Notes de cours d'analyse numérique élémentaire: Des équations algébriques aux équations différentielles

Université de Bourgogne, Département de Mathématiques
Licence de Mathématiques, LM5

Année 2002

Il semble que tout l'effort industriel de l'homme, tous ses calculs, toutes ses nuits de veille sur les épures, n'aboutissent, comme signes visibles, qu'à la seule simplicité, comme s'il fallait l'expé rience de plusieurs générations pour dégager peu à peu la courbe d'une colonne, d'une carène, ou d'un fuselage d'avion, jusqu'à leur rendre la pureté élémentaire de la courbe d'un sein ou d'une épaule. Il semble que le travail des ingénieurs, des dessinateurs, des calculateurs du bureau d'études ne soit ainsi en apparence, que de polir et d'effacer, d'alléger ce raccord, d'é quilibrer cette aile, jusqu'à ce qu'il n'y ait plus une aile accroché e à un fuselage, mais une forme parfaitement épanouie, enfin dé gagée de sa gangue, une sorte d'ensemble spontané, mysté rieusement lié, et de la même qualité que celle du poème. Il semble que la perfection soit atteinte non quand il n'y a plus rien à ajouter, mais quand il n'y a plus rien à retrancher.

Terre des hommes

Antoine de Saint-Exupéry

Table des matières

Chapitre 1

Introduction

Ce cours a été donné dans le cadre de la Licence de Mathématiques, UV Analyse Numérique Elémentaire, durant les années 1996-2002. Cette UV a lieu au second semestre, et comprend 26 heures de cours et 37 heures de travaux dirigés.

Ces notes de cours, dont les démonstrations ne sont volontairement qu'esquissées, ne se substituent pas au cours complet tel qu'il est présenté en amphi. Elles ne dispensent pas non plus de consulter quelques uns des nombreux ouvrages consacrés à l'analyse numérique et dont une liste est fournie à la fin de ce cours. Enfin, ces notes ne constituent pas un document d'étude: elle ne sont qu'un rappel des principaux résultats (définitions et théorèmes) établis pendant le cours d'analyse numérique, et destinées à être *le seul document autorisé* pendant les examens.

Ce document a été rédigé par Hervé Leferrand et Eric Busvelle à l'aide du traitement de texte LaTeX. Les graphiques ont été réalisés avec le logiciel de calcul numérique Matlab. Une version électronique de ce document est disponible sur le serveur internet de l'université de Bourgogne à l'adresse http://www.u-bourgogne.fr/monge/e.busvelle/teaching.html. Les auteurs seront reconnaissant envers les lecteurs qui voudront bien leur signaler les erreurs qu'ils auront trouvées, ou même qui auront des commentaires sur ces notes (oralement ou par courrier éléctronique à l'adresse: busvelle@u-bourgogne.fr).

Chapitre 2

Résolution de systèmes linéaires

2.1 Introduction

2.1.1 Exemple

Considérons l'exemple d'un problème conduisant à la résolution d'un système linéaire. La méthode des différences finies pour un problème aux limites en dimension 1 ;

$$\left\{ \begin{array}{lcll} x''(t) & = & f(t, x(t), x'(t)) & \\ x(a) & = & \alpha & a \leq t \leq b \\ x(b) & = & \beta & \end{array} \right. \tag{2.1}$$

Discrètisons l'équation ci-dessus : $t_i = a + ih$, $i = 1, \ldots, n$,$h = \frac{b-a}{n+1}$

$$x_i \simeq x(t_i) \ ; \ x'(t_i) \simeq \frac{x_{i+1} - x_{i-1}}{2h} \ ; \ x''(t_i) \simeq \frac{x_{i-1} - 2x_i + x_{i+1}}{h^2} \tag{2.2}$$

d'où

$$x_{i-1} - 2x_i + x_{i+1} = h^2 f(t_i, x_i, \frac{x_{i+1} - x_{i-1}}{2h}), \ i = 1, \ldots, n \tag{2.3}$$

avec $x_0 = \alpha$, $x_{n+1} = \beta$.

Considérons le cas particulier

$$x''(t) = p(t)x'(t) + q(t)x(t) + r(t), \ x(a) = \alpha, \ x(b) = \beta \tag{2.4}$$

avec comme hypothèses $q \geq 0$ sur $[a, b]$ et $h|p(t)| < 2$ pour tout $t \in [a, b]$.

On a alors

$$x_{i-1} - 2x_i + x_{i+1} = h^2 \left(\frac{p_i}{2h} (x_{i+1} - x_{i-1}) + q_i x_i + r_i \right) \qquad i = 1, \ldots, n$$

où $p_i = p(t_i)$, $r_i = r(t_i)$, $q_i = q(t_i)$.

En regroupant les termes il vient :

$$-b_i x_{i-1} + a_i x_i - c_i x_{i+1} = -r_i h^2 \quad i = 1 \text{ à } n,$$

soit un système linéaire (n,n) (x_0, x_{n+1} sont connus!) où $b_i = 1 + \frac{1}{2}p_i h$, $a_i = (2 + q_i h^2)$, $c_i = 1 - \frac{1}{2}p_i h$. Ceci s'écrit matriciellement $AX = d$:

$$A = \begin{pmatrix} a_1 & -c_1 & & 0 \\ -b_2 & a_2 & \ddots & \\ & \ddots & \ddots & -c_{n-1} \\ 0 & & -b_n & a_n \end{pmatrix} \qquad d = \begin{pmatrix} -r_1 h^2 + b_1 \alpha \\ -r_2 h^2 \\ \vdots \\ -r_{n-1} h^2 \\ -r_n h^2 + c_n \beta \end{pmatrix} \tag{2.5}$$

A la lumière de cet exemple, mais pour tout système linéaire, les questions pertinentes sont les suivantes:

1. A est-elle inversible ?
2. en supposant que le problème admet une solution et une seule x, a-t-on

$$\lim_{h \to 0} x_i(h) = x(t_i) \ ?$$

On sous-entend que $t_i = a + ih$ est fixé, $h \to 0$, $i \to +\infty$.

A est inversible : on regarde le système $AX = 0$.

1ère étape : $b_i > 0$ $c_i > 0$ $b_i + c_i \leq a_i$

2ème étape : (principe du maximum-minimum) Si $\Gamma_i = a_i x_i - b_i x_{i-1} - c_i x_{i+1} \leq 0$ $i = 1$ à n et si $x_0 \leq 0$, $x_{n+1} \leq 0$ alors $x_i \leq 0$, $i = 1$ à n. Si $\Gamma_i \geq 0$ $i = 1$ à n et si $x_0 \geq 0$, $x_{n+1} \geq 0$ alors $x_i \geq 0$, $i = 1$ à n.

3ème étape :

$$X = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \tag{2.6}$$

En effet en posant $\alpha = \beta = 0$, et supposant $\Gamma_i = 0$ pour $i = 1$ à n d'après la *2ème étape* $x_i = 0, i = 1$ à n.

Montrons la deuxième étape. Si tel n'est pas le cas $0 < \max_{0 \leq i \leq n+1} x_i = x_k$ pour un $k \in \{1, ..., n\}$. Posons $\mu = \frac{b_k}{a_k}$ et $\eta = \frac{c_k}{a_k}$, ainsi $\mu + \eta \leq 1$ et

$$x_k = \frac{1}{a_k}[\Gamma_k + b_k x_{k-1} + c_k x_{k+1}] \overset{(\Gamma_k \leq 0)}{\leq} \mu x_{k-1} + \eta x_{k+1} \leq \max(x_{k-1}, x_{k+1})$$

d'où $x_k = \max(x_{k-1}, x_{k+1})$. Si par exemple $x_k = x_{k+1} > x_{k-1}$ alors

$$x_k \leq \mu x_{k-1} + \eta x_{k+1} \overset{(\mu > 0)}{<} \mu x_k + \eta x_k \leq x_k$$

Ainsi $x_{k+1} = x_k = x_{k-1}$. On recommence la procédure avec x_{k+1}, avec x_{k-1}, d'où $x_0 = ... = x_{n+1}$., i.e. si (x_i) atteint son maximum en un *point intérieur*, (x_i) est constante.

2.1.2 Le problème central

A une matrice $n \times n$ réelle ou complexe inversible et b un vecteur colonne (matrice $n \times 1$) réel ou complexe, on cherche à résoudre $Ax = b$, dont la solution exacte est $x = A^{-1}b$.

Remarque 1 *Si $A \in \mathcal{M}_n(\mathbb{C})$ et $b \in \mathbb{C}^n$*

$$\left\{ \begin{array}{rcl} A & = & A_1 + iA_2 \\ b & = & b_1 + ib_2 \end{array} \right. \tag{2.7}$$

A_1, A_2, b_1, b_2 sont à coefficients réels. $Ax = b$ équivaut au système réel

$$\begin{pmatrix} A_1 & -A_2 \\ A_2 & A_1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \tag{2.8}$$

Dans la suite du cours on parlera de matrice de $M_n(\mathbf{K})$ ($\mathbf{K} = \mathbb{R}$ ou $\mathbb{C}$), et au besoin on fera la distinction entre $\mathbb{R}$ et $\mathbb{C}$. On distingue deux types de méthodes

a) les méthodes directes : on obtiendrait la solution exacte x en un nombre fini d'opérations s'il n'y avait pas d'erreur d'arrondi.

b) les méthodes itératives : x est la limite d'une suite (x_k) de solutions approchées (ex : $A = M - N, \quad Mx_{k+1} = Nx_k + b, \quad x_0$ choisi)

Remarque 2 *Les formules de Cramer (cadre A)) sont à écarter (le calcul d'un déterminant, avec la formule classique* $\det(A) = \sum_{\sigma \in \sigma_n} \varepsilon(\sigma) a_{(1),1} ... a_{\sigma(n),n}$*, requiert $(n-1)n!$ multiplications, alors si n est un peu grand...)*

Remarque 3 *Pourquoi des méthodes itératives ? Une méthode de résolution approchée peut être préférable à une méthode directe rendue mauvaise par les caractéristiques du problèmes. Par exemple considérons la taille des matrices :*

- *pour la méthode de Gauss, la* ***taille critique*** *est $n = 400$ (stockage des données, temps de calcul, erreurs numériques etc...)*
- *grande matrice* ***creuse*** *(en EDP (équations aux dérivées partielles) on arrive facilement à $n = 10^4$ avec des matrices très creuses, i.e. beaucoup de zéros, et éventuellement à diagonale strictement dominante*

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}| \text{ pour tout } i).$$

Remarque 4 *A chaque type de matrice sa méthode : le choix d'une méthode dépend des propriétés de la matrice du système (pensons par exemple aux matrices symétriques).*

2.2 Généralités

2.2.1 Normes matricielles

Définition 5 *Une norme matricielle est une application, $\|.\|$ de $\mathcal{M}_n(\mathbf{K})$ dans $\mathbb{R}$ qui vérifie :*

1. $\|A\| \geq 0$

2. $\|A\| = 0$ *si et seulement si* $A = 0$
3. $\|\alpha A\| = |\alpha|\|A\|,\ \alpha \in \mathbf{K}$
4. $\|A + B\| \leq \|A\| + \|B\|$
5. $\|AB\| \leq \|A\|\|B\|$ *(*$\|.\|$ *est sous multiplicative)*

Exemple 6 ***Normes subordonnées .*** *On part d'une norme* $\|.\|$ *sur* $\mathbf{K}^n$ *est on pose*

$$\|A\| = \sup_{x \in \mathbf{K}^n \|x\|=1} \|Ax\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}$$

C'est bien une application de $\mathcal{M}_n(\mathbf{K})$ *dans* $\mathbb{R}$ *(continuité de A sur la sphère unité, compact de* $\mathbf{K}^n$*) qui possède les propriétés d'une norme matricielle.*

Remarque 7 $\|Ax\| \leq \|A\|\|x\|$.

Remarque 8 $\|A\|$ *est atteint.*

Remarque 9 *Si* $A \in \mathcal{M}_n(\mathbb{R})$*, on peut considérer soit d'une part* $\sup_{x \in \mathbb{C}^n \|x\|=1} \|Ax\|$*, soit d'autre part* $\sup_{x \in \mathbb{R}^n \|x\|=1} \|Ax\|$ *où* $\|.\|$ *est une norme sur* $\mathbb{C}^n$ *(donc sur* $\mathbb{R}^n$*). Il est clair que* $\sup_{x \in \mathbb{R}^n \|x\|=1} \|Ax\| \leq \sup_{x \in \mathbb{C}^n \|x\|=1} \|Ax\|$*, mais rien ne dit que l'égalité est réalisée. Ce sera le cas cependant pour* $\|.\|_1, \|.\|_\infty, \|.\|_2$ *que l'on définira.*

Remarque 10 $\|I\| = 1$

Exemple 11 ***Norme de Frobenius*** *La norme de Frobenius est un exemple de norme matricielle non subordonnée :*

$$\|A\|_F = (\sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}|^2)^{\frac{1}{2}} (= (trA^T A)^{\frac{1}{2}} \text{ si } A \text{ est réelle })$$

$$(\|AB\|_F^2 = \sum_{i,j} |\sum_k a_{ik} b_{kj}|^2 \leq \sum_{i,j} (\sum_k |a_{ik}|^2)(\sum_\ell |b_{\ell j}|^2) = \sum_i (\sum_k |a_{ik}|^2) \times \sum_j (\sum_\ell |b_{\ell j}|^2) ...)$$

On a $\|I\|_F = \sqrt{n} \neq 1$ *dès que* $n \geq 2$ *ce qui prouve que la norme de Frobenius n'est pas subordonnée* $(n \geq 2)$*.*

2.2.2 Calcul de $\|.\|_1$, $\|.\|_\infty$, $\|.\|_2$ sur $\mathbb{R}$

Proposition 12 *Si on considère sur* $\mathbb{R}^n$ *les trois normes bien connues,* $\|x\|_1 = \sum_{i=1}^n |x_i|$ *,* $\|x\|_2 = (\sum_{i=1}^n x_i^2)^{\frac{1}{2}}$*,* $\|x\|_\infty = \max_i(x_i)$ *(*$x = (x_1, ..., x_n)^T$*), on a :*

1. $\|A\|_1 \equiv \max_{\|x\|_1=1} \|Ax\| = \max_{1 \leq j \leq n} \sum_{i=1}^n |a_{ij}|$
2. $\|A\|_\infty \equiv \max_{\|x\|_\infty=1} \|Ax\| = \max_{1 \leq i \leq n} \sum_{j=1}^n |a_{ij}|$
3. $\|A\|_2 \equiv \max_{\|x\|_2=1} \|Ax\| = [\rho(A^T A)]^{\frac{1}{2}}$ *où* $\rho(-)$ *désigne le rayon spectral.*

Preuve. On vérifie que $\|A\|_1 \leq \max_{1\leq j\leq n} \sum_{i=1}^n |a_{ij}|$ puis il suffit de construire un vecteur $x \neq 0$ tel que l'on ait l'égalité. On procède de même pour $\|A\|_\infty$. Pour $\|A\|_2$, il suffit d'écrire $\|Ax\|_2^2 = x^T A^T Ax$ puis de diagonaliser la matrice symétrique $A^T A$ par un changement de base orthogonal. ■

Lemme 13 *Si $B \in M_n(\mathbb{R})$ symétrique a pour valeurs propres $\lambda_1 \leq ... \leq \lambda_n$ alors $\lambda_1 x^T x \leq x^T Bx \leq \lambda_n x^T x$ et ceci pour tout x dans $\mathbb{R}^n$.*

Preuve. il suffit d'appliquer le théorème classique sur la diagonalisation des matrices symétriques réelles : $B = P^T \text{ diag } (\lambda_1, ...\lambda_n)P$, P orthogonale. Ainsi $x^T Bx = x^T P^T DPx = (Px)^T DPx = \sum \lambda_i y_i^2 \leq \lambda_n y^T y = \lambda_n x^T x$ où $y = Px$. ■

Rappel : M une matrice réelle ou complexe.

$$\rho(M) = \max\{|\lambda|; \lambda \text{ valeur propre complexe de } M\}.$$

Soit $\mu = [\rho(A^T A)]^{\frac{1}{2}}$. Remarquons que $A^T A$ est une matrice symétrique positive (i.e. $x^T A^T Ax \geq 0$), donc toutes ses valeurs propres sont positives. En particulier μ^2 est une valeur propre de $A^T A$. D'après le lemme pour tout $x \in \mathbb{R}^n$, $\|Ax\|_2^2 = x^T A^T Ax \leq \mu^2 x^T x$, ainsi $\|A\|_2 \leq \mu$.
De plus il existe $u \neq 0$ tel que $A^T Au = \mu^2 u$ d'où $\|Au\|_2^2 = u^T A^T Au = \mu^2 u^T u$ puis $\|A\|_2 = \mu$!

Remarque 14 *Dans le cas complexe on remplace A^T par $A^H (A^H = \bar{A}^T)$*

Remarque 15 *A est symétrique $\|A\|_2 = [\rho(A^2)]^{\frac{1}{2}} = [\rho(A)^2]^{\frac{1}{2}} = \rho(A)$*

Remarque 16 *$\|A\|_2 \leq \|A\|_F \leq \sqrt{n}\|A\|_2$ car $\rho(A^T A) \leq tr(A^T A) \leq n\rho(A^T A)$.*

2.2.3 Rayon spectral et norme

Proposition 17 *Soit $\|.\|$ une norme matricielle quelconque (subordonnée ou non) sur $\mathbb{C}$ alors $\rho(A) \leq \|A\|$.*

Preuve. Soit $x \neq 0$ tel que $Ax = \lambda x$ avec $|\lambda| = \rho(A)$. Si $y \in \mathbf{C}^n$, $xy^T = (y_1 x, ..., y_n x)$, donc si $y \neq 0$, $xy^T \neq 0$. D'où :

$$\rho(A)\|xy^T\| = \|\lambda xy^T\| = \|Axy^T\| \leq \|A\|\|xy^T\|$$

■

Remarque 18 *la démonstration est encore plus simple si on suppose $\|.\|$ subordonnée ($\|\lambda x\| = \|Ax\| \leq \|A\|\|x\|$)*

Proposition 19 *Soit $A \in M_n(\mathbb{R})$ subordonnée, alors $\rho(A) \leq \|A\|$*

Preuve. Si λ est une valeur propre réelle de A, associée à un vecteur propre u, alors $Au = \lambda u$ entraine $\|Au\| = |\lambda| \, \|u\|$. Si toutes les valeurs propres de A sont réelles, on en déduit bien $\rho(A) \leq \|A\|$.

Si λ est une valeur propre complexe de A, associée à un vecteur propre $u = x + iy$, on trouve par des calculs simples sur la partie imaginaire et la partie réelle que $|\lambda| \leq 2 \, |\lambda| \, \|A\|$, donc $\rho(A) \leq 2 \, \|A\|$.

On applique ce résultat à la matrice A^p, $p \in \mathbb{N}$ en utilisant le fait que $\rho(A)^p = \rho(A^p)$:

$$\rho(A)^p \leq 2\|A^p\| \leq 2\|A\|^p$$

si bien que $\rho(A) \leq 2^{\frac{1}{p}}\|A\|$ ce qui donne le résultat escompté en passant à la limite sur p. ■

Proposition 20 *Soit $A \in M_n(\mathbb{C})$, pour tout $\varepsilon > 0$ il existe une norme subordonnée telle que*

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

De plus, il existe une constante positive C telle que $\|B\|_2 \leq C\|B\|$ où $\|B\|_2$ est la norme Euclidienne de B.

Preuve. A est semblable à une réduite de Jordan, $A = PJP^{-1}$. On considère la matrice

$$D = \text{ diag } (1, \varepsilon, \varepsilon^2, ..., \varepsilon^{n-1}) \text{ alors } \hat{J} = D^{-1}JD$$

est strictement comme J sauf que chaque sur-diagonale 1 est remplacée par une sur-diagonale ε, car $(J)_{i,j}$ est multiplié par $\varepsilon^{1-i} \times \varepsilon^{j-1} \ldots$

L'existence de C découle de l'équivalence des normes sur $\mathbb{C}^n$. ■

Exemple 21 $\begin{pmatrix} 1 & 0 \\ 0 & \varepsilon^{-1} \end{pmatrix} \quad \begin{pmatrix} \alpha & 1 \\ 0 & \alpha \end{pmatrix} \quad \begin{pmatrix} 1 & 0 \\ 0 & \varepsilon \end{pmatrix} \quad = \quad \begin{pmatrix} \alpha & \varepsilon \\ 0 & \alpha \end{pmatrix}$

Ainsi $\|\hat{J}\|_\infty \leq \rho(A) + \varepsilon$. Posons $Q = PD$ et $\|x\| = \|Q^{-1}x\|_\infty$ (c'est une norme). Alors :

$$\begin{array}{rcl} \|A\| & = & \sup_{\|x\|=1} \|Ax\| \\ & = & \sup_{\|Q^{-1}x\|_\infty=1} \|Q^{-1}Ax\|_\infty \\ & = & \sup_{\|y\|_\infty=1} \|Q^{-1}AQy\|_\infty \\ & = & \sup_{\|y\|_\infty} \|\hat{J}y\|_\infty = \|\hat{J}\|_\infty \leq \rho(A) + \varepsilon \end{array} \tag{2.9}$$

2.2.4 Suites de Matrices

Il est clair (sinon, à démontrer en exercice) que la convergence d'une suite de matrices $n \times n$ équivaut à la convergence des n^2 suites scalaires formées par les éléments de ces matrices. On a le résultat suivant, fondamental pour la *convergence des méthodes itératives* :

Théorème 22 *B une matrice carrée, les conditions suivantes sont équivalentes :*

1. *(a)* $\lim_{k\to+\infty} B^k = 0$
   (b) $\lim_{k\to+\infty} B^k x = 0$ *pour tout vecteur x*
   (c) $\rho(B) < 1$
   (d) $\|B\| < 1$ *pour au moins une norme matricielle subordonnée sur $\mathbb{C}$*

Preuve. $a \Rightarrow b$ si $\|.\|$ est une norme subordonnée quelconque $\|B^k x\| \leq \|B^k\|\|x\|$ etc... $b \Rightarrow c$ si $\rho(B) \geq 1$, il existe un vecteur $0 \neq x$ (éventuellement complexe) tel que :

$$Bx = \lambda x \quad \text{avec} \quad |\lambda| = \rho(B) \geq 1.$$

On a alors $B^k x = \lambda^k x$ ne tend pas vers 0 quand $k \to +\infty$. (**Remarque** : Si B est réelle, il revient au même de dire que $\lim_{k\to+\infty} B^k x = 0$ pour tout x à coefficients réels)

$c \Rightarrow d$ c'est la proposition précédente

$d \Rightarrow a$ car $\|B^k\| \leq \|B\|^k$ ■

Remarque 23 *L'équivalence "*$\lim_{k\to+\infty} B^k = 0 \Leftrightarrow \rho(B) < 1$*" se démontre directement en calculant les puissances d'une réduite de Jordan.*

Exemple 24 *on sait que si* $\alpha \in \mathbb{C}$, $|\alpha| < 1$, $\frac{1}{1-\alpha} = 1 + \alpha + \alpha^2 + \alpha^3 + \cdots$ *On a envie de remplacer* α *par une matrice* B.

Lemme 25 ***(Neumann)*** *Si* $B \in M_n(\mathbb{C})$, $\rho(B) < 1$, *alors d'une part* $(I - B)$ *est inversible et*

$$(I - B)^{-1} = \lim_{k\to+\infty} \sum_{i=0}^{k} B^i = \sum_{i=0}^{+\infty} B^i$$

Preuve. Si les valeurs propres de B sont $\lambda_i, i = 1$ à n, les valeurs propres de $I - B$ sont $1 - \lambda_i, i = 1$ à n. Comme $|\lambda_i| \leq \rho(B) < 1$, pour tout i on a $1 - \lambda_i \neq 0$, donc $I - B$ est inversible. On a l'identité $(I - B)(I + B + ... + B^k) = I - B^{k+1}$, d'où $I + B + ... + B^k = (I - B)^{-1} - (I - B)^{-1} B^{k+1}$. On passe alors à la limite. ■

2.2.5 Conditionnement

On dit qu'un problème est bien (mal) conditionné si une petite variation des données entraîne une petite (grande) variation des résultats.

Exemple 26 $\lambda^n = 0$, *racine de multiplicité* n. *Soit* $\varepsilon > 0$, $\lambda^n - \varepsilon = 0$ *a pour racines :* $\omega^j \varepsilon^{\frac{1}{n}}$, $j = 1, \ldots, n$, $\omega = \exp(\frac{i2\pi}{n})$. *Si* $n = 100$, $\varepsilon = 10^{-100}$, $\varepsilon^{1/n} = 10^{-1}$, *ce qui correspond à* $10^{99} \times \varepsilon$.

Exemple 27 *Même des racines simples peuvent être mal conditionnées comme le montre l'exemple ci-dessous :*

$$p(\lambda) = \prod_{k=1}^{20} (\lambda - k).$$

Soit $\tilde{p}(\lambda) = p(\lambda) - \varepsilon\lambda^{19}$ *avec* $\varepsilon = 2^{-23}$ *(le coefficient de* x^{19} *dans* $p(\lambda)$ *est* 210*). On vérifie que* $\tilde{p}$ *ne possède que* 10 *racines réellesT*

Remarque 28 *Cette notion est liée au problème mathématique lui-même et est indépendante de la méthode utilisée pour le résoudre.*

Etudions le conditionnement d'un système linéaire à travers un exemple:

Exemple 29 *Soit le système* (S)

$$\begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix} \quad \textit{de solution} \quad \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \tag{2.10}$$

On perturbe le second membre de (S)

$$\begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \begin{pmatrix} u_1 + \delta_{u_1} \\ u_2 + \delta_{u_2} \\ u_3 + \delta_{u_3} \\ u_4 + \delta_{u_4} \end{pmatrix} = \begin{pmatrix} 32,1 \\ 23,9 \\ 33,1 \\ 30,9 \end{pmatrix} \quad \textit{de solution} \quad \begin{pmatrix} 9,2 \\ -12,6 \\ 4,5 \\ -1,1 \end{pmatrix} \tag{2.11}$$

L'erreur relative sur le second membre vaut $\simeq \frac{0,1}{20} = \frac{1}{200}$*, tandis que l'erreur relative sur la solution vaut* $\simeq \frac{10}{1}$*. L'erreur relative est multipliée par 2000 ! On perturbe la matrice de* (S)

$$\begin{pmatrix} 10 & 7 & 8,1 & 7,2 \\ 7,08 & 5,04 & 6 & 5 \\ 8 & 5,98 & 9,89 & 9 \\ 6,99 & 4,99 & 9 & 9,98 \end{pmatrix} \begin{pmatrix} u_1 + \delta_{u_1} \\ u_2 + \delta_{u_2} \\ u_3 + \delta_{u_3} \\ u_4 + \delta_{u_4} \end{pmatrix} = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix} \quad \textit{de solution} \quad \begin{pmatrix} -81 \\ 137 \\ -34 \\ 22 \end{pmatrix} \tag{2.12}$$

on obtient le même type de conclusion.

Dans la résolution de $Ax = b$ l'étude du conditionnement se justifie, car les éléments de A et b sont généralement issues de calculs numériques, donc entachés d'erreurs d'arrondis. On verra aussi un peu plus tard la notion de stabilité numérique d'une méthode ou d'un algorithme.

Revenons à l'étude théorique

Remarque 30 *L'erreur relative* ρ *vaut par définition* $\frac{\|y-x\|}{\|x\|}$*,* y *étant une valeur approchée du vecteur* x*. En dimension 1,* $-\log_{10}\rho$ *correspond au nombre de chiffres exacts de* y*. En dimension* > 1*, la situation est plus compliquée. Considérons les vecteurs*

$$x = \begin{pmatrix} 1,0000 \\ 0,0100 \\ 0,0001 \end{pmatrix} \qquad y = \begin{pmatrix} 1,0002 \\ 0,0103 \\ 0,0002 \end{pmatrix}. \tag{2.13}$$

Avec $\|.\|_\infty$*,* $\rho = 3 \times 10^{-4}$*, mais les erreurs relatives sur chaque composante, sont respectivement,* 2×10^{-4}*,* 3×10^{-2} *et* 1*. L'erreur relative donne ici une bonne idée de ce qui ce passe pour les grandes composantes, mais pas pour les petites.*

Lemme 31 *(lemme de perturbation) A inversible et E tel que* $\|A^{-1}\|\|E\| < 1$ *alors* $A+E$ *est inversible et*

$$\|(A+E)^{-1}\| < \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\|E\|}.$$

Preuve. $B = -A^{-1}E$, $\|B\| < 1$ d'où $I - B$ est inversible et $(I-B)^{-1} = \sum_{i=0}^{+\infty} B^k$. Il vient :

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

Mais $A+E = A(I+A^{-1}E) = A(I-B)$, $(A+E)^{-1} = (I-B)^{-1}A^{-1}$ puis $\|(A+E)^{-1}\| \leq \|(I-B)^{-1}\|\,\|A^{-1}\|$. ■

Enonçons maintenant le résultat principal :

Théorème 32 *(**estimation de l'erreur relative**) On considère le système perturbé $(A+E)y = b+f$ et on note x la solution du système $Ax = b$. Pour toute norme matricielle subordonnée on a :*

$$\frac{\|x-y\|}{\|x\|} \leq \frac{\|A^{-1}\|\|A\|}{1-\|A^{-1}\|\|E\|}\left\{\frac{\|E\|}{\|A\|} + \frac{\|f\|}{\|b\|}\right\}$$

où on a supposé $\|A^{-1}\|\,\|E\| < 1$ et $b \neq 0$.

Preuve.

$$\begin{aligned} x - y &= A^{-1}b - (A+E)^{-1}(b+f) \\ &= (A+E)^{-1}(-(b+f) + (A+E)A^{-1}b) \\ &= (A+E)^{-1}(Ex - f) \end{aligned}$$

donc

$$\begin{aligned} \frac{\|x-y\|}{\|x\|} &\leq \frac{\|(A+E)^{-1}\|}{\|x\|}(\|E\|\|x\| + \|f\|) \\ &\leq \|(A+E)^{-1}\|(\|E\| + \frac{\|f\|}{\|x\|}) \end{aligned}$$

or $Ax = b$ entraîne $\|x\| \geq \dfrac{\|b\|}{\|A\|}$ d'où

$$\frac{\|x-y\|}{\|x\|} \leq \|(A+E)^{-1}\|(\|E\| + \|A\|\frac{\|f\|}{\|b\|}).$$

On applique pour finir le lemme de perturbation. ■

Définition 33 ***Le conditionnement de la matrice*** *A est le nombre $K(A) = \|A\|.\|A^{-1}\|$ (relativement à $\|\quad\|$).*

Ainsi l'inégalité du théorème s'écrit :

$$\frac{\|x-y\|}{\|x\|} \leq \frac{K(A)}{1-K(A)(\frac{\|E\|}{\|A\|})}\left\{\frac{\|E\|}{\|A\|} + \frac{\|f\|}{\|b\|}\right\}.$$

Conséquence : comme $K(A)\frac{\|E\|}{\|A\|} = \|A^{-1}\|\|E\| < 1$,

$$\frac{K(A)}{1-K(A)(\frac{\|E\|}{\|A\|})} \sim K(A).$$

On a toujours $K(A) \geq 1$ et pour les "grandes" valeurs de $K(A)$ on dit que A est *mal conditionnée.*

Cela signifie que si $A = (a_{ij})$, $\tilde{A} = A + E = (\tilde{a}_{ij})$ avec $\tilde{a}_{ij} = a_{ij}(1 + \varepsilon_{ij})$, $|\varepsilon_{ij}| \leq \varepsilon_M = 10^{-t}$ (t désigne le nombre de chiffres significatifs), alors on a : $\|E\|_\infty \leq \varepsilon_M \|A\|_\infty$ d'où $(A + E)y = b$, puis

$$\frac{\|y - x\|_\infty}{\|x\|_\infty} \leq K(A)\varepsilon_M.$$

Si $K(A) = 10^k$ ($k \geq 0$) l'erreur relative est de l'ordre de 10^{-t+k}, ce qui signifie que l'on perd au moins k chiffres significatifs en résolvant le système $Ax = b$

Définition 34 *Le conditionnement* 2*, $K_2(A)$, est par définition* $K_2(A) = \|A\|_2 \|A^{-1}\|_2$.

Supposons A symétrique (réelle),

$$\|A\|_2 = \rho(A) = \max_i |\lambda_i|$$

les λ_i étant les valeurs propres de A et

$$\|A^{-1}\|_2 = \max_i |\lambda_i|^{-1} = (\min_i |\lambda_i|)^{-1}.$$

On obtient donc dans les cas d'une matrice symétrique :

$$K_2(A) = \frac{\max_i |\lambda_i|}{\min_i |\lambda_i|} \tag{2.14}$$

Si A n'est pas symétrique, que se passe-t-il ? (voir les travaux dirigés).

Exemple 35

$$A = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \qquad K_2(A) \simeq 2984 \tag{2.15}$$

Remarque 36 *Mise en garde : une matrice non symétrique peut avoir un grand conditionnement* 2 *même si toutes ses valeurs propres sont égales.*

2.3 Un exemple de méthode directe : la méthode de Gauss

2.3.1 Systèmes triangulaires

Définition 37 *Une matrice L est triangulaire inférieure (lower triangular) si $\ell_{ij} = 0$ dès que $i < j$.*

Définition 38 *Une matrice U est triangulaire supérieure (upper triangular) si $u_{ij} = 0$ dès que $i > j$.*

Détaillons la résolution de $Lx = b$. Pour cela, considérons le système

$$\left\{ \begin{array}{rcl} b_1 & = & \ell_{11}x_1 \\ b_2 & = & \ell_{21}x_1 + \ell_{22}x_2 \\ & \vdots & \\ b_n & = & \ell_{n1}x_1 + ... + \ell_{nn}x_n \end{array} \right. \tag{2.16}$$

où $\det L = \ell_{11} \times \ell_{22} \times ... \times \ell_{nn} \neq 0$. On applique une méthode de descente :

$$\left\{ \begin{array}{rcl} x_1 & = & \frac{b_1}{\ell_{11}} \\ x_2 & = & \frac{1}{\ell_{22}}(b_2 - \ell_{21}x_1) \\ & \vdots & \\ x_n & = & \frac{1}{\ell_{nn}}(b_n - \ell_{n1}x_1 - \cdots - \ell_{n,n-1}x_{n-1}) \end{array} \right. \tag{2.17}$$

Comptons le *nombre d'opérations* nécessaires pour cette méthode :

$$\left\{ \begin{array}{l} 1 + 2 + ... + (n-1) \text{ soustractions} \\ 1 + 2 + ... + (n-1) \text{ multiplications} \\ n \text{ divisions} \end{array} \right. \tag{2.18}$$

soit n^2 opérations.

2.3.2 La méthode de Gauss pure

On considère le système $Ax = b$ où $A = (a_{ij})$. On pose $A^{(1)} = A$ et $b^{(1)} = b$, puis on *élimine*: on suppose $a_{11}^{(1)} = a_{11} \neq 0$, et en multipliant la première ligne du système par un bon coefficient et en la retranchant à la seconde on *élimine le terme en* x_1. On fait de même avec les autres lignes. On aboutit au système $A^{(2)}x = b^{(2)}$ où

$$A^{(2)} = \begin{pmatrix} a_{11}^{(2)} & \cdots & \cdots & a_{1n}^{(2)} \\ 0 & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(2)} & \cdots & a_{nn}^{(2)} \end{pmatrix} \quad \text{et} \quad b^{(2)} = \begin{pmatrix} b_1^{(2)} \\ \vdots \\ b_n^{(2)} \end{pmatrix} \tag{2.19}$$

avec

$$a_{ij}^{(2)} = \left\{ \begin{array}{ll} a_{ij}^{(1)} = a_{ij} \quad \text{si} & i = 1 \\ a_{ij}^{(1)} - m_{i1}a_{1j}^{(1)} & i = 2, \ldots, n \quad j = 2, \ldots, n \end{array} \right. \tag{2.20}$$

$$b_i^{(2)} = \left\{ \begin{array}{l} b_i^{(1)},\ i = 1 \\ b_i^{(1)} - m_{i1}b_1^{(1)}, \quad i = 2, \ldots, n \end{array} \right. \tag{2.21}$$

où

$$m_{i1} = \frac{a_{i1}}{a_{11}} \tag{2.22}$$

A l'étape k : on a obtenu $A^{(k)}x = b^{(k)}$,

$$A^{(k)} = \begin{pmatrix} a_{11}^{(k)} & \cdots & \cdots & \cdots & a_{1n}^{(k)} \\ & \ddots & & & \vdots \\ & & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} \\ & \bigcirc & \vdots & & \vdots \\ & & a_{nk}^{(k)} & \cdots & a_{nn}^{(k)} \end{pmatrix} = \begin{pmatrix} a_{11}^{(1)} & \cdots & \cdots & \cdots & \cdots & \cdots & a_{1n}^{(1)} \\ & a_{22}^{(2)} & \cdots & \cdots & \cdots & \cdots & a_{2n}^{(2)} \\ & & \ddots & & & & \vdots \\ & & & a_{k-1,k-1}^{(k-1)} & a_{k-1,k}^{(k-1)} & \cdots & a_{k-1,n}^{(k-1)} \\ & & & & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} \\ & & & & \vdots & & \vdots \\ & & & & a_{nk}^{(k)} & \cdots & a_{nn}^{(k)} \end{pmatrix}$$

$$b^{(k)} = \begin{pmatrix} b_1^{(k)} \\ \vdots \\ b_n^{(k)} \end{pmatrix} = \begin{pmatrix} b_1^{(1)} \\ \vdots \\ b_2^{(2)} \\ \vdots \\ b_{k-1}^{(k-1)} \\ b_k^{(k)} \\ \vdots \\ b_n^{(k)} \end{pmatrix} \tag{2.23}$$

Si $a_{kk}^{(k)} \neq 0$, on passe au rang $k+1$ en faisant :

$$a_{ij}^{(k+1)} = \begin{cases} a_{ij}^{(k)} \text{ si } i = k \text{ (les k premières lignes ne sont pas modifiées)} \\ a_{ij}^{(k)} - m_{ik} a_{kj}^{(k)} \quad i = k+1, \ldots, n \ ; \ j = k+1, \ldots, n \end{cases} \tag{2.24}$$

et des 0 partout ailleurs, puis

$$b_i^{(k+1)} = \begin{cases} b_i^{(k)} \text{ si } \ i \leq k \\ b_i^{(k)} - m_{ik} b_k^{(k)} \quad i = k+1, \ldots, n \end{cases} \tag{2.25}$$

$$m_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} \ i = k+1, \ldots, n \tag{2.26}$$

(les m_{ik} *sont appelés multiplicateurs*).

Remarquons que tous les systèmes $A^{(k)}x = b^{(k)}$ sont équivalents (pourquoi ?). De plus on voit que $A^{(n)}x = b^{(n)}$ est un système triangulaire, donc facile à résoudre.

Pour une interprétation matricielle, remarquons que

$$A^{(2)} = A - \begin{pmatrix} 0 & \cdots & 0 \\ m_{21}a_{11} & \cdots & m_{21}a_{1n} \\ \vdots & & \vdots \\ m_{n1}a_{11} & \cdots & m_{n1}a_{1n} \end{pmatrix} = M^{(1)}A \tag{2.27}$$

où

$$M^{(1)} = \begin{pmatrix} 1 & & \bigcirc \\ -m_{21} & 1 & \\ \vdots & & \ddots \\ -m_{n1} & \bigcirc & 1 \end{pmatrix} \tag{2.28}$$

$$b^{(2)} = \begin{pmatrix} b^{(1)} \\ b_2^{(1)} - m_{21}b_1^{(1)} \\ \vdots \\ b_n^{(1)} - m_{n1}b_1^{(1)} \end{pmatrix} = M^{(1)}b^{(1)} \tag{2.29}$$

Plus généralement on a :

Proposition 39 *Dans la méthode de Gauss "pure" (i.e. $a_{kk}^{(k)} \neq 0$ pour tout k) on a*

$$A^{(k)} = M^{(k-1)}...M^{(1)}A$$

et $b^{(k)} = M^{(k-1)}...M^{(1)}A$ pour $k = 2, \ldots, n$ où on a posé :

$$M^{(k)} = I - m_k e_k^T = \begin{pmatrix} 1 & & & & \bigcirc \\ 0 & \ddots & & & \\ \vdots & & 1 & & \\ \vdots & & -m_{k+1,k} & 1 & \\ \vdots & & \vdots & & \ddots \\ 0 & & \underbrace{-m_{n,k}}_{k^{ieme}colonne} & \bigcirc & 1 \end{pmatrix} \tag{2.30}$$

$$(m_k = (0, ..., 0, m_{k+1,k}, ..., m_{n,k})^T).$$

Posons $L = [M^{(n-1)}...M^{(1)}]^{-1}$ (ceci a un sens !) et $U = A^{(n)} = M^{(n-1)}M^{(n-2)} \cdots M^{(1)}A$. On sait déjà que U est triangulaire supérieure. On a de plus :

$$L = (M^{(1)})^{-1}...(M^{(n-1)})^{-1}.$$

On remarque que

$$(M^{(k)})^{-1} = \begin{pmatrix} 1 & & & & & \bigcirc \\ 0 & \ddots & & & & \\ \vdots & & 1 & & & \\ \vdots & & m_{k+1,k} & 1 & & \\ \vdots & & \vdots & & \ddots & \\ 0 & & m_{n,k} & \bigcirc & & 1 \end{pmatrix} \tag{2.31}$$

(le vérifier) puis

$$L = \begin{pmatrix} 1 & & & & \\ m_{21} & 1 & & & \\ m_{31} & m_{32} & 1 & & \\ \vdots & \vdots & & \ddots & \\ m_{n1} & m_{n2} & \cdots & m_{n,n-1} & 1 \end{pmatrix}, \tag{2.32}$$

D'un point de vue matriciel, la méthode de Gauss "pure" (i.e $a_{kk}^{(k)} \neq 0$ pour tout k) conduit à la factorisation LU de la matrice A. L'intérêt est que si l'on a à résoudre plusieurs systèmes linéaires avec la même matrice A, on peut appliquer la méthode suivante :

1. $A = LU$;
2. $Ly = b$;
3. $Ux = y$

$$(\star) \quad A = \begin{pmatrix} 1 & & \bigcirc \\ \vdots & \ddots & \\ & \cdots & 1 \end{pmatrix} \begin{pmatrix} \cdots & \cdots & \cdots \\ & \ddots & \vdots \\ \bigcirc & & \ddots \end{pmatrix} \tag{2.33}$$

Problème 40 *Quand est-on sûr que $a_{kk}^{(k)} \neq 0$ pour tout k ?*

Théorème 41 *(**Décomposition triangulaire**) A inversible admet une factorisation du type $(\star)$ si et seulement si toutes les sous-matrices principales*

$$\begin{pmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & & \vdots \\ a_{k1} & \cdots & a_{kk} \end{pmatrix} \tag{2.34}$$

$(k = 1, \ldots, n)$ sont inversibles. De plus la factorisation est unique

Corollaire 42 *La méthode de Gauss "pure" est applicable si et seulement si toutes les sous-matrices principales sont inversibles.*

Problème 43 *Que se passe-t-il si $a_{kk}^{(k)} = 0$ pour un certain k ?*

Comme A est inversible, il existe une ligne pour laquelle $a_{ik}^{(k)} \neq 0$, $i > k$. On permute alors la ligne k et cette ligne (*pivotage mathématique*).

Preuve. Démonstration du théorème de décomposition triangulaire. Si $A = LU$ avec $(\ell_{ii} = 1)$ on a $\det A = \det U = \prod_{i=1}^{n} u_{ii} \neq 0$ et

$$(a_{ij})_{1\leq i,j\leq k} = (\ell_{ij})_{1\leq i,j\leq k} \times (u_{ij})_{1\leq i,j\leq k} \tag{2.35}$$

(à vérifier, $a_{ij} = \sum_{s=1}^{n} \ell_{is} \times u_{sj}$, $1 \leq i, j \leq k$, mais si $s > k$ $\ell_{is} = 0$...)

Réciproquement :

$a_{11} = a_{11}^{(1)} = detA_1 \neq 0$ donc on peut mener la première étape de l'élimination de Gauss.

Supposons avoir obtenu $A^{(k)}$ (pour $k \leq m-1$)

$$A^{(k)} = M^{(k-1)}...M^{(1)}A$$

d'où (!) avec des notations évidentes

$$\begin{array}{rcl} detA_k^{(k)} & = & detM_k^{(k-1)} \times \cdots \times detM_k^{(1)} \times detA_k \\ & = & detA_k \neq 0 \end{array} \tag{2.36}$$

Or $detA_k^{(k)} = a_{11}^{(1)}...a_{kk}^{(k)} \neq 0$, on peut affirmer que $a_{kk}^{(k)} \neq 0$. On peut continuer ! ■

Remarque 44 *On peut aussi utiliser le produit par bloc (à faire !)*

$$A^{(k)} = \begin{bmatrix} A_{11}^{(k)} & A_{12}^{(k)} \\ A_{21}^{(k)} & A_{12}^{(k)} \end{bmatrix} = \begin{bmatrix} M_{11}^{(k-1)} & 0 \\ M_{21}^{(k-1)} & M_{22}^{(k-1)} \end{bmatrix} \cdots \begin{bmatrix} M_{11}^{(1)} & 0 \\ M_{21}^{(1)} & M_{22}^{(1)} \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \tag{2.37}$$

où $A_{11}^{(k)}$ est la sous-matrice principale d'ordre k de $A^{(k)}$ et les matrices $M^{(k-1)}, \ldots, M^{(1)}$, A sont partitionnées en conséquence.

Preuve. Montrons l'unicité de la factorisation LU : si $A = L_1U_1 = L_2U_2$, on a $L_2^{-1}L_1 = U_2U_1^{-1}$. La matrice $B = L_2^{-1}L_1 = U_2U_1^{-1}$ est à la fois triangulaire inférieure et triangulaire supérieure avec ses éléments diagonaux égaux à 1. Ceci signifie que $B = I$ et $L_1 = L_2$, $U_1 = U_2$. ■

2.3.3 Aspect numérique

Nous allons ici compter le nombre d'opérations nécessaires dans cette méthode.

divisions : pour passer de $A^{(1)}$ à $A^{(2)}$, $n-1$ divisions, de $A^{(2)}$ à $A^{(3)}$, $n-2$ divisions etc ..., donc au total

$$(n-1) + (n-2) + ... + 2 + 1 = \frac{n(n-1)}{2} \text{ divisions.} \tag{2.38}$$

additions et multiplications : nombre de termes à transformer $A^{(k)} \hookrightarrow A^{(k+1)}$, i varie de $k+1$ à n et j de $k+1$ à $n+1$ (en considérant le second membre) d'où $2(n-k)(n-k+1)$ *opérations*, et *au total*

$$\begin{array}{rcl} 2\sum_{k=1}^{n-1}(n-k)(n-k+1) & = & 2(\sum_{k=1}^{n-1} n^2 + n - 2k_n + k^2 - k) \\ & = & 2(n(n^2-1) - 2n\sum_{k=1}^{n-1} k + \sum_{k=1}^{n-1} k^2 - k) \\ & = & 2(n^2(n-1) + \frac{n(n-1)(2n-1)}{6} - \frac{n(n-1)}{2}) \\ & = & \frac{2n(n^2-1)}{3} \end{array} \tag{2.39}$$

remontée : n^2.

En conclusion et pour n grand, on a environ $\dfrac{2n^3}{3}$ opérations. Analysons maintenant de plus près la stratégie du choix du pivot.

$$\left\{ \begin{array}{rcl} 10^{-4}u_1 + u_2 & = & 1 \\ u_1 + u_2 & = & 2 \end{array} \right. \qquad (t = 3 \text{ chiffres significatifs}) \tag{2.40}$$

$u_1 = 1,00010... \simeq 1, u_2 = 0,99990... \simeq 1$. Si on prend 10^{-4} pour pivot, on obtient $u_1 = 0, u_2 = 1$. Si on prend 1 pour pivot, on obtient $u_1 = u_2 = 1$. Il ne faut pas prendre des pivots "trop petits". L'explication du phénomène est que l'on effectue pour le calcul de u_1 une soustraction qui est très mal conditionnée car on considère des nombres trop voisins. Il existe deux stratégies couramment employées:

- **stratégie du pivot partiel**

$$|a_{ik}^{(k)}| = \max_{k \le p \le n} |a_{pk}^{(k)}| \quad i \ge k$$

- **stratégie du pivot total**

$$|a_{ij}^{(k)}| = \max_{k \le p,q \le n} |a_{pq}^{(k)}|.$$

Il reste à faire l'analyse des erreurs d'arrondi (analyse à posteriori)

Exemple 45

$$A = \begin{pmatrix} 3 & 2 \\ 1 & 2 \end{pmatrix} \quad (t = 4) \tag{2.41}$$

$$m_{21} = fl(\frac{a_{21}}{a_{11}}) = 0,3333$$

Si $\tilde{a}_{21} = 0,9999$, $m_{21} = \frac{\tilde{a}_{21}}{a_{11}} = \frac{0,9999}{3} = 0,3333$ *(calcul exact) . Puis*

$$a_{22}^{(2)} = fl(a_{22} - m_{21}a_{12}) = fl(2,000 - 0,6666) = 1,333.$$

Si $\tilde{a}_{22} = 1,9996$, $a_{22}^{(2)} = \tilde{a}_{22} - m_{21}a_{12} = 1,9996 - 0,3333 \times 2 = 1,9996 - 0,6666 = 1,333$. *Soit*

$$\tilde{A} = \begin{pmatrix} 3 & 2 \\ 0,9999 & 1,9996 \end{pmatrix} \tag{2.42}$$

On voit que la méthode de Gauss appliquée à $\tilde{A}$ *sans arrondi, donne le même résultat que la méthode de Gauss appliquée à* A *avec arrondi.*

De façon générale, on peut montrer que *la solution effectivement calculée sur ordinateur* est *la solution exacte* obtenue à partir de $A+E$ où E est une matrice de perturbation. Si A est bien conditionnée, d'après les théorème de l'estimation relative, il suffit d'évaluer la quantité $\frac{\|E\|}{\|A\|}$ pour avoir une idée de l'erreur relative commise dans le calcul de la solution. C'est possible et nous vous renvoyons par exemple au livre de Brezinski (les résultats sont dus à Wilkinson).

En fait cette façon de procéder est appelé analyse d'erreur à posteriori. Précisons les choses :

Définition 46 *Un algorithme pour résoudre le problème $\mathcal{P}(x)$ est numériquement stable au sens de l'analyse à posteriori si le résultat numérique $\hat{y}$ peut être considéré comme le résultat exact sur des données perturbées $\hat{x} = (\hat{x}_1, \ldots, \hat{x}_2)$, i.e. $\hat{y} = \mathcal{P}(\hat{x})$ et si*

$$\frac{|\hat{x}_i - x_i|}{|x_i|} \leq \textit{constante} \times \textit{précision de l'ordinateur.}$$

Remarque 47 *Pour cette constante dans la méthode de Gauss, voir le livre de Brezinski par exemple.*

Remarque 48 *Un algorithme peut être stable même si le problème est mal conditionné.*

2.4 Introduction aux méthodes itératives

2.4.1 Motivation

Soit le système 3×3

$$\left\{ \begin{array}{rcl} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 & = & b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 & = & b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 & = & b_3 \end{array} \right. \tag{2.43}$$

pour lequel on suppose $a_{11} \neq 0$, $a_{22} \neq 0$ et $a_{33} \neq 0$. On peut le réécrire sous la forme suivante :

$$\left\{ \begin{array}{rcl} x_1 & = & (b_1 - a_{12}x_2 - a_{13}x_3)/a_{11} \\ x_2 & = & (b_2 - a_{21}x_1 - a_{23}x_3)/a_{22} \\ x_3 & = & (b_3 - a_{31}x_1 - a_{32}x_2)/a_{33} \end{array} \right. \tag{2.44}$$

Imaginons que $x^{(k)}$ soit une approximation de $x = A^{-1}b$. Pour obtenir une nouvelle approximation $x^{(k+1)}$ de x, il est naturel de calculer :

$$\left\{ \begin{array}{rcl} x_1^{(k+1)} & = & (b_1 - a_{12}x_2^{(k)} - a_{13}x_3^{(k)})/a_{11} \\ x_2^{(k+1)} & = & (b_2 - a_{21}x_1^{(k)} - a_{23}x_3^{(k)})/a_{22} \\ x_3^{(k+1)} & = & (b_3 - a_{31}x_1^{(k)} - a_{32}x_2^{(k)})/a_{33} \end{array} \right. \tag{2.45}$$

C'est la méthode de Jacobi. Remarquons que l'on peut écrire cela sous la forme :

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

où

$$D = \text{ diag } (a_{11}, a_{22}, a_{33})$$

2.4.2 Convergence

Soit à résoudre $(*)$ $Ax = b$. Supposons que l'on ait trouvé une matrice B (avec $I - B$ inversible) et un vecteur c tels que le système $(**)$ $x = Bx + c$ soit équivalent à $(*)$. Ceci suggère la *méthode itérative* suivante :

$$(I) \qquad x^{k+1} = Bx^{(k)} + c \quad (k \geq 0) \qquad x^{(0)} \text{ donné} \tag{2.46}$$

Définition 49 *On dit que la méthode de itérative (I) est convergente si et seulement si $\lim_{k\to+\infty} x^{(k)} = x$ pour tout vecteur initial $x^{(0)}$.*

Théorème 50 *Les propositions suivantes sont équivalentes:*

1. *la méthode est convergente*
2. *$\rho(B) < 1$*
3. *$\|B\| < 1$ pour au moins une norme matricielle subordonnée.*

Preuve.

$$\left\{ \begin{array}{rcl} x^{k+1} & = & Bx^{(k)} + c \\ x & = & Bx + c \end{array} \right. \tag{2.47}$$

$$x^{(k+1)} - x = B(x^{(k)} - x).$$

d'où $x^{(k)} - x = B^k(x^{(0)} - x)$. Dire que la méthode est convergente, revient à dire que $x^{(k)} - x \to 0$ pour tout choix de $x^{(0)}$, i.e. $B^k y \to 0$ pour tout choix de y... ■

Remarque 51 $\|x^k - x\|_2 = \|B^k(x^{(0)} - x)\|_2 \leq \|B^k\|_2 \|x^{(0)} - x\|_2$. *Supposons B symétrique, alors*

$$\|x^k - x\|_2 = \|B^k(x^{(0)} - x)\|_2 \leq (\rho(B))^k \|x^{(0)} - x\|,$$

est la meilleure inégalité possible car si on choisit pour $x^{(0)} - x (\neq 0)$ un vecteur propre de B pour la valeur propre la plus grande en module, on a l'égalité. Donc la méthode sera d'autant plus rapide que $\rho(B)$ est plus petit.

2.4.3 Vitesse de convergence

La convergence n'est pas un critère suffisant pour comparer plusieurs méthodes. Il faut pour cela introduire une mesure de la vitesse de convergence.

Définition 52 *Soit $x^{(k+1)} = Bx^{(k)} + c$, $x^{(0)}$ donnée, une méthode itérative convergente (vers x solution de $x = Bx + c$). Pour tout entier m tel que $\|B^m\|_2 < 1$, on défini la vitesse moyenne de convergence par*

$$R(B^m) \stackrel{\text{déf.}}{=} -\log\left(\|B^m\|_2^{\frac{1}{m}}\right) = \frac{-\log\|B^m\|_2}{m}$$

En particulier, si $x^{(k+1)} = B_1 x^{(k)} + c_1$ et $x^{(k+1)} = B_2 x^{(k)} + c_2$ sont deux méthodes itératives convergent vers le même x, on dit que la première converge plus vite que la seconde en m itérations si $R(B_1^m) > R(B_2^m)$.

Remarque 53 *1. Si la méthode converge, $\rho(B) < 1$ donc il existe un $m \in \mathbb{N}$ tel que $\|B^m\|_2 < 1$.*

2. Sur deux méthodes itératives, il n'est pas toujours une qui soit uniformément meilleure que l'autre pour toute valeur de m.

3. Puisque $\varepsilon^{(m)} = B^m \varepsilon^{(0)}$, on a

$$\left\|\varepsilon^{(m)}\right\|_2 = \left\|B^m \varepsilon^{(0)}\right\|_2 \leq \|B^m\|_2 \left\|\varepsilon^{(0)}\right\|_2$$

donc

$$\left(\frac{\left\|\varepsilon^{(m)}\right\|_2}{\left\|\varepsilon^{(0)}\right\|_2}\right)^{\frac{1}{m}} \leq \|B^m\|_2^{\frac{1}{m}} = \mathrm{e}^{-R(B^m)}$$

i.e. R mesure le taux de décroissance exponentiel d'un majorant de la réduction moyenne de l'erreur après m itérations.

Le théorème suivant justifie l'intérêt du rayon spectral pour l'étude des méthodes itératives (en plus de l'étude de leur convergence ou divergence):

Théorème 54 *Soit B la matrice d'une méthode itérative convergente, alors*

$$\lim_{m \to +\infty} R(B^m) = -\log \rho(B) \overset{\text{déf.}}{=} R_\infty(B)$$

Preuve. Il suffit de montrer que $\lim_{m\to+\infty} \|B^m\|_2^{\frac{1}{m}} = \rho(B)$.

On sait déjà que $\rho(B^m) \leq \|B^m\|_2$ donc $\rho(B) \leq \|B^m\|_2^{\frac{1}{m}}$.

En utilisant la proposition 20, on établi l'existence d'une norme telle que

$$\|B\| \leq \rho(B) + \varepsilon$$

avec

$$\|\cdot\|_2 \leq C \|\cdot\|$$

pour une constante C. Donc $\|B^m\|_2 \leq C \|B^m\| \leq C \|B\|^m \leq C (\rho(B) + \varepsilon)^m$ d'où

$$\|B^m\|_2^{\frac{1}{m}} \leq \underbrace{C^{\frac{1}{m}}}_{\to 1} (\rho(B) + \varepsilon)$$

donc $\|B^m\|_2^{\frac{1}{m}} \leq \rho(B) + 2\varepsilon$ ■

2.4.4 Décomposition ("éclatement")

On *décompose* A, $A = M - N$ où M est "facile à inverser", i.e. la résolution d'un système de matrice M est simple. On a $Ax = b$ qui équivaut à $Mx = Nx + b$. D'où le schéma :

$$Mx^{(k+1)} = Nx^{(k)} + b \qquad x^{(0)} \text{ donné .} \tag{2.48}$$

La méthode sera convergente si et seulement si $\rho(M^{-1}N) < 1$.

2.4.5 Description des méthodes de Jacobi et Gauss-Seidel

Jacobi (en dimension n)

$$x_i^{(k+1)} = (b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)})/a_{ii}$$

pour $i = 1$ à n.

Gauss-Seidel : On remarque que pour calculer $x_i^{(k+1)}$, la méthode de Jacobi ne tient pas compte du fait que $x_\ell^{(k+1)}$ est déjà calculé pour $\ell = 1$ à $i-1$. On peut améliorer la méthode de Jacobi en posant :

$$x_i^{(k+1)} = (b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)})/a_{ii}$$

On gagne ainsi de la place en mémoire.

Eclatement : posons

$$L = -\begin{bmatrix} 0 & \cdots & \cdots & 0 \\ a_{21} & 0 & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ a_{n1} & \cdots & a_{n,n-1} & 0 \end{bmatrix}, \; U = -\begin{bmatrix} 0 & a_{12} & \cdots & a_{1n} \\ 0 & \ddots & \ddots & \vdots \\ \vdots & & \ddots & a_{n-1,n} \\ 0 & \cdots & \cdots & 0 \end{bmatrix} \tag{2.49}$$

$$D = diag(a_{11}, ..., a_{nn})$$

Jacobi

$$a_{ii} x_i^{(k+1)} = -\sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} + b_i$$

$$Dx^{(k+1)} = Lx^{(k)} + Ux^{(k)} + b$$

$$M_J x^{(k+1)} = N_J x^{(k)} + b \tag{2.50}$$

où $M_J = D$, $N_J = (L + U)$

Gauss–Seidel

$$a_{ii} x_i^{(k+1)} + \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} = -\sum_{j=i+1}^{n} a_{ij} x_j^{(k)} + b_i$$

$$Dx^{(k+1)} - Lx^{(k+1)} = Ux^{(k)} + b$$

$$x^{(k+1)} = (D - L)^{-1} U x^{(k)} + (D - L)^{-1} b \tag{2.51}$$

$$M_G x^{(k+1)} = N_G x^{(k)} + b \tag{2.52}$$

où $M_G = D - L$, $N_G = U$.

2.4.6 Quelques résultats de convergence

Convergence de la méthode de Jacobi pour une matrice à diagonale strictement dominante

Proposition 55 *Supposons que A soit à diagonale strictement dominante, alors la méthode de Jacobi est convergente.*

Preuve. On a :

$$\rho(M_J^{-1}N_J) = \rho(D^{-1}(L+U)) \leq \|D^{-1}(L+U)\|_\infty$$

$$D^{-1}(L+U) = D^{-1}(D-A) = [\alpha_{ij}]$$

où $\alpha_{ii} = 0$, $\alpha_{ij} = \frac{-a_{ij}}{a_{ii}} \quad i \neq j$.

$$\|D^{-1}(L+U)\|_\infty = \max_i \sum_{j=1}^{n} |\alpha_{ij}| \tag{2.53}$$

$$= \max_i \sum_{\substack{j=1 \\ j\neq i}}^{n} |\frac{a_{ij}}{a_{ii}}| < 1 \tag{2.54}$$

■

Proposition 56 *Avec les mêmes hypothèses, la méthode de Gauss-Seidel converge.*

Preuve. Soit λ une valeur propre de la matrice de Gauss-Seidel $(D-L)^{-1}U$, et x un vecteur propre $(\neq 0)$ associé.
$(D-L)^{-1}Ux = \lambda x$ s'écrit $Ux = \lambda(D-L)x$ i.e.

$$-\sum_{j=i+1}^{n} a_{ij}x_j = \lambda(a_{ii}x_i + \sum_{j=1}^{i-1} a_{ij}x_j) \quad \forall_i.$$

Soit k tel que $|x_k| = \|x\|_\infty (\neq 0)$:

$$\lambda(a_{kk} + \sum_{j=1}^{k-1} a_{kj}\frac{x_j}{x_k}) = -\sum_{j=k+1}^{n} a_{kj}\frac{x_j}{x_k}$$

d'où

$$|\lambda| \quad |a_{kk} + \sum_{j=1}^{k-1} a_{kj}\frac{x_j}{x_k}| \leq \sum_{j=k+1}^{n} |a_{kj}|.$$

Or

$$|\lambda| \times |a_{kk} + \sum_{j=1}^{k-1} a_{kj}\frac{x_j}{x_k}| \geq |\lambda|(|a_{kk}| - \sum_{j=1}^{k-1} |a_{kj}|)$$

Ainsi

$$|\lambda| \leq \frac{\sum_{j=k+1}^{n} |a_{kj}|}{|a_{kk}| - \sum_{j=1}^{k-1} |a_{kj}|} < 1$$

car $|a_{kk}| - \sum_{j=1}^{k-1} |a_{kj}| > \sum_{j=k+1}^{n} |a_{kj}|$! ■

Exemple 57 *On considère le problème de Dirichlet*

$$\frac{\partial^2 u(x,y)}{\partial x^2} + \frac{\partial^2 u(x,y)}{\partial y^2} = 0 \text{ dans } \Gamma \text{ (équation de Laplace)}$$
$$u(x,y) = g(x,y) \text{ sur } \partial\Gamma$$

où Γ *est le carré unité* $\Gamma = \{(x,y) \in \mathbb{R}^2 \quad 0 \leq x, y \leq 1\}$ *et où* g *est une fonction donnée sur le bord du carré.*

Soient $(x_0, y_0) \in \Gamma$ *et* h *petit. Supposons* $u \in C^4(\Gamma)$,

$$u(x_0 + h, y_0) = u(x_0, y_0) + h\frac{\partial u(x_0, y_0)}{\partial x} + \frac{h^2}{2}\frac{\partial^2 u(x_0, y_0)}{\partial x^2} + \frac{h^3}{6}\frac{\partial^3 u(x_0, y_0)}{\partial x^3} + O\left(h^4\right)$$

donc

$$\frac{1}{2}\left(u(x_0 + h, y_0) + u(x_0 - h, y_0)\right) = u(x_0, y_0) + \frac{h^2}{2}\frac{\partial^2 u(x_0, y_0)}{\partial x^2} + O\left(h^4\right)$$

puis en utilisant l'équation de Laplace:

$$\frac{1}{4}\left(u(x_0 + h, y_0) + u(x_0 - h, y_0) + u(x_0, y_0 + h) + u(x_0, y_0 - h)\right) = u(x_0, y_0) + O\left(h^4\right)$$

Comme pour une équation différentielle ordinaire, on va discrétiser, en choisissant un pas $h = \frac{1}{3}$,*et avec les conventions de la figure 2.1, on obtient alors les équations*

$$\begin{aligned}
\frac{1}{4}(g_0 + g_2 + u_2 + u_3) &= u_1 \\
\frac{1}{4}(g_1 + g_3 + u_1 + u_4) &= u_2 \\
\frac{1}{4}(g_4 + g_6 + u_1 + u_4) &= u_3 \\
\frac{1}{4}(g_5 + g_7 + u_2 + u_3) &= u_4
\end{aligned}$$

qui s'écrit $Au = b$ *où* A *est à diagonale strictement dominante*

$$A = \begin{pmatrix} 1 & -\frac{1}{4} & -\frac{1}{4} & 0 \\ -\frac{1}{4} & 1 & 0 & -\frac{1}{4} \\ -\frac{1}{4} & 0 & 1 & -\frac{1}{4} \\ 0 & -\frac{1}{4} & -\frac{1}{4} & 1 \end{pmatrix} \text{ et } b = \frac{1}{4}\begin{pmatrix} g_0 + g_2 \\ g_1 + g_3 \\ g_4 + g_6 \\ g_5 + g_7 \end{pmatrix}$$

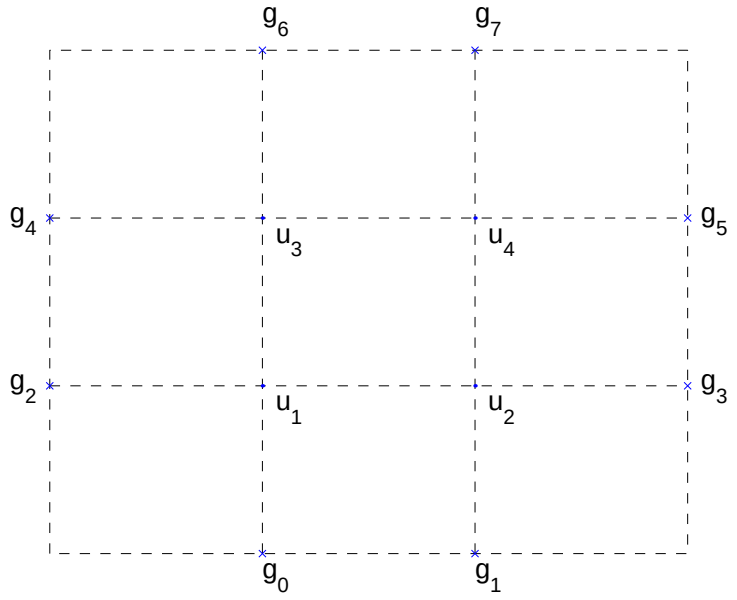


FIG. 2.1: Discrétisation de l'équation de Laplace dans le carré unité.

Convergence de la méthode de Gauss–Seidel pour une matrice symétrique définie positive

Théorème 58 *Si $A \in M_n(\mathbb{R})$ est symétrique définie positive, alors la méthode de Gauss-Seidel converge.*

Preuve. On a $A = D - L - U = M_G - N_G$ où $M_G = D - L$ et $N_G = U = L^T$ (à cause de la symétrie). On doit prouver que $\rho((D-L)^{-1}L^T) < 1$ ($G = (D-L)^{-1}L^T$). Posons :

$$\begin{aligned} G_1 \equiv D^{\frac{1}{2}} G D^{-\frac{1}{2}} &= D^{\frac{1}{2}}(D-L)^{-1}L^T D^{-\frac{1}{2}} \\ &= [(D-L)D^{-\frac{1}{2}}]^{-1}L^T D^{-\frac{1}{2}} \\ &= [D^{\frac{1}{2}} - LD^{-\frac{1}{2}}]^{-1}L^T D^{-\frac{1}{2}} \\ &= [D^{\frac{1}{2}}[I - D^{-\frac{1}{2}}LD^{-\frac{1}{2}}]]^{-1}L^T D^{\frac{1}{2}} \\ &= (I-L_1)^{-1}D^{-\frac{1}{2}}L^T D^{-\frac{1}{2}} \\ &= (I-L_1)^{-1}L_1^T \end{aligned} \tag{2.55}$$

($L_1 \equiv D^{-\frac{1}{2}}LD^{-\frac{1}{2}}$). En fait l'idée dans tout cela est de se ramener à une matrice symétrique définie positive avec des 1 sur la diagonale. Les matrices G et G_1 sont semblables, il suffit de voir que $\rho(G_1) < 1$. Soit λ une valeur propre de G_1, x un vecteur propre associé normé : $G_1 x = \lambda x$ $x^H x = 1$. Ce qui s'écrit encore : $L_1^T x = \lambda(I - L_1)x$. D'où

$$x^H L_1^T x = \lambda(1 - x^H L_1 x).$$

Posons $x^H L_1 x = a + ib$ (a, b réels). Or

$$x^H L_1 x = (x^H L_1 x)^T = x^T L_1^T \bar{x} = a + ib$$

ainsi $x^H L_1^T x = \overline{a+ib} = a - ib$, puis $|\lambda|^2 = |\frac{a-ib}{1-a-ib}|^2 = \frac{a^2+b^2}{1-2a+a^2+b^2}$. Mais la matrice $D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$ est symétrique définie positive, i.e. $I - L_1 - L_1^T$ est symétrique définie positive, ce qui permet d'écrire :

$$x^H (I - L_1 - L_1^T) x > 0.$$

On obtient $1 - x^H L_1 x - x^H L_1^T x > 0$, soit encore $1 - a - ib - a + ib = 1 - 2a > 0$, ce qui entraîne $|\lambda| < 1$. ■

Cas des matrices tridiagonales

Théorème 59

$$A = \begin{pmatrix} a_1 & b_1 & & 0 \\ c_2 & \ddots & \ddots & \\ & \ddots & \ddots & b_{n-1} \\ 0 & & c_n & a_n \end{pmatrix} \tag{2.56}$$

inversible tel que $a_i \neq 0$ *pour tout* i *alors* $\rho(B_G) = (\rho(B_J))^2$.

En conséquence, les deux méthodes convergent ou divergent simultanément.

Preuve.

$$D = \begin{pmatrix} a_1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & a_n \end{pmatrix}, \; L = -\begin{pmatrix} 0 & & \cdots & 0 \\ c_2 & \ddots & & \vdots \\ & \ddots & \ddots & \\ 0 & & c_n & 0 \end{pmatrix}, \; U = -\begin{pmatrix} 0 & b_1 & & 0 \\ & \ddots & \ddots & \\ \vdots & & \ddots & b_{n-1} \\ 0 & \cdots & & 0 \end{pmatrix} \tag{2.57}$$

On a $B_J = D^{-1}(L+U)$ et $B_G = (D-L)^{-1}U$. ■

Lemme 60 *Si*

$$A(\mu) = \begin{pmatrix} a_1 & \mu^{-1} b_1 & & \\ \mu c_2 & \ddots & \ddots & \\ & \ddots & \ddots & \mu^{-1} b_{n-1} \\ & & \mu c_n & a_n \end{pmatrix}$$

alors :

$$\det A(\mu) = det(D - \mu L - \mu^{-1} U) = \det A \; (\mu \neq 0).$$

Preuve. Il suffit de voir que $A(\mu) = Q(\mu) A Q(\mu)^{-1}$ où $Q(\mu) = diag(1, \mu^1, \ldots, \mu^{n-1})$ (se placer en i, j). ■

Terminons maintenant la preuve du théorème

Preuve. Les valeurs propres de B_J sont les racines de $p_J(\lambda) = det(D^{-1}(L+U) - \lambda I)$ qui sont aussi les racines du polynôme

$$q_J(\lambda) = \det(-D) p_J(\lambda) = \det(\lambda D - L - U).$$

De même, les valeurs propres de B_G sont les racines de $p_G(\lambda) = \det((D-L)^{-1}U - \lambda I)$, ou de

$$q_G(\lambda) = \det(\lambda D - \lambda L - U) = \det(L-D)p_G(\lambda).$$

Or

$$\begin{aligned} q_G(\lambda^2) &= \det(\lambda^2 D - \lambda^2 L - U) \\ &= \lambda^n \det(\lambda D - \lambda L - \tfrac{1}{\lambda}U) \\ &= \lambda^n \det(\lambda D - L - U) \end{aligned} \tag{2.58}$$

C'est à dire $q_G(\lambda^2) = \lambda^n q_J(\lambda)$.

Soit $s \in \mathbb{C}$ tel que $|s| = \rho(B_J)$: $q_G(s^2) = s^n q_J(s) = 0$, donc s^2 est valeur propre de B_G d'où $|s|^2 = (\rho(B_J))^2 \leq \rho(B_G)$.

Soit $s \in \mathbb{C}$ une valeur propre $\neq 0$ de B_G : $s^{\frac{1}{2}}$ et $-s^{\frac{1}{2}}$ sont dans le spectre de B_J d'où $(\rho(B_G))^{\frac{1}{2}} \leq \rho(B_J)$... ■

2.5 Méthode de relaxation

2.5.1 Présentation

La méthode de Gauss-Seidel est très intéressante par sa simplicité. Cependant *si* le rayon spectral de la matrice d'itération $M_G^{-1}N_G$ est proche de 1, la méthode est à écarter (car l'erreur se comporte en $\rho(M_G^{-1}N_G)^k$). L'idée est d'introduire un paramètre ω :

$$\bar{x}_i^{(k+1)} = \frac{1}{a_{ii}}(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}) \tag{2.59}$$

$$x_i^{(k+1)} = x_i^{(k)} + \omega(\bar{x}_i^{(k+1)} - x_i^{(k)}) \text{ (successive over relaxation)} \tag{2.60}$$

Remarque 61 *si $\omega = 1$ on retrouve Gauss-Seidel.*

Ecriture matricielle : on a

$$x_i^{(k+1)} = (1-\omega)x_i^{(k)} + \frac{\omega}{a_{ii}}(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}) \tag{2.61}$$

soit

$$a_{ii}x_i^{(k+1)} + \omega\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} = (1-\omega)a_{ii}x_i^{(k)} - \omega\sum_{j=1}^{n} a_{ij}x_j^{(k)} + \omega b_i.$$

On obtient :

$$Dx^{(k+1)} - \omega L x^{(k+1)} = (1-\omega)Dx^{(k)} + \omega U x^{(k)} + \omega b$$

puis

$$x^{k+1} = B_\omega x^k + \omega(D - \omega L)^{-1}b \tag{2.62}$$

où

$$B_\omega = (D - \omega L)^{-1}[(1-\omega)D + \omega U].$$

2.5.2 Condition nécessaire de convergence

Théorème 62 *(Kahan 1958) Soit $A \in \mathcal{M}_n(\mathbb{C})$ tel que $a_{ii} \neq 0$ pour tout i, alors $\rho(B_\omega) \geq |\omega - 1|$.*

Ainsi, pour que la méthode 2.60 converge il est nécessaire que $0 < \omega < 2$ si ω est réel.

Preuve. On a $detD^{-1} = det(D - \omega L)^{-1}$, puis

$$\begin{aligned} detB_\omega &= \text{produit des valeurs propres} \\ &= det(D-\omega L)^{-1} det\{(1-\omega)D + \omega U\} \\ &= det\{(1-\omega)I + \omega D^{-1}U\} \\ &= det((1-\omega)I) \ (D^{-1}U \text{ est strictement diagonale supérieure}) \\ &= (1-\omega)^n \end{aligned} \tag{2.63}$$

■

2.5.3 Le théorème d'Ostrowski-Reich

Théorème 63 *Soit $A \in M(\mathbb{R})$ symétrique défini positive et $\omega \in]0,2[$. Alors la méthode 2.60 converge.*

Définition 64 *A, B, C dans $M_n(\mathbb{R})$. Alors $A = B - C$ est un éclatement P-régulier de A, si B est inversible et $B + C$ est définie positive. (Attention, on ne fait aucune hypothèse de symétrie)*

Remarque 65 $x^T(B+C)x = x^T \frac{(B+C)+(B+C)^T}{2} x > 0 \ (x \neq 0)$.

Lemme 66 ***(de Stein)*** *$H \in M_n(\mathbb{R}), A$ symétrique définie positive telle que $A - H^T AH$ est définie positive. Alors $\rho(H) < 1$.*

Preuve. soit λ une valeur propre de H et $u \neq 0$ un vecteur propre associé. Les nombres $u^H Au$ et $u^H(A - H^T AH)u$ sont réels et positifs au sens strict. Ainsi : $u^H(A - H^T AH)u > 0$, puis

$$\begin{aligned} u^H Au > u^H H^T AHu &= (\bar{\lambda}\bar{u})^T A(\lambda u) \\ &= |\lambda|^2 \, u^H Au, \end{aligned} \tag{2.64}$$

soit $|\lambda|^2 < 1$. ■

Théorème 67 *(de l'éclatement P-Régulier) Soit A symétrique définie positive et $A = B - C$ un éclatement P-Régulier. Alors $\rho(B^{-1}C) < 1$.*

Preuve. D'après le lemme de Stein, il suffit de voir que

$$Q = A - (B^{-1}C)^T A(B^{-1}C)$$

est définie positive.
On a $B^{-1}C = B^{-1}(B - A) = I - B^{-1}A$ d'où

$$\begin{aligned} (B^{-1}C)^T A(B^{-1}C) &= (B^{-1}C)^T(A - AB^{-1}A) \\ &= (I - (B^{-1}A)^T)(A - AB^{-1}A) \\ &= A - AB^{-1}A - (B^{-1}A)^T A + (B^{-1}A)^T AB^{-1}A \end{aligned} \tag{2.65}$$

puis

$$\begin{array}{rcl} Q & = & (B^{-1}A)^T A + AB^{-1}A - (B^{-1}A)^T A(B^{-1}A) \\ & = & (B^{-1}A)^T(B + B^T - A)(B^{-1}A). \end{array} \tag{2.66}$$

Or $B + B^T - A = B^T + C$ est définie positive car $B + C$ l'est ($x^T B^T x = x^T Bx$!) donc Q l'est aussi (si S est symétrique définie positive, P inversible réelle alors $P^T SP$ est définie positive. En effet, $x^T P^T SPx = (Px)^T S(Px)$ etc...). ■

Pour démontrer le théorème de Ostrowski-Reich il suffit de prouver que

$$A = \omega^{-1}(D - \omega L) - \omega^{-1}[(1 - \omega)D + \omega L^T]$$

est un éclatement P régulier de A.

Preuve. $D - \omega L$ est inversible $(a_{ii} > 0)$;
ici

$$\begin{array}{rcl} (B + C) + (B + C)^T & = & \omega^{-1}[D - \omega L + (1 - \omega)D + \omega L^T] \\ & & +\omega^{-1}[D - \omega L^T + (1 - \omega)D + \omega L \\ & = & \omega^{-1}[4 - 2\omega)D] = 2\omega^{-1}(2 - \omega)D \end{array} \tag{2.67}$$

cette dernière matrice est définie positive. ■

Chapitre 3

Calcul des valeurs propres

3.1 Généralités

La recherche des valeurs propres d'une matrice A revient théoriquement à résoudre l'équation algébrique

$$\det(\lambda I - A) = 0$$

et il n'est donc pas possible de les calculer par une méthode directe (c'est le théorème de Galois). Nous devons donc développer des méthodes itératives pour résoudre ce problème. De façon générale, on verra au chapitre suivant des méthodes itératives pour résoudre les équations non-linéaires.

Avant de décrire une méthode pour calculer les valeurs propres d'une matrice, il peut-être utile d'avoir une idée de la localisation des valeurs propres. Le résultat suivant, très simple, en donne une approximation.

Théorème 68 *(Hadamard-Gershgorin) Les valeurs propres de* $A = (a_{ij})_{1 \leq i,j \leq n}$ *sont contenues dans l'union des disques*

$$\mathcal{D}_i = \left\{ z \in \mathbb{C} \quad ; \quad |z - a_{ii}| \leq \sum_{j \neq i} |a_{ij}| \right\} \quad i = 1, \ldots, n$$

Preuve. Soit λ une valeur propre de A et x un vecteur propre correspondant. On peut toujours supposer que $x_i \leq 1$ avec pour un $r \in \{1, \ldots, n\}$ l'égalité $x_r = 1$. On écrit alors la $r^{\text{ième}}$ composante de Ax

$$(\lambda - a_{rr}) = \sum_{j \neq r} a_{rj} x_j$$

d'où

$$|\lambda - a_{rr}| \leq \sum_{j \neq r} |a_{rj}|$$

c'est à dire que

$$\lambda \in \mathcal{D}_i$$

■

Ce théorème est valable pour une matrice à coefficients complexes.

3.2 Méthode de Jacobi

On applique la méthode de Jacobi lorsque l'on recherche toutes les valeurs propres et les vecteurs propres d'une matrice symétrique A. Elle consiste à transformer A en $B = \Omega A \Omega^T$ où Ω est une matrice de rotation (orthogonale)

$$\Omega = \begin{pmatrix} 1 & 0 & \cdots & & & & & & \cdots & 0 \\ 0 & \ddots & & & & & & & & \vdots \\ \vdots & & 1 & & & & & & & \\ & & & \cos(\theta) & 0 & \cdots & 0 & -\sin(\theta) & & \\ & & \vdots & 0 & 1 & & & 0 & & \\ & & & \vdots & & \ddots & & \vdots & & \\ & & & 0 & & & 1 & 0 & & \\ & & & \sin(\theta) & 0 & \cdots & 0 & \cos(\theta) & & \\ & & & & & & & & 1 & \vdots \\ \vdots & & & & & & & & \ddots & 0 \\ 0 & \cdots & & & & & & & \cdots\ 0 & 1 \end{pmatrix}$$

et l'angle θ est choisi de telle sorte que $b_{ij} = b_{ji} = 0$. Il suffit donc de prendre

$$b_{ij} = \cos(2\theta)\, a_{ij} + \frac{\sin(2\theta)}{2}(a_{jj} - a_{ii}) = 0$$

d'où

$$tg(2\theta) = \frac{2a_{ij}}{a_{ii} - a_{jj}}$$

De cette façon, la norme de A ne change pas *i.e.* $\|A\| = \|B\|$ car la matrice Ω est orthogonale. Mais la somme des carrés des termes diagonaux augmente dès que $a_{ij} \neq 0$ puisque

$$\sum_{i=1}^{n} b_{ii}^2 = \sum_{i=1}^{n} a_{ii}^2 + 2a_{ij}^2$$

comme on peut le vérifier. La méthode de Jacobi consiste à itérér la construction précédente afin d'augmenter la somme des carrés des termes diagonaux et à diminuer celle des termes extra-diagonaux. Attention, cette méthode n'est pas directe: un terme annulé au cours d'une itération peut réapparaître au cours d'une itération suivante.

Trois stratégies sont pratiquées en ce qui concerne les choix successifs des indices (i, j):

1. la méthode classique consiste à choisir à chaque étape le coefficient hors-diagonale le plus grand;
2. la méthode du balayage consiste à prendre successivement comme couples (i, j) tous les éléments hors-diagonaux par un balayage cyclique;

3. la méthode du balayage avec seuil qui se distingue de la précédente en omettant les éléments inférieurs à un certain seuil, seuil qui diminue à chaque balayage.

Nous admettrons ici la convergence de la méthode pour les trois stratégies ci-dessus.

Cette méthode présente l'avantage de donner simultanément les valeurs propres et les vecteurs propres de A. En effet, si $D = QAQ^T$ où D est diagonale et Q est orthogonale, les lignes de Q sont les vecteurs propres de A puisque $Q^T D = AQ^T$. On obtiendra donc les vecteurs propres de A en calculant

$$A^{(k+1)} = \Omega_k A^{(k)} \Omega_k^T$$
$$Q^{(k+1)} = \Omega_k Q^{(k)}$$

avec $A^{(0)} = A$ et $Q^{(0)} = I$.

Exemple 69 *Considérons l'exemple suivant, donné par la matrice symétrique*

$$A = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \\ 3 & 4 & 1 & 2 \\ 4 & 3 & 2 & 1 \end{bmatrix}$$

on aura alors la suite

$$A^{(1)} = \begin{bmatrix} -3 & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 \\ -\frac{\sqrt{2}}{2} & 1 & 4 & \frac{5\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & 4 & 1 & \frac{5\sqrt{2}}{2} \\ 0 & \frac{5\sqrt{2}}{2} & \frac{5\sqrt{2}}{2} & 5 \end{bmatrix} \qquad Q^{(1)} = \begin{bmatrix} \frac{\sqrt{2}}{2} & 0 & 0 & \frac{\sqrt{2}}{2} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -\frac{\sqrt{2}}{2} & 0 & 0 & \frac{\sqrt{2}}{2} \end{bmatrix}$$

$$A^{(2)} = \begin{bmatrix} -3 & -1 & 0 & 0 \\ -1 & -3 & 0 & 0 \\ 0 & 0 & 5 & 5 \\ 0 & 0 & 5 & 5 \end{bmatrix} \qquad Q^{(2)} = \begin{bmatrix} \frac{\sqrt{2}}{2} & 0 & 0 & \frac{\sqrt{2}}{2} \\ 0 & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 \\ 0 & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 \\ -\frac{\sqrt{2}}{2} & 0 & 0 & \frac{\sqrt{2}}{2} \end{bmatrix}$$

$$A^{(3)} = \begin{bmatrix} -3 & -1 & 0 & 0 \\ -1 & -3 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 10 \end{bmatrix} \qquad Q^{(3)} = \begin{bmatrix} \frac{\sqrt{2}}{2} & 0 & -\frac{1}{2} & \frac{1}{2} \\ 0 & \frac{\sqrt{2}}{2} & \frac{1}{2} & \frac{1}{2} \\ 0 & -\frac{\sqrt{2}}{2} & \frac{1}{2} & \frac{1}{2} \\ -\frac{\sqrt{2}}{2} & 0 & -\frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

$$A^{(4)} = \begin{bmatrix} -4 & 0 & 0 & 0 \\ 0 & -2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 10 \end{bmatrix} \qquad Q^{(4)} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

et donc on a obtenu à la fois les valeurs propres et les vecteurs propres associées qui sont les colonnes de la matrice $Q^{(4)}$. Il est néanmoins important de comprendre que tous les exemples ne sont pas aussi simples que celui-ci qui présente la particularité de converger en temps fini.

Chapitre 4

Résolution d'équations non linéaires

4.1 Introduction

Pourquoi des méthodes numériques pour "résoudre" des équations c'est à dire trouver des valeurs approchées des solutions ?

- On sait depuis Abel et Galois que les équations polynômiales de degré ≥ 5 sont non résolubles par radicaux ;
- Les méthodes implicites en EDO (voir un cours sur les équations différentielles) nécessitent la résolutions de systèmes non linéaires.

Remarque 70 *Bien que l'on s'intéresse à des équations, certaines démonstrations seront données en dimension* ≥ 2.

Principe
On considère $f : \mathbb{R}^n \to \mathbb{R}^n$, et on cherche à résoudre l'équation $f(x) = 0$. On se ramène à $G(x) = x$ (i.e. x *point fixe* de G). Une telle écriture conduit naturellement à la *méthode des approximations successives* : x^0 "donné" (x_0 sera une approximation raisonnable de la solution x^*), on calcule les itérées

$$x_{n+1} := G(x_n) \tag{4.1}$$

Remarque 71 *C'est une méthode à un point. Il existe des méthodes à plusieurs points, i.e. le calcul de* x_{n+1} *se fait en utilisant par exemple* x_n *et* x_{n-1}.

Quelles sont les difficultés ?

(a) existence (localisation de la solution)

(b) convergence (point de vue locale, point de vue globale)

(c) choix de x_0 ((c) est bien sûr lié à (a) et (b))

(d) notion de vitesse de convergence

(e) propagations des erreurs numériques.

4.2 Convergence locale

4.2.1 Notion de point d'attraction

Définition 72 *Un point fixe x^* de $G : \mathbb{R}^n \to \mathbb{R}^n$ est un point d'attraction pour les itérations $x^{k+1} = G(x^k)$ si il existe un voisinage ouvert S de x^* tel que, à chaque fois que $x^0 \in S$, les itérées (x^k) existent et convergent vers x^*.*

Théorème 73 ***(Théorème de Ostrowski)*** *Soit x^* tel que $G(x^*) = x^*$. On suppose que G est différentiable en x^* et que $\rho(G'(x^*)) < 1$. Alors x^* est un point d'attraction.*

Preuve. Soit $\sigma = \rho(G'(x^*))$ et $\varepsilon > 0$. Il existe une norme sur $\mathbb{R}^n$ telle que $\|G'(x^*)\| \leq \sigma+\varepsilon$. Comme G est différentiable en x^*, il existe $\delta > 0$ tel que : si $x \in S = \{x/\|x-x^*\| < \delta\}$ alors $\|G(x) - G(x^*) - G'(x^*)(x - x^*)\| \leq \varepsilon\|x - x^*\|$. Ainsi pour $x \in S$:

$$\begin{array}{rcl} \|G(x) - G(x^*)\| & \leq & \|G(x) - G(x^*) - G'(x^*)(x - x^*)\| + \|G'(x^*)(x - x^*)\| \\ & \leq & (\sigma + 2\varepsilon)\|x - x^*\|. \end{array} \tag{4.2}$$

Comme $0 \leq \sigma < 1$, on peut toujours choisir ε tel que $\alpha \equiv \sigma + 2\varepsilon < 1$. Conclusion : si $x^{(0)} \in S$, on a $\|x^{(1)} - x^*\| = \|G(x^{(0)}) - G(x^*)\| \leq \alpha\|x^{(0)} - x^*\|$. Ainsi $x^{(1)} \in S$ et par récurrence, $x^{(k)} \in S$ avec

$$\|x^{(k)} - x^*\| \leq \alpha^k\|x^{(0)} - x^*\|.$$

■

Remarque 74 *Dans le cas $n = 1, \rho(G'(x^*)) = |G'(x^*)|$.*

$$\left\{ \begin{array}{ll} |G'(x^*)| < 1 & \textit{point fixe attractif} \\ |G'(x^*)| > 1 & \textit{point fixe répulsif} \\ |G'(x^*)| = 1 & \textit{cas difficile !} \end{array} \right. \tag{4.3}$$

Exercice 75 *Illustrer graphiquement les différents cas ci-dessus.*

4.2.2 Contractions

Définition 76 *$G : \mathbb{R}^n \to \mathbb{R}^n$ est une contraction sur l'ensemble D si il existe une constante $0 \leq \alpha < 1$ telle que :*

$$\|G(x) - G(y)\| \leq \alpha\|x - y\| \textit{ pour tout } x \textit{ et } y \textit{ dans } D.$$

Théorème 77 ***(de l'application contractante)*** *Soit $G : \mathbb{R}^n \to \mathbb{R}^n$ une contraction sur l'ensemble fermé D. On suppose que $G(D) \subset D$. Alors G a un unique point fixe x^* dans D et pour $x^{(0)} \in D$ les itérées $x^{(k+1)} = G(x^{(k)})$ convergent vers x^*. De plus on a :*

$$\|x^{(k)} - x^*\| \leq \frac{\alpha}{1 - \alpha}\|x^{(k)} - x^{(k-1)}\|.$$

($\alpha \equiv$ constante de contraction).

Preuve. Voir le cours de Topologie. ■

En conséquence, la proposition suivante permet de majorer la précision d'une méthode basée sur des approximations successives en fonction du choix de $x^{(0)}$.

Proposition 78 *Soit $G : \mathbb{R}^n \to \mathbb{R}^n$ une contraction sur D de constante α, et supposons que $x^{(0)} \in D$ soit tel que*

$$S \equiv \{x / \|x - G(x^{(0)})\| \leq \frac{\alpha}{1-\alpha} \|G(x^{(0)}) - x^{(0)}\|\} \subset D.$$

Alors les itérées convergent vers l'unique point fixe x^* de G dans S (donc aussi dans D).

Preuve. $x \in S$ (S est fermé)

$$\begin{aligned} \left\|G(x) - G\left(x^{(0)}\right)\right\| &\leq \alpha \left\|x - x^{(0)}\right\| && (4.4)\\ &\leq \alpha \left(\left\|x - G\left(x^{(0)}\right)\right\| + \left\|G\left(x^{(0)}\right) - x^{(0)}\right\|\right) && (4.5)\\ &\leq \left(\frac{\alpha^2}{1-\alpha} + \alpha\right) \left\|G\left(x^{(0)}\right) - x^{(0)}\right\| = \frac{\alpha}{1-\alpha} \left\|G\left(x^{(0)}\right) - x^{(0)}\right\| && (4.6) \end{aligned}$$

i.e. $G(S) \subset S$ etc... ■

4.2.3 Ordre de convergence sur $\mathbb{R}$

Voyons pour commencer une définition de l'ordre de convergence d'une suite numérique

Définition 79 *Soit $(x_k) \in \mathbb{R}^n$ avec $x_k \to x^*$*

1. *si il existe $\rho (0 < |\rho| < 1)$ tel que*

$$\lim_{k \to +\infty} \frac{x_{k+1} - x^*}{x_k - x^*} = \rho$$

on dit que la suite (x_k) converge linéairement.

2. *Si $\rho = 0$ on dit que la convergence est super-linéaire. Certains types de convergence super-linéaire peuvent être caractérisés de la façon suivante : si il existe $p > 1$ et $C > 0$ tels que*

$$\lim_{k \to +\infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|^p} = C$$

On dit que la suite est d'ordre p.

Exemple 80 – *$p = 2$, convergence quadratique*
– *$p = 3$, convergence cubique*

En raisonnant sur le nombre de chiffres exacts, on voit l'importance de cette notion d'ordre. Il est important de noter que p peut ne pas être entier !

Nous allons pour finir nous intéresser à l'ordre d'un point attractif. Soit $f(x^*) = x^*$ avec $|f'(x^*)| < 1$, $x_{k+1} = f(x_k)$.

Proposition 81 *Si $f'(x^*) \neq 0$, la suite (x_k) des itérées converge linéairement. Si $0 = f'(x^*) = \ldots = f^{(r-1)}(x^*) \neq f^{(r)}(x^*)$ la convergence est d'ordre r.*

Preuve. Il suffit d'écrire la formule de Taylor-Young, ce qui donne

$$f(x) - f(x^*) = \frac{1}{n!} f^{(r)}(x^*)(x - x^*)^r + 0((x - x^*)^r)$$

etc... ■

4.3 La méthode de Newton

4.3.1 Description

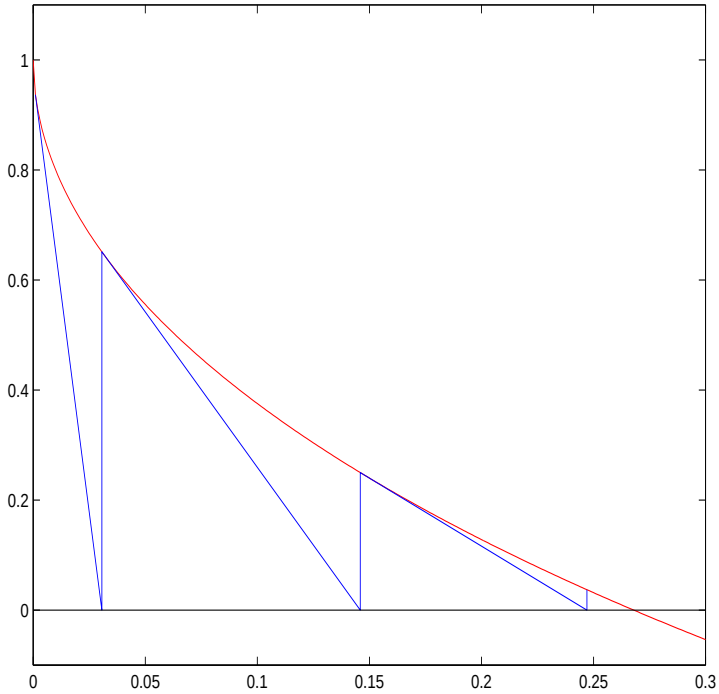


FIG. 4.1: Méthode de Newton

approche géométrique :

$$x_{k+1} := x_k - \frac{f(x_k)}{f'(x_k)}$$

approche analytique :

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f"(\xi_0)(x - x_0)^2, \ \xi_0 \in (x, x_0).$$

Ainsi $\hat{f}(x) = f(x_0) + f'(x_0)(x - x_0)$ peut être considérée comme une "bonne" approximation de $f(x)$ au voisinage de x^* (si x_0 est proche de x^*). On regarde l'équation $\hat{f}(x) = 0$ dont la solution est $x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$. L'intérêt de cette approche est d'envisager d'autres méthodes, par exemple la méthode qui consisterait à considérer la suite $x_{k+1} = x_k - \frac{f(x_k)}{f'(x_0)}$.

4.3.2 Convergence

Aspect local

Il est clair que la méthode de Newton consiste à itérer la fonction $\varphi(x) = x - \frac{f(x)}{f'(x)}$. Supposons que $f'(x^*) \neq 0$ et f de classe C^2 au voisinage de x^*. Alors on obtient :

$$\varphi'(x) = 1 - \frac{[f'(x)]^2 - f(x)f''(x)}{[f'(x)]^2} = \frac{f(x)f"(x)}{[f'(x)]^2}$$

d'où $\varphi'(x^*) = 0$, ce qui signifie que x^* est super attractif.

Remarque 82 *1. si f est trois fois différentiable au voisinage de x^* on peut calculer $\varphi''(x^*)(= \frac{f''(x^*)}{f'(x^*)})$, d'où la convergence quadratique de la méthode.*

2. on peut voir aussi que $x_{k+1} - x^* = x_k - x^* - \frac{f(x_k)}{f'(x_k)}$ d'où

$$\begin{array}{rcl} f'(x_k)(x_{k+1} - x^*) & = & f(x^*) - f(x_k) - f'(x_k)(x^* - x_k) \\ & = & \frac{1}{2}f"(\xi_k)(x^* - x_k)^2 \quad \xi_k \in (x^*, x_k) \end{array} \tag{4.7}$$

et finalement

$$\frac{x_{k+1} - x^*}{(x^* - x_k)^2} \sim \frac{1}{2}\frac{f''(x^*)}{f'(x^*)}.$$

Aspect global : un exemple.

Proposition 83 *On suppose F de classe C^2 sur $[a, b]$ vérifiant :*

1. $F(a) < 0, F(b) > 0$
2. $F'(x) \neq 0$ *sur* $[a, b]$
3. $F'' \leq 0$ *sur* $[a, b]$
4. $\frac{F(b)}{F'(b)} \leq b - a$

Alors la méthode de Newton converge vers l'unique solution sur $[a, b]$, x^*, de $F(x) = 0$, pour tout choix x_0 dans $[a, b]$.

Preuve. Graphiquement Soit x^* l'unique solution de $F(x) = 0$ (ceci a un sens d'après i) (existence) et ii) (unicité due à la strict monotonie)).
Supposons que $a \leq x_0 \leq x^*$:

$$x_1 = x_0 - \frac{F(x_0)}{F'(x_0)} \geq x_0$$

car $F'(x_0) > 0$ et $F(x_0) \leq 0$.
Montrons que $x_n \leq x^*, x_{n+1} \geq x_n$ pour tout n (en particulier $a \leq x_n \leq x^*$) :

- $n = 0$ c'est vrai

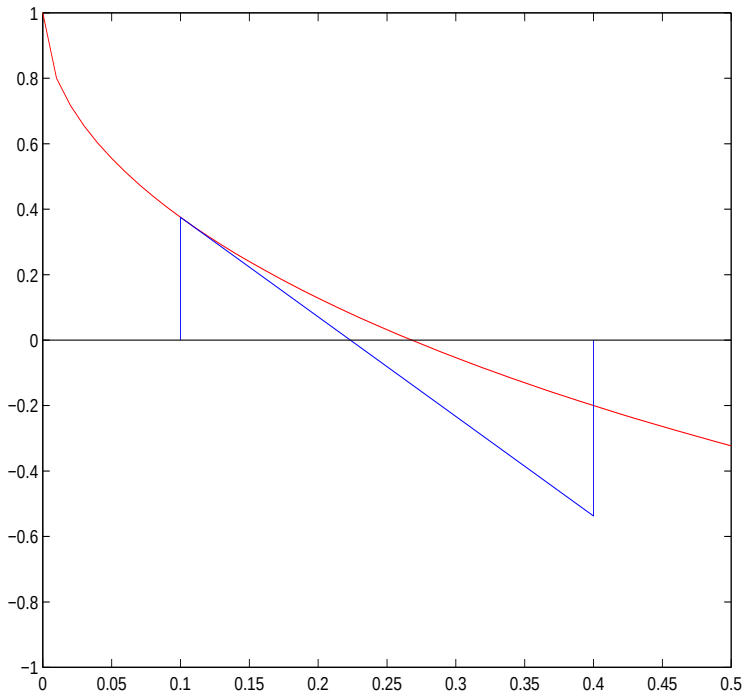


FIG. 4.2: Convergence de la méthode de Newton

– supposons la propriété vraie jusqu'au rang n, on a :

$$-F(x_n) = F(x^*) - F(x_n) = (x^* - x_n)F'(\xi_n) \quad \xi_n \in]x_n, x^*[$$

d'où $-F(x_n) \leq (x^* - x_n)F'(x_n)$ $(F' \searrow)$ puis

$$x_{n+1} = x_n - \frac{F(x_n)}{F'(x_n)} \leq x_n + (x^* - x_n) = x^* \ (F' \geq 0).$$

On en déduit que $F(x_{n+1}) \leq 0$ et $x_{n+2} = x_{n+1} - \frac{F(x_{n+1})}{F'(x_{n+1})} \geq x_{n+1}$. En conclusion, la suite (x_n) est croissante bornée, donc converge vers ℓ avec $\ell = \ell - \dfrac{F(\ell)}{F'(\ell)}$, ainsi $F(\ell) = 0$ soit $\ell = x^*$.

A présent si $x^* \leq x_0 \leq b$ (on voit sur le graphique que $x_1 \in [a, x^*] \dots$) on a

$$F(x_0) - F(x^*) = F'(\xi_0)(x_0 - x^*) \quad \xi_0 \in [x^*, x_0].$$

Comme $F' \searrow$, il vient $F(x_0) \geq F'(x_0)(x_0 - x^*)$ puis

$$x_1 = x_0 - \frac{F(x_0)}{F'(x_0)} \leq x_0 - (x_0 - x^*) = x^*.$$

A-t-on $x_1 \geq a$? De

$$\begin{array}{rcll} F(x_0) & = & F(b) - (b - x_0)F'(\eta_0) & \eta_0 \in [x_0, b] \\ F(x_0) & \leq & F(b) - (b - x_0)F'(b) & \end{array} \tag{4.8}$$

il vient :

$$\begin{array}{rcl} x_1 = x_0 - \frac{F(x_0)}{F'(x_0)} \geq x_0 - \frac{F(x_0)}{F'(b)} & \geq & x_0 - \frac{F(b)}{F'(b)} + (b - x_0) \\ & \geq & x_0 - (b-a) + (b - x_0) = a \end{array} \tag{4.9}$$

■

Exercice 84 *Etudier le cas* $F(x) = x^2 - c (c > 0)$.

4.4 La méthode de la sécante

La méthode de la sécante (dite aussi méthode "régula-falsi") est un exemple de méthode à plusieurs points.

4.4.1 Présentation de la méthode

Un des problèmes de la méthode de Newton est l'évaluation de la dérivée f'. Il existe des méthodes pour contourner cette difficulté pratique:

a) Whittaker

$$x_{k+1} = x_k - \frac{f(x_k)}{m} \quad m \equiv \text{ constante}$$

b) Régula-falsi

$$f'(x_k) \approx \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}$$

$$\begin{array}{rcl} x_{k+1} & = & x_k - \frac{(x_k - x_{k-1}) f(x_k)}{f(x_k) - f(x_{k-1})} \\ & = & \frac{x_{k-1} f(x_k) - x_k f(x_{k-1})}{f(x_k) - f(x_{k-1})} \end{array} \tag{4.10}$$

L'interprétation de cette dernière méthode est montrée graphiquement sur la figure 4.3

$$\begin{array}{rcl} y - f(x_0) & = & (x - x_0) \frac{f(x_1) - f(x_0)}{x_1 - x_0} - f(x_0) \\ & = & (x_2 - x_0) \frac{f(x_1) - f(x_0)}{x_1 - x_0} \end{array} \tag{4.11}$$

$$x_2 = x_0 - \frac{f(x_0)}{\frac{f(x_1) - f(x_0)}{x_1 - x_0}} = \cdots$$

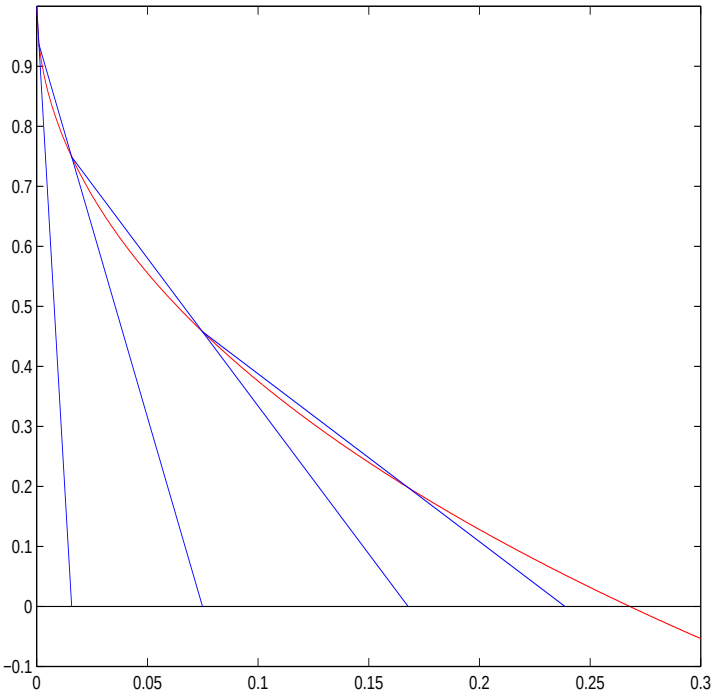


FIG. 4.3: Méthode de la sécante

4.4.2 Convergence

Proposition 85 *La méthode de la sécante est d'ordre* $\frac{1+\sqrt{5}}{2}$.

Remarque 86 *Pour comparer la méthode de Newton et la méthode de la sécante, il faudrait introduire la notion d'indice d'efficacité (voir les travaux dirigés).*

Preuve. On trouvera aussi dans le livre de Demailly [4], ainsi que dans "Calcul Infinitésimal" de J. Dieudonné, une démonstration de ce résultat.
Posons :

$$\varphi(u,v) = u - \frac{f(u)(u-v)}{f(u)-f(v)} = \frac{vf(u)-uf(v)}{f(u)-f(v)} \quad u \neq v \tag{4.12}$$

$$\varphi(u,u) = u - \frac{f(u)}{f'(u)} \quad (\varphi(x^*,x^*) = x^*) \tag{4.13}$$

Remarquons que $\varphi(u,x^*) = \varphi(x^*,v) \equiv x^*$ d'où

$$\begin{array}{rcl} \frac{\partial\varphi}{\partial u}(u,x^*) & = & \frac{\partial\varphi}{\partial y}(x^*,v) \equiv 0 \\ \frac{\partial^2\varphi}{\partial u^2}(u,x^*) & = & \frac{\partial^2\varphi}{\partial v^2}(x^*,v) \equiv 0 \end{array} \tag{4.14}$$

Ainsi :

$$\begin{array}{rcl} \varphi(x^*+p,x^*+q) & = & \varphi(x^*,x^*) + D\varphi(x^*,x^*)(p,q) \\ & & +\frac{1}{2}D^2\varphi(x^*+\theta p,x^*+\theta q)((p,q),(p,q)) \end{array} \tag{4.15}$$

avec $\theta \in]0,1[$ dépendant de p et q.
En utilisant le fait que $\varphi(x^*,x^*) = x^*$ et $\frac{\partial\varphi}{\partial u}(x^*,x^*) = \frac{\partial\varphi}{\partial u}(x^*,x^*) = 0$ on obtient

$$\varphi(x^*+p,x^*+q) = x^* + \frac{1}{2}\Big[\frac{\partial^2\varphi}{\partial u^2}(x^*+\theta p,x^*+\theta q)p^2 + 2\frac{\partial^2\varphi}{\partial u\partial v}(\quad)pq + \frac{\partial^2\varphi}{\varphi v^2}(\quad)q^2\Big]$$

Or $\frac{\partial^2\varphi}{\partial u^2}(x^*+\theta p,x^*) = 0$ d'où

$$\frac{\partial^2\varphi}{\partial u^2}(x^*+\theta p,x^*+\theta q) = \frac{\partial^2\varphi}{\partial u^2}(x^*+\theta p,x^*) + \theta q\frac{\partial^3\varphi}{\partial u^2\partial v}(x^*+\theta p,x^*+\tau\theta q)\ \tau\in]0,1[$$

d'où

$$\begin{aligned}\varphi(x^*+p,x^*+q) &= x^* + \tfrac{1}{2}pq\Big[\theta\tfrac{\partial^3\varphi}{\partial u^2\partial v}(x^*+\theta p,x^*+\tau\theta q)p \\ &\quad +2\tfrac{\partial^2\varphi}{\partial u\partial v}(x^*+\theta p,x^*+\theta q) \\ &\quad +\theta\tfrac{\partial^3\varphi}{\partial v^2\partial u}(x^*+\tau'\theta p,x^*+\theta q)q\Big]\end{aligned} \tag{4.16}$$

Attention, $\theta,\tau',\tau \in [0,1]$ dépendent de p et q! On pose : $e_k = x_k - x^*$ et

$$\varphi(x^*+p,x^*+q) - x^* = \frac{1}{2}\,pq\,r(p,q).$$

Ainsi

$$e_{k+1} = \frac{1}{2}e_k e_{k-1} r(e_k,e_{k-1})$$

Remarquons que

$$r(0,0) = 2\frac{\partial^2\varphi}{\partial u\partial v}(x^*,x^*).$$

Il existe donc $\delta > 0$, tel que si $|u| \le \delta$, $|v| \le \delta$ alors $|vr(u,v)| \le C < 1$ (continuité de r en $(0,0)$!).
Si $|e_0|$ et $|e_1|$ sont inférieurs à δ alors

$$|e_2| = \frac{|e_1,e_0|}{2}|r(e_1,e_0)| \le C|e_1| \le \delta,$$

puis $|e_1 r(e_2,e_1)| \le C < 1$ d'où $|e_3| \le C|e_2| \le C^2|e_1|$ et plus généralement

$$|e_k| \le C^{k-1}|e_1|$$

i.e. $e_k \to 0$. On a la convergence. ■

4.4.3 Ordre de convergence

Soit $p = \frac{1+\sqrt{5}}{2}$ ($p^2 - p - 1 = 0$). Montrons que $s_k = \frac{|e_{k+1}|}{|e_k|^p}$ a une limite non nulle. On a :

$$\left\{ \begin{array}{lcl} |e_k| & = & s_{k-1}|e_{k-1}|^p \\ |e_{k+1}| & = & s_k|e_k|^p = s_k s_{k-1}^p |e_{k-1}|^{p^2} \end{array} \right. \tag{4.17}$$

donc

$$\underbrace{|r(e_k, e_{k-1})|}_{2r_k} = 2\frac{|e_{k+1}|}{|e_k||e_{k-1}|}$$

$$\begin{aligned} &= 2s_k \quad s_{k-1}^{p-1}|e_{k-1}|^{p^2-p-1} \\ &= 2s_k s_{k-1}^{p-1} \end{aligned}$$

Si $\sigma_k = \ln s_k$, alors $\sigma_k = \log|r_k| - (p-1)\sigma_{k-1}$. Si $\varphi_{u,v}(x^*, x^*) \neq 0$ [1] ($\ln|r_k|$) a une limite l. Posons $\sigma^* = \ell - (p-1)\sigma^*$. Il vient

$$\alpha_k = \beta_k - (p-1)\alpha_{k-1}$$

où $\beta_k = k\ln|r_k| - l$ et $\alpha_k = \sigma_k - \sigma^*$.

Preuve. A-t-on $\alpha_k \to 0$? La réponse est oui (à prouver !). on peut alors conclure ! ■

Exercice 87 *Ne peut-on pas prouver la convergence en se ramenant à des itérations vectorielles ?*

[1] vrai si $f'(x^*) \neq 0$

Chapitre 5

Interpolation

5.1 Introduction

Il y a trois grandes classes d'approximation fonctionnelle d'une fonction numérique $f : [a, b] \to \mathbb{R}$:

1. *L'interpolation.* On approche f par une fonction $\widehat{f}$ appartenant à une classe de fonctions approximantes, par exemple des polynômes, et qui coïncide avec f en n points $x_1, \ldots, x_n \in [a, b]$.
2. *L'approximation aux moindres carrés.* On minimise la somme des carrés des erreurs aux points $x_1, \ldots, x_n$, i.e. on cherche $\widehat{f}$ qui minimise

$$\sum_{i=1}^{n} \left\| f(x_i) - \widehat{f}(x_i) \right\|^2$$

ou plus généralement

$$\int \left\| f(x) - \widehat{f}(x) \right\|^2 d\mu(x)$$

où μ peut être une mesure discrète ou absolument continue avec $d\mu(x) = w(x)\,dx$ auquel cas $w(x)$ est une fonction de poids sur $[a, b]$ (*i.e.* une fonction positive bornée). Dans le dernier cas, on peut aussi considérer une fonction de poids sur $\mathbb{R}$ et approcher une fonction continue de $\mathbb{R}$ dans $\mathbb{R}$. Ce type d'approximation nous conduit à nous placer sur l'espace de Hilbert $\mathcal{L}^2([a, b])$ (ou $\mathcal{L}^2(\mathbb{R})$).

3. *L'approximation uniforme.* On minimise le maximum de l'amplitude de l'erreur (éventuellement pondérée) entre $f(x)$ et son approximation sur $[a, b]$. On travaille donc sur l'espace de Banach des fonctions continues muni de la norme de la convergence uniforme

$$\|f\|_\infty = \sup_{[a,b]} |f(x)|$$

Pour chacune de ces trois classes de méthodes, on doit considérer une famille de fonctions approximantes. Les plus utilisées sont

1. Les polynômes, à cause du théorème de Stone-Weierstrass
2. Les fractions rationnelles, qui conduisent aux approximants de Padé;
3. Les fonctions trigonométriques, qui conduisent aux séries de Fourier.

Dans tous les cas, la famille doit être assez simple pour que l'on puisse facilement établir des résultats et que l'on puisse manipuler les fonctions approximantes, mais aussi assez riche pour pouvoir approcher correctement une fonction quelconque. On a le célèbre résultat suivant:

Théorème 88 ***(Stone-Weierstrass)*** *Soit E un espace métrique compact. Soit A une sous-algèbre de l'espace de Banach $\mathcal{C}_{\mathbb{R}}(E)$ tel que A contient les fonctions constantes et A sépare les points de E, i.e.*

$$\forall x, y \in E \qquad x \neq y \Rightarrow \exists f \in A \quad f(x) \neq f(y)$$

alors A est dense dans $\mathcal{C}_{\mathbb{R}}(E)$.

Ce théorème fondamental admet deux applications immédiates importantes:

Corollaire 89 *Toute fonction numérique continue définie sur un compact $E \subset \mathbb{R}^n$ est limite d'une suite de polynômes qui converge uniformément sur E.*

Corollaire 90 *Toute fonction continue à valeurs complexes définie sur $\mathbb{R}$ périodique de période 2π est limite d'une suite de polynômes trigonométriques*

$$p_n(x) = \sum_{j=-n}^{n} z_j e^{\overline{ijx}}$$

qui converge uniformément sur $\mathbb{C}$.

5.2 Interpolation de Lagrange

On notera $\mathcal{P}_n$ l'ensemble des polynômes à coefficients réels à une variable de degré inférieur ou égal à n. C'est un espace vectoriel de dimension $n+1$. Soit $[a,b] \subset \mathbb{R}$ et f une fonction numérique continue sur $[a,b]$. Soient $\{x_0, \ldots x_n\}$ un ensemble de $n+1$ points distincts de $[a,b]$.

Théorème 91 *Il existe un unique polynôme $p_n \in \mathcal{P}_n$ tel que $p_n(x_i) = f(x_i)$ pour tout $i = 0, \ldots, n$ et ce polynôme peut s'écrire*

$$p_n(x) = \sum_{i=0}^{n} f(x_i)\, l_i(x) \tag{5.1}$$

où

$$l_i(x) \stackrel{\text{def}}{=} \prod_{j=0, j\neq i}^{n} \frac{x - x_j}{x_i - x_j} \tag{5.2}$$

Preuve. Il suffit de vérifier que $l_i(x_j) = \delta_{ij}$ et l'unicité est triviale. ■

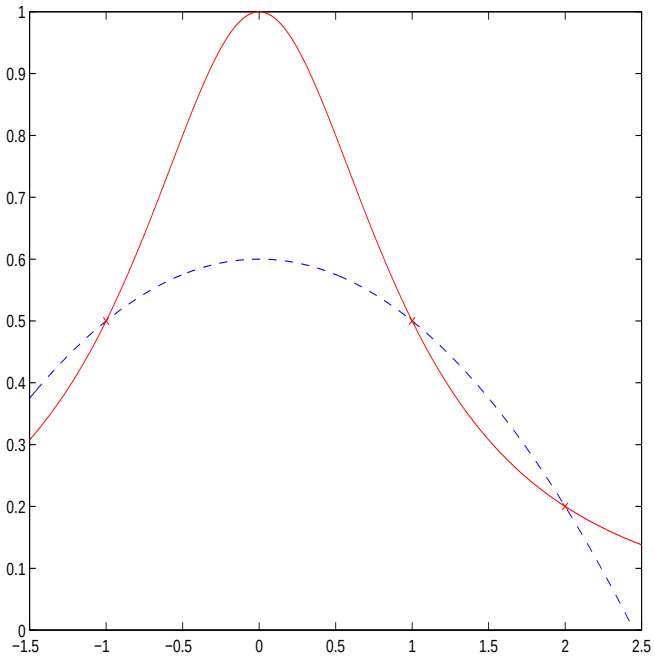


FIG. 5.1: Interpolation de Lagrange de la fonction $f(x) = \frac{1}{1+x^2}$ aux points $-1, 1$ et 2.

Définition 92 *Le polynôme p_n défini par 5.1 et 5.2 s'appelle le polynôme d'interpolation de Lagrange de f aux points $x_0, \ldots, x_n$.*

On remarque que l'application $L_{x_0,\ldots,x_n}$ qui à f associe son polynôme d'interpolation aux points $x_0, \ldots, x_n$ fixés est une application linéaire.

En introduisant

$$\Pi(x) \stackrel{\text{def}}{=} \prod_{j=0}^{n} (x - x_j)$$

on remarque que l_i peut s'écrire plus commodément

$$l_i(x) = \frac{\Pi(x)}{(x - x_i)\,\Pi'(x_i)}$$

Théorème 93 *Supposons $f \in \mathcal{C}^{n+1}_{\mathbb{R}}([a,b])$, alors*

$$\forall x \in [a,b] \quad \exists \xi \in [a,b] \qquad f(x) - p_n(x) = \frac{1}{(n+1)!}\Pi(x)\, f^{(n+1)}(\xi)$$

Preuve. On considère la fonction $F(z) = f(z) - p_n(z) - (f(x) - p_n(x))\frac{\Pi(z)}{\Pi(x)}$ et on déduit du théorème de Rolle que $F^{(n+1)}$ possède au moins un zéro qui sera ξ. ■

5.2.1 Approximation uniforme

Nous avons jusqu'à présent supposé les points $x_0, \ldots, x_n$ fixés. En l'absence de toute information supplémentaire sur f, le théorème 93 nous suggère de choisir $x_0, \ldots, x_n$ de

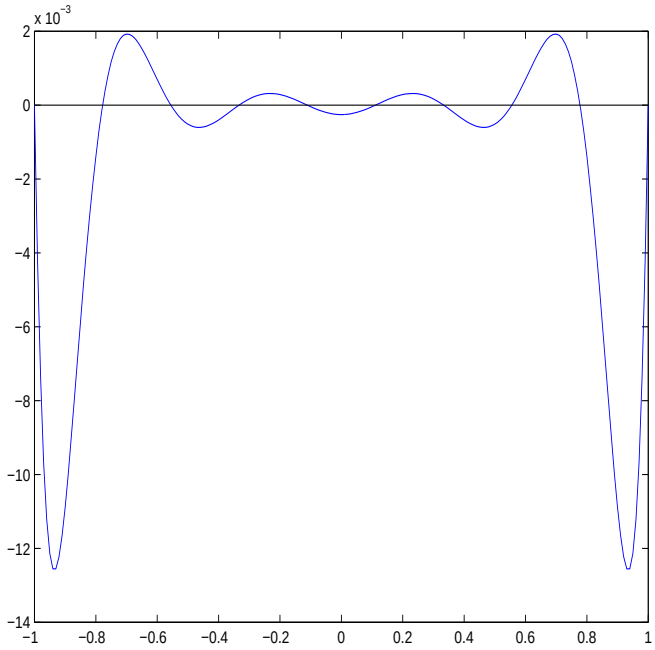


FIG. 5.2: Polynôme de degré 10 dont les racines sont régulièrement espacées.

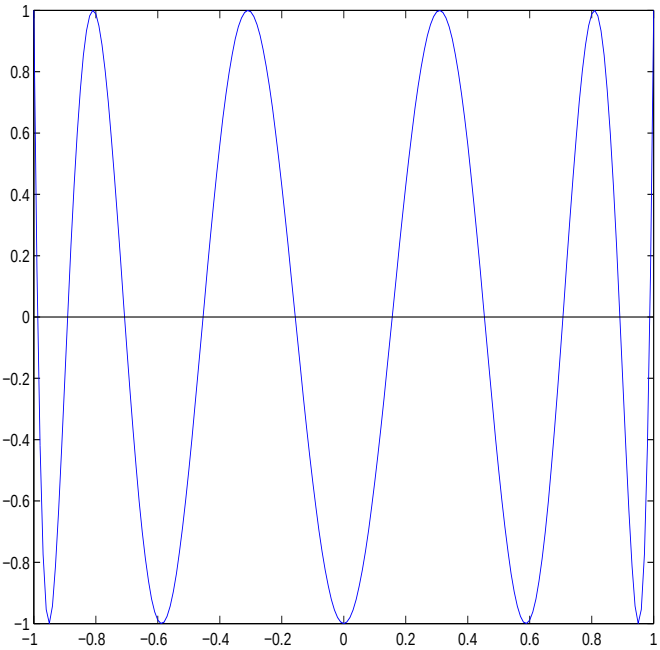


FIG. 5.3: Polynôme de Chebyshev de degré 10.

façon à minimiser $\max_{[a,b]} |\Pi(x)|$. Cette remarque conduit à la notion d'approximation uniforme. Les polynômes qui minimisent la norme de la convergence uniforme sont, à un coefficient près, les polynômes de Chebyshev. On peut les définir sur l'intervalle $[-1,1]$, quitte ensuite à se ramener à un intervalle quelconque par translation/dilatation.

Définition 94 *Les polynômes de Chebyshev sont les polynômes de la forme*

$$T_n(x) \stackrel{\text{def}}{=} \cos(n \arccos(x))$$

Ainsi, $T_0(x) = 1$, $T_1(x) = x$, $T_2(x) = \cos(2t) = 2\cos^2 t - 1 = 2x^2 - 1$ en posant $t = \arccos(x)$, ... La formule de Moivre, qui dit que

$$(\cos a + \underline{i} \sin a)^n = \cos na + \underline{i} \sin na$$

et donc

$$\cos na = \cos^n a - C_n^2 \cos^{n-2} a \sin^2 a + C_n^4 \cos^{n-4} a \sin^4 a - \cdots$$

nous assure que les T_n sont bien des polynômes que l'on peut du reste calculer par récurrence en remarquant que

$$\begin{aligned} T_{n+1}(x) + T_{n-1}(x) &= \cos(n+1)t + \cos(n-1)t \\ &= 2\cos nt \cos t = 2xT_n(x) \end{aligned}$$

On en profite pour remarquer que le polynôme T_n est bien de degré n et par récurrence que le coefficient du terme de plus haut degré est 2^{n-1}. On a aussi le théorème suivant

Théorème 95 ***(Chebyshev)*** *Soit p un polynôme de degré n monique*[1]*,*

$$\sup_{[0,1]} |p(x)| \geq \frac{1}{2^{n-1}}$$

Preuve. Par l'absurde, en posant $q(x) = p(x) - \frac{1}{2^{n-1}} T_n(x)$ et en raisonnant sur le nombre de changements de signe de $q(x)$. ■

Ainsi, la norme $\frac{1}{2^{n-1}}$ est atteinte par le $n^{ième}$ polynôme de Chebyshev divisé par 2^{n-1}. Par conséquent, les $n+1$ points à choisir sur l'intervalle $[-1,1]$ pour l'interpolation polynômiale sont définis par

$$\Pi(x) = \frac{1}{2^n} T_{n+1}(x)$$

et sont donc les racines de T_{n+1}

$$x_i = \cos\left(\frac{(2i+1)\pi}{2n+2}\right) \qquad i = 0, \ldots, n$$

soit sur un intervalle $[a,b]$

$$x_i = \frac{a+b}{2} + \frac{b-a}{2} \cos\left(\frac{(2i+1)\pi}{2n+2}\right) \qquad i = 0, \ldots, n$$

[1] *i.e.* le coefficient du terme de plus haut degré de p est 1.

Les polynômes d'interpolation de Lagrange de f en ces points forment une suite de polynômes qui converge uniformément vers f puisque d'après le théorème 93 et la borne uniforme obtenue sur Π,

$$|f(x) - p_n(x)| \leq \frac{1}{2^n (n+1)!} \sup_{\xi \in [a,b]} \left| f^{(n+1)}(\xi) \right|$$

5.2.2 Calcul des différences divisées

La forme de Lagrange du polynôme d'interpolation n'est pas propice aux calculs. Nous allons voir une autre façon d'exprimer p_n permettant par exemple d'évaluer sa valeur en tout point par la méthode de Horner.

Il est possible de construire les polynômes d'interpolation de Lagrange p_n d'une fonction f aux points $x_0, \ldots, x_n$ par récurrence sur n. Naturellement, $p_0(x) = f(x_0)$. Remarquons d'autre part que pour tout $k \geq 1$, $p_k - p_{k-1}$ est un polynôme de degré k qui s'annule aux points $x_0, \ldots, x_{k-1}$ donc qui peut s'écrire

$$p_k(x) - p_{k-1}(x) = f[x_0, \ldots, x_k](x - x_0) \cdots (x - x_{k-1})$$

où $f[x_0, \ldots, x_k]$ est le coefficient du terme en x^k dans $p_k(x)$. On obtient donc

$$\begin{array}{ll} p_n(x) & = (p_n(x) - p_{n-1}(x)) + (p_{n-1}(x) - p_{n-2}(x)) + \cdots + (p_1(x) - p_0(x)) + p_0(x) \\ & = f[x_0] + \sum_{k=1}^{n} f[x_0, \ldots, x_k](x - x_0) \cdots (x - x_{k-1}) \end{array}$$

Définition 96 *La formule précédente est par définition la forme de Newton du polynôme d'interpolation de Lagrange. Les coefficients $f[x_0, \ldots, x_k]$ sont appelés les différences divisées d'ordre k de f aux points $x_0, \ldots, x_k$.*

Théorème 97 *On a les relations de récurrence*

$$\begin{cases} f[x_i] = f(x_i) \qquad i = 0, \ldots, n \\ f[x_0, \ldots, x_k] = \dfrac{f[x_1, \ldots, x_k] - f[x_0, \ldots, x_{k-1}]}{x_k - x_0} \end{cases}$$

Preuve. On pose

$$q_k(x) = \frac{(x - x_0)\, \widetilde{p}_{k-1}(x) + (x_k - x)\, p_{k-1}(x)}{x_k - x_0}$$

où $\widetilde{p}_{k-1}(x)$ est le polynôme d'interpolation de $f(x)$ aux points $x_1, \ldots, x_k$ et on vérifie que $q_k(x)$ est bien le polynôme d'interpolation de $f(x)$ aux points $x_0, \ldots, x_k$. ■

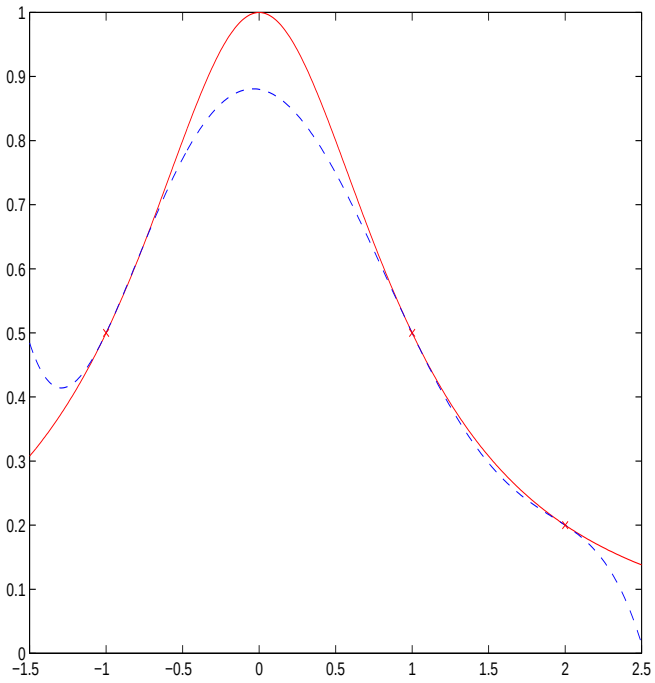
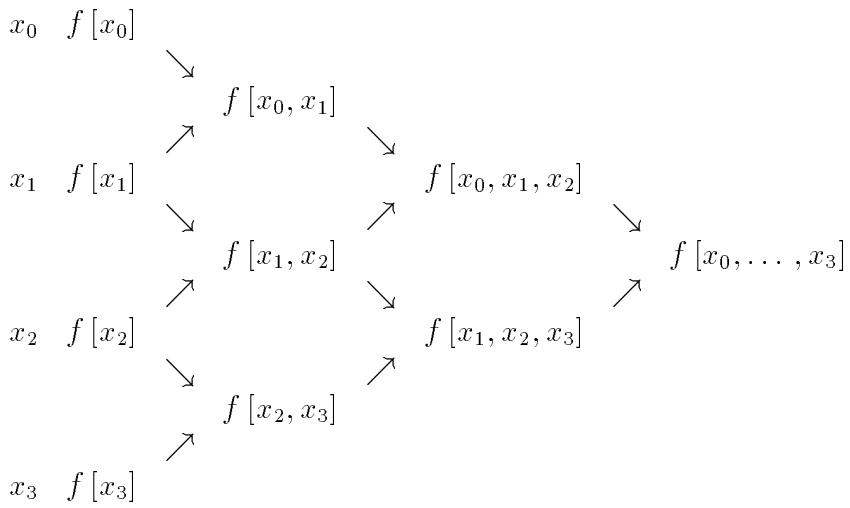


FIG. 5.4: Interpolation de Hermite de la fonction $f(x) = \frac{1}{1+x^2}$ aux points $-1, 1$ et 2.

Ce théorème conduit au schéma de récurrence suivant (pour $n = 3$):



5.3 Interpolation de Hermite

Plutôt que de faire coïncider f et p_n aux points $x_i \in [a, b]$, $i = 0, \ldots, n$, on peut chercher à faire aussi coïncider les dérivées de f et de p_n en ces points, jusqu'à un ordre fixé que l'on notera α_i au point x_i. On a donc

Théorème 98 *Etant donnés $n+1$ points $x_0, \ldots, x_n$ dans $[a, b]$ et $n+1$ entiers naturels $\alpha_0, \ldots, \alpha_n$, posons $N = n + \alpha_0 + \cdots + \alpha_n$. Soit f une fonction continue sur $[a, b]$ admettant*

des dérivées d'ordre α_i aux points x_i. Il existe un unique polynôme $p_N \in \mathcal{P}_N$ tel que

$$\forall i = 0, \ldots, n \quad \forall l = 0, \ldots, \alpha_i \qquad p_N^{(l)}(x_i) = f^{(l)}(x_i)$$

Preuve. On considère le système linéaire $\left\{p_N^{(l)}(x_i) = f^{(l)}(x_i)\right\}$ de $N+1$ équations à $N+1$ inconnues et il suffit de montrer que son système homogène associé $\left\{p_N^{(l)}(x_i) = 0\right\}$ admet 0 comme unique solution. ■

Définition 99 *Le polynôme p_N défini d'après le Théorème 98 est appelé polynôme d'interpolation de Hermite de f aux points $x_0, \ldots, x_n$ aux ordres $\alpha_0, \ldots, \alpha_n$.*

Les polynômes d'interpolation généralisent les polynômes de Lagrange qui correspondent à $\alpha_i = 0$ et le développement de Taylor d'ordre α qui correspond à $n = 0$ et $\alpha_0 = \alpha$.

De même que dans le cas des polynômes de Lagrange, on peut majorer l'erreur entre f et son polynôme d'interpolation.

Théorème 100 *Supposons $f \in \mathcal{C}_{\mathbb{R}}^{N+1}([a,b])$, alors*

$$\forall x \in [a,b] \quad \exists \xi \in [a,b] \qquad f(x) - p_N(x) = \frac{1}{(N+1)!}\overline{\Pi}(x) f^{(N+1)}(\xi)$$

où

$$\overline{\Pi}(x) \stackrel{\text{def}}{=} \prod_{i=0}^{n} (x - x_i)^{\alpha_i + 1}$$

Preuve. La démonstration est une extension de celle du théorème 93. ■

Exemple 101 *Soit $f \in \mathcal{C}_{\mathbb{R}}^{2n}([a,b])$ et $x_1, \ldots, x_n \in [a,b]$. Choisissons $\alpha_1 = \cdots = \alpha_n = 1$. On montre que le polynôme d'interpolation de Hermite est donné par*

$$p_{2n-1}(x) = \sum_{i=1}^{n} \left(1 - 2l_i'(x_i)(x - x_i)\right) l_i^2(x) f(x_i) + \sum_{i=1}^{n} (x - x_i) l_i^2(x) f'(x_i)$$

et la majoration est donc

$$f(x) - p_{2n-1}(x) = \frac{1}{(2n)!}\Pi^2(x) f^{(2n)}(\xi)$$

Corollaire 102 *Le théorème 93*

Corollaire 103 *Posons $M_{N+1} = \sup_{\xi \in [a,b]} \left|f^{(N+1)}(\xi)\right|$, alors*

$$|f(x) - p_N(x)| \leq \frac{M_{N+1}}{(N+1)!} \left|\overline{\Pi}(x)\right|$$

On retrouve une borne uniforme en prenant le sup sur x à droite et en choisissant les points x_i judicieusement.

5.4 Approximation aux moindres carrés

5.4.1 Rappels sur les espaces de Hilbert

Soit E un espace vectoriel de fonctions à valeurs dans $\mathbb{R}$. On suppose E muni d'un produit scalaire, *i.e.* une opération binaire de $E \times E$ dans $\mathbb{R}$ notée $< \cdot, \cdot >$ qui soit
- bilinéaire (linéaire par rapport à chacune de ses variables);
- symétrique ($\langle f, g \rangle = \langle g, f \rangle$);
- définie positive ($\langle f, f \rangle > 0 \quad \forall f \neq 0$).

On vérifie que $\|f\| \stackrel{\text{def}}{=} \langle f, f \rangle^{\frac{1}{2}}$ est une norme sur E i.e.
- $\|f + g\| \leq \|f\| + \|g\|$
- $\|\lambda f\| \leq |\lambda| \, \|f\|$
- $\|f\| = 0 \Leftrightarrow f = 0$

Un tel espace est dit préhilbertien. Si E est complet pour la topologie induite par la norme $\|\cdot\|$, on dit que E est un espace de Hilbert. Un espace de Hilbert est donc un espace de Banach dont la norme provient d'une forme bilinéaire symétrique définie positive.

Deux fonctions telles que $\langle f, g \rangle = 0$ sont dites orthogonales, ce que l'on note $f \perp g$. On a les propriétés classiques dans les espaces préhilbertiens
- L'égalité (Pythagore si f et g sont orthogonales):

$$\|f + g\|^2 = \|f\|^2 + 2 \langle f, g \rangle + \|g\|^2 \tag{5.3}$$

- L'inégalité de Minkowski :

$$\|f + g\| \leq \|f\| + \|g\| \tag{5.4}$$

- L'inégalité de Schwarz :

$$\langle f, g \rangle \leq \|f\| \, \|g\| \tag{5.5}$$

Si E est un espace de Hilbert et que F est une partie fermée convexe non vide de E (par exemple un sous espace vectoriel de dimension finie) alors chaque point de E admet une projection et une seule sur F *i.e.*

$$\exists! g \in F \qquad \|f - g\| = \inf_{h \in F} d \|f - h\|$$

Exemple 104 *$L^2_\mu([a, b])$ est un espace de Hilbert*

Exemple 105 *$\mathcal{C}_{\mathbb{R}}([a, b])$ est préhilbertien pour le produit scalaire $\int_a^b f(x) \, g(x) \, dx$*

Exemple 106 *$L^1_\mu([a, b])$ et $L^\infty_\mu([a, b])$ ne sont que des espaces de Banach (et donc il n'y a pas unicité du polynôme de meilleure approximation dans ces espaces)*

5.4.2 Meilleure approximation dans un espace de Hilbert

On se donne une fonction $w(x)$ définie, continue, strictement positive sur un intervalle $]a, b[\subseteq \mathbb{R}$ telle que pour tout entier n, $x^n w(x) \in L^1([a, b])$. On appelle ce type de fonction une fonction de poids.

Soit $E = \{f : [a,b] \to \mathbb{R}; \quad f\sqrt{w} \in L^2([a,b])\} \stackrel{\text{def}}{=} L^2_w([a,b])$ que l'on munit du produit scalaire

$$\langle f, g \rangle_w = \int_a^b f(x)\, g(x)\, w(x)\, dx$$

et de la norme induite $\|\cdot\|_w$ qui seront notés plus simplement $\langle \cdot, \cdot \rangle$ et $\|\cdot\|$.

Théorème 107 *Il existe une unique suite de polynômes p_0, p_1, ..., p_n, ... telle que*

$$\begin{aligned} \partial^\circ p_n &= n \\ \text{coeff}(p_n, x^n) &= 1 \\ p_n &\perp \mathcal{P}_{n-1} \end{aligned}$$

et cette suite vérifie l'équation de récurrence suivante

$$p_n(x) = (x - \lambda_n)\, p_{n-1}(x) - \mu_n p_{n-2}(x)$$

où

$$\mu_n = \frac{\|p_{n-1}\|^2}{\|p_{n-2}\|^2}$$

et

$$\lambda_n = \frac{\langle x p_{n-1}, p_{n-1} \rangle}{\|p_{n-1}\|^2}$$

Preuve. On applique le procédé d'orthonormalisation de Gram-Schmidt puis on vérifie que la base ainsi obtenue vérifie la récurrence annoncée en constatant que $xp_{n-1}(x) - p_n(x)$ est de degré $n-1$ donc s'écrit $\sum_{i=0}^{n-1} \alpha_i p_i(x)$ et il ne reste plus qu'à identifier les α_i en formant les produits scalaires $\langle xp_{n-1} - p_n, p_i \rangle$. ■

Définition 108 *Les polynômes $(p_n)_{n\in\mathbb{N}}$ s'appellent les polynômes orthogonaux sur $[a,b]$ pour la fonction de poids $w(x)$.*

Exemple 109 *Il existe de nombreux cas remarquables dont les suivants:*

Intervalle	***Fonction de poids***	***Dénomination***
$[-1,1]$	$(1-x)^\alpha (1+x)^\beta$, $\alpha, \beta > 1$	*Polynômes de Jacobi*
$[-1,1]$	1	*Polynômes de Legendre*
$[-1,1]$	$1/\sqrt{1-x^2}$	*Polynômes de Chebyshev $1^{ére}$ espèce*
$[-1,1]$	$\sqrt{1-x^2}$	*Polynômes de Chebyshev 2^{nde} espèce*
$[0,+\infty[$	$x^\alpha e^{-x}$, $\alpha > 0$	*Polynômes associés de Laguerre*
$[0,+\infty[$	e^{-x}	*Poynômes. de Laguerre*
$\mathbb{R}$	e^{-x^2}	*Polynômes d'Hermite*

Proposition 110 *Les n racines du polynôme p_n sont réelles, distinctes et intérieures à l'intervalle $]a,b[$.*

Preuve. On suppose par l'absurde qu'il n'y a que $k < n$ racines réelles distinctes dans $]a,b[$,$x_1,\ldots,x_k$, et on pose $q(x) = (x-x_1)\cdots(x-x_k)$. Il ne reste plus qu'à considérer $\int_a^b p_n(x)\,q(x)\,w(x)\,dx$ qui est nulle alors que $p_n(x)\,q(x)$ garde un signe constant. ■

Théorème 111

$$\exists! q_n \in \mathcal{P}_n; \quad \|f-q_n\| = \inf_{q\in P_n} \|f-q\| \tag{5.6}$$

et c'est aussi l'unique polynôme $q_n \in \mathcal{P}_n$ tel que

$$\forall q \in \mathcal{P}_n \quad \langle q_n, q\rangle = \langle f, q\rangle \tag{5.7}$$

et il est donné par

$$q_n(x) = \sum_{i=0}^{n} \frac{\langle f, p_i\rangle}{\langle p_i, p_i\rangle} p_i(x) \tag{5.8}$$

Preuve. (5.7) signifie que q_n est la projection orthogonale de f sur $\mathcal{P}_n$, l'écriture (5.8) en découle. L'égalité (5.6) résulte de (5.3). ■

Chapitre 6

Intégration numérique

6.1 Introduction

Nous allons maintenant étudier le calcul numérique des intégrales

$$\int_a^b f(x)\, w(x)\, dx$$

où w est une fonction de poids sur l'intervalle $]a,b[\subseteq \mathbb{R}$ que l'on supposera maintenant seulement de signe constant sur l'intervalle $]a,b[$, usuellement positive, et où f est une fonction telle que fw soit intégrable au sens de Riemann sur $[a,b]$.

Nous étudierons trois méthodes:

1. *Les méthodes composées (ou formules de quadrature):* L'intervalle $[a,b]$ est borné et $w(x) = 1$. On subdivise $[a,b]$ en n sous intervalles par le choix de $n+1$ points $a_0, a_1, \ldots, a_n$ dans $[a,b]$ et on approche chaque intégrale $\int_{a_i}^{a_{i+1}} f(x)dx$ en remplaçant f par un polynôme d'interpolation sur $[a_i, a_{i+1}]$.
2. *La méthode de Romberg:* on utilise la méthode d'accélération de la convergence de Richardson pour améliorer la convergence d'une méthode composée: la méthode des trapèzes.
3. *Les méthodes de Gauss:* on approche f par un polynôme d'interpolation sur l'intervalle $[a,b]$ en des points $x_0, x_1, \ldots, x_n$ choisis de façon à obtenir une formule exacte pour des polynômes de degré le plus élevé possible. Nous verrons que cela revient à prendre les points x_i comme racines des polynômes orthogonaux associés aux poids $w(x)$ sur $[a,b]$.

6.2 Les formules de quadrature élémentaires et composées

Soient $a \leq x_0 < x_1 < \cdots < x_n \leq b$ où a et b sont les bornes finies de l'intervalle d'intégration. On supposera dans ce paragraphe que $w(x) = 1$ sur $[a,b]$. Ainsi, en remplaçant

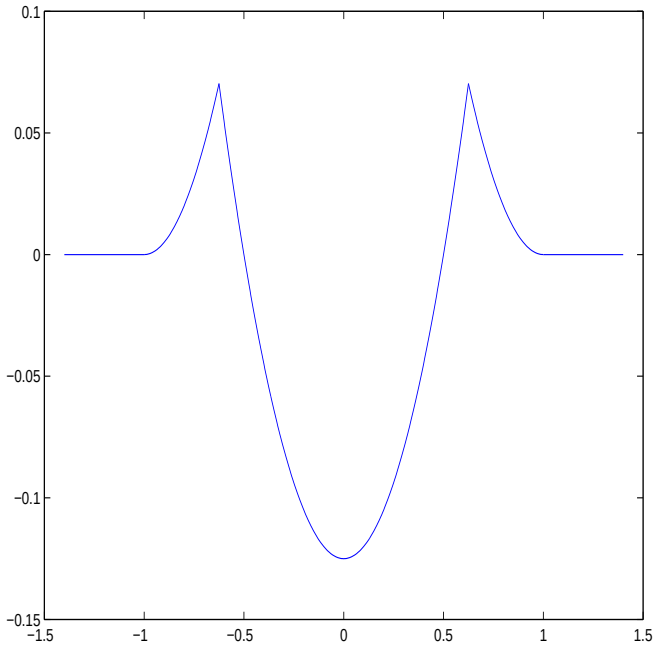


FIG. 6.1: Noyau de Péano d'une méthode d'ordre 1.

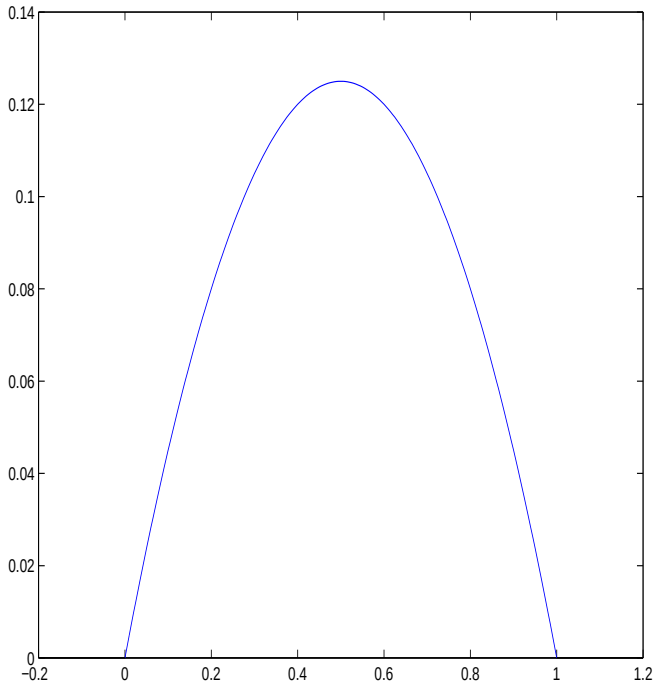


FIG. 6.2: Noyau de Péano correspondant à la méthode des trapèzes.

f par son polynôme d'interpolation de Lagrange et si $f \in \mathcal{C}_{\mathbb{R}}^{n+1}([a,b])$ (cf. Théorème 93)

$$\begin{aligned}\int_a^b f(x)\,dx &= \int_a^b \sum_{i=0}^{n} f(x_i)\,l_i(x) + \frac{1}{(n+1)!}\Pi(x)\,f^{(n+1)}(\xi_x)\,dx \\ &= \sum_{i=0}^{n} \lambda_i f(x_i) + \frac{1}{(n+1)!}\int_a^b \Pi(x)\,f^{(n+1)}(\xi_x)\,dx\end{aligned}$$

où

$$\lambda_i \overset{\text{déf.}}{=} \int_a^b l_i(x)\,dx \text{ et } \Pi(x) = \prod_{i=0}^{n}(x - x_i)$$

ξ_x étant un réel dépendant de x compris entre a et b. Nous voyons que si f est un polynôme de degré inférieur ou égal à n, la formule d'intégration est exacte. Néanmoins, même si $n \to +\infty$, on ne peut pas garantir que pour toute fonction f, ce type de quadrature élémentaire va converger. En pratique, on utilise plutôt des quadratures composées.

Soient $a = a_0 < a_1 < \cdots < a_N = b$, on écrira

$$\int_a^b f(x)\,dx = \sum_{k=0}^{N-1}\int_{a_k}^{a_{k+1}} f(x)\,dx$$

et on applique des formules de quadrature élémentaires pour chaque intégrale de cette somme. Cela revient à subdiviser chaque intervalle $[a_k, a_{k+1}]$ en $(x_i)_{i=0,\ldots,n}$. Supposons donc que l'on ait une subdivision de l'intervalle $[0,1]$ (que l'on translatera) *i.e.* $0 \leq x_0 < x_1 < \cdots < x_n \leq 1$ et la formule de quadrature élémentaire correspondante

$$\int_0^1 \varphi(x)\,dx \simeq \sum_{i=0}^{n} \lambda_i \varphi(x_i)$$

La formule de quadrature composée s'écrira donc

$$\begin{aligned}\int_a^b f(x)\,dx &= \sum_{k=0}^{N-1}\int_{a_k}^{a_{k+1}} f(x)\,dx = \sum_{k=0}^{N-1}(a_{k+1}-a_k)\int_0^1 f(a_k + t(a_{k+1}-a_k))\,dt \\ &\simeq \sum_{k=0}^{N-1}(a_{k+1}-a_k)\sum_{i=0}^{n}\lambda_i f(a_k + x_i(a_{k+1}-a_k))\,dt \qquad (6.1)\end{aligned}$$

Avant d'étudier de plus près ce type de méthode, remarquons que plusieurs méthodes classiques entrent dans cette catégorie:

1. La méthode du rectangle à gauche ($n = 0$ et $x_0 = 0$)
2. La méthode du point milieu ($n = 0$ et $x_0 = \frac{1}{2}$)
3. La méthode du rectangle à droite ($n = 0$ et $x_0 = 1$)
4. La méthode des trapèzes ($n = 1$, $x_0 = 0$ et $x_1 = 1$)
   $l_0(x) = 1 - x \Rightarrow \lambda_0 = \frac{1}{2}$
   $l_1(x) = x \Rightarrow \lambda_1 = \frac{1}{2}$
   $\Rightarrow \int_0^1 f(x)\,dx \simeq \frac{1}{2}(f(0) + f(1))$

5. La méthode de Simpson ($n = 2$, $x_0 = 0$, $x_1 = \frac{1}{2}$ et $x_2 = 1$)
   $l_0(x) = (2x-1)(x-1) \Rightarrow \lambda_0 = \frac{1}{6}$
   $l_1(x) = 4x(1-x) \Rightarrow \lambda_1 = \frac{2}{3}$
   $l_2(x) = x(2x-1) \Rightarrow \lambda_2 = \frac{1}{6}$
   $\Rightarrow \int_0^1 f(x)\,dx \simeq \frac{1}{6}\left(f(0) + 4f\left(\frac{1}{2}\right) + f(1)\right)$
6. La méthode de Boole-Villarceau ($n = 4$)
7. La méthode de Weddle et Hardy ($n = 6$)

A partir de $n = 8$, les méthodes deviennent très sensibles aux erreurs d'arrondi, car on trouve $\lambda_2 = \lambda_4 = \lambda_6 < 0$.

Dans toute ces méthodes (sauf quand $n = 0$), on a choisi des noeuds équidistants. Ce sont les méthodes de Newton-Cotes fermées (on les appelle ouvertes quand on exclu les bornes, ou de Steffensen). On les utilise pour des valeurs paires de n (car elles sont plus efficaces pour des raisons de parité, justement) et pour $n < 8$ car elles deviennent ensuite numériquement sensibles.

Voyons l'efficacité de ces méthodes.

Définition 112 *Une formule de quadrature élémentaire ou composée*

$$\int_a^b f(x)\,dx \simeq \sum_{i=0}^{n} \alpha_i f(x_i)$$

est d'ordre m si elle est exacte pour tout polynôme de degré inférieur ou égale à m et fausse pour au moins un polynôme de degré plus grand que m.

Notation 113 *On pose $E(f) = \int_a^b f(x)\,dx - \sum_{i=0}^{n} \alpha_i f(x_i)$. Notons que E est une fonctionnelle linéaire.*

Notation 114 *Pour toute fonction $u : \mathbb{R} \to \mathbb{R}$, on note*

$$u_+(x) = \max(u(x), 0)$$

et pour tout entier m,

$$u_+^m(x) = \begin{cases} 1 \text{ si } m = 0 \text{ et } u(x) > 0 \\ 0 \text{ si } m = 0 \text{ et } u(x) \leq 0 \\ (u_+(x))^m \text{ sinon} \end{cases}$$

Définition 115 *Le noyau de Péano d'une méthode d'ordre m est défini par*

$$G_m(t) = E\left(u_+^m\right)$$

où $u(x) = x - t$.

Théorème 116 *Pour une méthode d'ordre m et pour toute fonction $f \in C^{m+1}([a,b])$*

$$E(f) = \frac{1}{m!} \int_a^b G_m(t)\, f^{(m+1)}(t)\,dt$$

Preuve. On applique la formule de Taylor à f,

$$f(x) = p_m(x) + \int_a^x \frac{(x-t)\,m}{m!} f^{(m+1)}(t)\,dt$$

puis on calcule $E(f)$ ■

Remarque 117 *Un examen attentif de la démonstration de ce théorème montre que seule la linéarité de E est essentielle.*

Corollaire 118 *Sous les hypothèses du théorème précédent*

$$|E(f)| \leq \frac{1}{m!} \max_{x\in[a,b]} \left|f^{(m+1)}(x)\right| \int_a^b |G_m(t)|\,dt$$

Corollaire 119 *Sous les hypothèses du théorème précédent et si le noyau de Péano garde un signe constant, il existe ξ dans $[a,b]$ tel que*

$$E(f) = \frac{1}{(m+1)!} f^{(m+1)}(\xi)\, E\left(x^{m+1}\right)$$

Preuve. Utiliser la formule de la moyenne. ■

Finallement, il reste à analyser les méthodes composées.

Proposition 120 *Soient $0 \leq x_0 < x_1 < \cdots < x_{n-1} < x_n \leq 1$ et la formule de quadrature élémentaire correspondante*

$$\int_0^1 f(x)dx \simeq \sum_{i=0}^{n} \lambda_i f(x_i)$$

Soit $g_m(t)$ le noyau de Péano associé à cette méthode élémentaire supposée d'ordre m. Soit $[a,b]$ un intervalle de $\mathbb{R}$ et une subdivision de l'intervalle $[a,b]$ en N sous intervalles $a = a_0 < a_1 < \cdots < a_N = b$. On considère la formule de quadrature composée du type (6.1) déduite de la formule de quadrature élémentaire. Alors cette méthode est d'ordre au moins m et si on note $E(f)$ l'erreur entre la formule de quadrature composée et la valeur de l'intégrale pour une fonction donnée f alors

$$E\left(x \rightsquigarrow (x-t)_+^m\right) = \sum_{j=0}^{N-1} (a_{j+1}-a_j)^{m+1}\, g_m\left(\frac{t-a_j}{a_{j+1}-a_j}\right) \tag{6.2}$$

i.e. si elle est d'ordre m alors son noyau de Péano $G_m(t)$ est donné par (6.2).

Preuve. La formule (6.2) s'obtient par un simple calcul, alors que le fait que la méthode composée est au moins celui de la méthode élémentaire est facile à déduire. ■

On tire de cette proposition le théorème important suivant, car c'est lui qui établit la convergence et la vitesse de convergence des formules de quadratures en fonction de l'ordre. C'est de ce théorème que l'on déduit que la méthode des trapèzes converge en $O(h^2)$ puisqu'elle est d'ordre 1 alors que la méthode de Simpson élémentaire étant d'ordre 3, les formules comosées de Simpson convergent en $O(h^4)$ (à vérifier).

Théorème 121 *On se place dans le cadre de la proposition précédente et on note désigne le diamètre de la subdivision par* $h_N = \max_{i=0\ldots N-1}(a_{i+1} - a_i)$. *Soit* $f \in C^{m+1}([a,b])$ *alors*

$$|E(f)| \leq C(b-a)h^{m+1} \left\| f^{(m+1)} \right\|_{C^0([a,b])}$$

où $\|g\|_{C^0([a,b])} = \max_{x\in[a,b]} |g(x)|$, *et* C *est une constante qui ne dépend que de la méthode de quadrature élémentaire.*

Preuve. Il suffit d'appliquer le corollaire 118 en remarquant que le théorème 116 reste vrai même si l'ordre de la méthode est en fait plus grand que m auquel cas le corollaire dit que

$$|E(f)| \leq \frac{1}{m!} \max_{x\in[a,b]} \left| f^{(m+1)}(x) \right| \int_a^b \left| E\left(x \rightsquigarrow (x-t)_+^m\right) \right| dt$$

■

6.3 Méthode de Romberg

Cette méthode très populaire est basée sur une méthode d'accélération de la convergence (l'extrapolation à la limite de Richardson) d'une méthode de quadrature telle que la méthode des trapèzes. Elle est rendue possible grâce à la formule d'Euler-MacLaurin que nous allons démontrer maintenant.

6.3.1 La formule d'Euler-MacLaurin

Soit $f \in \mathcal{C}^n_{\mathbb{R}}([0,1])$. Si $n \geq 1$, une intégration par partie donne

$$\int_0^1 f(x)\,dx = \left[\left(x - \frac{1}{2}\right) f(x)\right]_0^1 - \int_0^1 \left(x - \frac{1}{2}\right) f'(x)\,dx$$

Posons $B_1(x) \overset{\text{def}}{=} x - \frac{1}{2}$, la constante $\frac{1}{2}$ ayant été arbitrairement choisie pour que $\int_0^1 B_1(x)\,dx = 0$. On peut bien sûr recommencer. On défini par récurrence sur n

$$\begin{cases} B_n'(x) = nB_{n-1}(x) \\ \int_0^1 B_n(x)\,dx = 0 \end{cases}$$

et B_n est ainsi entièrement déterminé. C'est un polynôme monique à coefficients rationnels de degré n qui vérifie pour tout $n \geq 2$

$$B_n(1) - B_n(0) = \int_0^1 B_n'(x)\,dx = n\int_0^1 B_{n-1}(x)\,dx = 0$$

et on convient de noter $b_n \overset{\text{def}}{=} B_n(0)$ pour $n \geq 1$ $(= B_n(1)$ pour $n \geq 2)$.

Définition 122 *Les polynômes* B_n *sont appelés polynômes de Bernoulli et les nombres rationnels* b_n *sont appelés nombres de Bernoulli .*

Posons par convention $b_0 = 1$, on a

Théorème 123 $\forall n \geq 1$

$$B_n(x) \overset{(i)}{=} \sum_{p=0}^{n} C_n^p b_p x^{n-p} \overset{(ii)}{=} (-1)^n B_n(1-x)$$

Preuve. On montre (i) par récurrence sur n, en utilisant simplement

$$B_n(x) = B_n(0) + \int_0^x B_n(\xi)\, d\xi = b_n + n \int_0^x B_{n-1}(\xi)\, d\xi$$

L'égalité (ii) se démontre aussi par récurrence, on vérifie que

$$\frac{d}{dx}\left[(-1)^n B_n(1-x)\right] = \frac{d}{dx}\left[B_n(x)\right]$$

et comme $(-1)^n B_n(1-x)$ est aussi d'intégrale nulle sur $[0,1]$, les deux polynômes coïncident. ■

Corollaire 124 $\forall n \geq 2$

$$b_n \overset{(i)}{=} \sum_{p=0}^{n} C_n^p b_p$$

et $b_n = 0$ *si* n *est impair* ≥ 3.

Preuve. La première affirmation découle de l'égalité (i) du théorème en prenant $x = 1$. La seconde est obtenue en posant $x = 0$ dans (ii) ce qui donne $-B_n(1) = B_n(0)$ si n est impair et ≥ 3, soit $b_n = -b_n = 0$. ■

Ceci permet de calculer les premiers nombres de Bernoulli. On a donc

b_0	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	$\cdots$
1	$-\frac{1}{2}$	$\frac{1}{6}$	0	$-\frac{1}{30}$	0	$\frac{1}{42}$	0	$-\frac{1}{30}$	$\cdots$

On décide de prolonger B_n sur $\mathbb{R}$ par

$$\widetilde{B}_n(x) = B_n(x - \lfloor x \rfloor)$$

où $\lfloor x \rfloor$ désigne la partie entière de x. Par abus de notation, on notera toujours $\widetilde{B}_n = B_n$. On a le théorème suivant.

Théorème 125 *(**Formule d'Euler-MacLaurin**) Soit* $f \in \mathcal{C}_{\mathbb{R}}^{2n}([0,m])$ *où* m *est un entier naturel strictement positif et notons*

$$T_1(f) = \frac{1}{2}f(0) + f(1) + \cdots + f(m-1) + \frac{1}{2}f(m)$$

la somme des trapèzes associés à f.

$$T_1(f) = \int_0^m f(x)\, dx + \sum_{p=1}^{n} \frac{b_{2p}}{(2p)!}\left(f^{(2p-1)}(m) - f^{(2p-1)}(0)\right) - \int_0^m \frac{B_{2n}(x)}{(2n)!} f^{(2n)}(x)\, dx$$

Preuve. En itérant les intégrations par parties, on obtient si $n \geq 2$

$$\left[\left(x-\frac{1}{2}\right)f(x)\right]_0^1 = \frac{1}{2}(f(0)+f(1)) = \int_0^1 f(x)\,dx + \int_0^1 B_1(x)\,f'(x)\,dx$$
$$= \int_0^1 f(x)\,dx + \left[\frac{1}{2}B_2(x)\,f'(x)\right]_0^1 - \int_0^1 \frac{B_2(x)}{2}f''(x)\,dx$$

puis si n est assez grand

$$\frac{1}{2}(f(0)+f(1)) = \int_0^1 f(x)\,dx + \sum_{p=2}^{n}(-1)^p\frac{b_p}{p!}\left(f^{(p-1)}(1)-f^{(p-1)}(0)\right)$$
$$+(-1)^{n+1}\int_0^1 \frac{B_n(x)}{n!}f^{(n)}(x)\,dx$$

On applique cette formule sur chaque intervalle $[k,k+1]$, en tenant compte de la périodicité des polynômes de Bernoulli, qui n'en sont plus vraiment, et du fait que b_p est nul pour tout p impair ≥ 3. Il ne reste plus qu'a sommer pour obtenir le résultat annoncé. ■

La formule d'Euler-MacLaurin relie donc l'évaluation par la méthode des trapèzes et la vraie valeur de l'intégrale, à travers un développement limité de f. On peut bien sûr appliquer cette formule sur une fonction $f \in \mathcal{C}_{\mathbb{R}}^{2n}([a,b])$, où $[a,b]$ est maintenant un intervalle quelconque de $\mathbb{R}$.

Corollaire 126 *Soit $f \in \mathcal{C}_{\mathbb{R}}^{2n+2}([a,b])$ et notons*

$$T_h(f) = \frac{h}{2}(f(a)+2f(a+h)+\cdots+2f(b-h)+f(b))$$

la somme des trapèzes associés à f lorsque l'intervalle $[a,b]$ est divisé en m sous-intervalles de longueur $h=\frac{b-a}{m}$.

$$T_h(f) = \int_a^b f(x)\,dx + \sum_{p=1}^{n}\frac{b_{2p}h^{2p}}{(2p)!}\left(f^{(2p-1)}(b)-f^{(2p-1)}(a)\right)+O\left(h^{2n+2}\right)$$

Preuve. Il suffit de poser $\widetilde{f}(x) = f(a+xh)$ pour $x \in [0,m]$, et d'appliquer la formule d'Euler-MacLaurin à $\widetilde{f}$ entre 0 et m à l'ordre $n+1$. On obtient le développement annoncé avec même une erreur en $o(h^{2n+1})$. ■

6.3.2 L'extrapolation à la limite de Richardson

On suppose que l'on a un algorithme dépendant d'un paramètre h qui permet de calculer une quantité $A(h)$ et on cherche $a_0 \stackrel{\text{def}}{=} \lim_{h\to 0} A(h)$. Si l'algorithme est assez régulier, on peut écrire

$$A(h) = a_0 + a_1 h + a_2 h^2 + \cdots + a_n h^n + O(h^{n+1})$$

donc connaissant $A(h)$ pour plusieurs valeurs de h, on devrait pouvoir extrapoler sa valeur en $h=0$.

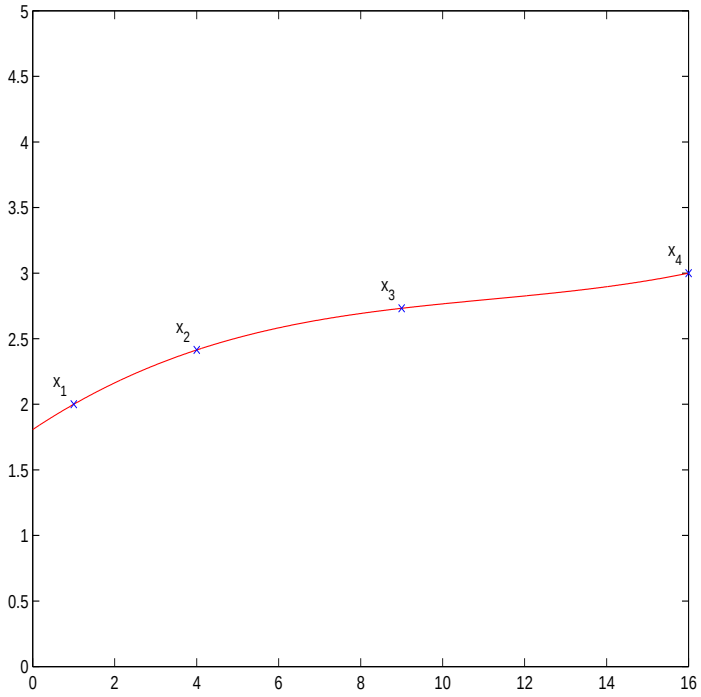


FIG. 6.3: Principe de l'extrapolation à la limite de Richardson

Il nous faut pour cela éliminer les a_k. Remarquons que pour tout $r > 1$

$$A(rh) = a_0 + a_1 rh + a_2 r^2 h^2 + \cdots + a_n r^n h^n + O(h^{n+1})$$

donc en multipliant par r^m, $m \in \{0, \ldots, n\}$, on élimine facilement le terme en h^m:

$$\frac{r^m A(h) - A(rh)}{r^m - 1} = a_0 + a_1 \frac{r^m - r}{r^m - 1} h + \cdots + a_m \frac{r^m - r^m}{r^m - 1} h^m + \cdots + a_n \frac{r^m - r^n}{r^m - 1} h^n + O(h^{n+1})$$

Ainsi, en calculant successivement

$$\begin{aligned} A_0(h) &= A(h) \\ A_1(h) &= \frac{r A_0(h) - A_0(rh)}{r - 1} \\ &\ldots \text{etc} \ldots \\ A_n(h) &= \frac{r^n A_{n-1}(h) - A_{n-1}(rh)}{r^n - 1} \end{aligned}$$

on élimine tous les coefficients des termes de degré $\leq n$ et donc finalement

$$A_n(h) = a_0 + o(h^n)$$

En regardant de plus près, on voit que pour calculer $A_n(h)$, il faut connaître $A_{n-1}(h)$ et $A_{n-1}(rh)$ donc $A_{n-2}(h)$, $A_{n-2}(rh)$ et $A_{n-2}(r^2 h)$, etc...

Notons $A_k^m \overset{\text{def}}{=} A_k\left(hr^{-m}\right)$, les formules de récurrence sont donc

$$\begin{aligned} A_k^m = A_k\left(hr^{-m}\right) &= \frac{r^k A_{k-1}\left(hr^{-m}\right) - A_{k-1}\left(hr^{-m+1}\right)}{r^k - 1} \\ &= \frac{r^k A_{k-1}^m - A_{k-1}^{m-1}}{r^k - 1} \text{ pour } k = 1, \ldots, n \text{ et } m = k, \ldots, n \end{aligned}$$

qui vaut donc $a_0 + O\left(h^{m+1}\right)$ et la quantité qui nous intéresse est A_k^m pour m et k les plus grands possibles (correspondant ainsi à l'erreur minimum pour un pas minimum). Les $A_0^m = A\left(hr^{-m}\right)$ sont donnés par l'algorithme des trapèzes. L'ordre des calculs est donc donné par le tableau suivant

$$\begin{array}{ccccccc} A_0^0 & \rightarrow & A_0^1 & \rightarrow & A_0^2 & \rightarrow & A_0^3 \\ & \searrow & \downarrow & \searrow & \downarrow & \searrow & \downarrow \\ & & A_1^1 & & A_1^2 & & A_1^3 \\ & & & \searrow & \downarrow & \searrow & \downarrow \\ & & & & A_2^2 & & A_2^3 \\ & & & & & \searrow & \downarrow \\ & & & & & & A_3^3 \end{array}$$

6.3.3 Application

Soit $f \in C_{\mathbb{R}}^{2n+2}\left([a,b]\right)$. Le corollaire 126 nous dit que

$$T_h(f) = \int_a^b f(x)\,dx + \sum_{p=1}^{n} \frac{b_{2p} h^{2p}}{(2p)!}\left(f^{(2p-1)}(b) - f^{(2p-1)}(a)\right) + O\left(h^{2n+2}\right)$$

donc

$$T_h(f) = a_0 + \sum_{p=1}^{n} a_p h^{2p} + O\left(h^{2n+2}\right)$$

avec les notations *ad hoc* et en particulier $a_0 = \int_a^b f(x)\,dx$. Mais du fait que seuls les carrés de h interviennent, il convient de prendre $r = 4$ plutôt que $r = 2$. En effet, on choisit logiquement de diminuer le pas de 2 à chaque itération et donc de calculer $T_{b-a}(f)$, $T_{\frac{b-a}{2}}(f)$, etc, et on pose donc

$$A_0^m = T_{\frac{b-a}{2^m}}(f)$$

mais puisque les coefficients impaires sont déjà nuls, il suffit de tuer les coefficients paires donc pour tuer le premier terme par exemple, on calculera

$$A_1^1 = \frac{2^2 T_{\frac{b-a}{2}}(f) - T_{b-a}(f)}{2^2 - 1} = \frac{4A_0^1 - A_0^0}{4 - 1}$$

donc on appliquera en fait la méthode de Richardson avec $r = 4$.

Ainsi, la méthode des trapèzes appliquée à une fonction suffisamment régulière se prête très bien au procédé d'accélération de la convergence de Richardson. La méthode qui vient d'être décrite est la méthode de Romberg. On peut l'illustrer sur un petit exemple académique: soit à estimer

$$\int_0^{\pi} \sin(x)\, dx = 2$$

par le méthode des trapèzes: on calcule successivement

$$\begin{aligned}
T_{\pi}(\sin) &= \frac{\pi}{2}(\sin(0) + \sin(\pi)) = 0 \\
T_{\pi/2}(\sin) &= \frac{\pi}{4}\left(\sin(0) + 2\sin\left(\frac{\pi}{2}\right) + \sin(\pi)\right) = \frac{\pi}{2} \\
T_{\pi/4}(\sin) &= \frac{\pi}{8}\left(\sin(0) + 2\sin\left(\frac{\pi}{4}\right) + 2\sin\left(\frac{\pi}{2}\right) + 2\sin\left(\frac{3\pi}{4}\right) + \sin(\pi)\right) \\
&= \frac{\pi}{4}\left(1+\sqrt{2}\right) \simeq 1.8961 = 2 \pm 5\,\%
\end{aligned}$$

puis on applique la méthode de Richardson. On pose donc

$$A_0^0 = 0,\ A_0^1 = \frac{\pi}{2} \text{ et } A_0^2 = \frac{\pi}{4}\left(1+\sqrt{2}\right)$$

puis

$$\begin{aligned}
A_1^1 &= \frac{4A_0^1 - A_0^0}{4-1} = \frac{2}{3}\pi \\
A_1^2 &= \frac{4A_0^2 - A_0^1}{4-1} = \frac{1}{6}\pi + \frac{1}{3}\pi\sqrt{2}
\end{aligned}$$

et enfin

$$\begin{aligned}
A_2^2 &= \frac{16A_1^2 - A_1^1}{16-1} \\
&= \frac{2}{15}\pi + \frac{16}{45}\pi\sqrt{2} \simeq 1.99857
\end{aligned}$$

La méthode des trapèzes conduit donc à une erreur relative de plus de 5 % alors que l'erreur relative a l'issu de la méthode de Romberg est inférieure à $0,1\,\%$ sans faire plus d'évaluations de la fonction à intégrer.

6.4 Méthodes de Gauss

Pour une fonction de poids $w(x)$ donnée, on a vu (cf. Théorème 107) comment calculer une suite $(p_n)_{n\in\mathbb{N}}$ de polynômes orthogonaux. On peut les normaliser si on ne les suppose plus moniques: on note alors a_n le coefficient choisi strictement positif de x^n dans $p_n(x)$, et $x_1, \ldots, x_n$ les n racines réelles distinctes de p_n dans $[a,b]$ (cf. Théorème 110).

Théorème 127 *Il existe n constantes $\lambda_1, \ldots, \lambda_n$ strictement positives telles que pour tout polynôme q de degré au plus $2n-1$, on ait*

$$\int_a^b q(x)\, w(x)\, dx = \sum_{i=1}^{n} \lambda_i q(x_i)$$

Preuve. On fait la division Euclidienne de q par p_n, $q = \alpha p_n + \beta$ puis on calcule $\int_a^b q(x)\, w(x)\, dx$ en tenant compte du fait que α et β sont tous deux de degré au plus $n-1$. On pose $\lambda_i = \int_a^b l_i(x)\, w(x)\, dx$ qui sont bien des constantes strictement positives: calculer $\int_a^b (l_i(x))^2\, w(x)\, dx$. ■

Théorème 128 *Si $f \in L^2_w([a,b]) \cap \mathcal{C}^{2n}_{\mathbb{R}}([a,b])$ alors il existe ξ dans $[a,b]$ tel que*

$$\int_a^b f(x)\, w(x)\, dx = \sum_{i=1}^{n} \lambda_i f(x_i) + \frac{1}{(2n)! a_n^2} f^{(2n)}(\xi)$$

Preuve. Utiliser l'exemple 101, intégrer, et remarquer que $p_n(x) = a_n \Pi(x)$. ■

Chapitre 7

Équations différentielles

7.1 Introduction

Remarque 129 *Cette partie du cours requiert quelques connaissances sur la théorie des équations différentielles, avec au minimum le théorème de Cauchy-Lipschitz dans le cas globalement Lipschitzien: il est illusoire d'espérer comprendre ce chapitre sans avoir compris ce théorème.*

7.1.1 la méthode d'Euler

Soit $f(t,x)$ une application de $[t_0, t_0+T]\times\mathbb{R}^d$ dans $\mathbb{R}^d$ continue en t et Lipschitzienne en x uniformément en t *i.e.* il existe une constante L telle que pour tout triplet $(t,x,y)\in [t_0,t_0+T]\times\mathbb{R}^d\times\mathbb{R}^d$

$$\|f(t,x)-f(t,y)\|\leq L\|x-y\|$$

($\|\cdot\|$ désigne la norme Euclidienne).

On considère le problème de Cauchy (où $\dot{x}(t)$ est une notation abrégée de $\frac{dx}{dt}(t)$, *i.e.* le point représente la dérivée par rapport au temps t)

$$\begin{cases} \frac{dx}{dt}(t) = f(t,x(t)) \\ x(t_0)=\eta \end{cases}$$

que l'on veut résoudre numériquement. On notera toujours $x(t)$ la solution de ce problème, qui existe et qui est unique avec nos hypothèses sur f: c'est une fonction dérivable de $[t_0,t_0+T]$ dans $\mathbb{R}^d$.

La méthode la plus simple est la méthode d'Euler: on subdivise $[t_0,t_0+T]$ en introduisant N points tels que $t_0<t_1<t_2<\cdots<t_N=t_0+T$.

Notation 130 *On pose*

$$\begin{aligned} h_n &= t_{n+1}-t_n \\ h_* &= \min_{n=0\ldots N-1} h_n \\ h^* &= \max_{n=0\ldots N-1} h_n \end{aligned}$$

et on suppose que la subdivision est choisie de telle façon que $h^* \leq H$, H *étant une constante fixée une fois pour toute. On dira que la subdivision tend uniformément vers zéro et on écrira simplement*

$$h \to 0$$

si il existe une constante $0 < \varrho \leq 1$ *telle que*

$$h^* \geq h_* \geq \varrho h^* \text{ et } h^* \underset{n\to\infty}{\longrightarrow} 0$$

i.e. "la subdivision tend vers 0 *uniformément".*

La méthode d'Euler consiste à approcher

$$x(t_{n+1}) = x(t_n) + \int_{t_n}^{t_{n+1}} f(s, x(s))\, ds$$

par la méthode du rectangle à gauche

$$x(t_{n+1}) \simeq x(t_n) + h_n f(t_n, x(t_n))$$

et à remplacer cette approximation par une égalité

$$x_{n+1} = x_n + h_n f(t_n, x_n)$$

ce qui s'interprète aussi comme une approximation par la tangente puisqu'alors

$$\frac{x_{n+1} - x_n}{t_{n+1} - t_n} = f(t_n, x_n)$$

On supposera $x_0 = \eta$. On se propose de montrer un résultat de convergence de la méthode d'Euler, *i.e.* $x_n \to x(t_n)$ quand $h \to 0$. Plus précisément,

Théorème 131

$$\sup_{n=0\ldots N-1} \|x_n - x(t_n)\| \underset{h\to 0}{\longrightarrow} 0$$

Preuve. La démonstration fait appel à deux outils importants que nous allons développer immédiatement: le module de continuité d'une fonction et le lemme de Gronwall. Il faut ensuite majorer $\varepsilon_n = \|x_n - x(t_n)\|$ uniformément en n par une fonction de la subdivision qui tend vers zéro quand $h \to 0$. On utilisera le lemme de Gronwall (Lemme 135) en posant $\delta_n = \left\| \int_{t_n}^{t_{n+1}} f(s, x(s))\, ds - h_n f(t_n, x(t_n)) \right\|$ qui est ce qu'on appelle l'erreur de consistance qu'il faudra majorer en utilisant le module de continuité de f. ■

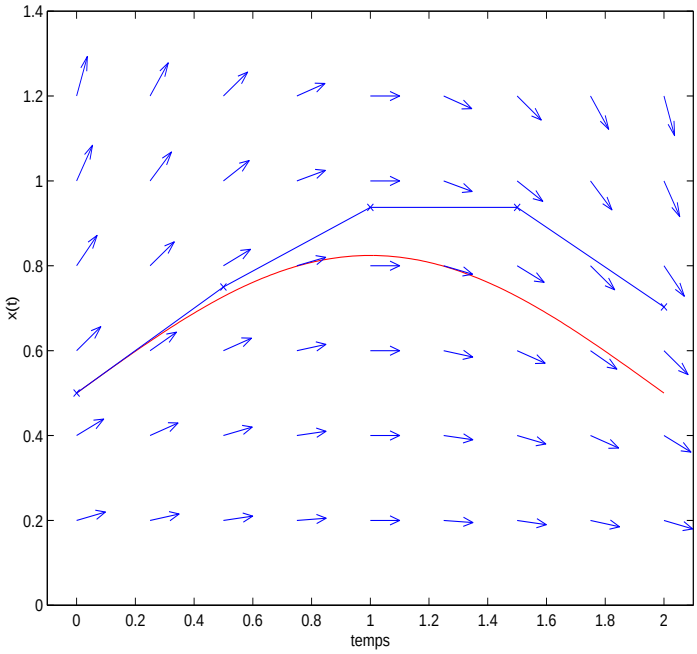


FIG. 7.1: Interprétation géométrique de la méthode d'Euler pour l'équation différentielle $\frac{dx}{dt}(t) = (1-t)\,x(t)$

7.1.2 Module de continuité

Définition 132 *Soit $f(t,x)$ une application de $[t_0, t_0+T]\times\mathbb{R}^d$ dans $\mathbb{R}^d$ continue en t et Lipschitzienne en x uniformément en t. On appelle module de continuité de f et on note ω_f la fonction de $\mathbb{R}^+$ dans $\mathbb{R}$ définie par*

$$\omega_f(\delta) \overset{\text{déf.}}{=} \sup_{(x,t,t')\in C} \{\|f(t,x) - f(t',x)\|\,;\,|t-t'|\leq\delta\}$$

où $C = [t_0, t_0+T]\times B(\eta, R)$ est un cylindre de sécurité pour la solution du problème de Cauchy i.e. un ensemble tel que pour tout $t\in[t_0,t_0+T]$, $(t,x(t))\in C$. L'existence d'un tel ensemble est une conséquence presque immédiate du caractère Lipschitzien de f.

Proposition 133 *Le module de continuité est une fonction continue croissante sous-additive, cette dernière propriété signifiant que pour tout $\alpha,\beta\geq 0$, $\omega_f(\alpha+\beta)\leq\omega_f(\alpha)+\omega_f(\beta)$.*

Preuve. On montre successivement que ω_f est croissante, sous-additive et continue en zéro. La continuité découle de ces trois propriétés. ■

Corollaire 134 *f est uniformément continue dans C.*

7.1.3 Lemme de Gronwall

Lemme 135 *(Gronwall) Soit $t_0 < t_1 < t_2 < \cdots < t_N = t_0+T$ et h_n, h_*, h^* définis comme précédemment (Notation 130). Soit $(\varepsilon_n)_{n=0\ldots N-1}$ une suite numérique vérifiant pour tout*

$n = 0, \ldots, N-1$

$$0 < \varepsilon_{n+1} < \varepsilon_n (1 + h_n L) + \delta_n$$

où $0 < \delta_n < D$ *alors*

$$\begin{aligned} \varepsilon_n &\leq \mathrm{e}^{L(t_n - t_0)} \varepsilon_0 + \sum_{i=0}^{n-1} \mathrm{e}^{L(t_n - t_{i+1})} \delta_i \\ &\leq \mathrm{e}^{L(t_n - t_0)} \varepsilon_0 + \frac{D}{L h_*} \left(\mathrm{e}^{L(t_n - t_0)} - 1 \right) \end{aligned}$$

Preuve. Par récurrence sur n en utilisant simplement le fait que $1 + x \leq \mathrm{e}^x$ pour la première majoration. La seconde est obtenue en remarquant que

$$h_* \mathrm{e}^{L(t_n - t_{i+1})} \leq h_i \mathrm{e}^{L(t_n - t_{i+1})} \leq \int_{t_i}^{t_{i+1}} \mathrm{e}^{L(t_n - s)} ds$$

■

7.2 Généralités sur les méthodes à un pas

On étudie toujours le problème de Cauchy mais cette fois dans $\mathbb{R}$:

$$\begin{cases} \frac{dx}{dt}(t) = f(t, x(t)) \\ x(t_0) = \eta \end{cases}$$

où $f : [t_0, t_0 + T] \times \mathbb{R} \to \mathbb{R}$ est lipschitzienne de constante de Lipschitz L *i.e.*

$$\forall (t, x, y) \in [t_0, t_0 + T] \times \mathbb{R}^2 \qquad |f(t, x) - f(t, y)| \leq L |x - y|$$

De même que pour une méthode d'Euler, on se donne $t_0 < t_1 < t_2 < \cdots < t_N = t_0 + T$, $N - 1$ points intermédiaires. On garde les notations 130 concernant h_n, h_* et h^* ainsi que la notation "$h \to 0$".

7.2.1 Définitions

Définition 136 *On appelle méthode à un pas tout schéma d'intégration numérique qui peut se mettre sous la forme*

$$\begin{cases} x_{n+1} = x_n + h_n \Phi(t_n, x_n, h_n) \\ x_0 = \eta \end{cases}$$

où Φ *est une fonction continue de* $[t_0, t_0 + T] \times \mathbb{R} \times [0, H]$ *dans* $\mathbb{R}$ *qui ne dépend que de la fonction* f.

Exemple 137 *La méthode d'Euler est l'exemple le plus simple avec* $\Phi(t, x, h) = f(t, x)$

Exemple 138 *Les méthodes de Runge-Kutta que l'on verra plus tard.*

Notre objectif sera de construire des méthodes d'ordre $p > 1$. La méthode d'Euler est d'ordre 1 car l'erreur est en $o(h)$ alors que certaines méthodes de Runge-Kutta sont d'ordre 4 (en $o(h^4)$). Voici trois notions fondamentales:

Définition 139 *(**Consistance**) Une méthode est dite consistante si*

$$\sum_{n=0}^{N-1} |x(t_{n+1}) - x(t_n) - h_n \Phi(t_n, x(t_n), h_n)| \underset{h\to 0}{\longrightarrow} 0$$

(ce qui se traduit par: "l'erreur d'approximation en remplaçant l'équation différentielle par le schéma numérique tend vers 0 quand le pas tend vers 0). Comme d'habitude, $x(t)$ est la solution du problème de Cauchy considéré.

Définition 140 *(**Stabilité**) Une méthode est dite stable s'il existe un réel M (indépendant de la subdivision) tel que si*

$$x_{n+1} = x_n + h_n \Phi(t_n, x_n, h_n),\ x_0 \text{ quelconque fixé} \tag{7.1}$$

et pour $(\varepsilon_n)_{n=0,\dots,N-1} \subset \mathbb{R}^N$

$$y_{n+1} = y_n + h_n \Phi(t_n, y_n, h_n) + \varepsilon_n,\ y_0 \text{ quelconque fixé} \tag{7.2}$$

on ait

$$\max_{0\le n<N} |x_n - y_n| \le M \left(|x_0 - y_0| + \sum_{n=0}^{N-1} |\varepsilon_n| \right)$$

(qui est une propriété de continuité, absolument nécessaire rien que pour les problèmes d'arrondi)

Définition 141 *(**Convergence**) Une méthode est dite convergente si*

$$\max_{0\le n<N} |x(t_n) - x_n| \underset{h\to 0}{\longrightarrow} 0$$

Il semble intuitif que les notions de consistance et de stabilité sont complémentaires: si l'erreur commise a chaque pas est faible, et si le cumul des erreurs ne fait pas diverger la méthode, elle doit être convergente. De fait, on a un premier résultat (presque) immédiat:

Théorème 142 *Toute méthode à un pas stable et consistante est convergente.*

Preuve. Il suffit de poser $\varepsilon_n = x(t_{n+1}) - x(t_n) - h_n \Phi(t_n, x(t_n), h_n)$ dans (7.2), de vérifier par récurrence que $y_n = x(t_n)$ puis de conclure par stabilité puis consistance. ■

7.2.2 Etude générale

Nous allons maintenant caractériser les méthodes stables et consistantes (donc... convergentes).

Théorème 143 *Une condition nécessaire et suffisante pour qu'une méthode à un pas soit consistante est que*

$$\forall (t,x) \in [t_0, t_0 + T] \times \mathbb{R} \qquad \Phi(t, x, 0) = f(t, x)$$

Preuve. On pose $\varepsilon_n = x(t_{n+1}) - x(t_n) - h_n \Phi(t_n, x(t_n), h_n)$ et on applique le théorème des accroissements finis: $\exists \tau_n \in]t_n, t_{n+1}[$

$$\begin{aligned} \varepsilon_n &= h_n (f(\tau_n, x(\tau_n)) - \Phi(t_n, x(t_n), h_n)) \\ &= h_n (f(\tau_n, x(\tau_n)) - \Phi(\tau_n, x(\tau_n), 0)) \qquad (7.3) \\ &+ h_n (\Phi(\tau_n, x(\tau_n), 0) - \Phi(t_n, x(t_n), h_n)) \qquad (7.4) \end{aligned}$$

On pose $\alpha_n = (7.3)$ et $\beta_n = (7.4)$. On vérifie que

$$\sum_{n=0}^{N-1} |\beta_n| \underset{h \to 0}{\longrightarrow} 0$$

alors que

$$\sum_{n=0}^{N-1} |\alpha_n| \underset{h \to 0}{\longrightarrow} \int_{t_0}^{t_0+T} |f(t, x(t)) - \Phi(t, x(t), 0)| \, dt$$

ce qui permet de conclure, dans les deux sens, en utilisant le théorème de Cauchy-Lipschitz. ■

Théorème 144 *Une condition suffisante pour que la méthode soit stable est qu'il existe une constante Λ telle que $\forall t \in [t_0, t_0 + T]$, $\forall x, y \in \mathbb{R}$, $\forall h \in [0, H]$*

$$|\Phi(t, x, h) - \Phi(t, y, h)| \leq \Lambda |x - y|$$

et la constante de stabilité est alors $e^{\Lambda T}$.

Preuve. Utiliser le lemme de Gronwall, page 71. ■

Définition 145 *Nous dirons qu'une méthode d'intégration numérique est d'ordre $p > 0$ s'il existe une constante K ne dépendant que de t_0, T, et Φ telle que toute solution $(t \mapsto x(t)) \in C^{p+1}_{\mathbb{R}}([t_0, t_0 + T])$ du problème de Cauchy*

$$\begin{cases} \frac{dx}{dt}(t) = f(t, x(t)) \\ x(t_0) = \eta \end{cases}$$

vérifie

$$\sum_{n=0}^{N-1} |x(t_{n+1}) - x(t_n) - h_n \Phi(t_n, x(t_n), h_n)| \leq K h^p$$

Théorème 146 *Si une méthode est stable et d'ordre p, et si $f \in C^p_{\mathbb{R}}([t_0, t_0 + T] \times \mathbb{R})$ alors*

$$\max_{0 \leq n < N} |x(t_n) - x_n| \leq O(h_*^p)$$

Preuve. Similaire à la démonstration du théorème 142. ■

Il nous faut maintenant caractériser les méthodes d'ordre p comme nous avons caractérisé la consistance. Pour cela, on définit $f^{(k)}(t,x)$ comme la $k^{\text{ième}}$ dérivée de f par rapport à t lorsque x est vu comme une solution dépendant du temps du problème de Cauchy. Plus formellement, on pose

$$\begin{cases} f^{(0)}(t,x) = f(t,x) \\ f^{(k)}(t,x) = \frac{\partial}{\partial t} f^{(k-1)}(t,x) + \frac{\partial}{\partial x} f^{(k-1)}(t,x) f(t,x) \end{cases}$$

Ainsi, on a le résultat de régularité de la solution des équations différentielles suivant:

Théorème 147 *Soit $x(t)$ une solution du problème de Cauchy $\frac{dx}{dt}(t) = f(t, x(t))$ avec condition initiale $x(t_0) = \eta$. Si f est p fois différentiable au point $(t, x(t))$ alors x est $p+1$ fois différentiable au point t et*

$$x^{(p+1)}(t) = f^{(p)}(t, x(t))$$

Théorème 148 *Supposons que f soit p fois continûment différentiable dans $[t_0, t_0+T] \times \mathbb{R}$ et que Φ, $\frac{\partial \Phi}{\partial h}$, $\ldots$, $\frac{\partial^p \Phi}{\partial h^p}$ existent et soient continûment différentiable dans $[t_0, t_0+T] \times \mathbb{R} \times [0, H]$. Alors, une condition nécessaire et suffisante pour que la méthode soit d'ordre p est que $\forall t \in [t_0, t_0+T]$, $\forall x, y \in \mathbb{R}$, $\forall h \in [0, H]$*

$$\begin{cases} \Phi(t,x,0) = f(t,x) \\ \frac{\partial \Phi}{\partial h}(t,x,0) = \frac{1}{2} f^{(1)}(t,x) \\ \\ \frac{\partial^{p-1} \Phi}{\partial h^{p-1}}(t,x,0) = \frac{1}{p} f^{(p-1)}(t,x) \end{cases}$$

Preuve. La démonstration ressemble à celle du théorème 143 en poussant le développement limité à l'ordre $p-1$. ■

Remarque 149 *(Définition équivalente à le définition 145)Sous les hypothèses du théorème précédent, il devient clair qu'une méthode est d'ordre p si et seulement si il existe une constante K ne dépendant que de t_0, T, et Φ telle que toute solution $(t \mapsto x(t)) \in C^{p+1}_{\mathbb{R}}([t_0, t_0+T])$ du problème de Cauchy vérifie*

$$|x(t_{n+1}) - x(t_n) - h_n \Phi(t_n, x(t_n), h_n)| \leq K h^{p+1}$$

En effet, cette condition clairement suffisante est aussi nécessaire, ce que l'on peut vérifier en développant $x(t_n + h_n)$ au voisinage de $h_n = 0$.

Enfin, on a le théorème aux multiples applications suivant, qui donne une estimation de l'erreur. On supposera que $h_n = h$.

Théorème 150 *Supposons que f soit p fois continûment différentiable dans $[t_0, t_0+T] \times \mathbb{R}$ et que la méthode à un pas utilisée soit stable d'ordre $p \geq 1$ telle que Φ soit aussi $p+1$ fois continûment différentiable dans $[t_0, t_0+T] \times \mathbb{R} \times [0, H]$. Alors*

$$x(t_n) - x_n = h^p \delta(t_n) + O\left(h^{p+1}\right)$$

où $\delta(t)$ est elle même solution de

$$\begin{cases} \frac{d\delta}{dt}(t) = \frac{\partial f}{\partial x}(t, x(t))\,\delta(t) + \Psi_p(t, x(t)) \\ \delta(t_0) = 0 \end{cases} \tag{7.5}$$

entre t_0 et $t_0 + T$ et où l'on a posé

$$\Psi_p(t,x) = \frac{1}{(p+1)!} f^{(p)}(t,x) - \frac{1}{p!}\frac{\partial^p \Phi}{\partial h^p}(t,x,0)$$

Preuve. On pose $\xi_n = x(t_n) - h^p \delta(t_n)$ et on montre que $\xi_n = x_n + O(h^{p+1})$. Pour cela, on écrit

$$\begin{aligned} \xi_{n+1} - \xi_n - h\Phi(t_n, \xi_n, h) = x(t_{n+1}) - x(t_n) - h\Phi(t_n, x(t_n), h) & \quad (7.6) \\ + h(\Phi(t_n, x(t_n), h) - \Phi(t_n, \xi_n, h)) & \quad (7.7) \\ - h^p(\delta(t_{n+1}) - \delta(t_n)) & \quad (7.8) \end{aligned}$$

puis on majore chaque terme (7.6) et (7.7) respectivement par $h^{p+1}\Psi_p(t_n, x(t_n)) + O(h^{p+2})$ et par $h^{p+1}\frac{\partial f}{\partial x}(t_n, x(t_n))\,\delta(t_n) + O(h^{p+2})$ et on reconnaît le schéma d'Euler appliqué à l'équation (7.5). ■

7.3 Les méthodes de Runge-Kutta

7.3.1 Introduction: méthode de Taylor

Une première idée pour construire des méthodes d'ordre supérieur est d'étendre la méthode d'Euler en poussant le développement limité à l'ordre p soit

$$x(t_n + h_n) = x(t_n) + \sum_{k=1}^{p} \frac{h_n^k}{k!} f^{(k-1)}(t_n, x(t_n)) + O\left(h_n^{p+1}\right)$$

qui conduit à la méthode à un pas où

$$\Phi(t,x,h) = \sum_{k=1}^{p} \frac{h^{k-1}}{k!} f^{(k-1)}(t,x)$$

C'est effectivement une méthode à un pas (appliquer le Théorème 148) qui est stable si par exemple toutes les dérivée de f sont bornées (appliquer le Théorème 144). Nous allons voir que les méthodes de Runge-Kutta sont basées sur cette approche mais évitent le recours explicite aux dérivées de f.

7.3.2 Méthodes de Runge-Kutta d'ordre quelconque

L'idée est d'introduire q points intermédiaires dans chaque intervalle de la subdivision $[t_n, t_{n+1}]$ que l'on notera $t_{n,1}, \ldots, t_{n,q}$, ceci de la façon suivante: si on se donne $c_1, \ldots, c_q$

dans l'intervalle $[0,1]$ alors on pose $t_{n,i} = t_n + c_i h_n$ pour $i = 1 \dots q$. On a

$$\begin{aligned} x(t_{n,i}) &= x(t_n) + \int_{t_n}^{t_{n,i}} f(t, x(t))\, dt \\ &= x(t_n) + h_n \int_0^{c_i} f(t_n + \tau h_n, x(t_n + \tau h_n))\, d\tau \end{aligned}$$

et bien sûr

$$x(t_{n+1}) = x(t_n) + h_n \int_0^1 f(t_n + \tau h_n, x(t_n + \tau h_n))\, d\tau$$

On choisi $q+1$ méthodes de quadrature à q points $c_1, \dots, c_q$ pour approximer ces intégrales, ce qui revient à se donner $(q+1)\,q$ paramètres $(a_{ij})_{1 \le i,j \le q}$ et $(b_i)_{1 \le i \le q}$ tels que

$$\int_0^{c_i} g(\tau)\, d\tau \simeq \sum_{j=1}^{q} a_{ij} g(c_j)$$

$$\int_0^1 g(\tau)\, d\tau \simeq \sum_{i=1}^{q} b_i g(c_i)$$

On obtient ainsi les approximations suivantes

$$x(t_{n,i}) \simeq x(t_n) + h_n \sum_{j=1}^{q} a_{ij} f(t_n + c_j h_n, x(t_n + c_j h_n))$$

$$x(t_{n+1}) \simeq x(t_n) + h_n \sum_{i=1}^{q} b_i f(t_n + c_i h_n, x(t_n + c_i h_n))$$

La méthode de Runge-Kutta consiste à remplacer ces approximations par des égalités, comme pour la méthode d'Euler page 70, donc on introduit $x_{n,i}\,(\simeq x(t_{n,i}))$ et on pose

$$x_{n,i} = x_n + h_n \sum_{j=1}^{q} a_{ij} f(t_{n,j}, x_{n,j})$$

$$x_{n+1} = x_n + h_n \sum_{i=1}^{q} b_i f(t_{n,i}, x_{n,i})$$

C'est une méthode à un pas où

$$\Phi(t, x, h) = \sum_{i=1}^{q} b_i f(t + c_i h, x_i)$$

$$\text{où } x_i = x + h \sum_{j=1}^{q} a_{ij} f(t + c_j h, x_j)$$

A cause de la façon dont les x_i sont définis, elle n'est pas nécessairement explicite, *i.e.*

$$\left\{ x_i = x + h \sum_{j=1}^{q} a_{ij} f\left(t + c_j h, x_j\right) \right\}_{i=1\ldots q} \tag{7.9}$$

constitue *a priori* un système de q équations non linéaires à q inconnues $x_1, \ldots, x_q$.

Ce type de méthode est parfaitement caractérisé par les paramètres $(a_{ij})_{1\leq i,j\leq q}$, $(b_i)_{1\leq i\leq q}$ et $(c_j)_{1\leq j\leq q}$ que l'on a coutume de représenter sous la forme d'un tableau

$$\begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \cdots & a_{1q} \\ c_2 & a_{21} & \ddots & & \\ \vdots & \vdots & & & \vdots \\ c_q & a_{q1} & & \cdots & a_{qq} \\ \hline & b_1 & b_2 & \cdots & b_q \end{array}$$

7.3.3 Etude des méthodes de Runge-Kutta

Notation 151 *On pose*

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1q} \\ a_{21} & \ddots & & \\ \vdots & & & \vdots \\ a_{q1} & & \cdots & a_{qq} \end{pmatrix} \qquad b = \begin{pmatrix} b_1 \\ \vdots \\ \vdots \\ b_q \end{pmatrix}$$

$$C = \begin{pmatrix} c_1 & 0 & \cdots & 0 \\ 0 & \ddots & & \\ \vdots & & \ddots & \vdots \\ 0 & & & c_q \end{pmatrix} \qquad e = \begin{pmatrix} 1 \\ \vdots \\ \vdots \\ 1 \end{pmatrix}$$

Lemme 152 *Supposons A triangulaire inférieure à diagonale nulle (auquel cas la méthode de Runge-Kutta associée est explicite). On note $\|.\|_\infty$ la norme infinie dans $\mathbb{R}^q$ et on rappelle que*

$$\|A\|_\infty = \sup_{x\neq 0} \frac{\|Ax\|_\infty}{\|x\|_\infty} = \max_{1\leq i\leq q} \sum_{j=1}^{q} |a_{ij}|$$

Alors

$$|x_i - y_i| \leq \sum_{j=1}^{i} \left(Lh \|A\|_\infty\right)^{j-1} |x - y|$$

où y_i est défini à partir de y de la même façon que x_i, i.e. $y_i = y + h\sum_{j=1}^{q} a_{ij} f\left(t + c_j h, y_j\right)$

Preuve. Par récurrence. ■

Théorème 153 *Les méthodes de Runge-Kutta explicites sont stables et leur constante de stabilité (cf Théorème 144) vaut $e^{\Lambda T}$ avec*

$$\Lambda = L\sum_{i=1}^{q}|b_i|\left(1+Lh\left\|A\right\|_\infty+\cdots+\left(Lh\left\|A\right\|_\infty\right)^{i-1}\right)$$

Preuve. On applique le Théorème 144. ■

Proposition 154 *Dès que $\sum_{i=1}^{q} b_i = 1$, la méthode de Runge-Kutta est consistante.*

Preuve. On applique le Théorème 143, c'est donc aussi une condition nécessaire. ■
Un méthode de Runge-Kutta, pour être consistante, doit donc vérifier

$$b^T e = 1 \qquad \text{(Ordre 1)}$$

Pour être d'ordre 2, elle doit en plus satisfaire

$$b^T Ce = b^T Ae = \frac{1}{2} \qquad \text{(Ordre 2)}$$

On le montre en utilisant le théorème 148. On peut aller encore plus loin mais les calculs deviennent très pénibles au delà de l'ordre 3: une méthode de Runge-Kutta d'ordre 4 satisfait les conditions précédentes Ordre 1 et Ordre 2, mais aussi les conditions Ordre 3 et Ordre 4 qui sont celles d'ordre 3 et 4 respectivement:

$$\left\{\begin{array}{l} b^T C^2 e = b^T CAe = b^T\left(Ae\right)^{\otimes 2} = \frac{1}{3} \\ b^T ACe = b^T A^2 e = \frac{1}{6} \end{array}\right. \qquad \text{(Ordre 3)}$$

(la notation $y^{\otimes k}$ signifie que toutes les composantes de y sont élevées à la puissance k)

$$\left\{\begin{array}{l} b^T C^3 e = b^T C^2 Ae = b^T C\left(Ae\right)^{\otimes 2} = b^T C\left(Ae\right)^{\otimes 3} = \frac{1}{4} \\ b^T CACe = b^T CA^2 e = b^T\left(Ae\right)\otimes\left(ACe\right) = b^T\left(Ae\right)\otimes\left(A^2 e\right) = \frac{1}{8} \\ b^T AC^2 e = b^T ACAe = b^T A\left(Ae\right)^{\otimes 2} = \frac{1}{12} \\ b^T A^2 Ce = b^T A^3 e = \frac{1}{24} \end{array}\right. \qquad \text{(Ordre 4)}$$

Cela fait 21 conditions pour 24 paramètres, mais ces conditions sont non-linéaires. La solution utilisée habituellement est la **méthode de Runge-Kutta d'ordre 4 classique** décrite par le tableau suivant:

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & \frac{1}{6} & \frac{2}{6} & \frac{2}{6} & \frac{1}{6} \end{array}$$

qui correspond aux équations

$$\left\{\begin{array}{l} x_{n,1} = x_n \\ x_{n,2} = x_n + \frac{h_n}{2} f\left(t_n, x_{n,1}\right) \\ x_{n,3} = x_n + \frac{h_n}{2} f\left(t_n + \frac{h_n}{2}, x_{n,2}\right) \\ x_{n,4} = x_n + h_n f\left(t_n + \frac{h_n}{2}, x_{n,3}\right) \\ x_{n+1} = x_n + h_n \left(\frac{1}{6} f\left(t_n, x_{n,1}\right) + \frac{2}{6} f\left(t_n + \frac{h_n}{2}, x_{n,2}\right) + \frac{2}{6} f\left(t_n + \frac{h_n}{2}, x_{n,3}\right) + \frac{1}{6} f\left(t_{n+1}, x_{n,4}\right)\right) \end{array}\right.$$

On a enfin le résultat suivant, sur les méthodes implicites (A triangulaire inférieure à diagonale nulle) ou semi-implicites (A triangulaire inférieure):

Théorème 155 *Si $hL\rho(|A|) < 1$, le système (7.9)*

$$\left\{ x_i = x + h \sum_{j=1}^{q} a_{ij} f\left(t + c_j h, x_j\right) \right\}_{i=1\ldots q}$$

admet une solution unique (pour toute matrice M, $\rho(M)$ désigne le rayon spectral de M et $|M| = (|M_{ij}|)$).

Preuve. On appelle Θ l'application de $\mathbb{R}^q$ dans $\mathbb{R}^q$ définie par ses composantes

$$\Theta_i(\xi) = x + h \sum_{j=1}^{q} a_{ij} f\left(t + c_j h, \xi_j\right)$$

de telle façon que le système 7.9 s'écrive $\xi = \Theta(\xi)$ et on montre que Θ^m est contractante pour m assez grand. Le théorème du point fixe permet de conclure. ■

chapter:8,page:69,theorem:157

Chapitre 8

Équations aux dérivées partielles

8.1 Dimension 1 (problème aux limites)

On considère le problème suivant

$$\begin{cases} -u''(x) = f(x) & x \in [0,1] \\ u(0) = u(1) = 0 \end{cases} \tag{8.1}$$

pour lequel on a déjà brièvement présenté une méthode de résolution approchée au chapitre 2: la méthode des différences finies, cf 2.1.1.

8.1.1 Méthode de Galerkin

Soit $v(x)$ une fonction $\mathcal{C}^1([0,1])$ sauf éventuellement en un nombre fini de points et nulle en 0 et en 1: 8.1 entraîne

$$-\int_0^1 u''(x)\, v(x)\, dx = \int_0^1 f(x)\, v(x)\, dx$$

soit en intégrant par parties

$$\int_0^1 u'(x)\, v'(x)\, dx - \underbrace{u'(0)\, v(1)}_{=0} + \underbrace{u'(0)\, v(0)}_{=0} = \int_0^1 f(x)\, v(x)\, dx$$

d'où

$$\int_0^1 u'(x)\, v'(x)\, dx = \int_0^1 f(x)\, v(x)\, dx \qquad \forall v \in V \tag{8.2}$$

en posant

$$V = \{v \in \mathcal{C}^0([0,1]); \quad v(0) = v(1) = 0,$$
$$v'(x) \text{ continue à droite et à gauche et continue par morceaux}\}$$

Une solution de 8.2 s'appelle solution faible de 8.1. Les deux problèmes 8.1 et 8.2 ne sont pas équivalents. Il est clair par construction que toute solution de 8.1 est une

solution de 8.2 mais la réciproque est fausse: une solution de 8.2 peut ne pas être deux fois continûment différentiable et donc ne pas être solution forte. Nous allons voir cependant qu'il existe une très bonne méthode pour résoudre 8.2 donc pour trouver une solution forte quant elle existe.

Pour cela, remarquons déjà que V est un sous-espace vectoriel de $\mathcal{C}^1([0,1])$de dimension infinie. Nous allons approcher une solution faible de 8.2 en cherchant $\widetilde{u} \in \widetilde{V}$ telle que

$$\int_0^1 \widetilde{u}'(x)\,\widetilde{v}'(x)\,dx = \int_0^1 f(x)\,\widetilde{v}(x)\,dx \qquad \forall \widetilde{v} \in \widetilde{V} \tag{8.3}$$

où $\widetilde{V}$ sera un s.e.v. de V de dimension finie.

Soient donc $\varphi_1, \ldots, \varphi_N$ N fonctions linéairement indépendantes de V et on pose alors $\widetilde{V} = \mathrm{Vect}_{\mathbb{R}}(\varphi_1, \ldots, \varphi_N)$. Ainsi, toute fonction $\widetilde{u} \in \widetilde{V}$ s'écrira

$$\widetilde{u}(x) = \sum_{j=1}^{N} u_j \varphi_j(x)$$

et résoudre 8.3 revient à chercher $u = (u_1, \ldots, u_N)^T \in \mathbb{R}^N$ vérifiant

$$\sum_{j=1}^{N} u_j \int_0^1 \varphi_j'(x)\,\widetilde{v}'(x)\,dx = \int_0^1 f(x)\,\widetilde{v}(x)\,dx \qquad \forall \widetilde{v} \in \widetilde{V}$$

soit encore par linéarité la condition suffisante (et évidemment nécessaire)

$$\sum_{j=1}^{N} u_j \int_0^1 \varphi_j'(x)\,\varphi_i'(x)\,dx = \int_0^1 f(x)\,\varphi_i(x)\,dx \qquad \forall \varphi_i \in \widetilde{V}$$

Posant $A = (a_{ij})_{1 \le i,j \le N}$ et $b = (b_i)_{1 \le i \le N}$ avec $a_{ij} = \int_0^1 \varphi_j'(x)\,\varphi_i'(x)\,dx$ $1 \le i,j \le N$ et $b_i = \int_0^1 f(x)\,\varphi_i(x)\,dx$ $1 \le i \le N$, on obtient le système linéaire à résoudre $Au = b$. La matrice A est symétrique. Avant de voir comment choisir les fonctions de base φ_i de sorte que A soit la plus pertinente possible (inversible mais aussi de sorte que $Au = b$ soit simple à résoudre numériquement quand N est grand), nous allons donner une première estimation de l'erreur commise en résolvant 8.3 a la place de 8.2.

Définition 156 *Pour tout $v \in V$, on note*

$$|v|_1 = \left(\int_0^1 (v'(x))^2\,dx \right)^{\frac{1}{2}}$$

Proposition 157 *$|\cdot|_1$ est une norme sur V.*

Théorème 158 *Soit u une solution de 8.2 et $\widetilde{u}$ une solution de 8.3, alors*

$$|u - \widetilde{u}|_1 = \min_{\widetilde{v} \in \widetilde{V}} |u - \widetilde{v}|_1$$

Preuve. On forme la différence entre 8.2 et 8.3 et on obtient l'orthogonalité de Galerkin

$$\int_0^1 \left(u'(x) - \widetilde{u}'(x)\right) v'(x)\, dx = 0$$

puis on utilise cette propriété conjointement avec l'inégalité de Cauchy–Schwartz pour calculer $\int_0^1 \left(u'(x) - \widetilde{u}'(x)\right)^2 dx$ et obtenir l'inégalité cherchée. ■

8.1.2 Méthode des éléments finis

Nous allons nous intéresser dans cette partie au choix des fonctions de base φ_i. On introduit N points intermédiaires dans l'intervalle $]0,1[$ notés $0 = x_0 < x_1 < \cdots < x_N < x_{N+1} = 1$ et on note pour $i = 1, \ldots, N$

$$\varphi_i(x) = \begin{cases} \frac{x - x_{i-1}}{x_i - x_{i-1}} & \text{si } x_{i-1} \leq x \leq x_i \\ \frac{x - x_{i+1}}{x_i - x_{i+1}} & \text{si } x_i \leq x \leq x_{i+1} \\ 0 & \text{sinon} \end{cases}$$

Il est clair que $\varphi_i \in V$ et que les fonctions $\varphi_1, \ldots, \varphi_N$ sont linéairement indépendantes dans V. Ces fonctions sont appelées les éléments finis de degré 1.Si $u \in V$, alors la fonction $\widehat{u}(x) = \sum_{i=1}^N u(x_i)\varphi_i(x)$ est une fonction d'interpolation. De plus, puisque toutes les fonctions φ_i sont affines par morceaux entre les points x_i et que en tout point x_j, $\widehat{u}(x_j) = \sum_{i=1}^N u(x_i)\varphi_i(x_j) = u(x_j)$, $\widehat{u}(x)$ n'est rien d'autre que l'interpolation linéaire de $u(x)$ entre les points x_i. On peut montrer le théorème suivant:

Théorème 159 *Si u est deux fois continûment différentiable alors*

$$|u - \widehat{u}|_1 \leq Ch$$

où $h = \max_{i=0,\ldots,N} |x_{i+1} - x_i|$ est le pas de la subdivision et $C = |u'|_1$

Preuve. Il suffit d'appliquer le théorème de Rolle entre les points x_i. ■

Corollaire 160 *Soit u une solution de 8.2 et $\widetilde{u}$ une solution de 8.3, alors*

$$|u - \widetilde{u}|_1 \leq Ch$$

Il reste à étudier le système linéaire obtenu avec ce choix des fonctions de base. Pour commencer, on peut regarder le cas plus simple est assez fréquent où les points x_i sont choisis équidistribués dans l'intervalle $[0,1]$ i.e. $x_i = ih$ avec $h = \frac{1}{N+1}$.

Exercice 161 *Montrer que dans ce cas,*

$$A = \frac{1}{h}\begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & -1 & 2 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{pmatrix}$$

et montrer que A est inversible (fait en TD!)

Nous verrons ultérieurement que l'intérêt de la méthode des éléments finis, comparativement à la méthode des différences finies, est justement le fait que l'on peut facilement adapter la subdivision, appelée aussi maillage, en fonction du problème. Lorsque les points ne sont pas nécessairement choisis équidistant, nous avons le résultat suivant:

Proposition 162 *A est tridiagonale symétrique inversible telle que $a_{ii} > 0$ et $a_{ij} < 0$ lorsque $|i-j| = 1$.*

Preuve. Le caractère symétrique tridiagonal de A se déduit immédiatement de sa définition. On vérifie aussi facilement les propriétés sur les signes des coefficients en raisonnant sur les signes des φ_i' sur chaque intervalle $[x_j, x_{j+1}]$.Reste la régularité de A. Elle s'obtient en considérant le système homogène associé en remarquant que pour tout $j = 2, \ldots, N-1$

$$\sum_{i=1}^{N} a_{ij} = \int_0^1 \varphi_j'(x) \left(\varphi_{j-1}'(x) + \varphi_j'(x) + \varphi_{j+1}'(x)\right) dx = 0$$

puisque $\varphi_{j-1} + \varphi_j + \varphi_{j+1}$ est constante sur $[x_{j-1}, x_{j+1}]$. ■

La méthode des éléments finis consiste donc à

1. Choisir N points entre 0 et 1;
2. Construire la matrice A;
3. Déterminer le vecteur b en calculant chacune de ses composantes par une méthode d'intégration comme celles que l'on a vu au chapitre 6;
4. Résoudre le système linéaire en utilisant l'une des méthodes du chapitre 2;
5. Retourner l'approximation de la solution faible $\widetilde{u}(x) = \sum_{i=1}^{N} u_i \varphi_i(x)$

Remarque 163 *Si les points sont choisis équidistants et si les composantes de b sont calculées par la méthode des trapèzes, la méthode obtenue est exactement la méthode des différences finies. A vérifier en exercice.*

Remarque 164 *La méthode des éléments finis décrite précédemment peut être généralisée en choisissant pour éléments finis des fonctions de degré 2 ou plus (interpolation quadratique, fonctions splines,...).*

8.2 Dimension 2 (problème elliptique))

On considère maintenant le problème suivant

$$\begin{cases} -\triangle u(x) = f(x) & x \in \Omega \\ u(x) = 0 & x \in \partial\Omega \end{cases} \tag{8.4}$$

où cette fois Ω est un domaine connexe de $\mathbb{R}^2$ et $x = (x_1, x_2)^T \in \overline{\Omega}$. L'opérateur $\triangle$ est le laplacien i.e.

$$\triangle u = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2}$$

Physiquement, la solution de ce problème correspondrait au déplacement vertical d'une membrane tendue fixée sur $\partial\Omega$ et soumise à un champ de force orthogonal d'intensité proportionnelle à f. Nous allons utiliser la même technique qu'en dimension 1. Soit $v : \Omega \longrightarrow \mathbb{R}$

$$-\iint_{\Omega} \triangle u\,(x)\, v\,(x)\, dx = \iint_{\Omega} f\,(x)\, v\,(x)\, dx$$

où $dx = dx_1 dx_2$. Ici, nous allons utiliser la formule d'intégration par partie en dimension 2 qui est la formule de Green :

$$\iint_{\Omega} \frac{\partial y}{\partial x_i}\, z\, dx = -\iint_{\Omega} y\, \frac{\partial z}{\partial x_i}\, dx + \int_{\partial\Omega} y\, z\, n_i\, ds \quad i = 1, 2$$

où n_i est la $\text{i}^{\text{ième}}$ composante de la normale unitaire extérieure à Ω et où s est un paramètrage de $\partial\Omega$ Alors

$$\begin{aligned} -\iint_{\Omega} \triangle u\, v\, dx &= -\sum_{i=1}^{2} \iint_{\Omega} \frac{\partial^2 u}{\partial x_i^2}\, v\, dx \\ &= \sum_{i=1}^{2} \iint_{\Omega} \frac{\partial u}{\partial x_i}\, \frac{\partial v}{\partial x_i}\, dx - \sum_{i=1}^{2} \int_{\partial\Omega} \frac{\partial u}{\partial x_i}\, v\, n_i\, ds \\ &= \iint_{\Omega} \nabla u\, \nabla v\, dx - \int_{\partial\Omega} (\nabla u.n)\, v\, ds \end{aligned}$$

qui est la formule de Stokes en notant ∇u le gradient de u. On a obtenu la formulation faible suivante

$$\iint_{\Omega} \nabla u\, \nabla v\, dx - \underbrace{\int_{\partial\Omega} (\nabla u.n)\, v\, ds}_{=0} = \iint_{\Omega} f\, v\, dx$$

en imposant $v = 0$ sur $\partial\Omega$ soit comme dans le cas mono–dimensionnel

$$\iint_{\Omega} \nabla u\, \nabla v\, dx = \iint_{\Omega} f\, v\, dx \text{ pour tout } v \in V \tag{8.5}$$

où $V = \left\{ v \in \mathcal{C}_{\mathbb{R}}\left(\overline{\Omega}\right),\ \left(\frac{\partial v}{\partial x_i}\right)_{i=1,2} \text{ continue par morceaux, nulle sur } \partial\Omega \right\}$. Ici, continue par morceaux signifie continue sur des parties de $\overline{\Omega}$ formant une partition de $\overline{\Omega}$.

Ainsi, procédant de même que dans la section précédente, on introduit N fonctions linéairement indépendantes $\varphi_1, \ldots, \varphi_N$ de V et on définit le sous–espace vectoriel $\widetilde{V}$ de V par $\widetilde{V} = \text{Vect}_{\mathbb{R}}\left(\varphi_1, \ldots, \varphi_N\right)$ puis nous cherchons $u = \sum_{j=1}^{N} u_j \varphi_j\,(x) \in \widetilde{V}$ tel que $\forall v \in \widetilde{V}$ ou de manière équivalente $\forall \varphi_i$, $i = 1, \ldots, N$

$$\sum_{j=1}^{N} u_j \iint_{\Omega} \nabla \varphi_i\, \nabla \varphi_j\, dx = \iint_{\Omega} f\, \varphi_i\, dx \tag{8.6}$$

qui est équivalent au système linéaire $Au = b$ où A est la matrice $(a_{ij})_{1\leq i,j\leq N}$ définie par

$$a_{ij} = \iint_{\Omega} \nabla\varphi_i \nabla\varphi_j \, dx$$

et où

$$b_i = \iint_{\Omega} f \, \varphi_i \, dx$$

soit à nouveau un problème linéaire, et là aussi, on a

Proposition 165 *A est symétrique définie positive*

Preuve. Soit $u = (u_1, \ldots, u_N) \in \mathbb{R}^N$, on constate que

$$u^T Au = \iint_{\Omega} |\nabla\psi(x)|^2 \, dx \geq 0$$

avec $\psi(x) = \sum_{i=1}^{N} u_i \varphi_i(x)$ et on conclut en utilisant l'hypothèse que les fonctions de V sont nulles sur $\partial\Omega$. ■

8.2.1 Triangulation

Nous allons voir comment étendre le choix des fonctions de base à la dimension 2 tout en gardant de bonnes propriétés.

Définition 166 *On dit que $\mathcal{T}$ est une triangulation admissible de Ω si $\mathcal{T}$ est un recouvrement de Ω en triangles tels que si T_1 et T_2 sont deux triangles de $\mathcal{T}$ alors $T_1 \cap T_2$ est soit vide, soit un point qui est alors un sommet de T_1 et de T_2 soit un segment qui est alors un coté de T_1 et un coté de T_2. De plus, on notera* $\operatorname{diam}(\mathcal{T})$ *la longueur du plus grand coté de tous les triangles de $\mathcal{T}$.*

On introduit alors autant d'applications φ_i, $i = 1, \ldots, N$ qu'il existe de sommets $P_1, \ldots, P_N$ de triangles de $\mathcal{T}$ intérieurs à Ω et on défini φ_i sur ces sommets par $\varphi_i(P_j) = \delta_{ij}$. On pose aussi $\varphi_i(P) = 0$ si P est un sommet de triangle de $\mathcal{T}$ situé sur la frontière $\partial\Omega$ de Ω, et on étend enfin φ_i sur Ω telle que φ_i soit affine sur chaque triangle ce qui revient à poser $\varphi_i(x) = a_{ij}x_1 + b_{ij}x_2 + c_{ij}$ sur chaque triangle $T_j \in \mathcal{T}$ en identifiant les coefficients (a_{ij}, b_{ij}, c_{ij}) de sorte que si on note P_j^1, P_j^2 et P_j^3 les trois sommets de T_j

$$\left\{ \begin{array}{rcl} a_{ij}P_{j1}^1 + b_{ij}P_{j2}^1 + c_{ij} & = & \varphi_i\left(P_j^1\right) \\ a_{ij}P_{j1}^2 + b_{ij}P_{j2}^2 + c_{ij} & = & \varphi_i\left(P_j^2\right) \\ a_{ij}P_{j1}^3 + b_{ij}P_{j2}^3 + c_{ij} & = & \varphi_i\left(P_j^3\right) \end{array} \right.$$

On vérifie facilement que $\varphi_i \in V$ et que la famille de fonctions $(\varphi_i)_{1,\ldots,N}$ est linéairement indépendante, donc répond aux exigences de la méthode. On a encore un théorème similaire au cas de la dimension 1:

Théorème 167 *Si on note* $|v|_1 = \left(\int_0^1 (\nabla v(x))^2 dx\right)^{\frac{1}{2}}$ *alors si* u *est une solution faible deux fois continûment différentiable sur* Ω *et si les angles des triangles ne tendent pas vers* 0 *quand* $\mathrm{diam}(\mathcal{T})$ *tend vers* 0 *alors*

$$|u - \widetilde{u}|_1 = O(\mathrm{diam}(\mathcal{T}))$$

où u *est la solution du problème faible 8.5 et* $\widetilde{u}$ *la solution du problème faible en dimension finie 8.6.*

Bibliographie

[1] ♣ Brezinski C., Algorithmique Numérique, Ellipses, Paris, 1988.

[2] ♣ Ciarlet P.G., Introduction à l'analyse matricielle et à l'optimisation, Masson, Paris, 1985.

[3] ♣ Crouzeix M., Mignot A.L, Analyse Numérique des Equations différentielles, Masson,Paris, 1984.

[4] ♣ Demailly J.P., Analyse Numérique et Equations Différentielles, Presses Universitaires de Grenoble, 1996.

[5] ♣ Déminovitch B., Maron I., Eléments de Calcul Numérique, Editions Mir, Moscou, 1979.

[6] Engel A., Mathématique et Informatique, Cedic-Nathan, Paris, 1985.

[7] Gastinel, N. Analyse numérique linéaire, Hermann, Paris, 1966

[8] Golub G.H., Van Loan C.F., Matrix Computations, The Johns Hopkins University Press, Baltimore, 1989.

[9] ♣ Rappaz J., Picasso M. Introduction à l'analyse numérique, Presses Polytechniques et Universitaires Romandes, 1998

[10] ♣ Sainsaulieu L., Calcul scientifique, Masson, 1996

[0]Le symbole ♣ indique que le livre se trouve à la Bibliothèque Universitaire des Sciences de Dijon.

Index