité importe peu.

[4] une norme $\|\cdot\|$ sur un $\mathbb{C}$-espace vectoriel est par définition une application à valeurs dans $\mathbb{R}^+$ telle que : 1) ($\|\mathbf{u}\| = 0 \iff \mathbf{u} = 0$) ; 2) $\|\lambda\,\mathbf{u}\| = |\lambda|\,\|u\|$ pour tout scalaire $\lambda \in \mathbb{C}$ et tout vecteur $\mathbf{u}$; 3) $\|\mathbf{u}+\mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|$ pour tous vecteurs $\mathbf{u}$ et $\mathbf{v}$.

Par ailleurs, $L^2(\mathbb{T})$ est un *espace de Hilbert* pour le produit scalaire hermitien[5]

$$\langle f\,,\,g\rangle \;=\; \int_0^1 \; f(x)\;\overline{g(x)}\;\mathrm{d}x$$

naturellement associé à la norme $\|\cdot\|_{L^2(\mathbb{T})}$:

$$\|f\|_{L^2(\mathbb{T})} \;=\; \langle f\,,\,f\rangle^{1/2}\,.$$

On observe alors que la famille $(f_n \;:\; x \;\mapsto\; \mathrm{e}^{2\,i\,\pi\,n\,x})_{n\in\mathbb{Z}}$ est orthonormale dans $L^2(\mathbb{T})$, ce qui signifie

$$\|f_n\|_{L^2(\mathbb{T})} \;=\; 1 \quad\text{et}\quad \langle f_n\,,\,f_k\rangle \;=\; 0 \quad \text{quels que soient } n \text{ et } k \text{ avec } n \neq k\,.$$

En utilisant de façon essentielle cette propriété, on peut montrer le

Théorème 6.1 *Pour tout $f \in L^2(\mathbb{T})$, on a :*

- **(Inégalité de Bessel)** $\|S_N(f)\|_{L^2(\mathbb{T})} \;\leq\; \|f\|_{L^2(\mathbb{T})}$ *quel que soit $N \in \mathbb{N}$.*
- $\lim_{N\to+\infty} \|S_N(f) \;-\; f\|_{L^2(\mathbb{T})} \;=\; 0\,.$
- **(Identité de Parseval)** *Les coefficients de Fourier c_n de f forment une famille de carré sommable et*

$$\sum_{n\in\mathbb{Z}} |c_n|^2 \;=\; \|f\|^2_{L^2(\mathbb{T})}\,.$$

Remarque. Inversement, toute suite $(c_n)_{n\in\mathbb{Z}}$ de carré sommable est la suite des coefficients de Fourier d'une application $f \in L^2(\mathbb{T})$: on montre en effet que les sommes partielles $\sum_{n=-N}^{N}\; c_n\;\mathrm{e}^{2\,i\,\pi\,n\,x}$ forment une suite convergente dans $L^2(\mathbb{T})$, et la limite f est telle que

$$\langle f, \mathrm{e}^{2\,i\,\pi\,k\,x}\rangle \;=\; \lim_{N\to+\infty} \langle \sum_{n=-N}^{N} \; c_n\;\mathrm{e}^{2\,i\,\pi\,n\,x}, \mathrm{e}^{2\,i\,\pi\,k\,x}\rangle \;=\; c_k \quad \text{quel que soit } k \in \mathbb{Z}\,.$$

6.2 Convergence ponctuelle des séries de Fourier.

La théorie précédente peut être raffinée en montrant que, sous certaines conditions, la série de Fourier d'une fonction converge *point par point* vers cette fonction.

Théorème 6.2 *Si la série $\sum_n c_n$ des coefficients de Fourier d'une fonction $f \in L^1(\mathbb{T})$ est absolument convergente, alors sa série de Fourier est uniformément convergente et*

$$\lim_{N\to+\infty}\; \sup_{[0,1]} |S_N(f) \;-\; f| \;=\; 0\,.$$

En particulier, la fonction f est nécessairement continue, comme limite uniforme d'une suite de fonctions continues.

[5]cette notion généralise celle de produit scalaire euclidien aux $\mathbb{C}$-espaces vectoriels. Un produit scalaire hermitien $\langle\cdot,\cdot\rangle$ doit par définition être tel que $\langle \mathbf{u},\mathbf{u}\rangle \;\in\; R^+$ pour tout vecteur $\mathbf{u}$, et 1) $\langle \mathbf{u},\mathbf{v}\rangle \;=\; \overline{\langle \mathbf{v},\mathbf{u}\rangle}$; 2) $\langle \mathbf{u}+\lambda\mathbf{w},\mathbf{v}\rangle \;=\; \langle \mathbf{u},\mathbf{v}\rangle \;+\; \lambda\,\langle \mathbf{w},\mathbf{v}\rangle$; 3) $\langle \mathbf{u},\mathbf{v}+\lambda\mathbf{w}\rangle \;=\; \langle \mathbf{u},\mathbf{v}\rangle \;+\; \overline{\lambda}\,\langle \mathbf{u},\mathbf{w}\rangle$ pour tous vecteurs $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$ et tout scalaire $\lambda \in \mathbb{C}$, et enfin $(\langle \mathbf{u},\mathbf{v}\rangle \;=\; 0 \quad \forall \mathbf{v}) \;\Longleftrightarrow\; \mathbf{u} \;=\; 0$.

Théorème 6.3 (Dirichlet) *Si $f \in L^1(\mathbb{T})$ admet une dérivée à gauche et une dérivée à droite en tout point, alors pour tout x*

$$\lim_{N\to+\infty} S_N(f)(x) \,=\, \frac{1}{2}\,(f(x+0) \,+\, f(x-0))\,,$$

où

$$f(x+0) \,=\, \lim_{\varepsilon \overset{>}{\to} 0} f(x+\varepsilon)\,, \quad f(x-0) \,=\, \lim_{\varepsilon \overset{>}{\to} 0} f(x-\varepsilon)\,.$$

Démonstration. On a par définition

$$S_N(f)(x) \,=\, \sum_{|n|\le N} \int_0^1 f(y)\, \mathrm{e}^{\,2\,i\,\pi\,n\,(x-y)}\,\mathrm{d}y \,=\, \int_0^1 f(x-y) \sum_{|n|\le N} \mathrm{e}^{\,2\,i\,\pi\,n\,y}\,\mathrm{d}y\,.$$

Notons

$$K_N(y) \,:=\, \sum_{|n|\le N} \mathrm{e}^{\,2\,i\,\pi\,n\,y}\,.$$

On montre facilement que

$$K_N(y) \,=\, \frac{\sin(\,(2N+1)\,\pi\,y\,)}{\sin(\pi\,y)} \quad \text{et} \quad \int_0^{1/2} K_N(y)\,\mathrm{d}y \,=\, \int_{-1/2}^0 K_N(y)\,\mathrm{d}y \,=\, \frac{1}{2}\,.$$

D'où

$$\begin{aligned} & S_N(f)(x) \,-\, \frac{f(x+0) \,+\, f(x-0)}{2} \\ =\; & \int_0^{1/2} (\,f(x-y) \,-\, f(x-0)\,)\; K_N(y)\,\mathrm{d}y \\ & + \int_{-1/2}^0 (\,f(x-y) \,-\, f(x+0)\,)\; K_N(y)\,\mathrm{d}y \\ =\; & \int_0^{1/2} \varphi_x(y)\; K_N(y)\,\mathrm{d}y\,, \end{aligned}$$

où

$$\varphi_x(y) \,:=\, f(x-y) \,-\, f(x-0) \,+\, f(x+y) \,-\, f(x+0) \quad \text{vérifie } \lim_{y \overset{>}{\to} 0} \varphi_x(y) \,=\, 0\,.$$

Mais attention,

$$\lim_{y \overset{>}{\to} 0} K_N(y) \,=\, (2N+1)\,\pi\,,$$

ce qui tend vers $+\infty$ avec N ! C'est pourquoi on a besoin de faire apparaître la dérivée de f. Considérons par exemple le premier morceau :

$$\begin{aligned}\int_0^{1/2} (\, f(x-y) \,-\, f(x-0)\,)\, K_N(y)\,\mathrm{d}y \;=\; &= \\ -\int_0^{1/2} \Big(\int_0^{y} f'(x-u)\,\mathrm{d}u\Big)\, K_N(y)\,\mathrm{d}y \;=\; &\\ -\int_0^{1/2} \Big(\int_u^{1/2} K_N(y)\,\mathrm{d}y\Big)\, f'(x-u)\,\mathrm{d}u\,. &\end{aligned}$$

On montre que $\int_u^{1/2} K_N(y)\,\mathrm{d}y$ est bornée indépendamment de N, et tend vers 0 lorsque N tend vers $+\infty$ dès que u est *strictement positif* (et inférieur à $1/2$). Donc on peut appliquer le théorème de convergence dominée : cela prouve que la double intégrale tend vers 0 lorsque N tend vers $+\infty$. L'autre morceau se traite de la même façon. □

Attention cependant, aux points de discontinuité, l'approximation d'une fonction par les sommes partielles de sa série de Fourier n'est pas très bonne : c'est ce qu'on appelle le *phénomène de Gibbs*.

Théorème 6.4 (Phénomène de Gibbs) *Soit*

$$C \;:=\; \int_0^1 \frac{\sin(\pi x)}{\pi x}\,\mathrm{d}x \;\approx\; 0.58948987\ldots$$

Si $f \in L^1(\mathbb{T})$ admet une dérivée à gauche et une dérivée à droite en tout point, alors pour tout x

$$\begin{aligned}\lim_{N\to+\infty} \Big(\, S_N(f) \,-\, f\Big)|_{x+\frac{1}{2N+1}} \;&=\; (C \,-\, \frac{1}{2})\,(f(x+0) \,-\, f(x-0))\,, \\ \lim_{N\to+\infty} \Big(\, S_N(f) \,-\, f\Big)|_{x-\frac{1}{2N+1}} \;&=\; -\,(C \,-\, \frac{1}{2})\,(f(x+0) \,-\, f(x-0))\,.\end{aligned}$$

Nous voulons approcher la fonction f par un polynôme trigonométrique. Ce polynôme s'écrit alors de la manière suivante

$$P_n(x) \;=\; \sum_{|k|\le n} c_k\, e^{2\,\pi\, i\, k\, x/T}$$

où les nombres c_k sont des nombres complexes tandis que n est un entier positif non nul. Un tel polynôme trigonométrique est dit de degré n. Il est exactement de degré n dès que le coefficient c_n est non nul. La suite des polynômes trigonométriques $(P_n)_n$ définie des polynômes orthogonaux pour le produit scalaire

$$(p,q) = \int_0^{2\,\pi} p(x)\,\overline{q(x)}\,dx.$$

Nous avons le résultat d'approximation suivant

Théorème 6.5 *Soit f une fonction définie sur l'intervalle $[0,T]$ à valeur dans $\mathbb{C}$ de classe C^1. Alors, il existe une meilleure approximation au sens des moindres carrés dans l'espace des polynômes trigonométriques de la forme*

$$P_n(x) = \sum_{|k|\leq n} c_k \, e^{2\,\pi\,i\,k\,x/T}$$

où les coefficients c_k sont donnés par

$$c_k = \frac{1}{T}\int_0^T f(x) e^{-2\,\pi\,i\,k\,x/T} dx \;=\; \hat{f}(k).$$

Démonstration. Nous appliquons simplement le Théorème 5.3. □

Dans la suite nous donnons une procédure qui permet un calcul précis et rapide des coefficients $(c_k)_{-n\leq k\leq n}$. L'algorithme s'appelle la transformée de Fourier discrète.

Plus généralement, la transformée de Fourier discrète (TFD) est un outil mathématique de traitement du signal, qui est l'équivalent discret de la transformée de Fourier.

Commençons en considérant une fonction périodique f sur l'intervalle $[0,T]$, nous définissons le polynôme trigonométrique

$$P_n(x) = \sum_{|k|\leq n} \hat{f}(k)\, e^{2\,\pi\,i\,k\,x/T},$$

avec le coefficient de Fourier $\hat{f}(k)$

$$\hat{f}(k) \;:=\; \frac{1}{T}\int_0^T f(x)\, e^{-2\,\pi\,i\,k\,x}\, dx, \quad -n \leq k \leq n.$$

Notre objectif est de fournir des algorithmes efficaces pour le calcul d'un polynôme trigonométrique ou pour une transformée de Fourier discrète. Pour commencer, nous approchons les coefficients de Fourier par la formule des rectangles à points équidistants (voir complément du chapitre) ; une approximation $\hat{f}^p(k)$ de $\hat{f}(k)$ s'écrit alors

$$\hat{f}^p(k) = \frac{1}{p}\sum_{j=0}^{p-1} f\left(\frac{j\,T}{p}\right)\, e^{-2\,\pi\,i\,k\,j/p}$$

Remarquons d'abord que cette formule produit au plus p nombres complexes distincts. En effet,

$$\begin{aligned}\hat{f}^p(k+p) &= \frac{1}{p}\sum_{j=0}^{p-1} f\left(\frac{j\,T}{p}\right)\, e^{-2\,\pi\,i\,(k+p)\,j/p} \\ &= \frac{1}{p}\sum_{j=0}^{p-1} f\left(\frac{j\,T}{p}\right)\, e^{-2\,\pi\,i\,k\,j\,/p} \;=\; \hat{f}^p(k).\end{aligned}$$

Il suffit donc de calculer p coefficients consécutifs pour calculer les coefficients c_k, $k = 0, \ldots, p-1$. En général, nous choisissons $p = 2\,n + 1$ pour un choix optimal puisque cela correspond au nombre de coefficients $(c_k)_{-n \leq k \leq n}$.

Nous définissons donc la transformation de Fourier discrète $\mathcal{F}_p$ comme l'opérateur linéaire qui associe à une suite finie $(f_j)_{0 \leq j \leq p-1}$ composée de nombres complexes, la suite des $(\hat{f}^p(k))_{0 \leq k \leq p-1}$ donnée par

$$\hat{f}^p(k) = \frac{1}{p} \sum_{j=0}^{p-1} f_j \, e^{-2\,\pi\,i\,k\,j/p}, \quad 0 \leq k \leq p-1.$$

Lemme 6.1 *Nous avons pour tout p les identités suivantes*

$$\mathcal{F}_p o \overline{\mathcal{F}}_p \; = \; p\, I_p \; = \; \overline{\mathcal{F}}_p o \mathcal{F}_p.$$

Démonstration. Montrons que $\mathcal{F}_p o \overline{\mathcal{F}}_p \; = \; p\, I_p$. Pour cela, nous considérons un ensemble $(f_j)_{0 \leq j \leq p-1}$

$$f_j = \sum_{k=0}^{p-1} \hat{f}^p(k)\, e^{2\,\pi\,i\,j\,k/p}.$$

Nous calculons alors le terme suivant

$$\sum_{j=0}^{p-1} f_j \, e^{-2\,\pi\,i\,j\,l/p}$$

et montrons qu'il vaut $n\, \hat{f}^p(k)$.

En effet, nous avons

$$\begin{aligned} \sum_{j=0}^{p-1} f_j \, e^{-2\,\pi\,i\,j\,l/p} &= \sum_{j=0}^{p-1} e^{-2\,\pi\,i\,j\,l/p} \sum_{k=0}^{p-1} \hat{f}^p(k)\, e^{2\,\pi\,i\,j\,k/p}. \\ &= \sum_{k=0}^{p-1} \left(\sum_{j=0}^{p-1} e^{-2\,\pi\,i\,j\,(k-l)/p} \right) \hat{f}^p(k). \end{aligned}$$

Or, nous savons que

$$\sum_{j=0}^{p-1} e^{2\,\pi\,i\,j\,(k-l)/p} = \begin{cases} p & \text{si } p \text{ divise } k-l \\ 0 & \text{sinon.} \end{cases}$$

Comme k et l varient entre 0 et $p-1$, alors p ne peut diviser $(k-l)$ que lorsque $k = l$. Par conséquent, il vient

$$\sum_{j=0}^{p-1} f_j \, e^{-2\,\pi\,i\,j\,k/p} = p\, \hat{f}^p(k),$$

ce qui montre bien que $\mathcal{F}_p o \overline{\mathcal{F}}_p \; = \; p\, I_p$.

La seconde égalité se déduit immédiatement de la première par conjugaison. □

6.3 Algorithme de Cooley-Tukey

C'est en 1965 que James Cooley et John Tukey publient cette méthode [4] mais il semblerait que l'algorithme avait déjà été proposé par Carl Friedrich Gauss [9] en 1805 et adapté à plusieurs reprises sous des formes différentes.

Nous décrivons maintenant le principe de la transformation de Fourier Rapide. Elle est basée sur une décomposition d'un entier p. Nous supposerons ici que $p = p_1\, p_2$, où p_1 et p_2 sont deux entiers supérieurs à 2. Nous notons

$$\omega_n = e^{-2\,\pi\, i/p}$$

Il s'agit donc de calculer le vecteur $\hat{f}^p \in \mathbb{C}^p$ de composantes $\hat{f}^p(k)$ pour $0 \leq k \leq p-1$ données par

$$\hat{f}^p(k) = \sum_{j=0}^{p-1} f_j\ \omega_p^{j\,k}.$$

D'abord, nous interprétons le vecteur $(f_j)_{0\leq j\leq p-1}$ comme une matrice à p_1 lignes et p_2 colonnes, pour $0 \leq j_1 < p_1$ et $0 \leq j_2 < p_2$

$$f_{j_1,j_2} = f_j, \quad 0 \leq j = j_2\, p_1 + j_1 < p,$$

puis nous écrivons $\hat{f}^p(k)$ en utilisant cette écriture

$$\begin{aligned}
\hat{f}^p(k) &= \sum_{j_1=0}^{p_1-1} \sum_{j_2=0}^{p_2-1} f_{j_1,j_2}\ \omega_p^{(j_2\,p_1+j_1)\,k} \\
&= \sum_{j_1=0}^{p_1-1} \left(\sum_{j_2=0}^{p_2-1} f_{j_1,j_2}\ \omega_{p_2}^{j_2\,k} \right) \omega_p^{j_1\,k},
\end{aligned}$$

puisque $\omega_p^{j_2\,p_1\,k} = e^{-2\,\pi\, i\, j_2\, p_1\, k/p} = e^{-2\,\pi\, i\, j_2\, k/p_2} = \omega_{p_2}^{j_2\,k}$.

Maintenant, il nous faut observer que la somme entre parenthèses est une fonction de k périodique de période p_2 car nous avons déjà vu que

$$\omega_{p_2}^{j_2\,(k+p_2)} = \omega_{p_2}^{j_2\,k}\, \omega_{p_2}^{j_2\,p_2} = \omega_{p_2}^{j_2\,k}.$$

Ainsi en décomposant k comme $k = k_1\, p_2 + k_2$ avec $0 \leq k_1 < p_1$ et $0 \leq k_2 < p_2$, nous voyons qu'il suffit de calculer la somme entre parenthèse seulement pour $0 \leq k_2 < p_2$

$$\omega_{p_2}^{j_2\,k} = \omega_{p_2}^{j_2\,(k_1\,p_2+k_2)} = \omega_{p_2}^{j_2\,k_2}.$$

L'algorithme pour calculer une transformée de Fourier discrète devient donc le suivant.

Algorithme 1. La transformation de Fourier discrète

Considérons une fonction f connue en $p = p_1\, p_2$ points est donnée par :

-Nous commençons par calculer la somme entre parenthèses

$k = k_1\, p_2 + k_2$ et $j = j_2\, p_1 + j_1$:

-Pour $k_2 = 0, ..., p_2 - 1$ et $j_1 = 0, ..., p_1 - 1$

$$S_{j_1,k_2} \;=\; \sum_{j_2=0}^{p_2-1} f_{j_1,j_2}\; \omega_{p_2}^{j_2\, k_2}.$$

Fin de pour k_2 et j_1.

-Nous calculons chaque composante $\hat{f}^p(k)$ avec $k = k_1\, p_2 + k_2$:

Pour $k_1 = 0, ..., p_1 - 1$ et $k_2 = 0, ..., p_2 - 1$

$$\hat{f}^p(k) = \sum_{j_1=0}^{p_1} S_{j_1,k_2}\; \omega_p^{j_1\,(k_1/,p_2+k_2)}.$$

Fin de pour k_1 et k_2.

La première étape nécessite $p\, p_2$ opérations tandis que la deuxième nécessite $p\, p_1$ opérations, ce qui représente un coût global de $p\,(p_1 + p_2)$ opérations ce qui est déjà bien moindre que le coût initial de n^2 opérations.

En effet, prenons $p \;=\; 1\,000$, le coût initial représente alors $p^2 = 1\,000\,000$ opérations. Alors que lorsque nous décomposons $p = 10 \times 100$, le coût devient de $p(p_1 + p_2) = 1\,000 \times 110 \;=\; 110\,000$ ou alors $p = 25 \times 40$, et le coût est alors de $p\,(p_1 + p_2) = 1000 \times 65 = 65000$!

Nous pouvons encore diminuer le nombre d'opérations en décomposant n sous la forme $p = p_1\, p_2\, ...\, p_k$ en répétant l'idée ci-dessus. Un cas fréquent utilisé est celui où $p = 2^k$.

7 Complément du Chapitre 5

7.1 Formules de quadratures classiques

Soit f une fonction continue définie de l'intervalle $[a, b]$ à valeurs dans $\mathbb{R}$. Il existe toute une famille d'algorithmes permettant d'approcher la valeur numérique de l'intégrale

$$I \;=\; \int_a^b f(x)\, dx.$$

Toutes consistent à approcher l'intégrale par une formule dite de "quadrature", du type

$$I(f) \;=\; \sum_{i=1}^{p} \omega_i \, f(x_i).$$

Le choix de p, des poids ω_i et des points x_i dépendent de la méthode employée. Il conviendra aussi de s'intéresser à la précision des formules utilisées. Ces méthodes utilisent l'interpolation des fonctions à intégrer. Généralement, les fonctions sont interpolées par des polynômes dont nous connaissons facilement la primitive.

Formules des rectangles et du point milieu

C'est la méthode la plus simple qui consiste à interpoler la fonction f à intégrer par une fonction constante (polynôme de degré 0). Soit ξ le point d'interpolation ; la formule devient alors :

$$I(f) \;=\; (b-a)\, f(\xi).$$

Le choix du point a de l'importance pour la détermination du terme d'erreur :

– Si $\xi = a$ ou $\xi = b$, l'erreur est

$$|I(f) - I| \;\leq\; \frac{(b-a)^2}{2} \sup_{y\in[a,b]} |f'(y)|.$$

C'est la méthode des rectangles ;

– Si $\xi = (a+b)/2$, alors l'erreur devient

$$|I(f) - I| \;\leq\; \frac{(b-a)^3}{6} \sup_{y\in[a,b]} |f''(y)|.$$

Il s'agit de la méthode du point milieu.

Ainsi, le choix du point milieu améliore le degré d'exactitude de la formule, qui est défini comme le plus haut degré des polynômes pour lesquels la formule est exacte (c'est-à-dire $E(f) = |I(f) - I| = 0$). La méthode des rectangles est exacte pour les fonctions constantes et celle du point milieu est exacte pour les polynômes de degré inférieur ou égal à un. Ceci s'explique par le fait que pour l'intégration de x, la méthode du point milieu donne lieu à deux erreurs d'évaluation, égales en valeur absolue et opposées en signe.

Formule des trapèzes

Si nous interpolons f par un polynôme de degré un (c'est une fonction affine), nous avons besoin de deux points d'interpolation, à savoir $(a, f(a))$ et $(b, f(b))$. L'intégrale est alors approchée par l'aire du polynôme d'interpolation, en l'occurrence un trapèze. Ceci justifie le nom de méthode des trapèzes :

$$I(f) \;=\; \frac{(b-a)}{2}\,(f(a) + f(b)).$$

L'erreur commise est

$$E(f) \;=\; |I(f) - I| \;\leq\; \frac{(b-a)^3}{12} \sup_{y\in[a,b]} |f''(y)|.$$

L'erreur s'annule pour tout polynôme de degré inférieur ou égal à un. Selon ce critère, la méthode des trapèzes est donc moins performante que celle du point milieu, étant donné que les degrés d'exactitude sont les mêmes et que le nombre d'évaluations est plus grand pour la méthode des trapèzes que pour celle du point milieu.

Formule de Simpson

La fonction f est maintenant remplacée par une parabole, qui nécessite trois points d'interpolation. Les extrémités a, b, et leur milieu m sont choisis. La méthode de Simpson consiste alors à remplacer l'intégrale par

$$I(f) \;=\; \frac{(b-a)}{6}\,(f(a) \;+\; 4\,f(m) \;+\; f(b)).$$

L'erreur est

$$E(f) \;=\; |I(f) - I| \;\leq\; \frac{(b-a)^5}{90} \sup_{y\in[a,b]} |f^{(4)}(y)|.$$

Le degré d'exactitude est de 3 pour cette méthode, pour 3 évaluations de f.

7.2 Formules de Newton-Cotes

Les formules de Newton-Cotes permettent de généraliser ces résultats sur des intervalles constants, où la fonction f est interpolée par des polynômes de degré de plus en plus élevé. Pour des questions de stabilité numérique, il est préférable de limiter le degré du polynôme d'interpolation en subdivisant l'intervalle en sous-intervalles, pour lesquels une interpolation linéaire est suffisante.

Pour chacune des méthodes précédentes, le terme d'erreur dépend de $b - a$. Si cette amplitude est trop élevée, nous pouvons réduire simplement l'erreur en découpant l'intervalle $[a, b]$ en n sous-intervalles, sur lesquels nous calculerons la valeur approchée de l'intégrale. Nous parlons alors de formule composite. La valeur sur l'intervalle $[a, b]$ sera la somme de la valeur sur chaque sous-intervalle.

Pour la méthode du point milieu, la formule devient

$$I(f) \;=\; h \sum_{k=0}^{n-1} f(m_k)$$

où m_k est le milieu du k-ième sous-intervalle. Puisque les n sous-intervalles sont identiques, ils sont de la forme $[a + kh, a + (k + 1)h]$, avec $h \;=\; (b - a)/n$ et $k \;=\; 0, 1, 2, ..., n - 1$. Ceci

entraîne finalement que $m_k = a + (k + 1/2)\, h$. Le terme d'erreur s'écrit

$$E(h) \leq h^2 \, \frac{(b-a)}{24} \sup_{y \in [a,b]} |f''(y)|.$$

La formule composite a un ordre un, comme précédemment. La somme a fait baisser d'une puissance le terme en $(a - b)$.

Pour la méthode des trapèzes, la formule composite est

$$I(f) = h \left(\frac{f(a) + f(b)}{2} + \sum_{i=1}^{n-1} f(x_i) \right)$$

Le terme d'erreur s'écrit .

$$E(f) \ \leq \ h^2 \frac{(b-a)}{24} \sup_{y \in [a,b]} |f''(y)|.$$

La formule composite de Simpson prend la forme

$$I(f) = \frac{h}{6} \left(f(a) + f(b) + 2 \sum_{i=1}^{n-1} f(x_i) + 4 \sum_{i=0}^{n-1} f(x_i + h/2) \right)$$

et l'erreur devient

$$E(f) \ \leq \ h^4 \frac{(b-a)}{2880} \sup_{y \in [a,b]} |f^{(4)}(y)|.$$

7.3 Méthode de Gauss

En général, nous remplaçons le calcul de l'intégrale par une somme pondérée prise en un certain nombre de points du domaine d'intégration. La méthode de quadrature de Gauss est une méthode de quadrature exacte pour un polynôme de degré $2n - 1$ avec n points pris sur le domaine d'intégration. Si ce dernier est (a, b), nous voulons calculer une approximation de

$$I \ = \ \int_a^b f(x)\, \omega(x)\, dx$$

où ω est une fonction de poids définie sur $[a, b]$, qui peut assurer l'intégrabilité de f. Les méthodes de Gauss sont de la forme

$$I(f) \ = \ \sum_{i=1}^{n} \omega_i \, f(x_i)$$

où ω_i sont appelés les coefficients de quadrature. Les points x_i sont réels, distincts, uniques et sont les racines de polynômes orthogonaux, choisis conformément au domaine d'intégration et à la fonction de poids.

Pour le problème d'intégration le plus classique, nous utilisons la méthode de Gauss-Legendre. Il s'agit d'intégrer la fonction f sur le segment $[-1,1]$. Les n points sont les racines du $n$ème polynôme de Legendre, $P_n(x)$, et les coefficients sont

$$\omega_i \;=\; \frac{2}{(n+1)\,P_{n+1}(x_i)\,P'(x_i)}$$

Nous pouvons aussi remarquer que la somme des coefficients est égale à 2. Le tableau suivant donne l'ensemble des informations pour réaliser le calcul approché de I pour les formules à un, deux et trois points.

Nombre de points	Poids ω_i	Points x_i	Polynôme de Legendre
1	2	0	x
2	(1,1)	$(-\sqrt{1/3},\sqrt{1/3})$	$\frac{1}{3}(3\,x^2-1)$
3	(5/9, 8/9, 5/9)	$(\sqrt{3/5},0,\sqrt{3/5})$	$\frac{1}{2}(5\,x^3-3\,x)$

7.4 Polynômes de Chebychev

Les polynômes de Chebychev de première espèce définis par

$$T_n(x) = \cos(n\,\arccos(x)), \quad x \in [-1,1]$$

sont orthogonaux pour le produit scalaire associé à

$$\omega(x) = \frac{1}{\sqrt{1-x^2}}.$$

Les propriétés des polynômes de Chebyshev sont les suivantes

Proposition 7.1 *Les fonctions $T_n(x)$ satisfont la formule de récurrence*

$$\begin{cases} T_0(x) = 1, \\ T_1(x) = x, \\ T_{n+1}(x) \;=\; 2\,x\,T_n(x) \;-\; T_{n-1}(x), \end{cases}$$

Par conséquent, $T_n(x)$ est un polynôme de degré n dont le coefficient de x^n est 2^{n-1}, c'est-à-dire

$$T_n(x) \;=\; 2^{n-1}\,x^n + \ldots$$

De plus,

- *nous avons la borne L^∞ : $|T_n(x)| \;\leq\; 1$ pour $x \;\in\; [-1,1]$,*
- *puis, T_n vérifie*

$$T_n\left(\cos\left(\frac{k\,\pi}{n}\right)\right) \;=\; (-1)^k$$

pour $k = 0, 1, \ldots, n$ et de plus

$$T_n\left(\cos\left(\frac{(2\,k\;+\;1)\,\pi}{2\,n}\right)\right) \;=\; 0$$

pour $k = 0, 1, \ldots, n-1$,

- *les polynômes $T_n(x)$ sont orthogonaux par rapport à la fonction poids*

$$\omega(x) \;=\; (1-x^2)^{-1/2}$$

et

$$\int_{-1}^{1} T_n(x)\;T_m(x)\,\frac{dx}{(1-x^2)^{1/2}} \;=\; \begin{cases} \pi & \textit{si } n = m = 0, \\ \pi/2 & \textit{si } n = m \neq 0, \\ 0 & \textit{si } n \neq m. \end{cases}$$

Démonstration. Posons $\theta(x) = \arccos(x)$, nous obtenons alors

$$T_{n+1}(x) = \cos\left((n+1)\theta\right) = \cos\left(n\,\theta\right)\,\cos\left(\theta\right) - \sin\left(n\,\theta\right)\,\sin\left(\theta\right)$$

et

$$T_{n-1}(x) = \cos\left((n-1)\theta\right) = \cos\left(n\,\theta\right)\,\cos\left(\theta\right) + \sin\left(n\,\theta\right)\,\sin\left(\theta\right)$$

Ainsi,

$$T_{n+1}(x) \;+\; T_{n-1}(x) \;=\; 2\,\cos\left(n\,\theta\right)\,\cos\left(\theta\right) = 2\,x\,T_n(x).$$

Aussi, par changement de variable $\theta = \arccos(x)$, nous avons

$$\int_{-1}^{1} T_n(x)\,T_m(x)\,\frac{dx}{\sqrt{1-x^2}} = \int_0^{\pi} \cos(n\theta)\;\cos(m\theta)\,d\theta = \begin{cases} 0, & \text{si } m \neq n \\ \pi, & \text{si } m = n = 0 \\ \pi/2, & \text{si } m = n \geq 0 \end{cases}$$

Montrons ensuite que T_n est bien un polynôme de degré n. Pour cela, nous prcédonc par récurrence : T_0 est un polynôme de degré zéro et T_1 est bien un polynôme de degré un. Nous supposons alors que pour $k \geq 1$, T_{k-1} (*resp.* T_k) est un polynôme de degré $k-1$ (*resp.* k) et le coefficient devant le terme de plus haut degré de T_k est donné par 2^{k-1} ; alors en utilisant la formule de récurrence

$$T_{k+1}(x) \;=\; 2\,x\,T_k(x) - T_{k-1}(x)$$

nous avons bien que T_{k+1} est un polynôme de degré $k+1$ et le coefficient devant le terme de plus haut degré est donné par $2 \times 2^{k-1} = 2^k$. □

La formule de l'erreur de l'interpolation de Lagrange en un point $x \in [a,b]$ (2.4) montre que l'erreur de l'interpolation d'une fonction f est un produit de la $(n+1)$-ème dérivée de f évaluée en un point avec l'expression $\prod_{i=0}^{n}(x-x_i)$ qui ne dépend que de la répartition de points sur l'intervalle de résolution. Un problème intéressant consiste donc à rechercher, pour un n donné, la localisation des points $\{x_0,\ldots,x_n\}$ de l'intervalle $[a,b]$ pour laquelle

$$\mathcal{E} = \max_{x\in[a,b]} \left| \prod_{i=0}^{n}(x-x_i) \right|$$

est minimal. Nous montrons alors la proposition suivante

Proposition 7.2 *Soit $P_n(x)$ un polynôme de degré n considéré sur l'intervalle $[-1,1]$ et dont le coefficient de x^n est 2^{n-1} (comme pour le polynôme de Chebyshev) et soit $P_n \neq T_n$. Alors, nous avons*

$$\max_{x\in[-1,1]} |P_n(x)| > \max_{x\in[-1,1]} |T_n(x)| = 1.$$

Démonstration. Supposons, par l'absurde, que

$$\max_{x\in[-1,1]} |P_n(x)| \leq \max_{x\in[-1,1]} |T_n(x)|\,.$$

et considérons la différence $d(x) = P_n(x) - T_n(x)$. Puisque le polynôme T_n atteint ses extrema $+1$ et -1 aux points

$$y_k = \cos\left(\frac{k\,\pi}{n}\right), \quad k = 0,\ldots,n$$

La fonction d fonction s'annule au moins une fois dans chacun des intervalles fermés

$$\left[\cos\left(\frac{(k+1)\,\pi}{n}\right), \cos\left(\frac{k\,\pi}{n}\right)\right], \quad k = 0,\ldots,n-1 \tag{7.10}$$

Alors, $d(x)$ possède n zéros dans $[-1,1]$ (si une racine $\alpha \in [-1,1]$ est à l'extrémité de l'intervalle (7.10), elle doit être comptée deux fois car en un tel point $T_n'(\alpha) = 0$ et $P_n'(\alpha) = 0$). D'autre part, puisque $d(x)$ est un polynôme de degré $n-1$ (le coefficient de x^n est le même pour P_n et T_n), ceci est une contradiction à $d \neq 0$. □

Chapitre 6

Les équations différentielles ordinaires

1 Motivation : le problème du pendule

Le mouvement d'un pendule de masse m, suspendu à un point O par un fil non pesant de longueur l, en rotation d'angle $\theta(t)$ autour de O est gouverné par l'équation :

$$\theta''(t) = -\frac{g}{l}\,\sin(\theta(t)).$$

L'angle $\theta(t)$ est mesuré par rapport à une verticale passant par O. Nous nous intéressons au mouvement entre les instants $t^0 = 0$ et t^{fin}. Les conditions initiales peuvent être :

$$\theta_0 = \frac{\pi}{3}\,rad., \qquad \theta'(0) \,=\, 0\ rad./s.$$

En général, nous introduisons le paramètre $\omega^2 = g/l$. Ce problème est un problème non linéaire et différentiel. L'équation différentielle s'écrit en fonction des dérivées d'ordre deux, nous disons donc qu'elle est d'ordre deux et les conditions en font un problème à valeurs initiales ; ceci est nécessaire pour que le problème soit bien posé c'est-à-dire pour qu'il admette une unique solution. Nous pouvons également le transformer en un système de deux équations différentielles du premier ordre. Nous posons $u_1(t) = \theta(t)$ et $u_2(t) = \theta'(t)$ et obtenons alors :

$$\begin{cases} u_1'(t) = u_2(t) \\ \\ u_2'(t) = -\omega^2 \sin(u_1(t)) \end{cases}$$

2 Rappel théorique

Nous rappelons d'abord les résultats classiques d'existence et unicité de sollutions pour les systèmes différentiels. Nous renvoyons au cours sur les equations différentielles pour les démonstrations.

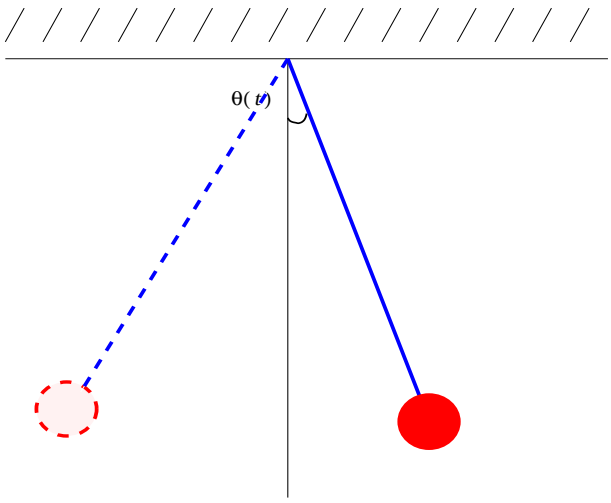

FIG. 6.1 – Étude du mouvement d'un pendule à l'aide d'une équation aux dérivées ordinaires.

Soit un système différentiel de la forme

$$\begin{cases} u(0) = u_0, \\ u'(t) = f(t, u(t)), \end{cases} \tag{2.1}$$

où la fonction $f : \Omega \subset I\!R^d \to I\!R^d$ et $u_0 \in \Omega$. Nous donnons d'abord un résultat d'existence locale

Théorème 2.1 (Théorème de Cauchy-Lipschitz) *Si $f : I\!R^+ \times I\!R^d \to I\!R^d$ est une fonction continue et localement Lipschitzienne en $x \in I\!R^d$, c'est-à-dire pour tout $x_0 \in I\!R^d$, il existe $L(x_0) > 0$ et un voisinnage V de x_0 tel que $\forall\, (x, y) \in V$*

$$\|f(t, x) - f(t, y)\| \ \leq \ L(x_0)\, \|x \ - \ y\|.$$

Alors, il existe une unique solution maximale $u \in \mathcal{C}([0, T], I\!R^d)$ de (2.1) pour $t < T$.

La démonstration de ce théorème repose sur le théorème de point fixe contractant (voir Chapitre 3[Théorème 2.1]). Nous définissons ensuite la notion de trajectoire

Définition 2.1 *Nous appelons trajectoire partant de u_0, l'ensemble défini par*

$$\mathcal{T}_{u_0} = \{u(t) \in \Omega; \quad t \ > \ 0\}$$

où $u(t)$ est la solution maximale correspondant à la donnée initiale u_0.

En appliquant le Théorème 2.1, nous avons immédiatement le résultat suivant

Corollaire 2.1 *Si $u_0 \neq u_1$ alors les trajectoires $\mathcal{T}_{u_0}$ et $\mathcal{T}_{u_1}$ sont disjinctes.*

Plaçons nous maintenant dans le cadre d'application du Théorème 2.1, le résultat suivant donne une condition suffisante pour que $T = \infty$

Théorème 2.2 *Soient $f : \mathbb{R}^+ \times \mathbb{R}^d \to \mathbb{R}^d$ est une fonction continue et localement Lipschitzienne en $x \in \mathbb{R}^d$ et u la solution maximale de (2.1) définie pour $t < T$. Si la solution u est bornée sur $[0, T]$ alors $T = \infty$.*

Après ces résultats préliminaires d'existence et d'unicité, rappelons quelques résultats sur le comportement qualitatif de la solution u.

Définition 2.2 *Soit $\bar{u} \in \Omega \subset \mathbb{R}^d$. Nous dirons que $\bar{u}$ est un point d'équilibre ou stationaire dès qu'il vérifie $f(\bar{u}) = 0$.*

À partir de cette définition, il se pose immédiatement le problème de stabilité de ces points d'équilibre.

Définition 2.3 *Soit $\bar{u} \in \Omega \subset \mathbb{R}^d$ un point d'équilibre ou stationaire. Nous dirons que $\bar{u}$ est stable dès qu'il vérifie*

$$\forall \varepsilon > 0, \quad \exists\, \eta > 0, \quad \|u_0 - \bar{u}\| \leq \eta \;\Rightarrow\; \|u(t) - \bar{u}\| \leq \varepsilon.$$

Si de plus

$$\lim_{t \to \infty} \|u(t) - \bar{u}\| = 0,$$

le point est dit asymptotiquement stable.

Dans le cas linéaire $f(x) = A\,x$, il suffit d'étudier le spectre de la matrice A ; En effet, le système différentiel s'écrit

$$\begin{cases} u(0) = u_0, \\ u'(t) = A\,u(t), \end{cases} \tag{2.2}$$

Nous nous intéressons au point d'équilibre $\bar{u} = 0$

Théorème 2.3 *Considérons la solution u du système differentiel (2.2). Alors, nous avons*

- *Le point 0 est asymptotiquement stable si et seulement si pour tout $\lambda \in Sp(A)$,*

$$Re(\lambda) < 0$$

- *Le point 0 est stable si et seulement si*
  - *pour tout $\lambda \in Sp(A)$, $Re(\lambda) < 0$,*
  - *pour tout $\lambda \in Sp(A)$, $Re(\lambda) = 0$ implique que λ n'est pas défective, c'est-à-dire $dim(Ker(A - \lambda\, I_n)) = p$, où p est la multiplicité de λ.*

3 Schémas à un pas explicites

Dans tout ce qui suit, nous supposons que la fonction vit sur l'intervalle $[0, T]$, que nous décomposons en N petits sous-intervalles $[t^n, t^{n+1}]$ avec $t^n = n\,\Delta t$ et $\Delta t = T/N$. L'objectif de l'analyse numérique des équations aux dérivées ordianires est de cnstruire des schémas qui permettent de calculer des valeurs approchées de la solution $u(t)$ de (2.1). Dans la suite le paramètre Δt va tendre vers zéro ce qui signifie que nous calculons une solution approchée en un nombre de points de plus en plus grand.

Ainsi, nous cherchons à mettre au point une méthode qui permette le calcul d'une solution approchée u^n aux points t^n, pour $n = 1, \ldots$ et telle que la solution approchée converge, en un sens à préciser, vers la solution exacte. Nous chercherons de plus à évaluer l'erreur de discrétisation $e^n = u(t^n) - u^n$, et plus précisément, à obtenir des estimations d'erreur de la forme

$$|e^n| = |u(t^n) - u^n| \leq C\,\Delta t^\alpha,$$

où C ne dépend que de la solution exacte, du temps final T mais surtout pas du pas de temps Δt ; tandis que α donne l'ordre de la convergence.

3.1 Les schémas de Runge-Kutta

Il nous faut d'abord décrire la manière de construire un schéma numérique pour l'équation (2.1). Nous intégrons cette équation sur l'intervalle $[t^n, t^n + \Delta t]$, nous avons alors

$$u(t^{n+1}) - u(t^n) = \int_{t^n}^{t^{n+1}} f(s, u(s))\,ds.$$

Nous utilisons alors une formule de quadrature pour approcher l'intégrale.

Par exemple, nous considérons simplement

$$\int_{t^n}^{t^{n+1}} f(s, u(s))\,ds \simeq \Delta t\, f(t^n, u(t^n));$$

nous obtenons le **schéma d'Euler explicite** en remplaçant $u(t^n)$ par son approximation u^n, il vient alors

$$\begin{cases} u^0 = u(0) \\ \\ u^{n+1} = u^n + \Delta t\, f(t^n, u^n), \quad \text{pour } n = 0, \ldots \end{cases} \tag{3.3}$$

Une autre approximation consiste à utiliser la formule du point milieux

$$\int_{t^n}^{t^{n+1}} f(s, u(s))\,ds \simeq \Delta t\, f(t^n + \frac{\Delta t}{2}, u(t^n + \frac{\Delta t}{2}))$$

mais au préalable, il nous faut construire une approximation de $u(t^n + \frac{\Delta t}{2})$. Pour cela, nous utilisons simplement le schéma d'Euler explicite présenté précédement sur un demi-pas de temps

$$u(t^n + \frac{\Delta t}{2}) \simeq u(t^n) + \frac{\Delta t}{2} f(t^n, u(t^n))$$

Nous obtenons alors en remplaçant $u(t^n)$ par sa valeur approchée u^n

$$\begin{cases} k_1 \;=\; f(t^n, u^n) \\ k_2 \;=\; f(t^n + \frac{\Delta t}{2}, u^n + \frac{\Delta t}{2}\, k_1) \\ u^{n+1} \;=\; u^n \;+\; \Delta t\, k_2 \end{cases}$$

C'est le **schéma de Runge explicite** qui peut aussi s'écrire sous la forme

$$u^{n+1} \;=\; u^n \;+\; \Delta t\, \phi(t^n, u^n, \Delta t)$$

avec

$$\phi(t^n, u^n, \Delta t) \;=\; f(t^n + \frac{\Delta t}{2}, u^n + \frac{\Delta t}{2} f(t^n, u^n)).$$

Plus généralement, nous définissons les schémas de Runge-Kutta explicites de la manière suivante

Définition 3.1 *Une méthode de Runge-Kutta à s étages est donnée par*

$$\begin{cases} k_1 \;=\; f(t^n, u^n) \\ k_2 \;=\; f(t^n \;+\; c_2\, \Delta t, u^n \;+\; \Delta t\, a_{2,1}\, k_1) \\ \vdots \\ k_s \;=\; f(t^n \;+\; c_s\, \Delta t, u^n \;+\; \Delta t\, (a_{s,1}\, k_1 + \ldots + a_{s,s-1}\, k_{s-1})) \\ u^{n+1} \;=\; u^n \;+\; \Delta t\, (b_1\, k_1 \;+\; b_2\, k_2 \;+\; \ldots \;+\; b_s\, k_s). \end{cases} \tag{3.4}$$

où c_i, $a_{i,j}$ et b_j sont des coefficients. Nous les représentons habituellement par le schéma

$$\begin{array}{c|c} c_i & a_{i,j} \\ \hline & b_j \end{array}$$

Exemple 3.1 *Les méthodes d'Euler et de Runge peuvent donc être représentées par les tableaux suivants :*

0	
1	

0		
1/2	*1/2*	
	0	*1*

Par la suite nous supposons toujours que les c_i satisfont

$$c_1 = 0, \quad c_i = \sum_{j=1}^{i-1} a_{i,j}, \quad i = 2, \ldots, s$$

Ceci signifie que $k_i = f(t^n + c_i\,\Delta t, u(t^n + c_i \Delta t)) \,+\, O(\Delta t^2)$

Nous donnons aussi l'exemple de la méthode de Runge-Kutta d'ordre 4. C'est une méthode excellente pour la plupart des problèmes de Cauchy, en tous cas souvent la première à essayer. Elle est à 4 étages, définie par la fonction $\phi(t, u, \Delta t) = \frac{1}{6}(k_1 \,+\, 2\,k_2 \,+\, 2\,k_3 \,+\, k_4)$ avec :

$$\begin{cases} k_1 = f(t, u) \\ k_2 = f(t + \frac{\Delta t}{2}, u + \frac{\Delta t}{2}\, k_1) \\ k_3 = f(t + \frac{\Delta t}{2}, u + \frac{\Delta t}{2}\, k_2) \\ k_4 = f(t + \Delta t, u + \Delta t\, k_3) \end{cases}$$

Cette méthode combine les formules de Simpson, du trapèze (voir complément du Chapitre 5), et les évaluations de $u(s)$ par la méthode d'Euler. Nous vérifions que son écriture matricielle est la suivante :

0				
1/2	1/2			
1/2	0	1/2		
1	0	0	1	
	1/6	1/3	1/3	1/6

Nous étudierons ici les méthodes de discrétisation des équations différentielles dits "schéma à un pas". Nous définissons un schéma à un pas pour la résolution numérique de (2.1) de la

manière suivante :

$$\begin{cases} u^0 = u(t^0) \\ u^{n+1} \;=\; u^n \;+\; \Delta t\ \phi(t^n, u^n, \Delta t), \end{cases} \tag{3.5}$$

où ϕ est une fonction de $\mathbb{R}^+ \times \mathbb{R}^d \times \mathbb{R}^+$ à valeur dans $\mathbb{R}^d$ et est obtenu en cherchant une approximation de $f(t^n, u(t^n))$. Le schéma numérique est défini par cette fonction ϕ. Bien sûr, nous parlons de méthode à un pas car u^{n+1} ne dépend que de u^n.

Ainsi l'algortithme sera satisfaisant dès lors que l'erreur $e^n = |u^n - u(t^n)|$ converge vers 0 pour tout $n \in \{0, \ldots, N\}$ lorsque le pas de temps Δt tend vers zéro. Pour cela, nous introduisons plusieurs notions : la consistance, la stabilité qui conduiront ensuite à la convergence de l'approximation vers la solution exacte.

3.2 Consistance, stabilité et convergence

Soit $u(t)$ la solution exacte de l'équation différentielle (2.1) et u^n la solution approchée donnée par le schéma à un pas (3.5) ; nous définissons l'erreur globale au temps t^n par la différence entre les solutions exacte et approchée :

$$e^n(\Delta t) \;=\; u(t^n) - u^n, \quad n \in \mathbb{N}$$

Définition 3.2 (Convergence) *Considérons l'équation différentielle ordinaire (2.1). Nous disons que le schéma (3.5) est convergent sur l'intervalle* $[0, T]$ *lorsque nous avons pour* $\Delta t = T/N$

$$\lim_{\Delta t \to 0} \max_{n=0,\ldots,N} \|e^n(\Delta t)\| \;=\; 0.$$

D'autre part, pour $p \in \mathbb{N}$*, le schéma est dit convergent d'ordre* p *s'il existe une constante* $C > 0$ *ne dépendant que de* f, T *,* u_0 *(et surtout pas de* Δt*) tel que*

$$\max_{n=0,\ldots,N} \|e^n(\Delta t)\| \;\leq\; C\, \Delta t^p.$$

Dans la suite, nous établissons des critères sur le schéma numérique permettant de démontrer la convergence de la solution approchée vers la solution exacte. Pour cela, nous introduisons deux notions importantes : la **consistance** (ou cohérence en français) et la **stabilité**.

Pour $n \in \mathbb{N}$, nous définissons l'erreur de "consistance" du schéma (3.5) au temps t, pour une solution u de (2.1) et un pas de temps Δt par :

$$R(t, u, \Delta t) \;=\; \frac{u(t + \Delta t) - u(t)}{\Delta t} \;-\; \phi(t, u(t), \Delta t). \tag{3.6}$$

Nous avons alors la définition suivante

Définition 3.3 (Consistance) *Considérons le schéma à un pas (3.5) associé à l'équation différentielle (2.1). Nous dirons que le schéma est "consistant" lorsque*

$$\lim_{\Delta t \to 0} \|R(t, u(t), \Delta t)\| \ = \ 0,$$

pour tout $t \in \mathbb{R}^+$ et toute solution $u(t) \in \mathbb{R}^d$ de (2.1).

De plus, pour $p \in \mathbb{N}$, le schéma est "consistant d'ordre p" s'il existe une constante $C > 0$ ne dépendant que de f, T, u_0 (et surtout pas de Δt) telle que

$$\|R(t, u, \Delta t)\| \ \leq \ C\, \Delta t^p, \ \textit{pour tout } t \geq 0$$

et toute solution u de (2.1).

Ainsi, la "consistance" nous donne une indication sur la cohérence de notre approximation ϕ. En effet, lorsque nous faisons tendre le paramètre de discrétisation Δt vers zéro dans le schéma numérique, nous devons retrouver la fonction f définissant l'équation différentielle ordinaire (2.1).

Nous proposons maintenant d'établir une condition nécessaire sur ϕ pour que le schéma (3.5) soit consistant.

Proposition 3.1 (Caractérisation de la consistance) *Considérons le schéma à un pas (3.5) associé à l'équation différentielle (2.1). Si la fonction $\phi \in \mathcal{C}(\mathbb{R}^+ \times \mathbb{R}^d \times \mathbb{R}^+, \mathbb{R}^d)$ et si*

$$\phi(t, u, 0) \ = \ f(t, u), \quad t \in [0, T].$$

Alors, le schéma (3.5) est consistant.

Démonstration. Prenons $u \in \mathcal{C}^1([0, T], \mathbb{R}^d)$ la solution exacte de (2.1), nous pouvons écrire que

$$u(t^{n+1}) - u(t^n) = \int_{t^n}^{t^{n+1}} u'(s)\, ds = \int_{t^n}^{t^{n+1}} f(s, u(s))\, ds.$$

Nous en déduisons alors que

$$\begin{aligned} R(t^n, u, \Delta t) \ &= \ \frac{u(t^{n+1}) - u(t^n)}{\Delta t} - \phi(t^n, u(t^n), \Delta t) \\ &= \ \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} \left[f(s, u(s)) - \phi(t^n, u(t^n), \Delta t) \right] ds. \end{aligned}$$

Soit $\varepsilon > 0$, puisque f est une fonction continue et $\phi(t^n, u(t^n), 0) = f(t^n, u(t^n))$, il existe $\eta_1 > 0$ tel que pour tout $\Delta t \ \leq \ \eta_1$

$$\|\phi(t^n, u(t^n), \Delta t) - f(t^n, u(t^n))\| \ \leq \ \frac{\varepsilon}{2}$$

Nous avons donc par inégalité triangulaire

$$\|R(t^n, u, \Delta t)\| \leq \frac{\varepsilon}{2} + \int_{t^n}^{t^{n+1}} \|f(s, u(s)) - f(t^n, u(t^n))\| ds.$$

La fonction $s \to f(s, u(s))$ est continue et donc uniformément continue sur $[t^n, t^{n+1}]$. Il existe donc $\eta_2 > 0$ tel que pour tout $\Delta t \leq \eta_2$, nous avons

$$\frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} \|f(s, u(s)) - f(t^n, u(t^n))\| \, ds \leq \frac{\varepsilon}{2}.$$

Nous avons ainsi montré que si $\Delta t \leq \eta = \min(\eta_1, \eta_2)$, alors

$$\|R(t^n, u, \Delta t)\| \leq \varepsilon,$$

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 u de (2.1) est dans $\mathcal{C}^p(\mathbb{R}^+, \mathbb{R}^d)$.

Définition 3.4 (Stabilité) *Nous disons que le schéma (3.5) est "stable" s'il existe $\Delta t^\star > 0$ et $R > 0$ tels que $u^n \in B(0, R)$ pour tout $n = 0, ..., N$ et pour tout $\Delta t \in [0, \Delta t^\star[$, où $B(0, R)$ désigne la boule de centre 0 et de rayon R.*

Nous disons que le schéma est "inconditionnellement stable" lorsque $\Delta t^\star = \infty$.

Nous donnons aussi une autre notion de stabilité souvent utilisée mais qui ne semble pas être la plus efficace en termes d'analyse d'erreur

Définition 3.5 (Stabilité par rapport aux erreurs) *Nous disons que le schéma (3.5) est "stable" par rapport aux erreurs s'il existe $\Delta t^\star > 0$ et $C > 0$ dépendant de u_0, f et T (mais pas de Δt) tels que si $\Delta t \leq \Delta t^\star$ et si*

$$u^{n+1} = u^n + \Delta t \, \phi(t^n, u^n, \Delta t),$$

et

$$v^{n+1} = v^n + \Delta t \, \phi(t^n, v^n, \Delta t) + \varepsilon^n,$$

pour $n = 0, ..., N-1$, et où $(\varepsilon^n)_{n \in \mathbb{N}} \subset \mathbb{R}^+$ est donnée, alors

$$\|u^n - v^n\| \leq C \left(\|u_0 - v_0\| + \sum_{i=0}^{n-1} \|\varepsilon^i\| \right),$$

pour tout $n = 0, ..., N-1$.

A partir de la notion de consistance et de stabilité, nous pouvons démontrer la convergence de la solution numérique vers la solution exacte de (2.1)

Théorème 3.1 (Consistance + Stabilité $\Leftrightarrow$ Convergence) *Nous supposons que le schéma (3.5) est consistant d'ordre p : il existe une constante $C > 0$ ne dépendant que de f, T, u_0 (et surtout pas de Δt) telle que*

$$\|R(t, u, \Delta t)\| \leq C\,\Delta t^p, \text{ pour tout } t \geq 0.$$

De plus, le schéma (3.5) est "stable" par rapport aux erreurs.

Alors, la solution numérique fournie par le schéma converge vers la solution exacte de (2.1). De plus, l'erreur vérifie l'estimation

$$\|e^n(\Delta t)\| \leq C\left[C\,T\,\Delta t^p + \|e^0(\Delta t)\|\right],$$

pour tout $n = 0, ..., N$.

Démonstration. Puisque le schéma est consistant, nous avons pour la solution exacte $u(t^n)$

$$u(t^{n+1}) = u(t^n) + \Delta t\,[\phi(t^n, u(t^n), \Delta t) + R(t^n, u, \Delta t)],$$

avec la condition suivante sur $R(t^n, u, \Delta t)$

$$\lim_{\Delta t \to 0} \|R(t, u(t), \Delta t)\| = 0,$$

et si le schéma est d'ordre p

$$\|R(t, u(t), \Delta t)\| \leq C\,\Delta t^p, \text{ pour tout } t \geq 0$$

D'autre part, la solution numérique est donnée par

$$u^{n+1} - u^n = \Delta t\,\phi(t^n, u^n, \Delta t).$$

En faisant la différence entre les deux égalités, nous devons estimer le terme $e^n(\Delta t) = u^n - u(t^n)$

$$e^{n+1}(\Delta t) = e^n(\Delta t) + \Delta t\,[\,\phi(t^n, u^n, \Delta t) - \phi(t^n, u(t^n), \Delta t) - R(t^n, u(t^n), \Delta t)\,].$$

Ainsi en appliquant directement la définition de la stabilité par rapport aux erreurs, nous obtenons

$$\|e^n(\Delta t)\| \leq C\left(\|e^0(\Delta t)\| + \Delta t \sum_{i=0}^{n-1} \|R(t^i, u(t^i), \Delta t)\|\right).$$

Puisque le schéma est consistant ; nous avons

$$\Delta t \sum_{i=0}^{n-1} \|R(t^i, u(t^i), \Delta t)\| \leq T \max_{0 \leq i \leq n} \|R(t^i, u(t^i), \Delta t)\|,$$

le terme de droite tend bien vers zéro lorsque Δt converge vers zéro, le schéma est donc convergent. De plus, si le schéma est consistant d'ordre p, l'erreur est du même ordre

$$\|e^n(\Delta t)\| \leq C \left[\|e^0(\Delta t)\| + \Delta t \sum_{i=0}^{n-1} \|R(t^i, u(t^i), \Delta t)\| \right] \leq C \left[\|e^0(\Delta t)\| + C\,T\,\Delta t^p \right].$$

□

Nous proposons aussi l'énoncé suivant, dont la démonstration est relativement simple.

Théorème 3.2 (Convergence) *Nous supposons que le schéma (3.5) est consistant d'ordre p : il existe une constante $C > 0$ ne dépendant que de f, T, u_0 (et surtout pas de Δt) telle que*

$$\|R(t, u, \Delta t)\| \leq C\,\Delta t^p, \text{ pour tout } t \geq 0.$$

De plus, la fonction ϕ est continue et Lipschitzienne par rapport à la variable $u \in \mathbb{R}^d$

$$\|\phi(t, u, \Delta t) - \phi(t, v, \Delta t)\| \leq \Gamma \|u - v\|$$

pour tout $t \geq 0$ et $\Delta t > 0$. Alors, la solution numérique fournie par le schéma converge vers la solution exacte de (2.1). De plus, l'erreur vérifie l'estimation

$$\|e^n(\Delta t)\| \leq \frac{C}{\Gamma} \left(\exp(\Gamma\, t^n) - 1\right) \Delta t^p + \|e^0(\Delta t)\| \exp(\Gamma\, t^n),$$

pour tout $n = 0, ..., N$.

Démonstration. Soit $u \in \mathcal{C}^2([0, T], \mathbb{R}^d)$ une solution de (2.1), en intégrant sur l'intervalle $[t^n, t^{n+1}]$, il vient d'une part

$$u(t^{n+1}) - u(t^n) = \int_{t^n}^{t^{n+1}} f(s, u(s))\, ds.$$

D'autre part, la solution numérique est donnée par

$$u^{n+1} - u^n = \Delta t\, \phi(t^n, u^n, \Delta t).$$

En faisant la différence entre ces deux dernières égalités, nous devons estimer le terme

$$\begin{aligned} \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} f(s, u(s))\, ds - \phi(t^n, u^n, \Delta t) &= \int_{t^n}^{t^{n+1}} f(s, u(s))\, ds - \phi(t^n, u(t^n), \Delta t) \\ &+ \phi(t^n, u(t^n), \Delta t) - \phi(t^n, u^n, \Delta t). \end{aligned}$$

En utilisant la définition de la consistance à l'ordre p, nous avons

$$\left\| \int_{t^n}^{t^{n+1}} f(s, u(s))\, ds - \phi(t^n, u(t^n), \Delta t) \right\| \leq C\,\Delta t^p.$$

D'autre part, puisque la fonction ϕ est Lipschtzienne par rapport à la variable y, il vient

$$\|\phi(t^n, u(t^n), \Delta t) - \phi(t^n, u^n, \Delta t)\| \ \leq \ \Gamma \, \|u(t^n) - u^n\|.$$

Nous obtenons alors en regroupant les termes

$$\|e^{n+1}(\Delta t)\| \leq \ \|e^n(\Delta t)\| + C \, \Delta t^{p+1} \ + \ \Gamma \, \|e^n(\Delta t)\| \, \Delta t.$$

Puis en remarquant que pour $x \in I\!R^+$, $1 + x \leq \exp(x)$; nous avons

$$\|e^{n+1}(\Delta t)\| \ \leq \ exp\,(\Gamma \, \Delta t) \ \|e^n(\Delta t)\| \ + \ C \, \Delta t^{p+1}.$$

En ré-itérant le procédé par récurrence, nous avons alors pour tout $n \geq 0$

$$\begin{array}{rcl} \|e^n(\Delta t)\| & \leq & \exp\,(\Gamma \, n \, \Delta t) \ \|e^0(\Delta t)\| \\ & + & C \, \Delta t^p \, (\Delta t + \Delta t \, \exp\,(\Gamma \, \Delta t) + \ldots + \Delta t \, \exp\,(\Gamma \, (n-1)\Delta t)) \, . \end{array}$$

Nous observons alors que sur chaque intervalle $[t^n, t^{n+1}]$

$$\Delta t \, \exp\,(\Gamma \, t^n) \ \leq \ \int_{t^n}^{t^{n+1}} \exp\,(\Gamma \, s) \ ds$$

nous en déduisons que

$$\begin{array}{rcl} & & C \ (\Delta t + \Delta t \, \exp\,(\Gamma \, \Delta t) + \ldots + \Delta t \, \exp\,(\Gamma \, (n-1)\Delta t)) \ \Delta t^p \\ & \leq & C \, \Delta t^p \displaystyle\int_0^{t^n} \exp\,(\Gamma \, s) \, ds \\ & = & \dfrac{C}{\Gamma} \ (\exp\,(\Gamma \, t^n) \ - \ 1) \ \Delta t^p. \end{array}$$

Nous avons alors le résultat l'erreur converge vers zéro lorsque Δt tend vers zéro et vérifie

$$\|e^n(\Delta t)\| \ \leq \ \exp\,(\Gamma \, t^n) \ \|e^0(\Delta t)\| \ + \ \frac{C}{\Gamma} \ (\exp\,(\Gamma \, t^n) \ - \ 1) \ \Delta t^p.$$

□

4 Schémas à un pas implicites

Reprenons la démarche utilisée pour construire les schémas à un pas : sur l'intervalle $[t^n, t^n + \Delta t]$, nous devons approcher l'intégrale de droite

$$u(t^{n+1}) \ - \ u(t^n) \ = \ \int_{t^n}^{t^{n+1}} f(s, u(s)) \, ds.$$

Pour mettre au point le schéma d'Euler explicite nous avons utilisé la valeur de $f(t^n, u(t^n))$ mais cette fois-ci prenons la valeur à droite $f(t^{n+1}, u(t^{n+1}))$, il vient alors

$$\int_{t^n}^{t^{n+1}} f(s, u(s))\, ds \simeq \Delta t\, f(t^{n+1}, u(t^{n+1}));$$

nous obtenons le **schéma d'Euler implicite** en remplaçant $u(t^n)$ par son approximation u^n, il vient alors

$$\begin{cases} u^0 = u(0) \\ \\ u^{n+1} = u^n \,+\, \Delta t\, f(t^{n+1}, u^{n+1}), \qquad \text{pour } n = 0, \ldots \end{cases} \tag{4.7}$$

Nous remarquons que dans le schéma d'Euler implicite, le calcul de u^{n+1} n'est hélàs pas explicite, il est donné de manière implicite et nous devons pour le calculer résoudre un problème non linéaire, ce qui n'est jamais très facile. D'ailleurs avant le calcul de u^{n+1}, la première question à se poser pour ce type de schéma est l'existence d'une solution u^{n+1}. Nous montrerons au Théorème 4.1 que si l'hypothèse suivante est vérifiée :

$$\|f(t, x) - f(t, y)\| \;\leq\; L\, \|x \,-\, y\|.$$

alors la valeur u^{n+1} fournie par le schéma (4.7) est bien définie en fonction de u^n, t^n, et Δt. Nous pouvons donc bien écrire le schéma (4.7) sous la forme (3.5)

$$u^{n+1} \;=\; u^n \,+\, \Delta t\; \phi(t^n, u^n, \Delta t),$$

et cela bien que la fonction ϕ ne soit définie ici qu'implicitement et non explicitement. En effet, elle vérifie

$$\phi(t^n, u^n, \Delta t) = f(t^{n+1}, v)$$

où v est l'unique solution de

$$v = u^n + \Delta t\, f(t^{n+1}, v).$$

Nous avons le résultat d'existence suivant

Théorème 4.1 *Si $f : I\!R^+ \times I\!R^d \to I\!R^d$ est une fonction continue et Lipschitzienne en $x \in I\!R^d$, c'est-à-dire il existe $L > 0$ tel que $\forall\, (x, y) \in I\!R^d \times I\!R^d$*

$$\|f(t, x) - f(t, y)\| \;\leq\; L\, \|x \,-\, y\|.$$

Alors, il existe une unique solution u^{n+1} fournie par le schéma (4.7) (schéma d'Euler implicite) dès lors que le pas de temps vérifie $\Delta t < 1/L$

Démonstration. Nous définissons l'application $g(z)$

$$z \in I\!R^d \longrightarrow\; g(z) \;=\; u^u + \Delta t\, f(t^n, z)$$

D'une part, l'espace $I\!R^d$ est bien un espace vectoriel normé complet et lorsque nous prenons Δt suffisament petit, l'application g est une contraction : pour tout $(x, y) \in I\!R^d$

$$\|g(x) - g(y)\| \ \leq \ L\,\Delta t \|x - y\|$$

L'application g qui est continue admet donc un unique point fixe, que nous notons $z = u^{n+1}$; ce qui prouve le théorème. □

Ce schéma entre donc bien dans le cadre des schémas à un pas (3.5) que nous avons précisemment étudiés dans la partie précédente. Néanmoins, une propriété supplémentaire dite de stabilité inconditionnelle, est vérifiée par ce schéma.

5 Equations différentielles raides

Considérons l'exemple suivant d'une équation différentielle, modélisant une réaction chimique

$$\begin{array}{rcll} A & \stackrel{\varepsilon}{\longrightarrow} & B & \text{lente} \\ B + B & \stackrel{1/\varepsilon^2}{\longrightarrow} & C + B & \text{très rapide} \\ B + C & \stackrel{1/\varepsilon}{\longrightarrow} & C + A & \text{rapide} \end{array}$$

ce qui donne un système d'équations différentielles raides

$$\left\{ \begin{array}{rcl} u_1'(t) & = & -u_1(t) \ + \ \frac{1}{\varepsilon}\, u_2(t)\, u_3(t) \\ u_2'(t) & = & +u_1(t) \ - \ \frac{1}{\varepsilon}\, u_2(t)\, u_3(t) \ - \ \frac{1}{\varepsilon^2}\ u_2^2(t) \\ u_3'(t) & = & +\frac{1}{\varepsilon^2}\ u_2^2(t) \end{array} \right.$$

Pour un tel système, l'utilisation d'une méthode de Runge-Kutta explicite est pratiquement impossible car nous serions obligé de prendre un pas de temps Δt très petits pour obtenir une approximation raisonnable. Une caractéristique de cette équation différentielle est que la solution cherchée est très lisse et les autres solutions s'approchent rapidement de celle-ci. Pour mieux comprendre le phénomène qui se produit dans cet exemple, considérons un problème beaucoup plus simple

$$\varepsilon\, u'(t) \ = \ -u(t) \ + \ \cos(t), \quad 0 < \varepsilon \ll 1$$

qui possède les mêmes caractéristiques.

Cette équation différentielle est linéaire inhomogène. Cherchons une solution particulière de la forme

$$u(t) \ = \ A\, \cos(t) \ + \ B\, \sin(t)$$

En introduisant cette fonction dans l'équation différentielle nous obtenons

$$-\varepsilon\, A\, \sin(t) \,+\, \varepsilon\, B\, \cos(t) \,=\, -A\, \cos(t) \,-\, B\, \sin(t) \,+\, \cos(t)$$

une comparaison des coefficients donne

$$A = 1/(1+\varepsilon^2), \quad B = -\varepsilon/(1+\varepsilon^2).$$

Comme la solution générale de cette équation différentielle est la somme de la solution générale de l'équation homogène et d'une solution particulière, nous obtenons

$$u(t) \,=\, \left(u_0 - \frac{1}{1+\varepsilon^2}\right)\, e^{-t/\varepsilon} \,+\, \frac{1}{1+\varepsilon^2}\, \cos(t) \,+\, \frac{\varepsilon}{1+\varepsilon^2}\, \sin(t).$$

Observons maintenant ce qu'il se passe lorsque nous appliquons la méthode d'Euler explicite à une telle équation

$$u^{n+1} - u^n = \Delta t\, f(t^n, u^n)$$

Nous appliquons ce schéma au problème actuel avec un pas Δt constant, ce qui donne avec $t^n \,=\, n\, \Delta t$

$$u^{n+1} \,=\, (1 - \frac{\Delta t}{\varepsilon})\, u^n \,+\, \frac{\Delta t}{\varepsilon}\, \cos(t^n).$$

Ceci est une équation aux différences finies, linéaire et inhomogène. La solution est obtenue comme pour une équation différentielle. Nous cherchons d'abord une solution particulière de la forme

$$u^n \,=\, A \cos(t^n) \,+\, B\, \sin(t^n).$$

Comme dans le cas continue, nous substituons u^n dans le schéma d'Euler explicite et, en utilisant $t^{n+1} = t^n + \Delta t$ et les propriétés de $\sin$ et $\cos$ nous obtenons ainsi

$$\begin{aligned}
&A\, (\cos(t^n)\, \cos(\Delta t) \,-\, \sin(t^n)\, \sin(\Delta t)\,) \,+ \\
&B\, (\sin(t^n)\, \cos(\Delta t) \,+\, \cos(t^n)\, \sin(\Delta t)\,) \\
&= \left(1 - \frac{\Delta t}{\varepsilon}\right)\, (A\, \cos(t^n) \,+\, B\, \sin(t^n)) \,+\, \frac{\Delta t}{\varepsilon}\, \cos(t^n).
\end{aligned}$$

En comparant les coefficients de $\cos(t^n)$ et $\sin(t^n)$, nous obtenons deux équations linéaires pour A et B dont la solution est

$$A \,=\, 1 \,+\, O(\varepsilon\, \Delta t), \quad B \,=\, \varepsilon \,+\, O(\varepsilon \Delta t^2).$$

En ajoutant la solution générale de l'équation homogène à la solution particulière, nous obtenons une suite de la forme

$$u^n = \left(1 - \frac{\Delta t}{\varepsilon}\right)^n C + \cos(t^n) + \varepsilon \sin(t^n) + O(\varepsilon\, \Delta t).$$

Nous observons que la solution numérique u^n est proche de la solution exacte seulement lorsque

$$|1 - \Delta t/\varepsilon| \; < \; 1,$$

c'est-à-dire si $\Delta t < 2\,\varepsilon$. Ainsi, lorsque ε est très petit une telle restriction est inacceptable.

Nous envisageons alors l'utilisation d'une méthode d'Euler implicite, le même calcul donne alors

$$\left(1 + \frac{\Delta t}{\varepsilon}\right) u^{n+1} = u^n + \frac{\Delta t}{\varepsilon} \cos(t^{n+1}),$$

dont la solution peut être écrite sous la forme

$$u^n = \left(1 + \frac{\Delta t}{\varepsilon}\right)^{-n} \; C + \cos(t^n) + \varepsilon \, \sin(t^n) \; + \; O(\Delta t\, \varepsilon).$$

Cette fois-ci nous n'avons pas de restriction sur la longueur du pas, car $|(1 + \varepsilon\, \Delta t)|^{-1} < 1$ pour tout $\Delta t \; > \; 0$. Dans ce cas, le schéma implicite donne une bonne approximation même si Δt est très grand. Le calcul précédent a montré que ce n'est pas la solution particulière qui pose des difficultés à la méthode explicite, mais c'est l'approximation de la solution de l'équation homogène $\varepsilon\, u' = -u$. Nous considérons donc le problème un peu plus général

$$u' = \lambda\, u$$

dont la solution exacte est $u(t) = C\, e^{\lambda\, t}$ et elle reste bornée pour $t \geq 0$ dès lors que $Re(\lambda) \leq 0$. La solution numérique d'une méthode de Runge-Kutta appliquée avec un pas de temps Δt constant au problème actuel, ne dépend que du produit $\lambda\, \Delta t$. Il est alors intéressant d'étudier pour quelle valeur de $\lambda\, \Delta t$ la solution numérique reste bornée.

Définition 5.1 (A-stabilité) *Considérons une méthode dont la solution numérique pour l'équation (9.7) est une fonction de $z = \lambda \Delta t$. Alors, l'ensemble :*

$$S := \{z \in \mathbb{C}; \quad (u^n)_{n \in \mathbb{N}} \textit{ est borné}\}$$

s'appelle domaine de stabilité de la méthode. Nous disons que la méthode est A-stable si

$$\mathbb{C}^- \subset S, \quad \textit{où} \quad \mathbb{C}^- = \{z \in \mathbb{C}, \quad Re(z) \leq 0\}.$$

Pour la méthode d'Euler explicite le domaine de stabilité est :

$$S = \{z; \quad |1 + z| \leq 1\},$$

le disque de rayon 1 et de centre -1. Pour la méthode d'Euler implicite il est :

$$S = \{z; \quad |z - 1| \geq 1\},$$

c'est l'extérieur du disque de rayon 1 et de centre $+1$. Seulement, la méthode implicite est A-stable.

Pour une méthode de Runge-Kutta, la solution numérique est de la forme

$$u^{n+1} = R(\lambda\,\Delta t)\,u^n,$$

où la fonction $R(z)$ s'appelle fonction de stabilité et le domaine de stabilité est alors

$$S \;=\; \{z \in \mathbb{C};\quad |R(z)| \leq 1\}.$$

Nous pouvons vérifier que la fonction $R(z)$ est en fait un polynôme, l'ensemble S est borné et la condition de stabilité $\lambda\,\Delta t \;\in S$ impose une restriction sévère à Δt. Ces méthodes ne sont donc pas recommandées pour la résolution des équations différentielles raides.

Bien sûr, nous pouvons appliquer le schéma d'Euler implicite mais celui-ci est seulement d'ordre un ; nous proposons alors des méthode d Runge-Kutta implicites.

Comme pour la dérivation des méthodes de Runge-Kutta explicites, nous partons de la formule intégrée

$$u(t^n + \Delta t) \;=\; u(t^n) \;+\; \int_{t^n}^{t^n+\Delta t} f(s, u(s))\,ds$$

de l'équation différentielle. Nous appliquons une formule de quadrature avec $c_s = 1$ dans la formule de Runge-Kutta (3.4) ayant l'ordre maximal $2s-1$. Par exemple, pour $s = 2$ nous avons

$$\begin{aligned} u(t^n + \Delta t) &= u(t^n) \;+\; \frac{\Delta t}{4}\left(3f(t^n + \frac{\Delta t}{3}, u(t^n + \frac{\Delta t}{3})) + f(t^n + \Delta t, u(t^n + \Delta t))\right) \\ &+ \;\; O(\Delta t^4). \end{aligned}$$

Pour approcher la valeur $u(t^n + \frac{\Delta t}{3})$ nous intégrons l'équation différentielle de t^n à $t^n + \frac{\Delta t}{3}$ et nous appliquons une formule de quadrature qui utilise les mêmes évaluations de f que la formule précédente :

$$\begin{aligned} u(t^n + \frac{\Delta t}{3}) &= u(t^n) \;+\; \frac{\Delta t}{12}\left(5f(t^n + \frac{\Delta t}{3}, u(t^n + \frac{\Delta t}{3})) - f(t^n + \Delta t, u(t^n + \Delta t))\right) \\ &+ \;\; O(\Delta t^3). \end{aligned}$$

En supprimant les termes du reste et en notant k_1 et k_2 les deux évaluations de f, nous arrivons à

$$\left\{\begin{array}{l} k_1 = f(t^n + \frac{\Delta t}{3}, u^n + \frac{\Delta t}{12}(5k_1 - k_2)) \\ \\ k_2 = f(t^n + \Delta t, u^n + \frac{\Delta t}{4}(3k_1 + k_2)) \\ \\ u^{n+1} = u^n + \frac{\Delta t}{4}\,(3k_2 + k_2) \end{array}\right.$$

qui est une méthode de Runge-Kutta. Les deux premières équations constituent un système non linéaire pour k_1 et k_2, qu'il faut résoudre avec les techniques du Chapitre3 (méthode de Newton).

Lemme 5.1 *La méthode de Runge-Kutta est d'ordre trois et elle est A-stable.*

Démonstration. L'ordre trois est une conséquence des formules de construction. Il suffit que la deuxième formule soit d'ordre deux car le terme correspondant dans la première est mulitplié par Δt. Il reste donc à démontrer la A-stabilité. Pour cela, nous appliquons la méthode à l'équation $u' = \lambda u$ et nous obtenons avec $z = \lambda\,\Delta t$

$$\Delta t\,k_1 \;=\; z\,u^n \;+\; \frac{z}{12}\,(5\,\Delta t\,k_1 \;-\; \Delta t\,k_2)\,, \quad \Delta t\,k_2 \;=\; z\,u^n \;+\; \frac{z}{4}\,(3\,\Delta t\,k_1 \;+\; \Delta t\,k_2)\,.$$

Nous résolvons ce système linéaire pour $k_1\,\Delta t$ et $k_2\,\Delta t$, et introduisons la solution dans la troisi'eme formule. Ceci donne $u^{n+1} \;=\; R(z)\,u^n$ avec comme fonction de stabilité $R(z)$

$$R(z) \;=\; \frac{P(z)}{Q(z)} \;=\; \frac{1 \;+\; z/3}{1 - 2\,z/3 \;+\; z^2/6}$$

Sur l'axe imaginaire, nous avons

$$|Q(i\,y)|^2 \;-\; |P(i\,y)|^2 \;=\; \left(1 \;-\; \frac{y^2}{6}\right)^2 \;+\; \frac{4}{9}y^2 \;-\; \left(1 \;+\; \frac{y^2}{9}\right) \;=\; \frac{y^4}{36} \;\geq\; 0.$$

Ceci implique $|Q(i\,y)| \;\geq\; |P(i\,y)|$ et aussi $|R(i\,y)| \;\leq\; 1$. Les singularités de $R(z)$ (les zéros de $Q(z)$) sont $z_\pm \;=\; 2 \pm i\sqrt{2}$ dans le demi-plan droit. Donc $R(z)$ est analytique dans le demi-plan gauche et, par le principe du maximum, $R(z)$ est majoré par un pour $Re(z) \leq 0$. □

Cette contruction peut être généralisée pour obtenir des méthodes de Runge-Kutta implicites qui sont A-stables et d'ordre $2s - 1$.

Après la publication remarquable de Dahlquist en 1963, la recherche sur la résolution des équations différentielles raides a rapidement pris une place importante en analyse numérique. Des nouvelles méthodes d'intégration, des nouvelles théories (par exemple, étoiles d'ordre) et des programmes informatiques performants ont été développé. Plus de détails peuvent être trouvés dans [10, 11].

6 Une incursion dans les schémas multi-pas

Pour construire un schéma multi-pas, notre point de départ est toujours la relation

$$u(t^{n+1}) \;-\; u(t^n) \;=\; \int_{t^n}^{t^{n+1}} f(s, u(s))\,ds.$$

Il s'agit ensuite de remplacer la fonction $f(s, u(s))$ par un polynôme $p(s)$ qui sera facile à intégrer. Pour cela, nous construisons un polynôme d'interpolation de la fonction $s \;\longrightarrow\; f(s, u(s))$ aux points t^j, pour $j = n - k + 1, \ldots, n$. Nous obtenons alors

$$u(t^{n+1}) \;-\; u(t^n) \;\simeq\; \int_{t^n}^{t^{n+1}} p(s)\,ds.$$

Prenons par exemple $p_0(t)$ le polynôme de degré zéro interpolant f au point t^n sur l'intervalle $[t^n, t^{n+1})$, il vient alors le schéma d'Euler explicite à un pas

$$u^{n+1} \;=\; u^n \;+\; \Delta t\, f(t^n, u^n).$$

Ensuite pour obtenir une meilleure approximation sur l'intervalle $[t^n, t^{n+1})$, nous calculons l'interpolé de f aux points t^{n-1} et t^n ; nous avons

$$p_1(t) = f(t^{n-1}, u^{n-1}) \;+\; (t - t^{n-1})\frac{f(t^n, u^n) - f(t^{n-1}, u^{n-1}}{\Delta t}$$

et donc le schéma d'ordre deux d'Adams-Balsforth

$$u^{n+1} \;=\; u^n \;+\; \frac{\Delta t}{2}\,\left(3\, f(t^n, u^n) - f(t^{n-1}, u^{n-1})\right).$$

Bien sûr rien ne nous empêche d'utiliser également le point t^{n+1} ce qui permet de construire des schémas d'Adams-Balsforth implicites. Par exemple à l'ordre deux

$$u^{n+1} \;=\; u^n \;+\; \frac{\Delta t}{2}\,\left(f(t^n, u^n) + f(t^{n+1}, u^{n+1})\right).$$

7 Complément du Chapitre 6

7.1 Tracer un cercle en approchant une EDO

Pour tracer un cercle $C(0,1)$ de centre $(0,0)$ et de rayon $r = 1$, nous pouvons tracer les courbes paramétrées $(x(t) = \cos(t), y(t) = \sin(t))$, avec $t \in [0, 2\pi]$ en utilisant un grand nombre de point t^n, $n = 0, \ldots, N$.

À partir des équations paramétrées de $(x(t), y())$, nous pouvons obtenir un cercle en résolvant le système différentiel

$$\begin{cases} x'(t) = -y(t), \quad x(0) = 1, \\ \\ y'(t) = x(t), \quad y(0) = 0. \end{cases} \tag{7.8}$$

Un schéma d'Euler explicite. Nous posons alors $h = 2\pi/N$. En appliquant le schéma d'Euler explicite (3.3) au système (7.8), cela nous conduit à calculer les points $P^n = (x^n, y^n)$ de coordonnées

$$\begin{bmatrix} x^n \\ \\ y^n \end{bmatrix} \;=\; A^n \;\begin{bmatrix} 1 \\ \\ 0 \end{bmatrix}, \quad \text{avec } A = \begin{bmatrix} 1 & -h \\ \\ h & 1 \end{bmatrix} \tag{7.9}$$

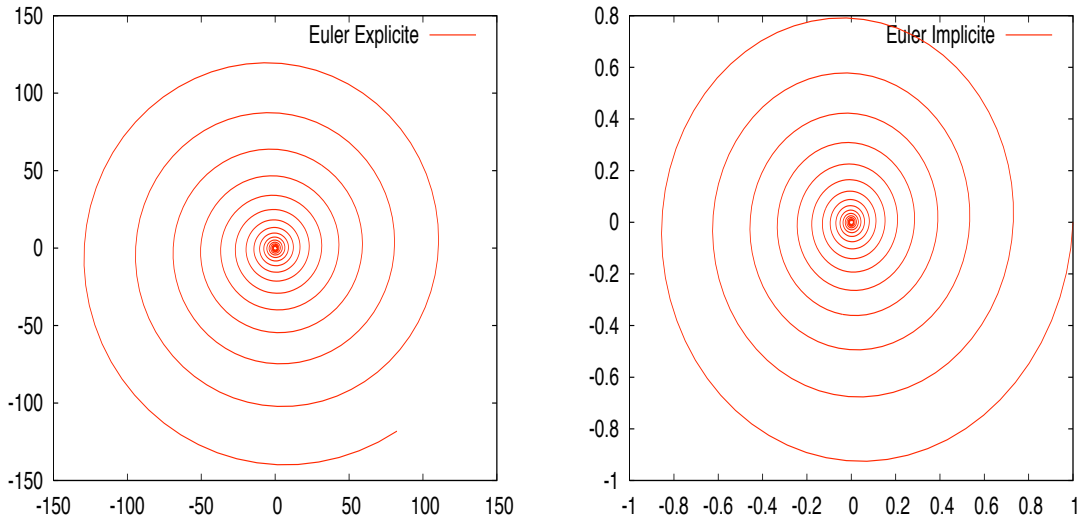

FIG. 6.2 – Evolution de la solution (x^n, y^n) en fonction de n pour un schéma d'Euler explicite et implicite

Par récurrence, nous calculons

$$A \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ h \end{bmatrix}$$

Puis

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

En effectuant le produit scalaire de ce système algébrique avec le vecteur (x^n, y^n), nous obtenons

$$|x^n|^2 + |y^n|^2 = (1+h^2)^n$$

Nous en déduisons alors que les points P^n vérifient $|x^n|^2+|y^n|^2 = (1+h^2)^n$. Ce qui signifie que les points P^n sont sur le cercle $C(0, r^n)$ avec $r = (1+h^2)$.

Un schéma d'Euler implicite. Dans un deuxième temps, nous appliquons le schéma d'Euler implicite (4.7) au système (7.8). Ceci conduit à calculer cette fois-ci les points $Q^n = (x^n, y^n)$ dont les coordonnées vérifient $|x^n|^2 + |y^n|^2 = 1/(1+h^2)^n$.

Une combinaison des schémas explicite et implicite. Soit $f : \mathbb{R}^+ \times \mathbb{R}^2 \to \mathbb{R}^2$ une fonction aussi régulière que nous le souhaitons. Nous pouvons définir un nouveau schéma implicite pour

résoudre le problème

$$u'(t) = f(t, u(t)), \quad u(0) = u_0,$$

en posant

$$u^{n+1} = u^n + \frac{h}{2}\left(f(t^n, u^n) + f(t^{n+1}, u^{n+1})\right)$$

Nous montrons comme nous l'avons fait pour le schéma d'euler implicite que ce dernier schéma peut écrire ce schéma comme un schéma à un pas

$$u^{n+1} \;=\; u^n \;+\; h\,\phi(t^n, u^n, h),$$

où u^{n+1} est bien défini et de manière unique. Nous pouvons alors appliquer directement la Proposition 3.1 pour montrer que ce schéma est consistant puis en appliquant le Théorème 3.2, la méthode est bien convergente.

En utilisant la définition de la consitance, nous pouvons même démontrer que ce schéma est d'ordre deux, c'est-à-dire

$$\|R(t^n, h)\| = \left\| \frac{u(t^n + h) - u(t^n)}{h} - \frac{1}{2}\left(f(t^n, u(t^n)) + f(t^n + h, u(t^n + h))\right) \right\| \leq Ch^2$$

Pour cela, nous utilisons la formule des trapèzes et vérifions qu'elle est bien d'ordre trois : pour une fonction g aussi régulière que nous le souhaitons

$$\left\| \int_a^b g(x)dx - \frac{b-a}{2}\left(g(a) + g(b)\right) \right\| \leq C(b-a)^3.$$

Finalement, cette méthode conduit à calculer des points $M^n = (x^n, y^n)$ dont les coordonnées vérifient

$$\begin{bmatrix} x^n \\ \\ y^n \end{bmatrix} \;=\; \frac{1}{(1 + (h/2)^2)^n} A^{2n} \begin{bmatrix} 1 \\ \\ 0 \end{bmatrix}, \tag{7.10}$$

avec

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

et de plus $|x^n|^2 + |y^n|^2 = 1$.

7.2 Vers le système de Lotka-Volterra

Les modèles de Lotka-Volterra sont des modèles généraux qui décrivent la croissance de deux populations (animales, espèces chimiques, etc) en interaction.

Avant de décrire les intéractions de populations, intéressons nous d'abord à des modèles de croissance de population plus simples.

Les premiers modèles de dynamique des populations. Nous considérons une population de $N(t)$ individus au temps t et souhaitons décrire l'évolution de cette population. Pour cela nous adoptons une démarche de modélisation en plusieurs étapes :

- Hypothèses sur la reproduction en se donnant un taux de reproduction ou de croissance et la prédation en se donnant un taux de mortalité.
- Mise en équations et autant que possible étude mathématique du modèle proposé
- Simulations numériques
- Discussions du modèle en fonction des résultats obtenus et propositions d'amélorations en prenant en compte de nouveaux phénomènes.

Pour commencer nous proposons d'étudier l'évolution de la population $N(t)$ en considérant seulement les naissances et les décés. Ce modèle fût proposé par L. Euler

(H1) Le nombre de naissances est proportionnel à la population : Naissances $= a\,N(t)$.

(H2) Le nombre de décés est également proportionnel à la population : Décés $= bN(t)$.

Nous obtenons ainsi une équation différentielle ordinaire pour la population $N(t)$

$$\begin{cases} \dfrac{dN}{dt}(t) = (a-b)N(t), \\ N(0) = N_0. \end{cases}$$

Ce modèle est très simple et peut d'ailleurs être résolu de manière explicite

$$N(t) = N_0\, \exp((a-b)t).$$

D'une part, nous en déduisons que

- lorsque $a > b$, la natalité est plus importante que la mortalité et la population se met donc à croître de manière exponentielle au cours du temps et tend vers l'infini lorsque le temps part à l'infini,
- lorsque $a < b$, la natalité est plus faible que la mortalité, la population de met alors à décroître de manière exponentielle au cours du temps, la population tend alors vers zéro lorsque t tend vers l'infini

D'autre part, nous constatons qu'à partir d'hypothèses volontairement simplificatrices, nous pouvons mettre au point un modèle simple à résoudre. Ce modèle reste valable sur des temps courts (t proche de zéro) tnt que la population totale rese peu élevée mais lorsque la population devient trop importante, elle atteint un seuil qu'elle ne peut pas dépasser. En effet, les ressources comme la nourriture et les conditions de vie en général se dégradent et la population se stabilise. Ce modèle là n'en tient pas compte.

Il faut en effet introduire une hypothèse supplémentaire comme les limitations dues au milieu ambiant. Pour cela, nous pouvons par exemple faire dépendre les taux de croissance a et de mortalité b de la population $N(t)$. Pour cela, nous formulons une hypothèse supplémentaire dans la mise en équations, ce qui permettra de décrire un comportement plus réaliste.

(H3) La population est limitée dans un milieu fini.

Nous obtenons alors le modèle logistique dû à Verhulst (1836). Dans ce cas nous tenons compte de l'hypothèse de "milieu limité", c'est-à-dire que le milieu peut nourrir au maximum K individus. Nous avons alors différentes situations :

- lorsque $N(t) < K$, il y a suffisamment de ressources pour permettre à la population de croître car la natalité est supérieure à la mortalité des individus,
- lorsque $N(t) > K$, il n'y a pas assez de ressources pour que la population puisse augmenter et des individus meurent de faim. Alors, la mortalité devient supérieure à la natalité,
- lorsque $N(t)$ est très petit par rapport à K, nous sommes dans le cas du modèle d'Euler et la croissance est proportionnelle à la population $N(t)$.

À partir de ces trois hypothèses, nous pouvons écrire une nouvelle équation pour l'évolution de la population $N(t)$

$$\begin{cases} \dfrac{dN}{dt}(t) = f(N(t)), \\ N(0) = N_0 \end{cases}$$

où

$$\begin{cases} f(N) > 0 & \text{si } N < K, \\ f(N) < 0 & \text{si } N > K, \\ f(N) \simeq c\,N & \text{si } N \text{ est petit par rapport } K \end{cases}$$

De plus, il n'y a pas de création spontanée d'individus, c'est-à-dire que $f(0) = 0$.

Bien sûr, nous avons plusieurs choix pour la fonction f, la plus simple qui satisfait ces hypothèses est

$$f(N) = r\,N\,(1 - N/K) \tag{7.11}$$

Parfois la fonction $f(N)$ est donnée par des mesures expérimentales. Aussi, lorsque nous connaissons bien la dynamique de reproduction/mort, nous pouvons en déduire une fonction f explicite, mais il faut pour cela des hypothèses supplémentaires. La fonction f proposée est la plus simple qui fonctionne : la constante K représente la capacité du milieu, c'est-à-dire le nombre d'individus qu'il peut nourrir.

Notion de stabilité. Dans le cas de la dernière EDO avec f donnée par (7.11), nous pouvons trouver des solutions explicites... Cependant, dans le cas général les calculs ne sont plus explicites et nous ne connaissons pas la solution exacte. À défaut, nous nous interessons aux propriétés qualitatives de la solution de cette équation, c'est-à-dire nous étudions par exemple

- les **états d'équilibres** : N^* tel que

$$f(N^*) = 0,$$

ce qui nous informe sur le comportement en temps long de la solution.

– la **stabilité** des états d'équilibres :
  – un équilibre est dit stable, lorsqu'une petite perturbation de l'équilibre, n'induit pas un changement important de la solution et le système revient à la position d'équilibre N^*.
  – un équilibre est dit instable, lorsqu'une petite perturbation déstabilise le système et la solution s'éloigne de l'équilibre N^*.

Nous notons $N(t)$ la solution obtenue à partir d'une perturbation $N^* + u_0$ de la donnée initiale correspondant à l'état d'équilibre N^*. Nous avons alors

$$\frac{dN}{dt}(t) = f(N(t)), \quad \text{avec} \quad N(0) = N^* + u_0$$

et

$$0 = \frac{dN^*}{dt} = f(N^*),$$

d'où en introduisant $u(t) = N(t) - N^*$, nous avons

$$\frac{du}{dt}(t) = f(N^* + u(t)), \quad \text{avec} \quad u(0) = u_0.$$

En supposant qu'initalement u_0 est suffisament petit nous pouvons remplacer $f(N^* + u)$ par son développemet de Taylor $f'(N^*)\,u + O(u^2)$ et négliger les termes de la taille de u^2, il vient alors

$$\frac{d\tilde{u}}{dt} = f'(N^*)\,\tilde{u}, \quad \text{avec} \quad \tilde{u}(0) = u_0,$$

ce qui donne

$$\tilde{u}(t) = u_0\ \exp(f'(N^*)\,t)$$

et donc
– lorsque $f'(N^*) > 0$, la solution $\tilde{u}$ est une fonction du temps croissante, ce qui signifie que N^* est un équilibre instable,
– lorsque $f'(N^*) < 0$, la solution $\tilde{u}$ est à décroissance exponentielle, ce qui signifie que N^* est un équilibre stable.

Pour compléter la modélisation, nous pouvons ajouter un nouveau phénomène comme celui de la prédation. Par exemple, nous considérons une population de vers $N(t)$ nichés dans des arbres et mangés par des oiseaux. Dans un premier temps, nous fabriquons un modèle simple en négligeant les variations des prédateurs au cours du temps en fonction de la population de proies c'est-à-dire que nous ne considérons pas d'équation sur le nombre de prédateurs.

Nous notons $P(N)$ le nombre d'individus morts par prédation par unité de temps, il vient alors

$$\begin{cases} \dfrac{dN}{dt}(t) = f(N(t)) - P(N(t)), \\ N(0) = N_0 \end{cases}$$

Pour définir le taux de prédation, nous devons trouver la fonction P en fonction des hypothèses sur la prédation. Par exemple :

- lorsque $N(t)$ est petit, la prédation est proportionnelle au nombre de vers $N(t)$,
- lorsqu'il y a beaucoup de vers ; il se produit un effet de saturation et les oiseaux se gênent entre eux,
- lorsqu'il y a trop peu de vers, les oiseaux ne se déplacent pas et donc ne nuisent pas à la population de vers :

$$P(N) = \frac{B\,N^2}{A^2 + N^2}$$

L'équation différentielle ordinaire obtenue est alors

$$\begin{cases} \dfrac{dN}{dt}(t) = r\,N\,\left(1\,-\,\dfrac{N}{K}\right) - \dfrac{B\,N^2}{A^2\,+\,N^2} \\ N(0) = N_0 \end{cases}$$

Nous étudions aussi les états d'équilibres de ce modèle :

$$r\,N^*\,\left(1\,-\,\frac{N^*}{K}\right)\,-\,\frac{B\,N^{*2}}{A^2 + N^{*2}}\,=\,0,$$

ce qui signifie que soit $N^* = 0$ ou alors N^* est tel que

$$r\,\left(1\,-\,\frac{N^*}{K}\right)\,\left(A^2\,+\,N^{*2}\right)\,-\,B\,N^*\,=\,0.$$

Ainsi, N^* est solution d'une équation polynômiale de degré trois qui a soit trois racines réelles ou bien une racine réelle et deux racines complexes conjuguées.

Populations en intéractions. La prochaine étape va consister à étudier des populations en interaction (le système de Lotka-Volterra). Nous considérons deux populations : une de proies, une de prédateurs et notons $N(t)$ le nombre de proies et $P(t)$ le nombre de prédateurs et formulons alors les hypothèses de modélisation suivantes

- la naissance des proies est proportionnelle à la population des proies N,
- la mort par prédation est à la fois proportionnelle aux nombres de proies N et au nombre de prédateurs P,
- la naissance des prédateurs est proportionnelle à la population des prédateurs P et des proies N,
- la mort des prédateurs est proportionnelle à la population des prédateurs P (mort naturelle).

La mise en équations donne alors

$$\begin{cases} \dfrac{dN}{dt}\,=\,a\,N(t)\,-\,b\,N(t)\,P(t) \\ \dfrac{dP}{dt}\,=\,c\,N(t)\,P(t)\,-\,d\,P(t) \end{cases} \tag{7.12}$$

où tous les paramètres a, b, c et d sont strictement positifs.

Notre premier travail va consister à faire un simple changement de variables pour écrire notre système en fonction d'un seul paramètre ; ce qui simplifie l'étude sans perte de généralité. Nous posons $u = cN/d$, $v = bP/a$, $s = a\,t$, puis enfin $\alpha = d/a$ nous obtenons alors le système

$$\left\{ \begin{array}{l} \dfrac{du}{dt} \;=\; u(t)\,(1 \,-\, v(t)) \\ \dfrac{dv}{dt} \;=\; \alpha\, v(t)\,(u(t) \,-\, 1) \end{array} \right. \tag{7.13}$$

ou encore

$$\left\{ \begin{array}{l} \dfrac{d\ln(u)}{dt}(t) \;=\; (1 \,-\, v(t)) \\ \dfrac{d\ln(v)}{dt}(t) \;=\; \alpha\,(u(t) \,-\, 1) \end{array} \right. \tag{7.14}$$

Nous constatons alors que si nous posons

$$H(x,y) \;=\; \alpha\, e^x \,+\, e^y \,-\, \alpha\, x \,-\, y,$$

le couple $(\ln(u), \ln(v))$ est solution du système Hamiltonien suivant

$$\left\{ \begin{array}{l} \dfrac{dx}{dt} \;=\; -\dfrac{\partial H}{\partial y}(x(t), y(t)) \\ \dfrac{dy}{dt} \;=\; +\dfrac{\partial H}{\partial x}(x(t), y(t)), \end{array} \right. \tag{7.15}$$

ce qui signifie que $H(\ln(u(t)), \ln(v(t)))$ est constant au cours du temps

$$\frac{dH}{dt}(\ln(u(t)), \ln(v(t))) = 0.$$

Interessons nous ensuite aux points stationaires du système différentiel (7.14) ; nous avons

$$(u,v) = (0,0), \quad \text{ou bien} \quad (u,v) = (1,1)$$

Un étude au voisinnage du point $(0,0)$ donne en linéarisant le système autour de cet équilibre

$$\left\{ \begin{array}{l} \dfrac{du}{dt} \\ \dfrac{dv}{dt} \end{array} \right. \;=\; \left(\begin{array}{cc} 1 & 0 \\ 0 & -\alpha \end{array} \right) \left(\begin{array}{c} u \\ v \end{array} \right).$$

En appliquant le Théorème 2.3, puisque les parties réélles des valeurs propres ne sont pas toutes strictement négatives, nous avons donc affaire à un point col qui est instable.

D'autre part, au voisinnage du point $(1,1)$ nous calculons le linéarisé en ce point et le système autour de cet équilibre est donné par

$$\left\{ \begin{array}{l} \dfrac{du}{dt} \\ \dfrac{dv}{dt} \end{array} \right. = \left(\begin{array}{cc} 0 & -1 \\ 1 & 0 \end{array} \right) \left(\begin{array}{c} u \\ v \end{array} \right)$$

Dans ce cas, les valeurs propres (i et $-i$) sont imaginaires pures et nous ne pouvons pas conclure directement sur la stabilité de cet équilibre en appliquant le Théorème 2.3 ; il faut réaliser une étude spécifique ou effectuer des simulations numériques.

Cependant, un simple schéma d'Euler qu'il soit iplicite ou explicite ne fonctionne pas toujours très bien. En effet, ces schéma ne préservent pas l'invariance de l'Hamiltonien H et ne sont pas suffisamment précis car ils sont seulement d'ordre un. Nous devons alors envisager l'utilisation de schéma d'ordre élevé du style Runge-Kutta d'ordre quatre. Une autre démarce consiste à construire de schémas symplectiques.

Dans la suite nous allons mettre au point des méthodes spécifiques pour la discrétisation de systèmes Hamiltoniens. Nous écrivons le Hamiltonien sous la forme

$$H(x,y) = T(x) + U(y)$$

et le système différentiel s'écrit alors

$$\left\{ \begin{array}{l} \dfrac{dx}{dt} = -\nabla U(y) \\ \dfrac{dy}{dt} = +\nabla T(x) \end{array} \right. \tag{7.16}$$

Ce système possède deux propriétés intéressantes.

La conservation de l'énergie ou de l'Hamiltonien. Un calcul direct montre que l'énergie totale $H(x,y) = T(x) + U(y)$ reste constante le long des trajectoires $(x(t), y(t))$. En effet,

$$\frac{d}{dt} H(x(t), y(t)) = \nabla T(x(t)) \frac{dx}{dt} + \nabla U(y(t)) \frac{dy}{dt} = 0.$$

La conservation de l'aire. Soit $A \subset I\!R^2$, nous notons par

$$\varphi_A(t) = \big\{ (x(t), y(t)) \in I\!R^2, \quad (x(0), y(0)) \in A \big\}$$

et nous avons alors

$$aire(\varphi_A(t)) = \int_{\varphi_A(t)} dxdy = \int_A J_{\varphi_A(t)} dx_0 dy_0$$

avec

$$J_{\varphi_A(t)} = \left| \begin{pmatrix} \frac{\partial x(t)}{\partial x_0} & \frac{\partial x(t)}{\partial y_0} \\ \frac{\partial y(t)}{\partial x_0} & \frac{\partial y(t)}{\partial y_0} \end{pmatrix} \right|.$$

Nous voulons démontrer que $J_{\varphi_A(t)} = 1$. Pour cela nous allons démonter que $J_{\varphi_A(t)}$ est constant au cour du temps et vaut $J_{\varphi_A(t)} = J_{\varphi_A(0)} = 1$. En effet, d'une part nous avons

$$\frac{\partial x'(t)}{\partial x_0} = -\nabla^2 U(y(t)) \frac{\partial y(t)}{\partial x_0}, \qquad \frac{\partial y'(t)}{\partial x_0} = \nabla^2 T(x(t)) \frac{\partial x(t)}{\partial x_0},$$

et d'autre part

$$\frac{\partial x'(t)}{\partial y_0} = -\nabla^2 U(y(t)) \frac{\partial y(t)}{\partial y_0}, \qquad \frac{\partial y'(t)}{\partial y_0} = \nabla^2 T(x(t)) \frac{\partial x(t)}{\partial y_0},$$

Ainsi,

$$\frac{d}{dt} J_{\varphi_A(t)} = \frac{d}{dt} \left(\frac{\partial x(t)}{\partial x_0} \frac{\partial y(t)}{\partial y_0} - \frac{\partial x(t)}{\partial y_0} \frac{\partial y(t)}{\partial x_0} \right) = 0,$$

ce qui montre bien que le système Hamiltonien préserve bien l'aire au cours du temps.

Dans une simulation numérique d'un système Hamiltonien nous souhaiterions que les propriétés géométriques du flot exact soient préservées aussi bien que possible. Hélàs avec un schéma traditionnel de type Euler explicite, nous observons que la solution numérique possède une énergie qui va croître et l'aire d'un ensemble augmente au cours du temps. Pour une méthode d'Euler implicite, c'est exactement l'inverse. Nous pouvons observer ce phénomène sur le cas très simple $H(x, y) = x^2/2 + y^2/2$

En définitive, aucune de ces deux méthodes ne donne une approximation acceptable de la solution. Nous proposons alors d'étudier une méthode d'Euler symplectique.

Méthode d'Euler symplectique. Nous traitons une équation du système Hamiltonien par la méthode d'Euler explicite et l'autre par la méthode implicite. Ceci donne

$$\begin{cases} x^{n+1} = x^n - \Delta t \, \nabla U(y^n), \\ y^{n+1} = y^n + \Delta t \, \nabla T(x^{n+1}). \end{cases}$$

ou alors

$$\begin{cases} x^{n+1} = x^n - \Delta t \, \nabla U(y^{n+1}), \\ y^{n+1} = y^n + \Delta t \, \nabla T(x^n). \end{cases}$$

Les deux méthodes sont parfaitement explicites. Pour la première nous calculons d'abord x^{n+1}et ensuite y^{n+1} ; pour la deuxième dans l'ordre inverse.

Proposition 7.1 *Si ψ représente une des méthodes ci-dessus, alors ψ préserve l'aire, c'est-à-dire*

$$aire(\psi(A)) = aire(A).$$

Nous disons que cette méthode est symplectique.

Démonstration. Nous décomposons le membre droit du système Hamiltonien comme suit

$$\begin{cases} \dfrac{dx}{dt} = 0, \\ \dfrac{dy}{dt} = \nabla T(x), \end{cases} \quad \text{et} \quad \begin{cases} \dfrac{dx}{dt} = -\nabla U(y) \\ \dfrac{dy}{dt} = 0 \end{cases}$$

Les deux systèmes peuvent être résolus de manière exacte et tous les deux proviennent d'un Hamiltonien, ils préservent donc l'aire (par contre nous remarquons que cette décomposition modifie l'énergie car elle n'est plus définie de la même manière pour les deux systèmes)

$$\begin{cases} x(t) = x(0) \\ y(t) = y(0) + t\,\nabla T(x(0)) \end{cases} \quad \text{et} \quad \begin{cases} x(t) = x(0) - t\,\nabla U(y(0)) \\ y(t) = y(0) \end{cases}$$

Le schéma symplectique n'est donc que la composition des solutions exactes des deux systèmes ci-dessus. Il préserve donc l'aire. □

Chapitre 7

Les équations aux dérivées partielles

Dans ce chapitre, nous allons aborder la résolution approchée des équations aux dérivées partielles. Ces équations concernent les fonctions de plusieurs variables. Elles sont déduites de modèles de la physique ou de la mécanique.

1 Motivation

Nous nous intéressons d'abord à la dimension une. Dans ce cas, un problème aux limites est un problème composé de :

- Une équation différentielle du second ordre, sur un intervalle $]a, b[$.
- Une condition en a et une en b.

Par exemple,

$$\begin{cases} -u''(x) + c(x)\,u(x) = f(x), \quad a < x < b, \\ u(a) = \alpha, \quad u(b) = \beta. \end{cases} \tag{1.1}$$

L'inconnue est une fonction $u : x \to u(x)$ définie sur $[a, b]$. Les conditions aux limites sont $u(a) = \alpha$, $u(b) = \beta$. Elles portent ici sur la valeur de la solution au bord du domaine : nous parlons dans ce cas de conditions de Dirichlet. Lorsque les conditions aux limites portent sur les valeurs de la dérivée de la solution au bord, ce sont des conditions de Neumann. Nous parlons de conditions mixtes lorsque nous avons une combinaison de ces conditions.

Le problème (1.1) modélise par exemple le fléchissement d'une poutre de longueur $b - a$, représentée par le segment $[a, b]$, étirée selon son axe par une force P et soumise à une force transversale $f(x)$. Le moment fléchissant $u(x)$ est solution du problème (1.1) avec $c(x) = \frac{P}{E\,I(x)}$ où E est le module de Young, $I(x)$ le moment principal d'inertie et avec $\alpha = \beta = 0$. Nous avons ici $c(x) \geq 0$, et nous supposerons dans la suite pour simplifier que $c(x) = 1$.

Concernant l'existence et l'unicité de solutions, nous pouvons considérer par exemple le problème de Dirichlet homogène ($\alpha = \beta = 0$) et écrire le problème (1.1) sous une forme plus

faible ne faisant intervenir que des dérivées d'ordre un, nous parlons de **formulation variationnelle** : nous définissons l'espace $H_0^1(]a,b[)$ par

$$H_0^1(]a,b[) = \left\{u \in L^2(]a,b[), \quad u' \in L^2(]a,b[) \quad \text{et} \quad u(a) = u(b) = 0\right\}$$

puis recherchons $u \in H_0^1(]a,b[)$ tel que

$$\int_a^b (u'(x)\, v'(x) \,+\, u(x)\, v(x))dx = \int_a^b f(x)\, v(x)\, dx, \quad \forall v \in H_0^1(]a,b[).$$

Nous démontrons alors l'existence et l'unicité d'une solution de (1.1) à l'aide du Théorème de Lax-Milgram

Théorème 1.1 (Théorème de Lax-Milgram) *Soient V un espace de Hilbert (espace de Banach muni d'un produit scalaire et d'une norme induite $\|.\|$), $a(.,.)$ une forme bilinéaire, continue et coercive et $l(.)$ une forme linéaire continue. Alors le problème variationnel : trouver $u \in V$*

$$a(u,v) = l(v), \quad \forall v \in V,$$

admet une solution unique.

2 La méthode des différences finies

Nous supposons que $f \in C^2([a,b], I\!R)$. La solution u est alors dans l'espace $C^4([a,b], I\!R)$ car $u'' = f - u$ est de classe C^2. Soit $N_x \in I\!N$ un entier fixé, $N_x \geq 1$, nous posons $h = (b-a)/(N_x+1)$, et définissons les points :

$$x_0 = a,\ x_1 = a + h,\ \ldots, x_i = a + ih,\ \ldots,\ x_{N_x+1} = b.$$

Pour un point x_i quelconque intérieur à l'intervalle $]a,b[$, c'est-à-dire tel que $1 \leq i \leq N_x$, nous écrivons les développements de Taylor :

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

et

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

Ainsi,

$$u(x_{i+1}) \,+\, u(x_{i-1}) \,=\, 2\, u(x_i) \,+\, h^2\, u''(x_i) \,+\, \frac{h^4}{12}\, u^{(4)}(x_i) \,+\, h^4\, \varepsilon(h)$$

où $\varepsilon(h)$ désigne une fonction qui tend vers zéro lorsque h tend vers zéro. Comme nous l'avons fait précédemment, nous adopterons la notation $h^p \varepsilon(h) = o(h^p)$ et avons donc :

$$u''(x_i) \,=\, \frac{u(x_{i+1}) \,-\, 2\, u(x_i) + u(x_{i-1})}{h^2} \,-\, \frac{h^2}{12} u^{(4)}(x_i) \,+\, o(h^2),$$

de sorte que

$$-u''(x_i) \,+\, u(x_i) \,=\, \frac{-u(x_{i-1}) \,+\, 2\,u(x_i) \,-\, u(x_{i+1})}{h^2} \,+\, u(x_i) \,+\, \frac{h^2}{12}\,u^{(4)}(x_i) \,+\, o(h^2).$$

Comme u est solution de (1.1), nous aurons donc pour tout $1 \leq i \leq N_x$,

$$\frac{-\,u(x_{i-1}) \,+\, 2\,u(x_i) \,-\, u(x_{i+1})}{h^2} \,+\, u(x_i) \,+\, \frac{h^2}{12}u^{(4)}(x_i) \,+\, o(h^2) \,=\, f(x_i). \tag{2.2}$$

Pour définir un schéma de discrétisation du problème par la méthode des différences finies, nous allons considérer que les termes d'ordre supérieur à un dans le développement (2.2) sont négligeables. Nous chercherons des réels u_0, u_1, …, u_{N_x+1} tels que pour tout $1 \leq i \leq N_x$,

$$f(x_i) = \frac{-\,u_{i-1} \,+\, 2\,u_i \,-\, u_{i+1}}{h^2} \,+\, u_i$$

L'erreur de consistance $R(h)$ du schéma au point x_i sera définie par :

$$R_i(h) = \frac{-\,u(x_{i-1}) \,+\, 2\,u(x_i) \,-\, u(x_{i+1})}{h^2} \,+\, u(x_i) \,-\, f(x_i), \tag{2.3}$$

où $u(.)$ est la solution exacte du problème aux limites (1.1). L'erreur de consistance représente la distance entre le schéma de discrétisation et le problème différentiel vérifié par u. Le développement (2.2) fournit l'estimation :

$$R_i(h) = \frac{h^2}{12}u^{(4)}(x_i) \,+\, o(h^2). \tag{2.4}$$

Nous avons négligé les erreurs de consistance et espérons trouver des u_i voisins des valeurs $u(x_i)$ prises par la solution $u(x)$ aux points x_i. Le système d'équations que nous avons écrit ci-dessus comporte N_x équations pour $N_x + 2$ inconnues ; il faut éliminer des inconnues ou rajouter des équations. Pour cela, nous allons interpréter les conditions aux limites. Ici, comme ce sont des conditions de Dirichlet, nous pouvons éliminer deux inconnues. Nous poserons :

- $u_0 = 0$ puisque $u(x_0) = 0$,
- $u_{N_x+1} = 0$ puisque $u(x_{N_x+1}) = 0$.

Le système à résoudre peut alors s'écrire :

$$\left\{\begin{array}{lcl}
(2+h^2)\,u_1 \,-\, u_2 & = & h^2 f(x_1), \\
\vdots & \vdots & \vdots \\
-u_{i-1} \,+\, (2+h^2)\,u_i \,-\, u_{i+1} & = & h^2 f(x_i), \\
\vdots & \vdots & \vdots \\
-u_{N_x-1} \,+\, (2+h^2)\,u_{N_x} & = & h^2 f(x_{N_x}),
\end{array}\right. \tag{2.5}$$

Ce système linéaire s'écrit matriciellement

$$A\,U \;=\; h^2\,F,$$

où A est une matrice carrée de taille $N_x \times N_x$

$$A = \begin{pmatrix} 2+h^2 & -1 & 0 & \dots & 0 \\ -1 & 2+h^2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -1 \\ 0 & \dots & 0 & -1 & 2+h^2 \end{pmatrix}$$

et U est le vecteur des inconnues et F la donnée

$$U = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{N_x} \end{pmatrix}, \quad F = \begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_{N_x}) \end{pmatrix}.$$

Nous remarquerons que la matrice A peut s'écrire

$$A \;=\; M \;+\; h^2 I_{N_x}.$$

La matrice $M \in \mathcal{M}_{N_x,N_x}(I\!R)$ définie par cette égalité est la matrice tridiagonale dont les seuls coefficients non nuls sont :

$$\begin{cases} m_{i,i} = 2, \quad 1 \le i \le N_x \\ m_{i,i+1} = m_{i+1,i} = -1, \quad 1 \le i \le N_x - 1 \end{cases} \tag{2.6}$$

Théorème 2.1 *La matrice M est symétrique définie positive. De plus, les valeurs propres sont données par*

$$\lambda^{(k)} = 4\,\sin^2\left(\frac{k\,\pi}{2(N_x+1)}\right)$$

et le vecteur propre associé à $\lambda^{(k)}$

$$v_j^{(k)} = \sin\left(\frac{j\,k\,\pi}{(N_x+1)}\right), \quad j = 1, ..., N_x.$$

Démonstration. Pour tout vecteur $x = (x_1, ..., x_{N_x})^T$ de $I\!R^{N_x}$, nous avons

$$\begin{aligned} x^T\, M\, x &= 2\, x_1^2 \,-\, 2\, x_1\, x_2 \,+\, 2\, x_2^2 \,+\, \ldots \,-\, 2\, x_{N_x-1}\, x_{N_x} \,+\, 2\, x_{N_x}^2, \\ &= x_1^2 \,+\, \sum_{i=1}^{N_x-1} (x_{i+1} \,-\, x_i)^2 \,+\, x_{N_x}^2. \end{aligned}$$

Ainsi, $x^T\, M\, x$ est donc toujours positif, et ne peut s'annuler que si $x = 0$. Il reste à trouver les éléments propres de la matrice M. Tout d'abord nous appliquons le Théorème 2.1 du Chapitre 2 :

$$\lambda \in [0, 4], \quad \forall\, \lambda \in Sp(M);$$

d'où il existe $\theta \in]0, \pi[$ tel que

$$\lambda \,=\, 2 - 2\cos(\theta).$$

Nous notons alors $v \in I\!R^{N_x+2}$ le vecteur propre associé à la valeur propre λ et obtenons

$$-v_{i-1} \,+\, 2\, v_i \,-\, v_{i+1} \,=\, \lambda\, v_i \,=\, (2 - 2\cos\theta)\, v_i, \quad i = 1, \ldots, N_x$$

ou encore $v_0 = v_{N_x+1} = 0$ et

$$v_{i-1} \,-\, 2\cos\theta\, v_i \,+\, v_{i+1} \,=\, 0, \quad i = 1, \ldots, N_x.$$

Nous recherchons alors la solution sous la forme $v_i = r^i$, où r est la solution de

$$(1 \,-\, 2\, \cos\theta\, r \,+\, r^2)\, r^{i-1} \,=\, 0,$$

c'est-à-dire $r_1 = e^{i\,\theta}$ et $r_2 = e^{-i\theta}$. Nous avons alors

$$v_i \,=\, \alpha\, r_1^i \,+\, \beta\, r_2^i, \quad i = 1, \ldots, N_x.$$

et $v_0 = v_{N_x+1} = 0$. D'une part, nous avons

$$\alpha \,+\, \beta \,=\, v_0 \,=\, 0$$

et donc $\alpha = -\beta$. D'autre part, nous avons

$$\alpha\, \left(e^{i(N_x+1)\theta} \,-\, e^{-i(N_x+1)\theta}\right) \,=\, 0$$

ce qui implique $\sin\left((N_x + 1)\,\theta\right) = 0$ ou encore

$$\theta^{(k)} \,=\, \frac{k\,\pi}{N_x + 1}, \quad 0 \,<\, k \,<\, N_x + 1.$$

Finalement, nous obtenons alors pour $k = 1, \ldots, N_x$

$$\lambda^{(k)} \,=\, 2 - 2\cos(\theta^{(k)}) \,=\, 4\, \sin^2\left(\frac{k\pi}{2(N_x+1)}\right)$$

et

$$v_i^{(k)} \,=\, \sin\left(\frac{ik\pi}{N_x+1}\right), \quad i = 1, \ldots, N_x.$$

□

2.1 Étude de l'erreur

Dans cette partie, nous voulons démontrer que la méthode aux différences finies est bien convergente. Pour cela, nous reconstruisons une approximation $u_h(x)$ pour tout $x \in [a,b]$ de la manière suivante

$$u_h(x) = \begin{cases} u_0 & x \in [a, a+h/2[\\ u_i & x \in [x_i - h/2, x_i + h/2[, \quad i = 0, \ldots, N_x \\ u_{N_x+1} & x \in [b-h/2, b] \end{cases}$$

Puis, nous démontrons que le terme d'erreur

$$\varepsilon_h := \left(\int_a^b |u_h(x) - u(x)|^2 dx \right)^{1/2}$$

converge vers zéro lorsque h tend vers zéro où bien sûr u désigne la solution exacte de (1.1).

D'une part puisque la méthode de Newton-Cotes présentée dans le complément du Chapitre 5 est d'ordre deux pour une fonction $u^2 \in C^2([a,b], I\!R)$ nous observons que

$$\left| \int_a^b u^2(x)dx - \sum_{i=1}^{N_x} h u^2(x_i) - h\,\frac{u^2(a)+u^2(b)}{2} \right| \leq \frac{(b-a)\|u''\|_{L^\infty}^2}{24}\, h^2$$

et puisque $u(a) = u(b) = 0$, nous avons

$$\left| \int_a^b u^2(x)dx - \sum_{i=1}^{N_x} h u^2(x_i) \right| \leq \frac{(b-a)\|u''\|_{L^\infty}}{24}\, h^2.$$

Nous remarquons ensuite qu'en prolongeant si besoin la fonction u sur $[a-h/2, b+h/2]$ et puisque $u(a) = u(b) = 0$, nous obtenons finalement

$$\int_a^{a+h/2} |u(x)|^2 dx \;+\; \sum_{i=1}^{N_x} \int_{x_i-h/2}^{x_i+h/2} |u(x) - u(x_i)|^2 dx \;+\; \int_{b-h/2}^{b} |u(x)|^2 dx$$
$$\leq\; 2 \left| \int_a^b u^2(x)dx - \sum_{i=0}^{N_x+1} h u^2(x_i) \right| \leq\; C_1\, h^2,$$

où $C_1 > 0$ est une constante qui ne dépend que de u. Il vient alors

$$\varepsilon_h^2 \;\leq\; 2\, C_1\, h^2 \;+\; 2 \sum_{i=1}^{N_x} h |u(x_i) - u_i|^2.$$

Nous définissons alors le vecteur $\mathcal{U}$ par la solution du problème aux limites (1.1) prise aux points $(x_i)_{1\leq i\leq N_x}$

$$\mathcal{U} = (u(x_1), \ldots, u(x_{N_x}))^T \in I\!R^{N_x},$$

et le vecteur des erreurs

$$e(h) \;=\; U - \mathcal{U} \;=\; (e_1, \ldots, e_{N_x})^T.$$

Ces composantes du vecteur $e(h)$ sont les scalaires $e_i = u_i - u(x_i)$. Nous posons alors

$$\|e(h)\|_h := \left(\sum_{i=1}^{N_x} h|e_i|^2 \right)^{1/2} \tag{2.7}$$

et souhaitons montrer que $\|e(h)\|$ converge vers zéro lorsque h tend vers zéro, ce qui démotrera d'après ce qui précède que ε_h tend lui aussi vers zéro.

Commençons d'abord par démontrer un lemme connu sous le nom d'inégalité de Poincaré dans le cas discret. En effet dans le cas continu, nous avons pour $u \in H_0^1(]a,b[)$

$$\int_a^b |u(x)|^2\,dx \;\leq\; \int_a^b |u'(x)|^2\,dx.$$

Lemme 2.1 *Pour la matrice modèle $M \in \mathcal{M}_{N_x,N_x}(I\!R)$, en posant $h = (b-a)/(N_x+1)$, nous avons pour tout vecteur $u \in I\!R^{N_x}$ et en posant $u_0 = u_{N_x+1} = 0$*

$$u^T\,M\,u \;=\; \sum_{i=0}^{N_x} |u_{i+1} - u_i|^2 \;\geq\; \frac{h^2}{(b-a)^2}\,\|u\|_2^2,$$

où $\|.\|_2$ définit la norme Euclidienne de $I\!R^{N_x}$

$$\|u\|_2^2 \;=\; u^T\,u.$$

Démonstration. Dans le cas continu, la preuve est relativement simple et nous allons nous en inspirer pour démonter le résultat dans le cas discret. En effet, pour $u \in H_0^1(]a,b[)$, nous avons puisuqe $u(a) = 0$

$$u(x) = \int_a^x u'(s)\,ds$$

et donc

$$\int_a^b |u(x)|^2\,dx \;=\; \int_a^b \left| \int_a^x u'(s)\,ds \right|^2 dx \;\leq\; \int_a^b \left(\int_a^b |u'(s)|^2\,ds \right) (x-a)\,dx$$

ce qui donne l'inégalité de Poincaré

$$\int_a^b |u(x)|^2\,dx \;\leq\; \frac{(b-a)^2}{2} \int_a^b |u'(x)|^2\,dx.$$

Passons maintenant au cas discret : de la même manière nous avons puisque $u_0 = 0$

$$u_i \;=\; (u_i - u_{i-1}) \;+\; (u_{i-1} - u_{i-2}) \;+\; (u_{i-2} - u_{i-3}) \ldots \;+\; u_0,$$

d'où

$$u_i \;\leq\; \sum_{k=0}^{i-1} |u_{k+1} - u_k| \;\leq\; \left(\sum_{k=0}^{i-1} |u_{k+1} - u_k|^2 \right)^{1/2} \left(\sum_{k=0}^{i-1} 1^2 \right)^{1/2}.$$

Ainsi, en prenant le carré et en faisant la somme pour $i = 1, \ldots, N_x$ nous obtenons

$$\sum_{i=1}^{N_x} |u_i|^2 \;\leq\; \sum_{i=1}^{N_x} \left(\sum_{k=0}^{N_x} |u_{k+1} - u_k|^2 \right) \, (i+1)\,.$$

ou encore

$$\sum_{i=1}^{N_x} |u_i|^2 \;\leq\; \frac{N_x\,(N_x+1)}{2} \sum_{k=0}^{N_x} |u_{k+1} - u_k|^2 \;=\; \frac{(b-a)^2}{h^2} \sum_{k=0}^{N_x} |u_{k+1} - u_k|^2.$$

Pour conclure, nous observons que pour tout $u \in I\!R^{N_x}$ et en posant $u_0 = u_{N_x+1} = 0$

$$u^T\, M\, u = \sum_{i=0}^{N_x} |u_{i+1} - u_i|^2 \;\geq\; \frac{h^2}{(b-a)^2}\, u^T\, u.$$

□

Puis nous démontrons le théorème de convergence suivant

Théorème 2.2 (Théorème de convergence) *Supposons que la fonction $f \in C^2([a,b], I\!R)$ (ce qui implique que la solution exacte $u \in C^4([a,b], I\!R)$). Alors, le vecteur des erreurs $e(h)$ tend vers zéro lorsque h tend vers zéro et de plus*

$$\|e(h)\|_h \;\leq\; C\, h^2,$$

où C dépend de la dérivée quatrième de la solution exacte u et $\|.\|_h$ est la norme introduite dans (2.7).

Démonstration. D'une part, nous avons résolu

$$A\, U \;=\; h^2\, F,$$

et (2.3) entraîne $A\,\mathcal{U} \;=\; h^2\, F \;+\; h^2 R(h)$. Nous obtenons donc par linéarité :

$$A\, e(h) \;=\; h^2\, R(h). \tag{2.8}$$

Nous multiplions (2.8) à gauche par $e^T(h)$:

$$e^T(h)\,A\,e(h) \;=\; h^2\,e^T(h)\,R(h),$$

ou encore

$$e^T(h)\,M\,e(h) \;+\; h^2\,e^T(h)\,I_{N_x}\,e(h) \;=\; h^2\,e^T(h)\,R(h),$$

puis,

$$e^T(h)\,M\,e(h) \;+\; h^2\,\|e(h)\|_2^2 \;=\; h^2\,e^T(h)\,R(h).$$

Nous pouvons majorer le terme de droite en appliquant l'inégalité de Cauchy-Schwartz :

$$e^T(h)\,R(h) \;=\; (e(h)\,,\,R(h)) \;\leq\; \|e(h)\|_2\,\|R(h)\|_2.$$

Il reste à minorer le terme de gauche, ce que nous faisons en appliquant le Lemme 2.1, nous obtenons :

$$\left(\frac{1}{(b-a)^2}+1\right)\,h^2\,\|e(h)\|_2^2 \;\leq\; e^T(h)\,A\,e(h) \;\leq\; h^2\,\|e(h)\|_2\,\|R(h)\|_2,$$

puis

$$\|e(h)\|_2 \;\leq\; \frac{(b-a)^2}{1+(b-a)^2}\,\|R(h)\|_2.$$

Finalement, puisque $f \in C^2([a,b],I\!R)$, ceci implique que la solution exacte $u \in C^4([a,b],I\!R)$ et nous pouvons alors utiliser l'estimation de l'erreur de consistance (2.4) pour conclure

$$\begin{aligned}\|e(h)\|_h \;=\; \left(\sum_{i=0}^{N_x} h\,|e_i|^2\right)^{1/2} \;=\; h^{1/2}\,\|e(h)\|_2 \;&\leq\; \frac{(b-a)^2}{1+(b-a)^2}\,\|R(h)\|_h \\ &\leq\; \frac{h^2}{24}\,\frac{(b-a)^3}{1+(b-a)^2}\,\|u^{(4)}\|_\infty.\end{aligned}$$

□

Pour ce problème, nous pouvons aussi exprimer l'erreur à l'aide de la norme $\|e(h)\|_\infty = \max_{1\leq i\leq N_x}(|e_i|)$ et montrer que $\|e(h)\|_\infty \;\leq\; \|R(h)\|_\infty$.

2.2 Conditions aux limites de Dirichlet

Le problème suivant est légèrement différent de (1.1).

$$\begin{cases} -u''(x) \;=\; f(x), \quad a<x<b, \\ \\ u(a)=\alpha, \quad u(b)=\beta. \end{cases} \tag{2.9}$$

Pour ce problème, le schéma de discrétisation s'écrit

$$M\,U \;=\; h^2\,F.$$

La matrice M est la matrice modèle, et le second membre F ne diffère de celui de la partie précédente que par sa première et sa dernière composante :

$$F_1 = f(x_1) + \frac{\alpha}{h^2}, \quad F_{N_x} = f(x_{N_x}) + \frac{\beta}{h^2}.$$

Nous pouvons aussi trouver la solution en ajoutant deux équations au système établi pour les points à l'intérieur du domaine et obtenons alors :

$$\left\{\begin{array}{lcl} u_0 & = & \alpha \\ -u_0 + 2u_1 - u_2 & = & h^2 f(x_1) \\ \vdots & \vdots & \vdots \\ -u_{N_x-1} + 2u_{N_x} - u_{N_x+1} & = & h^2 f(x_{N_x}) \\ u_{N_x+1} & = & \beta. \end{array}\right.$$

Nous perdons la symétrie de la matrice, mais gagnons en simplicité de mise en oeuvre.

2.3 Conditions aux limites mixtes

Nous considérons, pour $f \in C^2([a,b], I\!R)$, le problème :

$$\left\{\begin{array}{l} -u''(x) \;=\; f(x), \quad a < x < b, \\ u'(a) = \alpha, \quad u(b) = \beta. \end{array}\right. \tag{2.10}$$

Nous souhaitons discrétiser (2.10) par la méthode des différences finies, et pour cela nous procédons comme dans la partie précédente.

Pour définir le schéma, nous recherchons des u_i tels que :

$$-u_{i-1} \;+\; 2\,u_i \;-\; u_{i+1} \;=\; h^2\,f(x_i)$$

Le système ainsi défini possède $N_x + 2$ inconnues pour N_x équations. Nous pouvons naturellement poser $u_{N_x+1} = \beta$, ce qui élimine une inconnue. Mais, nous ne pouvons pas éliminer u_0 de cette façon car nous ne connaissons que $u'(a)$. Nous allons donc rajouter une équation qui interprète la condition $u'(a) = \alpha$. En s'inspirant de la formule

$$u'(x) = \frac{u(x+h) - u(x)}{h} + O(h), \tag{2.11}$$

nous poserons :

$$\frac{u_1 - u_0}{h} = \alpha$$

Nous ajoutons donc l'équation $u_0 - u_1 = -h\,\alpha$ à notre système qui s'écrit maintenant, de façon à préserver la symétrie dans $\mathbb{R}^{N_x+1}$:

$$\begin{cases} u_0 - u_1 & = & -h\,\alpha \\ -u_0 + 2\,u_1 - u_2 & = & h^2 f(x_1) \\ \vdots & \vdots & \vdots \\ -u_{N_x+1} + u_{N_x} & = & h^2 f(x_{N_x}) + \beta, \end{cases}$$

ou encore

$$A'\,U = h^2\,F$$

avec la matrice A' et le second membre F définis par :

$$A' = \begin{pmatrix} 1 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \dots & 0 & -1 & 1 \end{pmatrix}$$

et U est le vecteur des inconnues et F la donnée

$$U = \begin{pmatrix} u_0 \\ u_1 \\ \vdots \\ u_{N_x} \end{pmatrix}, \quad F = \begin{pmatrix} -\frac{\alpha}{h} \\ f(x_1) \\ f(x_2) \\ \vdots \\ f(x_{N_x}) + \frac{\beta}{h^2} \end{pmatrix}.$$

Pour $1 \leq i \leq N_x$, par un raisonnement analogue à celui que nous avons fait précédemment, nous montrons que les erreurs de consistance vérifient :

$$R_i(h) \,=\, \frac{-u(x_{i-1}) \,+\, 2\,u(x_i) \,-\, u(x_{i+1})}{h^2} \,-\, f(x_i) = \frac{h^2}{12}u^{(4)}(x_i) \,+\, o(h^2).$$

Par ailleurs, $R_0(h)$ est donnée par :

$$R_0(h) = \frac{1}{h^2}(u(a) - u(a+h)) + \frac{u'(a)}{h},$$

et vérifie

$$R_0(h) = -\frac{1}{2}u''(a) - \frac{h}{6}u^{(3)}(a) + o(h).$$

Pour $f \in C^2([a,b], I\!R)$, nous avons $u \in C^4([a,b], I\!R)$ et par continuité

$$-u''(a) = f(a)$$

de sorte que $R_0(h) \simeq f(a)/2$ ne tend pas vers zéro lorsque h tend vers zéro.

Nous pouvons supposer que la solution u se prolonge sur $]a-h, b]$ et discrétiser le problème au point a :

$$-u''(a) = f(a) \Rightarrow \frac{1}{h^2}(-u_{-1} \,+\, 2\,u_0 \,-\, u_1) \,=\, f(a). \tag{2.12}$$

Cette discrétisation de la dérivée seconde en a a introduit une inconnue supplémentaire. Elle sera virtuelle, et nous l'appelons quelque fois l'inconnue "fantôme" u_{-1}, car nous allons l'éliminer immédiatement. Nous interprétons pour cela la condition limite $u'(a) = \alpha$ en nous inspirant de la formule

$$u'(x) = \frac{u(x+h) - u(x-h)}{2h} \,+\, \frac{h^2}{3}u^{(3)}(x) \,+\, o(h^3), \tag{2.13}$$

La formule (2.11) était d'ordre un ; celle-ci est d'ordre deux. Nous pouvons poser :

$$\frac{u_1 - u_{-1}}{2h} = \alpha.$$

Cette équation nous fournit une valeur $u_{-1} \,=\, u_1 - 2h\alpha$, que nous reportons dans l'équation (2.12), pour obtenir :

$$\frac{1}{h^2}(u_0 - u_1) = \frac{f(a)}{2} - \frac{\alpha}{h}$$

Nous pouvons alors vérifier que $R_0(h) \,=\, -\frac{h}{6}f'(a)$. La matrice A' de notre schéma est symétrique définie positive. Nous pouvons faire l'analyse d'erreur en norme deux comme nous l'avons fait pour le problème avec conditions aux limites de Dirichlet en adaptant la démonstration du Lemme 2.1 à notre problème aux limite (2.10).

2.4 Ordre d'un schéma

Nous dirons qu'un schéma est d'ordre p si pour tout $n \geq 1$, et $1 \leq i \leq n$, avec $h = (b-a)/(N_x+1)$, nous avons :

$$|R_i(h)| = O(h^p).$$

Les schémas que nous avons vus sont d'ordre deux. Nous pouvons estimer l'ordre d'un schéma par l'analyse mathématique comme nous l'avons fait. Nous pouvons aussi le faire de façon empirique.

Supposons qu'un schéma d'ordre p soit appliqué à un problème stable ; nous pouvons nous attendre à obtenir

$$e_i = O(h^p), \quad 1 \leq i \leq N_x,$$

ce qui signifie ici que nous pouvons espérer qu'il existe une fonction $\phi(x)$ telle que

$$\frac{e_i}{h^p} \simeq \phi(x_i)$$

au voisinage de $h = 0$.

En calculant une solution approchée du problème pour un N_x donné, nous obtenons une solution $U = (u_1, \ldots, u_{N_x})^T$ telle que $u_i \simeq u(x_i) + \phi(x_i)h^p$. Nous faisons alors un nouveau calcul pour la discrétisation définie par $l = 2N_x + 1$; nous obtenons un vecteur $V = (v_1, \ldots, v_{2N_x+1})^T$ tel que $1 \leq j \leq l = 2N_x + 1$, nous ayons :

$$v_j \simeq u(a + jh/2) + \phi(a + jh/2) \left(\frac{h}{2}\right)^p,$$

de sorte que pour $j = 2i$, nous aurons :

$$v_{2i} \simeq u(a + ih) + \phi(x_i) \left(\frac{h}{2}\right)^p,$$

En effectuant un nouveau calcul pour $m = 2l + 1 = 4N_x + 3$, nous obtenons un vecteur $W = (w_1, \ldots, w_{4N_x+3})$ tel que $1 \leq k \leq m$, nous ayons

$$w_k \simeq u(a + kh/4) + \phi(a + kh/4) \left(\frac{h}{4}\right)^p,$$

de sorte que pour $k = 4i$, nous aurons :

$$w_{4i} \simeq u(x_i) + \phi(x_i) \left(\frac{h}{4}\right)^p,$$

Nous pouvons extraire du vecteur $V \in I\!R^{2N_x+1}$ le vecteur $\overline{V} = (v_2, \ldots, v_{2i}, \ldots, v_{2n})^T \in I\!R^{N_x}$, et du vecteur W le vecteur $\overline{W} = (w_4, \ldots, w_{4i}, \ldots, w_{4N_x})^T \in I\!R^{N_x}$. Les composantes de ces trois vecteurs vérifient alors $1 \leq i \leq N_x$:

$$u_i - \phi(x_i)h^p \simeq v_{2i} - \phi(x_i) \left(\frac{h}{2}\right)^p \simeq w_{4i} - \phi(x_i) \left(\frac{h}{4}\right)^p.$$

Il vient alors $1 \leq i \leq N_x$:

$$\frac{v_{2i} - u_i}{w_{4i} - v_{2i}} \simeq \frac{\left(\frac{1}{2}\right)^p - 1}{\left(\frac{1}{4}\right)^p - \left(\frac{1}{2}\right)^p} = 2^p.$$

2.5 Problèmes elliptiques plus généraux

Ces problèmes sont de la forme :

$$\begin{cases} -(a(x)u'(x))' = f(x), \quad a \leq x \leq b, \\ u(a) = \alpha, \quad u(b) = \beta. \end{cases}$$

Nous supposons que $f \in C^2([a,b], I\!R)$ et que $a \in C^3([a,b], I\!R)$ est telle qu'il existe $\eta > 0$ tel que pour tout $x \in [a,b]$, nous avons $a(x) \geq \eta > 0$.

Nous pouvons faire des hypothèses moins sévères, mais la seule qui soit indispensable est que $a(x)$ ne s'annule pas. En particulier, nous rencontrons souvent le cas d'une fonction $a(x)$ constante par morceaux. Les conditions aux limites données ici sont les conditions de Dirichlet.

Lorsque nous connaissons la fonction $a(x)$, la règle de dérivation d'un produit de fonctions permet d'écrire :

$$-(a(x)u'(x))' = -a'(x)u'(x) - a(x)u''(x).$$

Aussi, lorsque la fonction $a(x)$ est donnée par une formule mathématique explicite, nous pourrons proposer le schéma suivant :

$$\frac{a(x_i)}{h^2}(-u_{i-1} + 2u_i - u_{i+1}) + \frac{a'(x_i)}{2h}(u_{i-1} - u_{i+1}) = f(x_i). \tag{2.14}$$

Nous avons utilisé une formule centrée d'ordre deux pour discrétiser le terme $u'(x_i)$, de façon à avoir globalement une formule d'ordre deux.

Lorsque nous ne connaissons pas la fonction $a(x)$, nous ne pourrons évidemment pas utiliser (2.14). Comme nous ne connaissons pas les valeurs exactes de $a'(x)$, il faudra les remplacer par des formules de différences finies. En utilisant la formule centrée d'ordre deux pour écrire le schéma de discrétisation au point x_i, nous pourrons poser :

$$\frac{a(x_i)}{h^2}(-u_{i-1} + 2u_i - u_{i+1}) - \frac{1}{4h^2}(a(x_{i+1}) - a(x_{i-1}))(u_{i+1} - u_{i-1}) = f(x_i).$$

Pour alléger les notations, nous noterons $a_i = a(x_i)$. Notez bien que les a_i désigneront des valeurs exactes de la fonction $a(x)$ alors que les u_i désignent les valeurs numériques inconnues que nous cherchons à calculer pour approcher les valeurs $u(x_i)$. Nous pouvons réorganiser cette équation en regroupant les coefficients de chaque inconnue

$$\frac{1}{4h^2}((a_{i+1} - 4a_i - a_{i-1})u_{i-1} + 8a_iu_i + (a_{i-1} - 4a_i - a_{i+1})u_{i+1}) = f(x_i).$$

Ce schéma est assez compliqué, il est préférable de construire le schéma qui suit. En notant $v(x) = a(x)u'(x)$, une première étape de la discrétisation de (2.5) peut s'écrire, au point x_i :

$$-\frac{1}{2h}(v(x_{i+1}) - v(x_{i-1})) \simeq f(x_i).$$

Maintenant, nous observons que pour tout j, nous aurons

$$v(x_j) = a(x_j)u'(x_j) \simeq \frac{a(x_j)}{2h}(u(x_{j+1}) - u(x_{j-1})).$$

Dans une seconde étape, nous utilisons cette dernière formule pour évaluer les expressions $v(x_{i+1})$ et $v(x_{i-1})$:

$$-\frac{1}{2h}\left(a_{i+1}\left(\frac{u_{i+2} - u_i}{2h}\right) - a_{i-1}\left(\frac{u_i - u_{i-2}}{2h}\right)\right) = f(x_i).$$

En regroupant les coefficients de chaque inconnue, nous obtenons :

$$\frac{1}{4h^2}\left(-a_{i-1}u_{i-2} + (a_{i-1} + a_{i+1})u_i - a_{i+1}u_{i+2}\right) = f(x_i).$$

Cette fois, le schéma est symétrique, mais il fera perdre de la précision car les équations ne relient pas les inconnues associées aux points voisins. Un moyen de remédier à cet inconvénient consiste à utiliser un pas de $h/2$ pour écrire le schéma. Nous obtenons alors, en notant $a_{i\pm1/2} = a(x_i \pm h/2)$:

$$\frac{1}{h^2}\left(-a_{i-1/2}u_{i-1} + (a_{i-1/2} + a_{i+1/2})u_i - a_{i+1/2}u_{i+1}\right) = f(x_i).$$

Il faut disposer des valeurs de la fonction $a(x)$ aux points $x_i \pm h/2$. Si nous n'en disposons pas, il faudra remplacer les expressions $a_{i\pm1/2}$ par des expressions interpolées. Nous remplacerons $a_{i\pm1/2}$ par $\frac{a_i + a_{i\pm1}}{2}$ Nous pouvons vérifier que le schéma reste d'ordre deux.

Nous pouvons montrer que lorsque a et f sont suffisamment régulières, ce schéma est d'ordre deux. En outre, si $R(h)$ désigne le vecteur des erreurs de consistance et e celui des erreurs sur la solution, la majoration de l'erreur est donnée par :

$$\|e\|_h \leq \frac{1}{\alpha}\|R(h)\|_h$$

Nous voyons qu'elle est d'autant plus mauvaise que α est petit.

3 La méthode des éléments finis

Dans une première partie nous présentons la méthode de Galerkine puis nous nous concentrons sur la méthode des éléments finis en dimension une pour la construction du système linéaire à résoudre.

3.1 Méthodologie générale

Nous considérons le problème type

$$\begin{cases} -\Delta u = f, \text{ dans } \Omega \\ \\ u = 0, \text{ sur } \partial\Omega \end{cases}$$

où Ω est un ouvert de $\mathbb{R}^d$ Nous écrivons la formulation variationnelle en utilisant des intégrations par parties : nous cherchons $u \in H_0^1(\Omega)$ tel que

$$\int_\Omega \nabla u(x) \cdot \nabla v(x)\, dx \;=\; \int_\Omega f(x) v(x)\, dx, \qquad \forall\, v \;\in\; H_0^1(\Omega).$$

De manière plus générale, nous pouvons considérer un problème plus abstrait de la forme

$$a(u, v) \;=\; l(v), \qquad \forall\, v \;\in\; V, \tag{3.15}$$

où $a(.,.)$ est une forme bilinéaire, continue et coercive tandis que $l(.)$ est une forme linéaire continue et V est un espace de Hilbert. Nous sommes donc sous les hypothèses du Théorème de Lax-Milgram qui assure l'existence et l'unicité d'une solution $u \in V$.

La méthode de Galerkine conforme consiste alors à rechercher une solution u_h dans un sous-espace vectoriel de dimension finie $V_h \subset V$. Le problème discret s'écrit alors : trouver $u_h \in V_h$ tel que

$$a(u_h, v_h) \;=\; l(v_h), \qquad \forall\, v_h \;\in\; V_h, \tag{3.16}$$

Supposons que V_h est de dimension N_x, nous pouvons alors trouver une base $\{\varphi_1,\ldots,\varphi_{N_x}\}$ de V_h. Ensuite, en écrivant le vecteur solution u_h dans cette base, nous sommes amenés à rechercher $(u_1, \ldots, u_{N_x}) \in \mathbb{R}^{N_x}$ tel que

$$u_h(x) = \sum_{i=1}^{N_x} u_i\, \varphi_i(x).$$

Ainsi, en prenant successivement $v_h = \varphi_i$ pour $i = 1, \ldots, N_x$ nous avons

$$\sum_{j=1}^{N_x} u_j\, a(\varphi_j, \varphi_i) = l(\varphi_i).$$

La formulation variationnelle conduit donc à résoudre le système

$$AU = F \tag{3.17}$$

avec la matrice $A \in \mathcal{M}_{N_x,N_x}(\mathbb{R})$, composée des coefficients

$$a_{i,j} = a(\varphi_j, \varphi_i), \quad (i,j) \in \{1, \ldots, N_x\}^2,$$

le vecteur $U = (u_1, \ldots, u_{N_x})^T \in \mathbb{R}^{N_x}$ et $F = (l(\varphi_1), \ldots, l(\varphi_{N_x}))^T \in \mathbb{R}^{N_x}$.

Théorème 3.1 *Supposons que $a(.,.)$ soit une forme bilinéaire continue, coercive sur un espace de Hilbert de dimension finie V_h et $l(.)$ une forme linéaire continue sur V_h. Alors, le système (3.17) est équivalent à la formulation variationnelle discrète (3.16) et admet une solution unique.*

L'étude de la convergence de la méthode de Galerkine dans le cas d'une approximation conforme découle du Lemme de Céa suivant

Lemme 3.1 (Lemme de Céa) *Soit $u \in V$ la solution de (3.15) et $u_h \in V_h$ le solution de (3.16) avec $V_h \subset V$. Alors, nous avons*

$$\|u - u_h\| \leq C \inf_{v_h \in V_h} \|u - v_h\|$$

Démonstration. Nous avons d'une part

$$a(u, v) = l(v), \qquad \forall\, v \in V,$$

et d'autre part

$$a(u_h, v_h) = l(v_h), \qquad \forall\, v_h \in V_h.$$

Comme $v_h \subset V$, nous pouvons prendre $v = v_h$ dans la première égalité et faire la différence. Il vient alors

$$a(u - u_h, v_h) = 0, \qquad \forall\, v_h \in V_h.$$

Ensuite, nous avons pour tout $v_h \in V_h$

$$a(u - u_h, u - u_h) = a(u - u_h, u - v_h + v_h - u_h) = a(u - u_h, u - v_h)$$

puisque $u_h - v_h \in V_h$ et $a(u - u_h, u_h - v_h) = 0$. Enfin, en utilisant la coercivité et la continuité de $a(.,.)$; nous obtenons

$$\begin{aligned} \alpha \|u - u_h\|^2 &\leq a(u - u_h, u - u_h) \\ &\leq a(u - u_h, u - v_h) \\ &\leq M \|u - u_h\| \|u - v_h\|. \end{aligned}$$

D'où

$$\|u - u_h\| \leq \frac{M}{\alpha} \|u - v_h\|, \qquad \forall\, v_h \in V_h.$$

□

3.2 Cas de la dimension une

Une des idées essentielles pour la constrction d'une approximation de Galerkine efficace est d'avoir une matrice A la plus creuse possible, c'est-à-dire avec $a(\varphi_i, \varphi_j) = 0$ pour un grand nombre de couples (i, j), ce qui permettra de diminuer le nombres d'operations à effectuer pour la résolutin du système linéaire. Nous essayons de prendre des focntions de bases avec un petit support. Nous revenons au problème

$$\begin{cases} -u''(x) \;=\; f(x), & x \in]0,1[\\ \\ u(0) \;=\; u(1) \;=\; 0, \end{cases} \tag{3.18}$$

dont la formulation variationnelle s'écrit : trouver $u \in H_0^1(]0,1[)$ tel que

$$\int_0^1 u'(x)\, v'(x)\, dx \;=\; \int_0^1 f(x)\, v(x)\, dx, \quad v \in H_0^1(]0,1[).$$

Nous recherchons alors un espace $V_h \subset H_0^1(]0,1[)$. Pour cela, nous introduisons un maillage $(x_0, x_1, \ldots, x_{N_x+1})$ de l'intervalle $[0,1]$ tel que $x_i = i\,h$, avec $h = 1/(N_x+1)$ et construisons alors pour $i = 0, \ldots, N_x + 1$

$$\varphi_i(x) \;=\; \begin{cases} \dfrac{x - x_{i-1}}{x_i - x_{i-1}}, & x \in [x_{i-1}, x_i] \\ \dfrac{x - x_{i+1}}{x_i - x_{i+1}}, & x \in [x_i, x_{i+1}] \\ \\ 0, \text{ sinon.} \end{cases}$$

Nous définissons alors $V_h = vect\{\varphi_1, \ldots, \varphi_{N_x}\}$.

Proposition 3.1 *L'espace V_h est un sous-espace vectoriel de $H_0^1(]0,1[)$ de dimension N_x*

Démonstration. Les fonctions $(\varphi_i)_{1\le i\le N_x}$ forment un systèmes libres de N_x vecteurs fonctions. En effet, supposons qu'il existe $(\lambda_i)_{1\le i\le n_x} \subset I\!R$ tels que

$$0 \;=\; \sum_{i=1}^{N_x} \lambda_i\, \varphi_i(x), \quad \forall\, x \in [0,1].$$

Alors en particulier au point x_i pour n'importe quel point $i \in \{1, ldots, N_x\}$ et donc $\lambda_i = 0$. Il en résulte que $(\varphi_i)_{1\le i\le N_x}$ forment une famille libre et génératrice de V_h, c'est donc une base et $dim(V_h) = N_x$.

Il nous reste à démontrer que V_h est un sous-espace vectoriel de $H_0^1(]0,1[)$. Nous vérifions facilement que pour tout $v_h \in V_h$, nous avons $v_h \in L^1(]0,1[)$, puisque

$$\int_0^1 v_h^2(x)\, dx \;\le\; \max_{x\in[0,1]} |v_h(x)|^2 \;<\; \infty.$$

Nous vérifions ensuite que $v_h'(x)$ appartient à l'espace $L^2(]0,1[)$. En effet, la fonction v_h est linéaire par morceau et n'est pas dérivable aux points x_i pour tout $0 \leq i \leq N_x + 1$; il faut donc calculer v_h' au sens des distribution : pour toute fonction $\varphi \in C^\infty([0,1], I\!R)$ à support compact

$$< v_h', \varphi >= - < v_h, \varphi' > .$$

Ainsi $v_h \in H^1(]0,1[)$.

Enfin, puisque $v_h(0 - v_h(1)$, nous avons donc le résultat $v_h \in H_0^1(]0,1[)$. □

Nous démontrons finalement la convergence de la méthode des éléments finis. Pour cea, nous introduisons l'opérateur de projection de $H_0^1(]0,1[)$ sur l'espace V_h : pour tout $v \in H_0^1(]0,1[)$

$$\Pi_h v = \sum_{i=1}^{N_x} v(x_i)\, \varphi_i(x).$$

Nous avons alors

Théorème 3.2 *Supposons que la solution exacte u de (3.18) appartient à l'epsace $H^2(]0,1[)$. Alors, nous avons l'estimation d'erreur suivante*

$$\|u - u_h\|_{H^1} \;\leq\; C\, h,$$

où $u_h \in V_h$ est la solution du problème approchée par la méthodes des éléments finis.

Démonstration. Nous avon d'abord

$$\|u - u_h\|_{H^1}^2 \;=\; \|u - u_h\|_{L^2}^2 \;+\; \|u' - u_h'\|_{L^2}^2.$$

En appliquant le Lemme de Céa (Lemme 3.1), nous obtenons

$$\|u - u_h\|_{H^1} \;\leq\; C \inf_{v_h \in V_h} \|u - v_h\|_{H^1} \;\leq\; C\, \|u - \Pi_h u\|_{H^1}.$$

Il suffit donc d'estimer $\|u - \Pi_h u\|_{H^1}$. Pour $x \in [x_i, x_{i+1}]$, nous avons donc

$$\begin{aligned} \Pi_h u(x) &= u(x_i)\, \varphi_i(x) \;+\; u(x_{i+1})\, \varphi_{i+1}(x), \\ &= u(x_i)\, \frac{x - x_{i+1}}{x_i - x_{i+1}} \;+\; u(x_{i+1})\, \frac{x - x_i}{x_{i+1} - x_i}, \end{aligned}$$

Or, nous avons

$$(u - \Pi_h u)(x_i) \;=\; 0$$

et donc

$$\begin{aligned} (u - \Pi_h u)(x) &= \int_{x_i}^{x} (u - \Pi_h u)'(y)\, dy \\ &= \int_{x_i}^{x} u'(y) - \frac{u(x_{i+1}) - u(x_i)}{x_{i+1} - x_i}\, dy. \end{aligned}$$

Ensuite, puisque $u \in H^2(]0,1[)$, cela implique que $u \in C^1([0,1], I\!R)$ et donc d'après la formule des accrissements finis, il existe $\xi \in [x_i, x_{i+1}]$ tel que

$$u'(\xi) \;=\; \frac{u(x_{i+1}) - u(x_i)}{x_{i+1} - x_i}.$$

et donc pour $y \in [x_i, x]$ et en utilisant l'inégalité de Cauchy-schwartz, nous obtenons

$$u'(y) \;-\; \frac{u(x_{i+1}) - u(x_i)}{x_{i+1} - x_i} \;=\; \int_\xi^y u''(s)\, ds \;\leq\; \left(\int_{x_i}^{x_{i+1}} |u''(s)|^2\, ds \right)^{1/2} h^{1/2}.$$

Ainsi, nous avons pour tout $x \in [x_i, x_{i+1}]$

$$|u'(x) \;-\; (\Pi_h u)'(x)|^2 \;\leq\; h \int_{x_i}^{x_{i+1}} |u''(s)|^2\, ds$$

puis

$$\int_{x_i}^{x_{i+1}} |u'(x) \;-\; (\Pi_h u)'(x)|^2 dx \;\leq\; h^2 \int_{x_i}^{x_{i+1}} |u''(s)|^2\, ds$$

et en sommant pour $i = 0, \ldots, N_x$

$$\int_0^1 |u'(x) - (\Pi_h u)'(x)|^2\, dx \;\leq\; h^2 \int_0^1 |u''(s)|\, dx,$$

ce qui signifie que

$$\|u' - \Pi_h u'\|_{L^2} \;\leq h\, \|u''\|_{L^2}.$$

À l'aide de l'inégalité de Poincaré, nous obtenons le résultat

$$\|u - \Pi_h u\|_{H^1} \;\leq\; C \|u''\|_{L^2}\, h.$$

□

4 Les équations d'évolution

Dans cette partie, nous nous intéressons à des équations aux dérivées partielles d'évolution, c'est-à-dire dépendant du temps $t \in [0,T]$, où T représente le temps final. Nous considérons l'équation d'évolution suivante pour $u : [0,T] \to H$ avec H un espace fonctionnel,

$$\left\{ \begin{array}{l} \dfrac{\partial u}{\partial t} \;=\; A\, u + f, \\ \\ u(t=0) \;=\; u_0 \end{array} \right. \tag{4.19}$$

En intégrant sur un intervalle de temps $[t, t+\Delta t]$, nous avons

$$u(t+\Delta t,.) \;=\; \mathcal{L}\, u(t,.),$$

avec $\mathcal{L}$ un opérateur de dimension infinie

$$\mathcal{L}\, u(t,.) = u(t,.) \;+\; \int_t^{t+\Delta t} A\, u(s,.)\, ds + \Delta t\, f(.).$$

Nous disons que ce problème est bien posé dans V lorsque

$$\|u(t)\|_V \;\leq\; C\, \|u_0\|_V \;\exp(C\, t),$$

c'est-à-dire que la solution dépend continûment de la donnée initiale.

Exemple 4.1 *Nous prenons $V = L^2(I\!R)$ et*

$$A\, u \;=\; \frac{\partial^2 u}{\partial x^2}, \quad f = 0$$

nous avons alors

$$\|u(t)\|_{L^2} \;\leq\; \|u_0\|_{L^2} \;+\; t\, \|f\|_{L^2}.$$

4.1 Notion de convergence, consistance et stabilité

Pour discrétiser cette équation d'évolution linéaire, nous commençons par discrétiser l'intervalle de temps $[0, T]$ et l'espace $[a, b]$ en subdivisions

$$t^n \;=\; n\, \Delta t, \quad n = 0, \ldots, N_T,$$

avec $\Delta t = T/N_T$ et

$$x_i \;=\; a \;+\; i\, h, \quad i = 0, \ldots, N_x + 1,$$

avec $h = (b-a)/(N_x+1)$.

Nous écrivons le schéma aux différences finis sous la forme

$$U^{n+1} \;=\; \mathcal{L}_{h,\Delta t}\, U^n, \tag{4.20}$$

où $U^n \in I\!R^{N_x}$ et construisons une solution approchée pour $(i,n) \in \{0, \ldots, N_x\} \times \{1, \ldots, N_T\}$

$$U_{h,\Delta t}(t,x) = U_i^n, \quad t \in [t^n, t^{n+1}[\times[x_i, x_i + h]$$

Nous voulons démontrer que la solution approchée se rapproche de la solution exacte lorsque les paramètres de discrétisation h et Δt convergent vers zéro. Pour cela, nous donnons d'abord la définition de la convergence, comme dans la partie précédente

Définition 4.1 (Convergence) *Considérons l'équation aux dérivées partielles (5.27). Nous disons que le schéma numérique est convergent lorsque nous avons pour* $\Delta t = T/N_T$

$$\lim_{\substack{\Delta t \to 0 \\ h \to 0}} \max_{0 \leq t \leq T} \|U_{h,\Delta t}(t,.) - u(t,.)\| = 0,$$

où $\|.\|$ *est une norme de* V.

D'autre part, pour $p, q \in \mathbb{N}$*, le schéma est dit convergent d'ordre* p *en temps et* q *en espace s'il existe une constante* $C > 0$ *ne dépendant pas de* Δt *et* h *telle que*

$$\max_{0 \leq t \leq T} \|U_{h,\Delta t}(t,.) - u(t,.)\| \leq C \left(\Delta t^p + h^q\right).$$

Comme nous l'avons vu pour l'analyse d'erreur des problèmes sationnaires, en utilisant la convergence des méthodes de quadratures, il suffit d'étudier l'erreur pour une norme discrète. Pour cela, nous formons le vecteur $\mathcal{U}(t)$ en fonction des valeurs de $u(t,.)$ aux points x_i,

$$\mathcal{U}(t^n) = (u(t^n, x_1), \ldots, u(t^n, x_{N_x}))^T. \tag{4.21}$$

Ensuite, nous posons $v = \mathcal{U}(t^n) - U^n$ et choisissons la norme

$$\|v\| := \|v\|_h = \left(\sum_{i=0}^{N_x} h\,|v_i|^2\right)^{1/2}$$

ou bien

$$\|v\| := \|v\|_\infty = \max_{0 \leq i \leq N_x} |v_i|.$$

Nous définissons ensuite l'**erreur de consistance** ou de troncature $R(t^n, \Delta t, h)$. Soit $u(.,.)$ la solution exacte du problème aux limites (4.19) ; nous posons alors l'erreur de consitance correspondant au schéma (4.20)

$$R(t^n, \Delta t, h) = \frac{\mathcal{U}(t^{n+1}) - \mathcal{L}_{h,\Delta t}(\mathcal{U}(t^n))}{\Delta t}, \tag{4.22}$$

Définition 4.2 (Consistance) *Le schéma (4.20) est dit consistant lorsque l'erreur de troncature* $R(t^n, \Delta t, h)$ *tend vers zéro lorsque* h *et* Δt *tendent vers zéro. De plus, nous disons que le schéma es consistant d'ordre* p *en temps et* q *en espace (où* $p, q \in \mathbb{N}$*) lorsqu'il existe une contant* $C > 0$ *tell que*

$$\max_{0 \leq n \leq N_T} \|R(t^n, \Delta t, h)\| \leq C \left(\Delta t^p + h^q\right)$$

La dernière notion importante pour l'étude théorique des schéma aux différeces finis (4.20) est la **stabilité** de la solution numérique.

Définition 4.3 *Nous disons que le schéma est stable s'il existe des constante* K *et* τ *indépendante de* Δt *et* h *telles que pour tout* $\Delta t \in]0, \tau[$

$$\|U^n\| \leq K\,\|U^0\|, \quad n \geq 0.$$

où $\|.\|$ *est une norme discrète* $\mathbb{R}^{N_x}$

4.2 La stabilité au sens de Von Neumann

L'espace $L^2(I\!R)$ des fonctions carrées fournit un bon cadre pour l'étude de nombreux problèmes modélisés par les équations aux dérivées partielles. La norme associée à cet espace a souvent une interprétation physique en terme d'énergie. Nous considérons donc l'espace $L^2(I\!R)$ des fonctions $u : I\!R \to \mathbb{C}$ muni du produit scalaire

$$(u, v) = \int_{I\!R} u(x)\, \overline{v}(x)\, dx.$$

Nous définissons la transformation de Fourier $\mathcal{F}$, une application de $L^2(I\!R)$ dans $L^2(I\!R)$ telle que

$$u \in L^2(I\!R) \quad \to \quad \hat{u} \in L^2(I\!R)$$

avec

$$\hat{u}(\xi) := \int_{I\!R} u(x)\, e^{2\pi\, i\, \xi\, x}\, dx.$$

Nous donnons les propriétés classiques pour toutes fonctions $u, v \in L^2(I\!R^d)$

$$(u, v) = (\hat{u}, \hat{v})$$

et la formule de Plancherel

$$\|u\|_{L^2} = \|\hat{u}\|_{L^2}.$$

Puis, pour $u \in L^2(I\!R)$

$$\hat{u'}(\xi) = i\, \xi\, \hat{u}(\xi), \qquad \xi \in I\!R.$$

De plus, nous avons pour l'opérateur de translation τ_a :

$$\begin{aligned} L^2(I\!R) &\to L^2(I\!R) \\ u &\to \tau_a(u) = u(. + a), \end{aligned}$$

d'après la propriété sur la définition de la transformée de Fourier, nous obtenons

$$\hat{\tau_a u}(\xi) = e^{-2\,\pi\, i\, \xi\, a}\, \hat{u}(\xi)$$

Après ces quelques rappels, nous considérons l'équation d'évolution suivante pour $u : [0, T] \to L^2(I\!R)$

$$\left\{ \begin{array}{l} \dfrac{\partial u}{\partial t} = A\, u, \\ u(t = 0) = u_0 \end{array} \right. \tag{4.23}$$

En multipliant l'équation (4.23) par $e^{2\pi\, i\, \xi\, x}$ et en intégrant par rapport à la variable d'espace $x \in I\!R$, nous obtenons

$$\frac{\partial \hat{u}}{\partial t}(t,\xi) \;=\; L(\xi)\, \hat{u}(t,\xi),$$

qui s'écrit alors

$$\hat{u}(t,\xi) \;=\; \hat{u_0}(\xi)\, \exp(L(\xi)\, t)$$

Nous voulons étudier les propriétés de stabilité de la transformée de Fourier de la solution, nous utiliserons pour cela la proposition suivante.

Proposition 4.1 *Nous considérons l'opérateur L suivant $\xi \in I\!R \;\rightarrow\; L(\xi) \in \mathcal{L}(I\!R, I\!R)$ tel que pour tout $x \in I\!R$, nous ayons $y = L(\xi)\, x$. Alors, l'opérateur $\mathcal{L}$:*

$$u \;\in\; L^2(I\!R^d) \;\;\rightarrow\;\; \mathcal{L}u \;\in\; L^2(I\!R^d)$$

tel que $\mathcal{L}u(\xi) = L(\xi)\, \hat{u}(\xi)$ a pour norme

$$\|\mathcal{L}\| = \sup_{\xi \in I\!R} |L(\xi)|.$$

Nous avons le résultat général suivant

Proposition 4.2 *L'équation d'évolution (4.23) où A est un opérateur linéaire est bien posé dans $L^2(I\!R)$ si et seulement si pour tout $T > 0$, il existe une constante $C_T > 0$ telle que*

$$|\exp(g(\xi)t)| \;\leq\; C_T, \quad \forall\, \xi \in I\!R, \quad \forall\, t \;\in\; [0,T].$$

Voyons maintenant comment adapter cette propriété pour l'étude des schémas de différences finies. Par exemple, lorsque nous considérons le schéma d'Euler explicite et un opérateur discret A_h approchant l'opérateur continue A, nous avons

$$U^{n+1} \;=\; \mathcal{L}_{h,\Delta t}\, U^n.$$

Nous considérons le vecteur $U^n = (u_1^n, \ldots, u_{N_x}^n)^T \in I\!R^{N_x}$ et construisons une fonction $U_{h,\Delta t}$ constante par morceau telle que pour $j = 1, \ldots, N_x$

$$U_{h,\Delta t}(t,x) = U_j^n, \qquad (t,x) \;\in\; [t^n, t^n + \Delta t[\times [x_j, x_{j+1}[$$

où $[x_0,\; x_{N_x+1}]$ est un compact de $I\!R$ et $U_{h,\Delta t}(t,x) = 0$ pour $x \notin [x_0, x_{N_x+1}]$.

Nous avons alors pour $h = (x_{N_x+1}, x_0)/(N_x+1)$ et $t \in [t^n, t^n + \Delta t[$

$$\|U_{h,\Delta t}(t,.)\|_{L^2}^2 = \sum_{j=1}^{N_x} h|U_j^n|^2 \;<\; \infty.$$

Nous pouvons donc définir la transformée de Fourier de $U_{h,\Delta t}$ que nous notons $\hat{U}_{h,\Delta t}$. De plus, nous observons que pour $t \in [t^n, t^n + \Delta t[$

$$\hat{U}_{h,\Delta t}(t, \xi + h) \;=\; \hat{\tau_h}\, U_{h,\Delta t}(\xi) \;=\; e^{-2\pi\, i\, h\, \xi}\, \hat{U}_{h,\Delta t}(t, \xi).$$

Ceci signifie que

$$\hat{U}^n_{j+1} \;=\; e^{-2\pi\, i\, h}\, \hat{U}^n_j$$

et

$$\hat{U}^n_{j-1} \;=\; e^{+2\pi\, i\, h}\, \hat{U}^n_j$$

Ainsi en appliquant une transformée de Fourier dans le schéma numérique

$$U^{n+1} \;=\; \mathcal{L}_{h,\Delta t}\, U^n,$$

il vient

$$\hat{U}^{n+1}_j \;=\; S(j)\, \hat{U}^n_j, \quad j = 1, \ldots, N_x, \tag{4.24}$$

où $S(j)$ est un scalaire que nous calculons en fonction du schéma $\mathcal{L}_{h,\Delta t}$

Définition 4.4 (Stabilité au sens de Von Neumann) *Le schéma (4.20) aux différences finis*

$$U^{n+1} \;=\; \mathcal{L}_{h,\Delta t}\, U^n$$

est stable au sens de Von Neumann s'il existe une constante $K > 0$

$$\sup_{1 \le j \le N_x} |S(j)| \le 1 + K\Delta t, \quad \forall\, \Delta t \in [0, \Delta t_0],$$

où $S(j)$ est tel que $\hat{U}^{n+1}_j \;=\; S(j)\, \hat{U}^n_j$.

4.3 Théorème d'équivalence de Lax

Nous énonçons maintenant le théorème d'équivalence de Lax qui permet de montrer l'équivalence entre la stabilité et la convergence pour un schéma consistant approchant une équation d'évolution linéaire.

Théorème 4.1 (Théorème de Lax) *Soit V un espace de Banach, nous considérons un problème d'équations aux dérivées partielles linéaire : $u \;:\; t \in [0, T] \to u(t) \in V$ solution de (4.19). Nous notons $(U^n)_{0 \le n \le N_T}$ la solution approchée de $u(t^n, .)$ dans $V_h \subset V$, où h désigne le pas de discrétisation en espace par un schéma $\mathcal{L}_{h,\Delta t}$ consistant. Alors, la stabilité est une condition nécessaire et suffisante pour la convergence.*

Démonstration. Montrons d'abord que la stabilité d'un schéma consistant entraîne la convergence. Pour cela, nous écrivons l'erreur $e^n(h, \Delta t)$ donnée par

$$e^n(h, \Delta t) \;=\; \mathcal{U}(t^n) \;-\; U^n,$$

où $\mathcal{U}(t^n)$ est donné par (4.21) et U^n est la solution approchée donnée par (4.20).

Puisque le schéma est consistant, il vérifie alors

$$R(t^n, \Delta t, h) = \frac{\mathcal{U}(t^{n+1}) \, - \, \mathcal{L}_{h,\Delta t}(\mathcal{U}(t^n))}{\Delta t},$$

avec

$$\|R(t^n, \Delta t, h)\| \longrightarrow 0, \text{lorsque} \quad h, \Delta t \longrightarrow 0.$$

D'autre part, par définition du schéma

$$\frac{U^{n+1} \, - \, \mathcal{L}_{h,\Delta t}(U^n)}{\Delta t} \, = \, 0$$

donc par linéarité nous avons

$$R(t^n, \Delta t, h) = \frac{e^{n+1}(h, \Delta t) \, - \, \mathcal{L}_{h,\Delta t}(e^n(h, \Delta t))}{\Delta t},$$

ce qui signifie par récurrence

$$e^{n+1}(h, \Delta t) \, = \, \left[\mathcal{L}_{h,\Delta t}\right]^n (e^0(h, \Delta t)) + \sum_{k=0}^{n} \Delta t \, R(t^k, \Delta t, h).$$

D'une part, la consistance entraîne que

$$\left| \sum_{k=0}^{n} \Delta t \, R(t^k, \Delta t, h) \right| \, \leq \, T \max_{0 \leq k \leq N_T} \|R(t^k, \Delta t, h)\| \, \rightarrow \, 0,$$

lorsque $(h, \Delta t) \rightarrow 0$.

D'autre part, montrons qu la stabilité du schéma assure que la norme

$$\| \left[\mathcal{L}_{h,\Delta t}\right]^n \| := \sup_{v \in I\!R^{N_x}} \frac{\| \left[\mathcal{L}_{h,\Delta t}\right]^n (v)\|}{\|v\|}$$

est uniformément bornée par rapport à h et Δt. En effet, pour $v \in I\!R^{N_x}$ avec $v \neq 0$, nous avons d'après le schéma numérique et la définition de la stabilité il existe une constante $K > 0$ telle que

$$\| \left[\mathcal{L}_{h,\Delta t}\right]^n (v)\| \leq K \|v\|$$

et donc en prenant le $\sup$, il vient

$$\| \left[\mathcal{L}_{h,\Delta t}\right]^n \| \leq K$$

Finalement, nous obtenons

$$\begin{aligned} \|e^{n+1}(h, \Delta t)\| \; &= \; \| \left[\mathcal{L}_{h,\Delta t}\right]^n (e^0(h, \Delta t)) + \sum_{k=0}^{n} \Delta t \, R(t^k, \Delta t, h)\| \\ &\leq \; K \, \|e^0(h, \Delta t)\| \, + \, T \max_{0 \leq k \leq N_T} \|R(t^k, \Delta t, h)\|. \end{aligned}$$

En passat à la limite lorsque h et Δt tendent vers zéro, nous obtenons la convergence du schéma.

Pour démontrer la récurrence, nous raisonnons par l'absurde. Supposons que le schéma est convergent et dans le même temps pour tout $K > 0$, il existe deux suites $(n_k)_{k\in I\!N} \subset I\!N$ et $(\Delta t_k)_{k\in I\!N} \subset I\!R^+$ telles que $n_k\,\Delta t_k \leq T$ et

$$\lim_{k\to\infty} \| \left[\mathcal{L}_{h,\Delta t_k}\right]^{n_k} \| = \infty. \tag{4.25}$$

Tout d'abord puisque $\Delta t_k < T$, la suite $(\Delta t_k)_{k\in I\!N}$ st forcément bornée. Ainsi, si la suite $(m_k)_{k\in I\!N}$ est bornée alors

$$\sup_{k\in I\!N} \| \left[\mathcal{L}_{h,\Delta t_k}\right]^{n_k} \| \leq \sup_{k\in I\!N} \|\mathcal{L}_{h,\Delta t_k}\|^{n_k} < \infty,$$

ce qui n'est pas possible d'après (4.25).

Supposons alors que la suite $n_k \to \infty$ et $\Delta t_k \to 0$, lorsque k tend vers l'infini. Nous avons alors pour tout $U_0 \in I\!R^{N_x}$

$$U^{n_k} = \mathcal{U}(t^{n_k}) - e^{n_k}(h, \Delta t_k)$$

Puisque le problème (4.19) est bien posé, il existe une constante $C > 0$ telle que

$$\|\mathcal{U}(t^{n_k})\| \leq C\,\|U^0\|.$$

De plus, puisque le schéma est convergent, nous avons également

$$\|e^{n_k}(h, \Delta t_k)\| \longrightarrow 0,$$

lorsque $(h, \Delta t_k)$ tend vers zéro. Ainsi, il existe une constante $K > 0$ telle que

$$\|U^{n_k}\| = \| \left[\mathcal{L}_{h,\Delta t_k}\right]^{n_k} U^0\| \leq C\,\|U^0\|, \qquad \forall U^0 \in I\!R^{N_x}.$$

Ainsi,

$$\sup_{U^0\in I\!R^{N_x}} \frac{\| \left[\mathcal{L}_{h,\Delta t_k}\right]^{n_k} U^0\|}{\|U^0\|} < C, \quad k \in I\!N$$

et donc

$$\sup_{k\in I\!N} \| \left[\mathcal{L}_{h,\Delta t_k}\right]^{n_k} \| < C$$

ce qui contredit l'hypothèse (4.25). □

5 L'équation de la chaleur

Plus concrètement, nous considérerons le cas d'un fil de fer dont la section S est constante, égale à un. En supposant que la conductivité thermique du fil est constante, l'équation de diffusion de la chaleur s'écrit alors simplement

$$\frac{\partial u}{\partial t}(t,x) - \kappa \frac{\partial^2 u}{\partial x^2}(t,x) = f(x) \tag{5.26}$$

Pour calculer l'évolution de la température jusqu'à l'équilibre, il faudra tenir compte du fait que les extrémités de la barre sont maintenues à une température constante, ici égale à zéro. Il faudra aussi indiquer la température initiale pour que le problème d'évolution soit bien posé, c'est-à-dire pour qu'il admette une solution unique. Nous supposons que le fil est à la température de zéro degré lorsque nous commençons à chauffer en apportant l'énergie $f(x)$. La fonction $u(t,x)$ qui donne la température au point x à l'instant t sera alors solution de

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x) - \kappa \dfrac{\partial^2 u}{\partial x^2}(t,x) \;=\; f(x), & t \geq 0, a \;<\; x \;<\; b, \\ u(t,a) \;=\; u(t,b) \;=\; 0, & t \;\geq\; 0. \\ u(0,x) \;=\; 0, & a \;\leq\; x \;\leq\; b. \end{cases} \tag{5.27}$$

La température à l'équilibre est solution du problème

$$\begin{cases} -\kappa\, u''(x) \;=\; f(x), & a \;<\; x \;<\; b, \\ u(a) \;=\; u(b) \;=\; 0. \end{cases}$$

Pour en calculer une valeur approchée, nous choisissons un pas $h = (b-a)/(N_x+1)$ où N_x est un entier positif et nous résolvons le système linéaire du type

$$\frac{\kappa}{h^2}\, M\, U \;=\; F.$$

Les coefficients de la matrice M et les composantes du vecteur F se retrouvent dans ces équations. Dans ce système, la solution U a pour composantes les u_i qui tendent vers $u(x_i)$ lorsque h tend vers 0. La solution $u(t,x)$ du problème (5.27) dépend des deux variables t et x. Nous allons considérer que nous pouvons approcher ses valeurs $u(t,x_i)$ par des fonctions $u_i(t)$. Ces fonctions seront les composantes d'une fonction $U(t)$ à valeurs dans $\mathbb{R}^{N_x}$. En utilisant la discrétisation du système linéaire du problème elliptique, et comme ici, la fonction $f(x)$ ne dépend pas du temps, la fonction $U(t)$ sera solution du système différentiel

$$\begin{cases} \dfrac{dU}{dt}(t) \;+\; \dfrac{\kappa}{h^2} M\, U(t) \;=\; F, \quad t > 0, \\ U(0) = U_0. \end{cases} \tag{5.28}$$

La condition initiale est ici $U(0) = U_0 \in \mathbb{R}^{N_x}$.

5.1 Discrétisation de l'équation de la chaleur

Nous présentons ici différents schémas de discrétisation : la méthode de richardson, d'Euler explicite et implicite.

La méthode de Richardson Le premier schéma numérique pour résoudre l'équation de la chaleur (5.27) a été celui de Richardson (1910). Il s'agit d'un schéma centré en espace et en temps. Nous introduisons T le temps final de la simulation et $\Delta t = T/N_T$, où N_T correspond au nombre d'intervalles de temps sur lesquels nous calculons une approximation numérique.

$$\frac{U^{n+1} - U^{n-1}}{2\,\Delta t} + \frac{\kappa}{h^2} M\,U^n = F.$$

La solution numérique U^{n+1} s'écrit alors

$$U^{n+1} = U^{n-1} - \frac{2\,\kappa\,\Delta t}{h^2} M\,U^n + 2\,\Delta t\,F. \tag{5.29}$$

Hélàs la solution numérique produite par ce schéma n'est pas bornée et les calculs ne convergent pas vers la solution de l'équation de la chaleur (5.27). Ce phénomène est connu sous le nom d'instablité numérique.

La méthode d'Euler explicite Nous proposons d'abord d'utiliser la méthode explicite. En effet elle peut sembler plus simple a priori. Pour chaque valeur de $n \geq 0$, il faudra calculer :

$$U^{n+1} = \left(I_{N_x} - \frac{\kappa\,\Delta t}{h^2}\,M\right) U^n + \Delta t\,F. \tag{5.30}$$

Observons les résultats qu'elle fournit pour le pas de temps que nous avons choisit, qui est de $\Delta t = 0.1$. Le moins que nous puissions dire est qu'il y a un problème ; en fait, ce problème est raide, et la méthode explicite ne peut passer que pour un pas de temps très petit.

La méthode d'Euler implicite Le système (5.28) définit un problème du type des équations différentielles que nous avons étudié d'une point de vue numérique dans le chapitre précédent. Pour ce type de problème, nous pouvons essayer de mettre en oeuvre les deux méthodes que nous avons introduites : les méthodes d'Euler explicite et implicite.

Comme il s'agit d'un problème linéaire, nous pouvons utiliser l'une ou l'autre. Commençons par la méthode d'Euler implicite. Nous pouvons supposer que l'équilibre est atteint à l'instant $t = T_{fin}$. En choisissant un pas de temps constant $\Delta t = T_{fin}/N_T$, nous chercherons des valeurs approchées des $U(t^n)$, avec $t^n = n\,\Delta t$. Nous notons U^n ces valeurs et la méthode d'Euler implicite consiste à discrétiser (5.28) selon

$$\frac{U^{n+1} - U^n}{\Delta t} + \frac{\kappa}{h^2} M\,U^{n+1} = F.$$

Il suffit donc de calculer, pour $0 \leq n \leq N_T - 1$, les vecteurs U^{n+1} qui vérifient pour chaque valeur de n le système précédent. Nous devrons donc résoudre un système linéaire (puisque nois n'inversons jamais des matrices de grande taille !)

$$U^{n+1} = \left(I_{N_x} + \frac{\kappa\,\Delta t}{h^2}\,M\right)^{-1} (U^n + \Delta t\,F)\,. \tag{5.31}$$

5.2 Etude de la convergence pour l'équation de la chaleur.

Dans cette partie, nous souhaitons démontrer deux théorèmes de convergence pour les schémas d'Euler explicite et implicite

Théorème 5.1 *Supposons que fonction f est bornée dans L^∞ et que la solution u de (5.27) vérifie $u \in C^2([0,T], C^4([a,b]))$. Alors, le schéma d'Euler explicite (5.30) est convergent d'ordre un en temps et deux en espace.*

et

Théorème 5.2 *Supposons que fonction f est bornée dans L^∞ et que la solution u de (5.27) vérifie $u \in C^2([0,T], C^4([a,b]))$. Alors, le schéma d'Euler implicite (5.31) est convergent d'ordre un en temps et deux en espace.*

Pour démontrer ces deux théorèmes, il suffit de prouver la consistance et la stabilité des schémas d'Euler puis d'appliquer le théorème d'équivalence de Lax (Théorème 4.1).

Montrons d'abord un résultat de consistance des schémas d'Euler explicite et implicite

Proposition 5.1 *Supposons que fonction f est bornée dans L^∞ et que la solution u de (5.27) vérifie $u \in C^2([0,T], C^4([a,b]))$. Alors, le schéma d'Euler explicite (5.30) est consistant d'ordre un en temps et d'ordre deux en espace : il existe une constante $C > 0$, ne dépendant que de u, telle que*

$$\|R(t^n, \Delta t, h)\|_\infty \leq C\,(h^2 + \Delta t).$$

Démonstration. Nous supposons que la fonction f est bornée dans L^∞ et que la solution u de (5.27) appartient à $C^2([0,T], C^4([a,b]))$. D'abord par définition de la consistance, nous avons

$$R(t^n, \Delta t, h) = \frac{\mathcal{U}(t^{n+1}) - \mathcal{U}(t^n)}{\Delta t} + \frac{\kappa}{h^2} M\,\mathcal{U}(t^n) - F.$$

Ensuite, d'après un développement de Taylor, nous avons d'abord

$$\frac{\partial u}{\partial t}(t^n, x_i) = \frac{u(t^{n+1}, x_i) - u(t^n, x_i)}{\Delta t} - \frac{\Delta t}{2}\left(\frac{\partial^2 u}{\partial t^2}\right)(t^n, x_i) + o(\Delta t)$$

et

$$\frac{\partial^2 u}{\partial x^2}(t^n, x_i) = \frac{u(t^n, x_{i+1}) - 2\,u(t^n, x_i) + u(t^n, x_{i-1})}{h^2} - \frac{h^2}{12}\frac{\partial^4 u}{\partial x^4}(t^n, x_i) + o(h^2).$$

Ainsi, en utilisant que

$$\frac{\partial u}{\partial t}(t^n, x_i) - \kappa\,\frac{\partial^2 u}{\partial x^2}(t^n, x_i) = f(x_i),$$

nous obtenons

$$R_i(t^n, \Delta t, h) \;=\; \kappa \frac{h^2}{12} \frac{\partial^4 u}{\partial x^4}(t^n, x_i) + \;\frac{\Delta t}{2}\left(\frac{\partial^2 u}{\partial t^2}\right)(t^n, x_i) + o(\Delta t) + \;o(h^2).$$

Lorsque Δt et h tendent vers zero, l'erreur de troncature $R(\Delta t, h)$ tend aussi vers zero ; le schéma est donc consistant. □

De la même manière, nous démontrons que le schéma d'Euler implicite est inconditionellement stable.

Proposition 5.2 *Supposons que fonction f est bornée dans L^∞ et que la solution u de (5.27) vérifie $u \in C^2([0,T], C^4([a,b]))$. Alors, le schéma d'Euler implicite (5.31) est consistant d'ordre un en temps et d'ordre deux en espace : il existe une constante $C > 0$, ne dépendant que de u, telle que*

$$\|R(t^n, \Delta t, h)\|_\infty \;\leq\; C\,(h^2 \;+\; \Delta t).$$

Nous nous intéressons ensuite à la stabilité des schémas d'Euler explicite et implicite. D'abord pour le schéma d'Euler explicite, nous avons

Proposition 5.3 *Si le pas de temps satisfait la condition de type CFL suivante*

$$\kappa\,\Delta t < \frac{h^2}{2}.$$

Alors, le schéma d'Euler explicite (5.30) est stable pour la norme L^∞

$$\|U^n\|_\infty \;\leq\; \|U^0\|_\infty + t^n\,\|f\|_{L^\infty}.$$

Démonstration. Pour tout $i \in \{1, \ldots, N_x + 1\}$, nous avons

$$u_i^{n+1} = \left(1 - \frac{2\,\kappa\,\Delta t}{h^2}\right) u_i^n + \frac{\kappa\,\Delta t}{h^2}\left(u_{i+1}^n + u_{i-1}^n\right) + \Delta t\, f(x_i)$$

Or, en prenant la valeur absolue et puisque sous la condition CFL,

$$0 < 1 - \frac{2\,\kappa\,\Delta t}{h^2} < 1,$$

il vient

$$|u_i^{n+1}| \leq \left(1 - \frac{2\,\kappa\,\Delta t}{h^2}\right)|u_i^n| + \frac{\kappa\,\Delta t}{h^2}\left(|u_{i+1}^n| + |u_{i-1}^n|\right) \;+\; \Delta t\,|f(x_i)|.$$

Puis en prenant le maximum $\max_{0\leq i\leq N_x+1}|u_i^n|$, nous obtenons que pour tout $i \in \{1, \ldots, N_x\}$

$$|u_i^{n+1}| \leq \max_{0\leq j\leq N_x+1}|u_j^n| \;+\; \Delta t\,\|f\|_{L^\infty}.$$

D'où le résultat

$$\|U^{n+1}\|_\infty \leq \|U^0\|_\infty + t^{n+1}\,\|f\|_{L^\infty}.$$

□

Remarque 5.1 *Dans le cas d'un schéma d'Euler explicite, nous pouvons aussi étudier la stabilité au sens de Von Neumann, nous montrons alors que le coefficient $S(j)$ défini par (4.24) est donné par*

$$S(j) = 1 - 4\,\lambda\,\sin^2\left(\frac{j\,h}{2}\right)$$

avec $\lambda = \kappa\,\Delta t/h^2$, ce qui signifie que le schéma est stable au sens de Von Neumann dès que $\lambda < 1/2$.

Ensuite, pour le schéma d'Euler implicite, nous avons

Proposition 5.4 *Le schéma d'Euler implicite (5.31) est inconditionellement stable pour la norme L^∞*

$$\|U^n\|_\infty \;\leq\; \|U^0\|_\infty + t^n\,\|f\|_{L^\infty}.$$

Démonstration. Pour tout $i \in \{1, \ldots, N_x + 1\}$, nous avons

$$-\frac{\kappa\Delta t}{h^2}\left(u_{i+1}^{n+1} + u_{i-1}^{n+1}\right) + \left(1 + \frac{2\,\kappa\Delta t}{h^2}\right) u_i^{n+1} = u_i^n \;+\; \Delta t\, f(x_i)$$

ou encore

$$\left(1 + \frac{2\,\kappa\Delta t}{h^2}\right) u_i^{n+1} = u_i^n \;+\; \Delta t\, f(x_i) \;+\; \frac{\kappa\Delta t}{h^2}\left(u_{i+1}^{n+1} + u_{i-1}^{n+1}\right).$$

Or, puisque $1+\frac{2\,\kappa\Delta t}{h^2} > 0$, nous pouvons prendre la valeur absolue et le maximum sur $(u_i^n)_{1\leq i\leq N_x}$ puis sur $(u_i^{n+1})_{1\leq i\leq N_x}$, il vient

$$\|U^{n+1}\|_\infty \leq \|U^n\|_\infty + \Delta t\|f\|_{L^\infty}$$

et donc pour tout $n \geq 0$

$$\|U^n\|_\infty \leq \|U^0\|_\infty + t^n\|f\|_{L^\infty}.$$

□

Remarque 5.2 *Pour un schéma d'Euler implicite, nous pouvons aussi étudier la stabilité au sens de Von Neumann, nous montrons alors que le coefficient $S(j)$ défini par (4.24) est donné par*

$$S(j) = \frac{1}{1 + 4\,\lambda\,\sin^2\left(\frac{j\,h}{2}\right)}$$

ce qui signifie que le schéma d'Euler implicite est inconditionnellement stable au sens de Von Neumann.

6 L'équation des ondes

6.1 Motivation

Pour commencer par un exemple simple, nous considérons un système de N_x billes de masse identique reliées entre elles par des ressorts identiques et au repos. Pour étudier les mouvements possibles de ces billes, nous notons u_1, u_2, ..., u_{N_x} la position de chaque bille relative a la position d'équilibre. Les équations du mouvement sont alors données par

$$m\frac{d^2u_i}{dt^2} \;=\; k\,(u_{i+1} - 2\,u_i + u_{i-1})$$

où $i = 1, \ldots, N_x$ et $u_0 = u_{N_x+1} = 0$ et m est la masse d'une bille et k la constante de rappel des ressorts. Nous posons alors $u = (u_1, \ldots, u_{N_x})^T \in I\!R^{N_x}$, ces équations s'écrivent alors

$$\frac{d^2u}{dt^2} \;=\; A\,u$$

avec

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

Nous constatons que la matrice A est symétrique définie négative et pouvons donc la diagonaliser en l'écrivant sous la forme $D \;=\; R^{-1}\,A\,R$ avec $D \;=\; diag(-\omega_i^2)$, nous obtenons donc pour $v = R^{-1}u$

$$\frac{d^2v_i}{dt^2} \;=\; -\,\omega_i^2\,v_i, \qquad i \;=\; 1, \ldots, N_x.$$

Ces équations peuvent être résolues de manière explicites pour v_i

$$v_i(t) \;=\; [\lambda_i\,\cos(\omega_i\,t) \;+\; \mu_i\,\sin(\omega_i\,t)]\,v_i^0$$

ou encore

$$v_i(t) = [\lambda_i\,\cos(\omega_i\,t + \varphi_i)]\,v_i^0.$$

Ainsi le mouvement général des billes s'obtient en superposant des mouvements particuliers, c'est-à-dire que l'on décompose la donnée initiale dans la base des vecteurs propres de la matrice A.

Maintenant, en laisant tendre le nombre de billes vers l'infini, nous obtenons une modèle pour décrire des petites variations d'une barre fixée aux extrémités. Cette fois-ci, nous étudions

les variations d'une fonction à deux variables $u(t,x)$ où t représente toujours le temps et $x \in [0,l]$ est la variable d'espace. À la limite, le terme $A\,u$ tend "formellement" vers $\frac{\partial^2 u}{\partial x^2}$ et nous obtenons l'**équation des ondes**

$$\frac{1}{c^2}\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}$$

avec les conditions aux limites $u(t,0) = u(t,l) = 0$. Pour étudier ce problème, nous pouvons aussi nous ramener, comme dans le cas de la dimension finie, à la résolution d'un problème aux valeurs propres en dimension infinie. Nous recherchons (v,λ) où v est une fonction de x et $\lambda \in \mathbb{R}$ telles que

$$\frac{d^2 v}{dx^2} = \lambda v$$

Compte tenu des conditions aux limites, les solutions non triviales (non nulles) s'écrivent

$$\lambda_k = -\left(\frac{k\,\pi}{l}\right)^2, \quad k \in \mathbb{N}$$

et $v_k(x) = \sin(k\,\pi\,x/l)$.

À chaque vecteur propre v_k, nous recherchons la solution $u(t,x)$ de l'équation des ondes sous la forme

$$u(t,x) \;=\; h_k(t)\,v_k(x),$$

ce qui donne

$$h_k(t) = A\,\cos\left(\frac{c\,k\,\pi\,t}{l}\right) \;+\; B\,\sin\left(\frac{c\,k\,\pi\,t}{l}\right),$$

c'est-à-dire les vibrations ont la fréquence propre $k/2l$.

Nous nous attendons alors par analogie avec la dimension finie que le mouvement général de la barre s'écrit comme la superposition des mouvements propres, c'est-à-dire en décomposant n'importe quelle donnée initiale $u_0(x)$ comme

$$u_0(x) = \sum_{k\in\mathbb{N}} b_k\,v_k(x).$$

6.2 Discrétisation de l'équation des ondes

Comme la corde est fixée à ses deux extrémités, nous pouvons facilement donner des conditions aux limites pour accompagner l'équation des ondes, mais nous ne pourrons calculer un déplacement $u(t,x)$ qu'à la condition que nous fournissions aussi une position initiale $u(0,x)$ et une vitesse initiale $\frac{\partial u}{\partial t}(0,x)$. Le problème à résoudre est donc

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2}(t,x) \;-\; c^2\,\dfrac{\partial^2 u}{\partial x^2}(t,x) \;=\; 0 & (t,x) \in \mathbb{R}^+ \times (a,b) \\ u(t,a) \;=\; u(t,b) \;=\; 0, & t \in \mathbb{R}^+ \\ u(0,x) \;=\; u_0(x),\; \frac{\partial u}{\partial t}(0,x) \;=\; v_0(x), & x \in (a,b) \end{cases} \tag{6.32}$$

La constante c qui apparaît dans cette équation désigne la vitesse de propagation de l'onde dans cette corde. Cette vitesse est déterminée par $c^2 = T/\rho$, où T désigne la tension et ρ la masse linéique de la corde.

Ayant fixé comme précédemment un pas $h = (b-a)/(N_x+1)$ et choisi les point $x_i = a+ih$ du segment $[a, b]$ pour chercher des valeurs approchées de $u(t, x)$ aux points x_i, nous pouvons poser $u(t, x_i) = u_i(t)$. Ces fonctions sont les composantes d'un vecteur $U(t)$ dans $\mathbb{R}^{N_x}$, et la fonction $U(t)$ est solution du système différentiel :

$$\left\{ \begin{array}{l} \dfrac{d^2U}{dt^2}(t) \,=\, M\,U(t) \,+\, F, \quad t > 0, \\ U(0) \,=\, U_0,\, \dfrac{dU}{dt}(0) \,=\, V_0. \end{array} \right. \tag{6.33}$$

La matrice M est la matrice de discrétisation de l'équation

$$-c^2\, u''(x) \,=\, f(x)$$

aux points x_i du segment $[a, b]$ avec les conditions de Dirichlet $u(a) = u(b) = 0$.

Il s'agit d'un système différentiel à valeurs initiales :

- U_0 est le vecteur des positions initiales, de composantes les $u_0(x_i)$,
- V_0 est le vecteur des vitesses initiales, de composantes les $v_0(x_i)$.

C'est un système du second ordre. Nous allons nous inspirer de ce que nous avons fait au chapitre précédent, par exemple pour le problème du pendule, pour le transformer en un système du premier ordre.

Nous posons alors $V(t) = U'(t)$ et introduisons la fonction

$$Y(t) = \left(\begin{array}{c} U(t) \\ \\ V(t) \end{array} \right).$$

La fonction $Y(t)$ est à valeurs dans $\mathbb{R}^{2n}$; elle est solution du système différentiel :

$$Y'(t) = \left(\begin{array}{cc} 0_{N_x} & I_{N_x} \\ \\ M & 0_{N_x} \end{array} \right) Y(t),$$

où M est la matrice discrétisant l'operateur de la dérivée seconde donnée par (2.6).

Mise en oeuvre de la méthode d'Euler explicite

Nous choisissons l'intervalle d'étude $[0, T]$ et fixons à N_T le nombre de pas de temps. Nous posons $\Delta t = T/N_T$ et pouvons alors écrire la méthode à l'instant t^n en posant :

$$Y^{n+1} = Y^n + \Delta t \begin{pmatrix} 0_{N_x} & I_{N_x} \\ M & 0_{N_x} \end{pmatrix} Y^n.$$

En séparant les blocs de taille N_x, nous obtenons

$$\begin{cases} U^{n+1} = U^n + \Delta t\, V^n \\ V^{n+1} = V^n - \Delta t\, M\, U^n. \end{cases}$$

Nous pouvons aussi l'écrire en explicitant la discrétisation en espace, et avons alors, pour $1 \leq i \leq N_x$

$$\begin{cases} u_i^{n+1} = u_i^n + \Delta t\, v_i^n \\ v_i^{n+1} = v_i^n + \dfrac{c^2\,\Delta t}{h^2} \left(u_{i+1}^n - 2\,u_i^n + u_{i-1}^n\right). \end{cases}$$

Méthode d'Euler implicite

Nous pouvons aussi l'écrire par blocs, sous la forme :

$$\begin{cases} U^{n+1} = U^n + \Delta t\, V^{n+1} \\ V^{n+1} = V^n - \Delta t\, M\, U^{n+1}. \end{cases}$$

Elle consiste donc à résoudre le système de taille $2N_x \times 2N_x$ dont les inconnues sont U^{n+1} et V^{n+1}

$$\begin{cases} U^{n+1} - \Delta t\, V^{n+1} = U^n \\ V^{n+1} + \Delta t\, M\, U^{n+1} = V^n. \end{cases}$$

Pour mettre en oeuvre ce schéma, nous utilisons cette forme matricielle, en procédant à une élimination de Gauss par blocs. Il s'agit de s'inspirer de ce que nous ferions pour un système de 2 équations linéaire à 2 inconnues. Cela consiste, tout simplement :

- à multiplier le premier blocs d'équations par $-\Delta t\, M$,
- à ajouter ce nouveau premier bloc au second, pour éliminer le vecteur inconnu U^{n+1} du second bloc d'équations.

Nous résolvons alors, en cascade :

$$\begin{cases} (I_{N_x} + \Delta t^2\, M)\; V^{n+1} = V^n - \Delta t^2\, M\, U^n \\ U^{n+1} = U^n + \Delta t\, V^{n+1}. \end{cases}$$

Nous n'avons plus de problème de stabilité.

Cependant, c'est un autre phénomène qui est mis en évidence. Il s'agit d'un amortissement de l'onde : cet amortissement est purement numérique, car l'équation des ondes ne conduit pas à cet amortissement (nous disons qu'elle est conservative). Si le mouvement d'une corde vibrante effectivement observé finit par s'amortir, c'est à cause de termes qui ne sont pas pris en compte par ce modèle simplifié : résistance du milieu ambiant, par exemple.

Nous pouvons aussi écrire ce schéma en explicitant la discrétisation en espace. Nous avons alors, pour $1 \leq i \leq N_x$:

$$\begin{cases} u_i^{n+1} = u_i^n + \Delta t\, v_i^{n+1} \\ v_i^{n+1} = v_i^n + \dfrac{c^2\, \Delta t}{h^2} \left(u_{i+1}^{n+1} - 2\, u_i^{n+1} + u_{i-1}^{n+1}\right). \end{cases}$$

Une méthode inconditionnellement stable sans amortissement

Pour discrétiser l'équation des ondes, nous pouvons discrétiser :

- $\frac{\partial^2 u}{\partial t^2}$ au point x_i à l'instant t^n en utilisant la forme

$$\frac{1}{\Delta t^2} \left(u_i^{n+1} - 2\, u_i^n + u_i^{n-1}\right).$$

- $c^2\, \frac{\partial^2 u}{\partial x^2}$ au point x_i à l'un des instants qui apparaissent ci-dessus.
  - si nous choisissons l'instant t^n, nous obtenons le schéma explicite,
  - si nous choisissons l'instant t^{n+1}, nous obtenons le schéma d'Euler implicite,

Nous pouvons aussi combiner ces expressions. Le schéma "saute-mouton" va prendre la moyenne des évaluations de $c^2 \frac{\partial^2 u}{\partial x^2}$ aux instants t^{n+1} et t^{n-1}. En explicitant la discrétisation en espace, nous obtenons, pour $1 \leq i \leq n$:

$$\begin{aligned} \frac{1}{\Delta t^2} \left(u_i^{n+1} - 2\, u_i^n + u_i^{n-1}\right) &= \frac{c^2}{2\, h^2} \left(u_{i+1}^{n+1} - 2\, u_i^{n+1} + u_{i-1}^{n+1}\right) \\ &+ \frac{c^2}{2\, h^2} \left(u_{i+1}^{n-1} - 2\, u_i^{n-1} + u_{i-1}^{n-1}\right). \end{aligned}$$

Sa mise en oeuvre dans les conditions précédentes fournit le résultat ilustré par la figure. En utilisant la notation matricielle, on résoudra les systèmes :

$$\begin{cases} U^{n+1} = U^n + \Delta t\, V^{n+1} \\ V^{n+1} = V^n - \dfrac{\Delta t}{2h^2}\, M\, U^{n+1} - \dfrac{\Delta t}{2h^2}\, M\, U^{n-1}. \end{cases}$$

Bibliographie

[1] G. Allaire Analyse numérique et optimisation *les Éditions de l'École Polytechnique,* (2005)

[2] K. Chemla et S. Guo, Les neuf chapitres : le classique mathématique de la Chine ancienne et ses commentaires, *Paris, Dunod*, (2004)

[3] Ph. G. Ciarlet Introduction à l'analyse matricielle et à l'optimisation numérique, *Masson* (1982)

[4] J. W. Cooley et J. W. Tukey, An algorithm for the machine calculation of complex Fourier series, *Math. Comput.* **19**, pp. 297–301 (1965).

[5] M. Crouzeix et A.L. Mignot, Analyse numérique des équations différentielles *Masson* (1984)

[6] L. Dumas Modélisation à l'oral de l'agrégation. Calcul scientifique. *Ellipses* (1999)

[7] G. Evans, Practical numerical integration, *John Willey & Sons* (1993)

[8] J.-P. Ferrier, Mathématiques pour la licence, variable complexe, calcul différentiel et tensoriel, espaces normés et calcul intégral, analyse de Fourier. *Masson* (1984)

[9] C. F. Gauss, Nachlass : Theoria interpolationis methodo nova tractata, *Werke band* **3***, pp. 265–327 (Königliche Gesellschaft der Wissenschaften, Göttingen)*, (1866). Voir aussi M. T. Heideman, D. H. Johnson, et C. S. Burrus, Gauss and the history of the fast Fourier transform, *IEEE ASSP Magazine* **1**, pp. 14–21 (1984).

[10] E. Hairer, S.P. Norsett et G. Wanner, Solving ordinary differential equations I. Nonstiff problems *Springer Series in Comput. Math. vol. 8* 2nd edition (1993)

[11] E. Hairer et G. Wanner, Solving ordinary differential equations II. Stiff and differential algebraic problems *Springer Series in Comput. Math. vol. 14* 2nd edition (1996)

[12] P. Lascaux et R. Théodor Analyse numérique matricielle appliquée à l'art de l'ingénieur. *Dunod* (1993)

[13] M. Schatzman, Analyse Numérique *InterEditions, Paris* (1991)

[14] G.W. Stewart, Introduction to matrix computations *Academic Press* (1973)

[15] A.M. Stuart et A.R. Humphries Dynamical systems and numerical analysis *Cambridge Univ. Press* (1996)

[16] J. Stoer and R. Bulirsch, Introduction to numerical analysis. *Translated from the German by R. Bartels, W. Gautschi and C. Witzgall. Springer-Verlag, New York-Heidelberg,* (1980).