Université de Strasbourg

ENSIIE

Analyse Numérique

Auteur :
Michel Mehrenberger

Collaborateurs :
Pierre Boutry
Vincent Huber

2014-2015

Table des matières

Chapitre 1

Rappels à propos des matrices

Théorème 1 (Théorème du rang)**.** *Pour toute matrice $A \in \mathbb{R}^{M\times N}$, on a*

$$\mathrm{rg}(A) + \dim(\mathrm{Ker}A) = M$$

avec :

$$KerA = \{x \in E / Ax = 0\}.$$

Changement de base

Soit $\mathcal{E} = (e_1, \ldots, e_N)$ une base de $\mathbb{R}^N$ et $\mathcal{E}' = (e'_1, \ldots, e'_N)$ une autre base. La matrice de passage de $\mathcal{E}$ à $\mathcal{E}'$ est donnée par $S = P_{\mathcal{E}\to\mathcal{E}'} = (S_{i,j})$ avec

$$e'_j = \sum_{i=1}^{N} S_{i,j} e_i.$$

Proposition 1 (Effet d'un changement de base sur les composantes d'un vecteur)**.**
Si $\sum_{i=1}^N x_i e_i = \sum_{i=1}^N x'_i e'_i$, et $x = (x_1, \ldots, x_N)^t$, $x' = (x'_1, \ldots, x'_N)^t$, on a

$$x = Sx'.$$

Démonstration. On écrit

$$\sum_{i=1}^{N} x_i e_i = \sum_{i=1}^{N} x'_i e'_i = \sum_{i=1}^{N} x'_i \sum_{j=1}^{N} S_{ji} e_i = \sum_{j=1}^{N}\sum_{i=1}^{N} S_{ji} x'_i e_i,$$

et donc $x_j = \sum_{i=1}^N S_{ji} x'_i$. □

Proposition 2 (Effet d'un changement de base sur les éléments d'une matrice)**.** *Si A est une matrice représentant une application linéaire de E dans F dans une base $\mathcal{E}$ (base de E) et $\mathcal{F}$ (base de F) et si A' est une matrice représentant cette même application dans une base $\mathcal{E}'$ (base de E) et $\mathcal{F}'$ (base de F), on a*

$$A' = T^{-1} A S,$$

avec $S = P_{\mathcal{E}\to\mathcal{E}'}$ et $T = P_{\mathcal{F}\to\mathcal{F}'}$. Pour mémoriser, on a

$$A_{\mathcal{F}'\to\mathcal{E}'} = P_{\mathcal{F}'\to\mathcal{F}} A_{\mathcal{F}\to\mathcal{E}} P_{\mathcal{E}\to\mathcal{E}'}.$$

Démonstration. On note g l'application correspondante et on a

$$g(e_j) = \sum_{i=1}^{N} A_{i,j} f_i,\ g(e'_j) = \sum_{i=1}^{N} A'_{i,j} f'_i.$$

On en déduit

$$g(e'_j) = \sum_{i=1}^{N} S_{i,j} g(e_i) = \sum_{i=1}^{N} S_{i,j} \sum_{k=1}^{N} A_{k,i} f_k = \sum_{i=1}^{N} S_{i,j} \sum_{k=1}^{N} A_{k,i} \sum_{\ell=1}^{N} (T^{-1})_{\ell,k} f'_\ell,$$

et donc

$$A'_{\ell,j} = \sum_{i=1}^{N} S_{i,j} \sum_{k=1}^{N} A_{k,i} (T^{-1})_{\ell,k} = \sum_{k=1}^{N} (T^{-1})_{\ell,k} \left(\sum_{i=1}^{N} A_{k,i} S_{i,j} \right).$$

□

Définition 1.
Matrice transposée : $A^t = (a_{j,i})$
Matrice adjointe (transposée de la conuguée) : $A^* = (\overline{a_{j,i}})$
Matrice symétrique : $A^t = A$
Matrice hermitienne (= auto adjointe) : $A^* = A$
Matrice anti-hermitienne : $A^* = -A$

Proposition 3. *On a*

$$(AB)^t = B^t A^t,\ (AB)^* = B^* A^*,\ (AB)^{-1} = B^{-1} A^{-1},\ (A^*)^{-1} = (A^{-1})^*,\ (A^t)^{-1} = (A^{-1})^t.$$

Lemme 1. *Soient $L^{(1)}$ et $L^{(2)}$ deux matrices d'ordre N triangulaires inférieures. Alors $L^{(3)} = L^{(1)} L^{(2)}$ est triangulaire inférieure et $\ell^{(3)}_{i,i} = \ell^{(1)}_{i,i} \ell^{(2)}_{i,i}$.*

Démonstration. Voir exercices. □

Lemme 2. *Soit L une matrice carrée régulière (i.e. inversible) et triangulaire inférieure. Alors L^{-1} est aussi triangulaire inférieure et*

$$\ell^{-1}_{i,i} = \frac{1}{\ell_{i,i}}.$$

Démonstration. Voir exercices. □

Proposition 4. *Soit A une matrice carrée triangulaire par blocs (les blocs diagonaux $A_{i,i}$ sont supposés être carrés). On a*

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

Définition 2. *Une matrice bande est une matrice telle que $a_{i,j} = 0$ pour $j < i - c$ et $j > i + c$; c est appelé demi-largeur de bande.*
Pour $c = 1$, on a une matrice tridiagonale ; pour $c = 2$, on a une matrice pentadiagonale.

Définition 3. *Produit scalaire hermitien : On dit qu'une application ϕ :*

$$\begin{array}{rccc} \phi & : & E \times E & \rightarrow & \mathbb{C} \\ & & (x,y) & \mapsto & (x|y) \end{array}$$

est un produit scalaire hermitien si elle est :
** sesquilinéaire à gauche*
— *Linéaire par rapport au premier argument :* $a(\lambda x + \beta y|z) = \lambda a(x,z) + \beta a(y,z)$
— *antilinéaire par rapport à la deuxième variable :* $a(x|\lambda y + \beta z) = \bar{\lambda} a(x|y) + \bar{\beta} a(x,z)$
**symétrique hermitienne :* $\forall x, y \in E^2, (x|y) = (y|z)$
** positive :* $\forall x \in E, (x|x) \in \mathbb{R}^+$
** définie :* $\forall x \in E, (x|x) = 0 \rightarrow x = 0$

Proposition 5. *Soit A une matrice rectangulaire (M,N). Alors pour tout $x \in K^N$ et $y \in K^M$, on a*

$$(Ax|y) = (x|A^*y)$$

avec

$$(x|y) = \sum_{k=1}^{N} x_k \bar{y_k}.$$

Démonstration. Voir exercices. □

Théorème 2. *Soit A une matrice rectangulaire (M,N). Alors*

$$KerA^* = (ImA)^\perp,\ ImA^* = (KerA)^\perp.$$

Théorème 3 (Orthogonalisation de Gram-Schmidt)**.** *A partir d'une suite de vecteurs linéairement indépendants de K^N $(f_1, f_2, \ldots, f_k)$, on peut construire une suite de vecteurs $(p_1, \ldots, p_k)$ 2 à 2 orthogonaux tels que*

$$Vect(f_1, \ldots, f_j) = Vect(p_1, \ldots, p_j),\ j = 1, \ldots, k.$$

Lemme 3. *Une matrice carrée hermitienne A est telle que $(Ax|x)$ est réel pour tout $x \in \mathbb{C}^N$.*

Définition 4. *Une matrice hermitienne est définie positive si pour tout $x \in \mathbb{C}^N \setminus \{0\}$, on a*

$$(Ax|x) > 0.$$

Une matrice hermitienne est semi-définie positive si pour tout $x \in \mathbb{C}^N \setminus \{0\}$, on a

$$(Ax|x) \geq 0.$$

Soit A une matrice rectangulaire de format (M,N). Alors A^*A est une matrice carrée d'ordre N qui est hermitienne. Elle est semi-définie positive. Si le rang de A vaut N alors A^*A est définie positive.

Proposition 6. *Les sous-matrices principales d'une matrice hermitienne et définie positive sont hermitiennes et définies positives. Les éléments diagonaux de A sont strictement positifs.*

Définition 5. *Une matrice Q de format (M,N) avec $M \geq N$ est unitaire si les colonnes de Q sont des vecteurs deux à deux orthogonaux et de norme unité, i.e. $Q^*Q = I$.*

Lemme 4. *Une matrice unitaire vérifie $\|Qx\|_2 = \|x\|_2$.*

Démonstration. Voir exercices. □

Lemme 5. *Le produit de deux matrices unitaires est unitaire.*

Une matrice unitaire à coefficients réels est dite orthogonale.

Proposition 7. *Soient A matrice de format (M,N) et B de format (N,M), avec $M > N$. Alors*

$$det(\lambda I - BA) = \lambda^{M-N} det(\lambda I - AB).$$

Les valeurs propres non nulles de deux matrices AB et BA sont les mêmes.

Démonstration. On part de

$$\begin{pmatrix} I_N & 0 \\ -B & \mu I_M \end{pmatrix} \begin{pmatrix} \mu I_N & A \\ B & \mu I_M \end{pmatrix} = \begin{pmatrix} \mu I_N & A \\ 0 & \mu^2 I_M - BA \end{pmatrix}$$

et

$$\begin{pmatrix} \mu I_N & -A \\ 0 & I_M \end{pmatrix} \begin{pmatrix} \mu I_N & A \\ B & \mu I_M \end{pmatrix} = \begin{pmatrix} \mu^2 I_N - AB & 0 \\ B & \mu I_M \end{pmatrix}.$$

□

Les valeurs propres de A^* sont les conjuguées des valeurs propres de A.
Les valeurs propres de A^t sont les mêmes que les valeurs propres de A.

Théorème 4. *Soit u_i un vecteur propre de A correspondant à λ_i. Soit v_j un vecteur propre gauche correspondant à λ_j. Alors, si $\lambda_i \neq \lambda_j$, on a*

$$(u_i | v_j) = 0.$$

En particulier si A est hermitienne, les vecteurs propres correspondant à des valeurs propres distinctes sont orthogonaux.

Proposition 8. *Les vecteurs propres associés à des valeurs propres distinctes sont linéairement indépendants.*

Théorème 5. *Soit A une matrice carrée d'ordre N.*

A est diagonalisable ssi A possède N vecteurs propres u_i linéairement indépendants. Alors A peut se factoriser sous la forme

$$A = SDS^{-1},$$

où D est la matrice diagonale formée des valeurs propres. La i-ième colonne de S est un vecteur propre u_i associé à la valeur propre λ_i. La j-ième colonne de $(S^{-1})^*$ est un vecteur propre à gauche v_j associé à la valeur propre λ_j.

Corollaire 1. *Si toutes les valeurs propres de A sont distinctes, alors A est diagonalisable.*

Définition 6. *Un bloc de Jordan $J_k(\lambda)$ d'ordre k est une matrice d'ordre k avec des 1 sur la sur-diagonale et λ sur la diagonale.*

Définition 7. *On appelle matrice de Jordan, une matrice diagonale par blocs, où chaque bloc est un bloc de Jordan $J_{k_i}(\lambda_i)$. Les λ_i ne sont pas nécessairement distincts.*

Théorème 6. *Toute matrice est semblable*[1] *à une matrice de Jordan.*

Théorème 7. *Une condition nécessaire et suffisante pour que deux matrices diagonalisables commutent est qu'elles aient les mêmes vecteurs propres.*

Démonstration. Pour un sens, voir exercices. □

Théorème 8 (Théorème de Schur)**.** *Toute matrice carrée A peut s'écrire*

$$A = UTU^*,$$

avec U unitaire et T triangulaire supérieure.

Définition 8. *On dit qu'une matrice est normale si $A^*A = AA^*$.*

Théorème 9. *A est une matrice normale ssi il existe une matrice U unitaire telle que*

$$A = UDU^*.$$

D est la matrice diagonale formée des valeurs propres. Ainsi, une matrice normale est diagonalisable et les vecteurs propres sont orthogonaux.

Corollaire 2. *Une matrice hermitienne est diagonalisable. Ses valeurs propres sont réelles et ses vecteurs propres sont orthogonaux.*

Corollaire 3. *Une matrice symétrique et réelle est diagonalisable. Ses valeurs propres sont réelles et ses vecteurs propres sont orthogonaux.*

Corollaire 4. *Une matrice unitaire est diagonalisable. Ses valeurs propres ont pour module* 1 *et ses vecteurs propres sont orthogonaux.*

Théorème 10. *Une matrice hermitienne (ou symétrique réelle) est définie positive ssi ses valeurs propres sont strictement positives.*

Théorème 11. *Une matrice hermitienne (ou symétrique réelle) est définie positive ssi ses mineurs principaux sont strictement positifs.*

Proposition 9. *Une matrice anti-hermitienne*[2] *est diagonalisable. Ses valeurs propres sont imaginaires pures et les vecteurs propres sont orthogonaux.*

Définition 9. *Soit A une matrice rectangulaire de format (M, N). On appelle valeurs singulières (μ_i) de A les racines carrées positives ou nulles des valeurs propres de la matrice A^*A d'ordre N.*

Théorème 12. *Soit A une matrice rectangulaire de format (M, N). Il existe deux matrices carrées unitaires U et V d'ordre respectivement M et N telles que*

$$U^*AV = \Sigma,$$

où Σ est une matrice rectangulaire de format (M, N), avec $\sigma_{i,i} = \mu_i$, $i = 1, \ldots, N$ et les autres termes sont nuls. Les μ_i sont les valeurs singulières de la matrice A.

1. Si $M = PAP^{-1}$, M et A sont semblables.
2. $A^* = -A$

Corollaire 5. *Le rang de A est égal au nombre de valeurs singulières non nulles.*

Corollaire 6. *Soit $A = U\Sigma V^*$ la décomposition en valeurs singulières de A. Appelons u_i les colonnes de U et v_i les colonnes de V. Alors*

$$A = \sum_{i=1}^{r} \mu_i u_i v_i^*, \quad A^*A = \sum_{i=1}^{r} \mu_i^2 v_i v_i^*, \quad AA^* = \sum_{i=1}^{r} \mu_i^2 u_i u_i^*.$$

Normes

Définition 10. *Définition de la norme*
Une norme est une application de E K-espace vectoriel dans $\mathbb{R}^+$ qui vérifie
1. $\|x\| = 0 \Leftrightarrow x = 0$,
2. $\|\lambda x\| = |\lambda|\|x\|$ *pour tout* $\lambda \in K$ *et* $x \in E$,
3. $\|x + y\| \leq \|x\| + \|y\|$ *pour tout* $(x, y) \in E^2$.

Théorème 13. *Dans $\mathbb{C}^N$ toutes les normes sont équivalentes.*

Définition 11. *Norme matricielle induite par la norme vectorielle $\|\cdot\|$:*

$$\|A\| = max_{\|x\|=1}\|Ax\|.$$

Lemme 6. *Pour toute norme matricielle induite, on a*

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

Théorème 14. *Une norme matricielle induite est bien une norme pour les matrices.*

Remarque 1. *Pour toute norme induite par une norme vectorielle, on a*

$$\|I\| = 1.$$

Proposition 10. *Pour deux matrices A et B de format (M, N) et (N, P), on a*

$$\|AB\| \leq \|A\|\|B\|.$$

Définition 12. *On appelle rayon spectral de A, la quantité*

$$\rho(A) = \max_{i=1,\ldots,N} |\lambda_i(A)|.$$

A est une matrice de format (N, N) et $\lambda_i(A),\ i = 1, \ldots, N$ sont les valeurs propres de A.

Théorème 15. *Pour que $\lim_{k\to\infty} A^k = 0$, il faut et suffit que $\rho(A) < 1$.*

Démonstration. Si A est une matrice élémentaire de Jordan J_p d'ordre N_p. On a

$$J_p = \lambda I_{N_p} + E.$$

On peut vérifier que $E^n = 0$, pour $n \geq N_p$. On a alors pour $k \geq N_p$

$$J_p^k = \sum_{i=0}^{N_p-1} C_k \lambda_p^{k-i} E^i.$$

Pour i fixé et $k \to \infty$, on a

$$\lim_{k\to\infty} |C_k^i \lambda_p^{k-i}| = 0,$$

si $|\lambda_p| < 1$ et vaut ∞ sinon. On voit alors que J_p^k tend vers 0 ssi $|\lambda_p| < 1$.

Dans le cas général, on écrit $A^k = SJ^kS^{-1}$. □

Théorème 16. *Pour toute norme matricielle, on a*

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

Démonstration. On a $\rho^k(A) = \rho(A^k) \leq \|A^k\|$ (cf exercices), ce qui donne

$$\rho(A) \leq \|A^k\|^{1/k}.$$

Soit $\epsilon > 0$. On considère la matrice suivante $A(\epsilon) = \frac{1}{\rho(A)+\epsilon}A$, dont le rayon spectral est $\rho(A(\epsilon)) = \frac{\rho(A)}{\rho(A)+\epsilon} < 1$. Donc $\lim_{k\to\infty} A^k(\epsilon) = 0$. Donc il existe $k(\epsilon)$ tel que pour tout $k \geq k(\epsilon)$, on ait

$$\|A^k(\epsilon)\| \leq 1.$$

Or, comme $\|A^k(\epsilon)\| = \frac{\|A^k\|}{(\rho(A)+\epsilon)^k}$, on obtient

$$\|A^k\|^{1/k} \leq \rho(A) + \epsilon$$

□

Théorème 17. *La série $I + B + B^2 + \ldots$ converge vers $(I - B)^{-1}$ ssi $\rho(B) < 1$.*

Démonstration. Si $\rho(B) < 1$, 1 n'est pas valeur propre de B donc $I - B$ est inversible. Posons

$$A_k = I + B + \ldots B^k.$$

On a alors

$$BA_k = B + \cdots + B^{k+1}.$$

En prenant la différence, on obtient

$$(I - B)A_k = I - B^{k+1},$$

soit $A_k = (I - B)^{-1}(I - B^{k+1})$, puis

$$\|A_k - (I - B)^{-1}\| \leq \|(I - B)^{-1}\|\|B^{k+1}\|,$$

qui tend vers 0 et donc

$$\lim_{k\to\infty} A_k = (I - B)^{-1}.$$

Réciproquement, si $\lim_{k\to\infty} A_k$ existe, on a $\lim_{k\to\infty} B^k = 0$ et donc $\rho(B) < 1$. □

Chapitre 2

Conditionnement

2.1 Introduction

Il est rare que la solution d'un système linéaire $Ax = b$ puisse être obtenue sans être entachée d'erreurs.

Les erreurs peuvent provenir des incertitudes sur A et b, mais aussi des erreurs d'arrondis. Ainsi au lieu de résoudre

$$Ax = b,$$

on résoud en fait

$$(A + \Delta A)y = (b + \Delta b).$$

On cherche alors à majorer la différence $x - y$ en fonction des majorations de ΔA et ΔB.

Exemple

$$A = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix}, \; b = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix}.$$

$$A + \Delta A = \begin{pmatrix} 10 & 7 & 8.1 & 7.2 \\ 7.08 & 5.04 & 6 & 5 \\ 8 & 5.98 & 9.89 & 9 \\ 6.99 & 4.99 & 9 & 9.98 \end{pmatrix}, \; b + \Delta b = \begin{pmatrix} 32.01 \\ 22.99 \\ 33.01 \\ 30.99 \end{pmatrix}.$$

La solution de $Ax = b$ est $x = (1, 1, 1, 1)^t$.

La solution de $Ay = (b+\Delta b)$ est $y = (1.82, -0.36, 1.35, 0.79)$ La solution de $(A+\Delta A)z = b$ est $z = (-81, 137, -34, 22)^t$

2.2 Conditionnement d'une matrice

Définition 13. *Soit $\|\cdot\|$ une norme matricielle ; le conditionnement d'une matrice régulière A associé à cette norme est le nombre*

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

parfois noté $K(A)$.E n particulier, on note $cond_p(A) = \|A\|_p\|A^{-1}\|_p$.

Théorème 18. *On a les propriétés suivantes.*

1. $cond(\alpha A) = cond(A)$, *pour toute matrice régulière* A *et tout scalaire* $\alpha \neq 0$.
2. $cond(A) \geq 1$, *si le conditionnement est calculé par une norme induite.*
3. $cond_2(A) = \frac{\mu_{\max}}{\mu_{\min}}$ *où* $\mu_{\max}$ *et* $\mu_{\min}$ *sont respectivement la plus grande et la plus petite des valeurs singulières de* A.
4. $cond_2(A) = 1$ *si et seulement si* $A = \alpha Q$, *avec* α *un scalaire et* Q *une matrice unitaire.*

Démonstration.

1. Voir exercices.

2. Voir exercices.

3. Voir exercices.

4. Pour toute matrice A, il existe deux matrices unitaires U et V et une matrice diagonale Σ, dont les coefficients diagonaux sont les valeurs singulières μ_i de A telles que

$$A = U\Sigma V^*$$

On a alors $cond_2(A) = 1$ ssi toutes les valeurs singulières sont égales entre elles. Soit α leur valeur. On a donc $\Sigma = \alpha I$ et $A = \alpha UV^* = \alpha Q$ où $Q = UV^*$ est une matrice unitaire. □

Remarque 2. *On dit qu'une matrice est "bien conditionnée", si son conditionnement n'est pas beaucoup plus grand que* 1. *On voit donc que les matrices unitaires sont les mieux conditionnées possibles.*

Remarque 3. *La valeur du déterminant ne donne pas d'indications sur le conditionnement. Voir exemples ci-après.*

Exemple 1. *A matrice diagonale* $a_{1,1} = 1$, $a_{i,i} = 0.1,\ i = 2, \ldots, N = 100$. *On a* $\|A\|_2 = 1$ *et* $\|A^{-1}\|_2 = 10$, *donc* $cond_2(A) = 10$, *bien que* $det(A) = 10^{-99}$.

Exemple 2. *A bidiagonale avec des* 1 *sur la diagonale et des* 2 *sur le surdiagonale. Voir exercices.*

2.3 Majoration des perturbations

Théorème 19. *Soit* A *une matrice inversible. Soient* x *et* $x + \Delta x$ *les solutions des systèmes linéaires :*

$$Ax = b,\ A(x + \Delta x) = b + \Delta b.$$

On a

$$\frac{\|\Delta x\|}{\|x\|} \leq cond(A)\frac{\|\Delta b\|}{\|b\|}.$$

Théorème 20. *Soit* A *une matrice inversible. Soient* x *et* $x + \Delta x$ *les solutions des systèmes linéaires :*

$$Ax = b,\ (A + \Delta A)(x + \Delta x) = b.$$

On a

$$\frac{\|\Delta x\|}{\|x + \Delta x\|} \leq cond(A)\frac{\|\Delta A\|}{\|A\|}.$$

Chapitre 3

Résolution numérique de systèmes linéaires

Soit $A \in \mathbb{R}^{N\times N}$ est une matrice carrée, supposée inversible ($\det A \neq 0$) et $b \in \mathbb{R}^N$ un vecteur donné.
On cherche à trouver le vecteur $x \in \mathbb{R}^N$ tel que $Ax = b$.

3.1 Cas des matrices triangulaires

Proposition 11. *Soit A une matrice triangulaire à termes diagonaux non nuls. Le système se résoud en environ N^2 opérations (descente ou remontée).*

Démonstration. Supposons A triangulaire supérieure. On a alors (algorithme de remontée) :

$$x_N = \frac{b_N}{a_{N,N}},\ x_j = \frac{b_j - \sum_{k=j+1}^{N} a_{j,k}x_k}{a_{j,j}},\ j = N-1,\dots,1.$$

□

3.2 Méthode de Gauss

On part d'une matrice A et on cherche à rendre A triangulaire supérieure en faisant des opérations sur les lignes.

Proposition 12. *On suppose les mineurs principaux de déterminant non nul. Alors on peut résoudre le système en faisant des opérations sur les lignes. Le nombre d'opérations est de l'ordre de $2N^3/3$.*

Démonstration.
Il s'agit de la méthode de Gauss sans pivot.
Etape 1 $L_j \leftarrow L_j - \frac{a_{j,1}}{a_{1,1}}L_1,\ j = 2,\dots,N$
Etape k $L_j \leftarrow L_j - \frac{a_{j,k}^{(k)}}{a_{k,k}^{(k)}}L_k,\ j = k+1,\dots,N$, pour $k = 2,\dots,N-1$.
Il reste à voir que le pivot $a_{k,k}^{(k)}$ est non nul. Les opérations sur les lignes ne changent pas le déterminant du k-ième mineur A_k qui vaut $\prod_{j=1}^{k-1} a_{j,j}^{(j)} a_{k,k}^{(k)} = \det(A_k) \neq 0$ et donc $a_{k,k}^{(k)}$ est lui même non nul par récurrence.

Nombre d'opérations pour calculer la matrice triangulaire supérieure : à l'étape k : $2(N-k)^2$.
□

Dans le cas où le pivot est nul ou petit en valeur absolue, la méthode de Gauss sans pivot ne fonctionne plus, et on choisit de faire un échange de ligne ou de colonne : il s'agit de la méthode de Gauss avec stratégie de pivot.

Proposition 13. *Soit A inversible. On peut alors résoudre le système par la méthode de Gauss avec stratégie de pivot.*

Démonstration. Si le pivot est nul, on sait que toute la ligne (ou colonne) n'est pas nulle et on peut donc faire un échange de colonnes (ou de lignes). □

Chaque opération sur les lignes correspond à une multiplication de la matrice par une matrice triangulaire inférieure avec des 1 sur la diagonale. On peut alors écrire

$$A = LU,$$

avec L triangulaire inférieure et U triangulaire supérieure. Dans le cas où il y a une stratégie de pivot, on a

$$A = PLU,$$

avec P matrice de permutation.

Il peut être intéressant de stocker L et U (que l'on peut stocker informatiquement dans la même matrice A), notamment si l'on veut résoudre le système pour une succession de seconds membres b.

Proposition 14. *Soit A une matrice dont tous les mineurs principaux sont non nuls. Alors A admet une unique factorisation LU et on peut stocker informatiquement L et U dans la même matrice A qui est successivement mise à jour.*

Démonstration. L'unicité s'obtient en écrivant $L_2^{-1}L_1 = U_2U_1^{-1}$ qui est triangulaire inférieure et supérieure avec des 1 sur la diagonale et qui est donc la matrice identité.
On a pour $k = 1, \dots, N-1$

$$a_{j,k} = \frac{a_{j,k}}{a_{k,k}},\ j = k+1, \dots, N,$$

puis pour $\ell, j = k+1, \dots, N$,

$$a_{\ell,j} = a_{\ell,j} - a_{\ell,k}a_{k,j}.$$

□

3.3 Exemples

3.3.1 Résumé pour la méthode de Gauss

La méthode de Gauss est une méthode directe.

On utilise un pivot que l'on change éventuellement.

Elle fonctionne sans changement de pivot pour toute matrice telle que les mineurs principaux sont de determinant non nul.

Avec pivot, elle fonctionne pour toute matrice inversible.

Le cout est en $O(n^3)$.

Sans pivot, on écrit

$$A = LU,$$

avec L triangulaire inférieure avec des 1 sur la diagonale et U triangulaire supérieure.

Avec pivot, on rajoute une matrice de permutation

$$A = PLU,$$

avec P matrice de permutation (permutation des lignes ou colonnes de la matrice identité).

Le coût peut etre reduit pour des matrices bandes et fait intervenir la largeur de bande

Par exemple, pour les matrices de tridiagonale, L et U sont bidiagonales.

3.3.2 Interpolation par splines cubiques

On considère l'interpolation par splines cubiques sur un intervalle $[a, b]$ divisé en N sous intevalles pas forcément de même longueur. On se donne

$$f(x_j),\ j = 0, \dots, N,\ f'(x_0),\ f'(x_N),$$

avec $a = x_0 < \cdots < x_N = b$.

Localement sur un intervalle $[x_j, x_{j+1}]$ la reconstruction par splines cubiques de f est un polynôme de degré ≤ 3.

On suppose la continuité C^2 en chacun des points $x_j,\ j = 1, \dots, N-1$.

Sur l'intervalle $[x_j, x_{j+1}]$, on note $P_{j+1/2}$ le polynôme de degré ≤ 3 qui est uniquement déterminé par $f_j, f_{j+1}, f'_j, f'_{j+1}$, avec $f'_j = \lim_{x \to x_j} P'_j(x) = \lim_{x \to x_j} P'_{j-1}(x)$. On peut écrire

$$P_{j+1/2}(x) = f_j + \int_{x_j}^{x} P'_{j+1/2}(x)dx.$$

Le polynôme $P'_{j+1/2}$ est de degré ≤ 2 et satisfait

$$P'_{j+1/2}(x_j) = f'_j,\ P'_{j+1/2}(x_{j+1}) = f'_{j+1},$$

ainsi que

$$\int_{x_j}^{x_{j+1}} P'_{j+1/2}(x)dx = f_{j+1} - f_j.$$

On peut donc écrire

$$P'_{j+1/2}(x) = f'_j + \frac{x - x_j}{x_{j+1} - x_j}(f'_{j+1} - f'_j) + A_j(x - x_j)(x - x_{j+1}),$$

avec A_j de telle sorte que

$$f_{j+1} - f_j = A_j \int_{x_j}^{x_{j+1}} (x - x_j)(x - x_{j+1})dx + \frac{f'_j + f'_{j+1}}{2}(x_{j+1} - x_j).$$

On a

$$\int_{x_j}^{x_{j+1}} (x - x_j)(x - x_{j+1})dx = -2\frac{x_{j+1} - x_j}{3}\left(\frac{x_{j+1} - x_j}{2}\right)^2 = -\frac{(x_{j+1} - x_j)^3}{6},$$

ce qui donne

$$A_j \frac{x_{j+1} - x_j}{6} = \frac{\frac{f'_j + f'_{j+1}}{2} - \frac{f_{j+1} - f_j}{x_{j+1} - x_j}}{x_{j+1} - x_j}$$

On obtient alors

$$P"_{j+1/2}(x) = \frac{f'_{j+1} - f'_j}{x_{j+1} - x_j} + 2A_j x - A_j(x_j + x_{j+1}),$$

et la continuité de la dérivée seconde donne

$$\frac{f'_{j+1} - f'_j}{x_{j+1} - x_j} + A_j(x_j - x_{j+1}) = \frac{f'_j - f'_{j-1}}{x_j - x_{j-1}} + A_{j-1}(x_j - x_{j-1}),$$

ce qui se réécrit pour $j = 1, \ldots, N-1$

$$\frac{f'_{j+1} - f'_j}{x_{j+1} - x_j} - 6\frac{\frac{f'_j + f'_{j+1}}{2} - \frac{f_{j+1} - f_j}{x_{j+1} - x_j}}{x_{j+1} - x_j} = \frac{f'_j - f'_{j-1}}{x_j - x_{j-1}} + 6\frac{\frac{f'_j + f'_{j-1}}{2} - \frac{f_j - f_{j-1}}{x_j - x_{j-1}}}{x_j - x_{j-1}}.$$

Si on suppose que le maillage est uniforme, on obtient

$$f'_{j+1} - f'_j - f'_j + f'_{j-1} - 3(f'_j + f'_{j+1} + f'_j + f'_{j-1}) = -6(f_{j+1} - f_j + f_j - f_{j-1})/\Delta x,$$

soit

$$f'_{j-1} + 4f'_j + f'_{j+1} = 3\frac{f_{j+1} - f_{j-1}}{\Delta x}. \tag{3.1}$$

On considère alors le système linéaire en $f'_0, \ldots, f'_N$, donné par (3.1) et $f'_0 = f'_0$, $f'_N = f'_N$, pour $j = 0$ et $j = N$. On a bien un système tridiagonal.

3.3.3 Résolution d'un système tridiagonal

Théorème 21. *On considère la matrice suivante :*

$$A = \begin{pmatrix} b_1 & c_1 & 0 & \cdots & 0 \\ a_2 & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & c_{n-1} \\ 0 & \cdots & 0 & a_n & b_n \end{pmatrix}.$$

On définit la suite suivante :

$$\begin{cases} \delta_0 = 1, \\ \delta_1 = b_1, \\ \delta_k = b_k\delta_{k-1} - a_k c_{k-1}\delta_{k-2}, \ \forall k \in \{2, \ldots, n\}. \end{cases}$$

1. *Alors* $\forall k \in \{1,\dots,n\}$, $\delta_k = \det(\Delta_k)$, *où*

$$\Delta_k = \begin{pmatrix} b_1 & c_1 & 0 & \cdots & 0 \\ a_2 & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & c_{k-1} \\ 0 & \cdots & 0 & a_k & b_k \end{pmatrix}.$$

2. *De plus, si* $\forall k \in \{1,\dots,n\}$, $\delta_k \neq 0$, *alors la factorisation LU de A est :*

$$A = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ a_2\frac{\delta_0}{\delta_1} & \ddots & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & a_n\frac{\delta_{n-2}}{\delta_{n-1}} & 1 \end{pmatrix} \begin{pmatrix} \frac{\delta_1}{\delta_0} & c_1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & c_{n-1} \\ 0 & \cdots & \cdots & 0 & \frac{\delta_n}{\delta_{n-1}} \end{pmatrix}.$$

Démonstration. 1. Pour $k = 1$, le résultat est évident. Pour $k \geq 2$, il suffit de développer le déterminant de Δ_k par rapport à la dernière ligne.

2. On suppose que $\forall k \in \{1,\dots,n\}$, $\delta_k \neq 0$. D'après le point précédent, le théorème de la décomposition LU s'applique. En particulier, il y a unicité de la décomposition. Il suffit juste de multiplier L par U données dans l'énoncé et verifier que le produit donne bien A (en utilisant la relation de récurrence de la suite (δ_k).

□

3.3.4 Equation de diffusion $1D$

On cherche à trouver la solution de $-u"(x) = f(x)$, $u(a) = u(b) = 0$. On considére un maillage uniforme. On a pour toute fonction $g \in C^4$

$$g(x_{j+1}) = g(x_j) + g'(x_j)h + g"(x_j)h^2/2 + g'''(x_j)h^3/6 + O(h^4),$$

$$g(x_{j-1}) = g(x_j) - g'(x_j)h + g"(x_j)h^2/2 - g'''(x_j)h^3/6 + O(h^4),$$

ce qui donne en sommant

$$g_{j+1} + g_{j-1} = 2g_j + g"_j h^2/2 + O(h^4).$$

On obtient alors une discrétisation de $-u"(x) = f(x)$ par

$$-u_{j-1} + 2u_j - u_{j+1} = h^2 f_j,$$

pour $j = 1,\dots,N-1$. En utilisant le fait que $u(a) = u(b) = 0$, on a $u_0 = 0$, $u_N = 0$ et on obtient bien un système tridiagonal en $u_1,\dots,u_{N-1}$.

3.4 La méthode de Cholesky

Cette méthode s'applique quand la matrice du système est symétrique définie positive. On sait qu'une telle matrice vérifie les hypothèses du théorème de décomposition LU, et donc admet une unique factorisation LU. Dans le cas d'une matrice symétrique définie positive, on peut en plus trouver une matrice B triangulaire inférieure telle que $A = BB^t$. Cette décomposition est la factorisation de Cholesky.

Théorème 22. *Soit $A \in \mathcal{M}_n$ une matrice symétrique définie positive. Il existe au moins une matrice B triangulaire inférieure, telle que $A = BB^t$. De plus, si on impose que tous les coefficients diagonaux de B soient positifs, alors cette factorisation est unique.*

Démonstration. On sait qu'une matrice symétrique définie positive vérifie les hypothèses du théorème de décomposition LU. Donc il existe $L \in \mathcal{M}_n(K)$ triangulaire inférieure à diagonale unité et U triangulaire inférieure telle que $A = LU$. Soit $k \in \{1, \dots, n\}$, on note A_k la sous-matrice de A constituée des k premières lignes et k premières colonnes. On utilise la même notation pour L et U. Alors, on vérifie facilement que $A_k = L_k U_k$. Alors, comme L_k est comme L à diagonale unité,

$$\forall k \in \{1, \dots, n\} \ \det(A_k) = \det(L_k)\det(U_k) = \det(U_k) = \prod_{i=1}^{k} u_{i,i} > 0.$$

On montre alors facilement par récurrence sur k que $\forall k \in \{1, \dots, n\}$ $u_{k,k} > 0$. On pose alors Δ la matrice diagonale dont les coefficients diagonaux sont les $\sqrt{u_{i,i}}$. Elle est inversible (les coefficients diagonaux sont non nuls, vi la remarque ci-dessus) et A peut s'écrire $A = (L\Delta)(\Delta^{-1}U)$, *i.e.* :

$$A = \begin{pmatrix} \sqrt{u_{1,1}} & 0 & \cdots & 0 \\ & \ddots & \ddots & \vdots \\ & * & \ddots & 0 \\ & & & \sqrt{u_{n,n}} \end{pmatrix} \begin{pmatrix} \sqrt{u_{1,1}} & & & \\ 0 & \ddots & * & \\ \vdots & \ddots & \ddots & \\ 0 & \cdots & 0 & \sqrt{u_{n,n}} \end{pmatrix}.$$

Posons $B = L\Delta$ et $C = \Delta^{-1}U$. $A = BC = A^t = C^tB^t$ car A est symétrique. Donc, $BC = C^tB^t \Leftrightarrow C(B^t)^{-1} = B^{-1}C^t$. Il est facile de montrer que $C(B^t)^{-1}$ est triangulaire inférieure à diagonale unité et $B^{-1}C^t$ est triangulaire supérieure à diagonale unité aussi. Ceci implique que $C(B^t)^{-1} = B^{-1}C^t = \mathrm{Id}$. Donc en prenant l'une de ces deux égalités, $B = C^t$.
Pour l'unicité, on considère Δ la matrice diagonale constituée des éléments diagonaux de B. On a vu ci-dessus que les éléments diagonaux de U sont tous non nuls, donc ceux de B aussi. Donc Δ est inversible et $A = (B\Delta^{-1})(\Delta B^t)$. Il est facile de vérifier que $B\Delta^{-1}$ est une matrice triangulaire inférieure à diagonale unité et ΔB^t est triangulaire supérieure. Donc $(B\Delta^{-1})(\Delta B^t)$ est une décomposition LU de A.
Maintenant on suppose que $A = B_1(B_1)^t = B_2(B_2)^t = (B_1\Delta_1^{-1})(\Delta_1(B_1)^t) = (B_2\Delta_2^{-1})(\Delta_2(B_2)^t)$. Les deux dernières égalités sont donc deux décomposition LU de A. Par unicité de la décomposition LU, $\Delta_1(B_1)^t = \Delta_2(B_2)^t$. Or la diagonale de chacune de ces matrices est respectivement composée des $(b_{1_{i,i}})^2$ et $(b_{2_{i,i}})^2$. Donc, $\forall i \in \{1, \dots, n\}$, $absb_{1_{i,i}} = absb_{2_{i,i}}$. Comme on a supposé que les coefficients diagonaux des B_i sont positifs ou nul, $\forall i \in \{1, \dots, n\}$, $b_{1_{i,i}} = b_{2_{i,i}}$. Donc $\Delta_1 = \Delta_2$ et comme $\Delta_1(B_1)^t = \Delta_2(B_2)^t$, on en déduit $B_1 = B_2$. □

Le coût de la méthode de Cholesky est
— $\frac{(n^3-n)}{6}$ additions,
— $\frac{(n^3-n)}{6}$ multiplications,
— $\frac{n(n-1)}{2}$ divisions
pour la factorisation et
— $n(n-1)$ additions,
— $n(n-1)$ multiplications,
— $2n$ divisions
pour la résolution des deux systèmes linéaires triangulaires.
Exemple : Avec $n = 10$, il y a 383 opérations.

Le principal intérêt de la méthode de Cholesky par rapport à la décomposition LU est le gain en mémoire lorsqu'on implémente les méthodes. En effet, avec la méthode de Cholesky, on ne stocke que B triangulaire, tandis qu'avec la décomposition LU on doit stocker deux matrices triangulaires.

Chapitre 4

Méthodes itératives

4.1 Introduction

En écrivant $Ax = b$, on a

$$\sum_{j=1}^{n} a_{i,j}x_j = b_i.$$

Pour la méthode de Jacobi, on a

$$x_i^{(k+1)} = (b_i - \sum_{j \neq i} a_{i,j}x_j^{(k)})/a_{i,i}.$$

Pour la méthode Gauss-Seidel, on met à jour les x_j au fur et à mesure qu'ils sont calculés.

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

Enfin, pour une méthode de relaxation, on prend

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

où la formule de $\hat{x}^{(k+1)}$ est donné par Gauss-Seidel (méthode SOR) ou Jacobi.

Plus précisément, pour SOR, on a

$$x_i^{(k+1)} = \frac{\omega}{a_{i,i}}\left(b_i - \sum_{j=1}^{i-1} a_{i,j}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{i,j}x_j^{(k)}\right) + (1-\omega)x_i^{(k)}.$$

On écrit aussi ce type de méthodes sous la forme $A = M - N$ et

$$x^{(k+1)} = M^{-1}Nx^{(k)} + M^{-1}b.$$

La matrice $B = M^{-1}N$ est appelée matrice de l'itération.

4.2 Tests d'arrêt

Comme on cherche à résoudre un système linéaire en construisant une suite définie par récurrence, il faut se donner un test pour arrêter la suite. En général, on utilise le test suivant :

— On se fixe $\epsilon > 0$ suffisamment petit (0.001 par exemple).

— On s'arrête dès que $\frac{\|Ax^{(k)}-b\|}{\|b\|} \leq \epsilon$.

En posant $e^{(k)} = x - x^{(k)}$, l'inégalité précédente implique que

$$\frac{\|e^{(k)}\|}{\|x\|} \leq \text{cond}(A)\epsilon.$$

En effet, $\|e^{(k)}\| = \|A^{-1}(Ax^{(k)} - b)\| \leq \|A^{-1}\|\|Ax^{(k)} - b\| \leq \|A^{-1}\|\|b\|\epsilon$. Comme $Ax = b$, $\|b\| \leq \|A\|\|x\|$ et on obtient facilement

$$\|e^{(k)}\| \leq \|A^{-1}\|\|A\|\|x\|\epsilon,$$

d'où l'inégalité sur $e^{(k)}$.
On utillise aussi un autre test : $\|x^{(k+1)} - x^{(k)}\| \leq \|x^{(k)}\|\epsilon$, mais ce test peut être vérifié sans que $x^{(k)}$ soit proche de x.
Pour ce qui est des normes, on utilise en général $\|\cdot\|_\infty$ ou $\|\cdot\|_2$.

4.3 Convergence

Théorème 23. *La suite définie par $x^{(0)} \in \mathbb{C}^N$ et $x^{(k+1)} = Bx^{(k)} + c$ converge vers $\bar{x} = (I - B)^{-1}c$ quelque soit $x^{(0)}$ si et seulement si $\rho(B) < 1$.*

Démonstration. On se ramène à $e^{(k)} = \bar{x} - x^{(k)}$ converge vers 0 et on a $\rho(B) < 1$ ssi B^k converge vers 0. □

Définition 14. *On note les matrices :*

$$\begin{aligned} D_{i,j} &= \begin{cases} a_{i,j} \text{ si } i = j, \\ 0 \text{ sinon.} \end{cases} \\ L_{i,j} &= \begin{cases} a_{i,j} \text{ si } i > j, \\ 0 \text{ sinon.} \end{cases} \\ U_{i,j} &= \begin{cases} a_{i,j} \text{ si } i < j, \\ 0 \text{ sinon.} \end{cases} \end{aligned} \tag{4.1}$$

Proposition 15. *Les matrices de l'itération sont données par*

1. *Jacobi $J = L + U$,*
2. *SOR $\mathcal{L}_\omega = (I - \omega L)^{-1}((1 - \omega)I + \omega U)$,*
3. *Gauss-Seidel $\mathcal{L}_1 = (I - L)^{-1}U$,*

avec $L = D^{-1}E$, $U = D^{-1}F$, $A = D - E - F$, avec D diagonale, E triangulaire inférieure et F triangulaire supérieure.

Démonstration. Pour Jacobi, on voit bien que $M = D$ (terme $a_{i,i}x_i^{(k+1)}$), ce qui donne $N = M - A = E + F$, puis $B = D^{-1}(E + F) = L + U$.

Pour Gauss-Seidel, on trouve $M = D-E$, ce qui donne $N = M-A = D-E-(D-E-F) = F$, puis $B = (D - E)^{-1}F = (I - L)^{-1}U$.

Enfin, pour SOR, la méthode s'écrit

$$x_i^{(k+1)} = \frac{\omega}{a_{i,i}} \left(b_i - \sum_{j=1}^{i-1} a_{i,j} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{i,jl} x_j^{(k)} \right) + (1-\omega) x_i^{(k)}$$

donc en mettant à gauche les termes pour x_i^{k+1}, on obtient

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

On a donc $(D-\omega E)x^{(k+1)} = (1-\omega)Dx^{(k)} + \omega(b + Fx^{(k)})$, puis $\tilde{M} = D - \omega E$. On trouve alors

$$\tilde{N} = \omega F + (1-\omega)D.$$

On a $\tilde{M} - \tilde{N} = D - \omega E - \omega F - (1-\omega)D = \omega A$ puis $M = \tilde{M}/\omega = \frac{1}{\omega}D - E$ et $N = \tilde{N}/\omega = F + (\frac{1}{\omega} - 1)D$ et enfin

$$B = (D - \omega E)^{-1}((1-\omega)D + \omega F) = (I - \omega L)^{-1}((1-\omega)I + \omega U).$$

□

Théorème 24. *Pour toute matrice A, on a $\rho(\mathcal{L}_\omega) \geq |\omega - 1|$. Une condition nécessaire de convergence de la méthode de relaxation est $0 < \omega < 2$.*

Démonstration. Comme $\det(I - \omega L) = 1$, on a

$$\det(\lambda I - \mathcal{L}_\omega) = \det(\lambda(I - \omega L) - (1-\omega)I - \omega U).$$

On a la relation coefficients-racines [1]

$$\prod_{i=1}^{N} \lambda_i = (-1)^N \det(-(1-\omega)I - \omega U) = (1-\omega)^N.$$

On en déduit la relation comme $\rho(\mathcal{L}_\omega) \geq |\lambda_i|$. □

4.4 Vitesses de convergence

On a vu que $\|e^{(k)}\| \leq \|B^k\| \|e^{(0)}\|$, donc $\frac{\|e^{(k)}\|}{\|e^{(0)}\|} \leq \|B^k\|$. $\left(\frac{\|e^{(k)}\|}{\|e^{(0)}\|}\right)^{\frac{1}{k}}$ est le facteur moyen de réduction et il est donc majoré par $\|B^k\|^{\frac{1}{k}}$. C'est ce nombre qui va permettre de définir les vitesses de convergence, vitesses qui permettent de savoir si $x^{(k)}$ converge rapidement vers x ou non.

Définition 15. *On appelle vitesse moyenne (ou taux moyen) de convergence pour k itérations de la matrice de l'itération B le nombre*

$$R_k(B) = -\ln\left(\|B^k\|^{\frac{1}{k}}\right).$$

1. Pour un polynôme $P(x)$ de degré n s'écrivant sous la forme : $P(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0$. On définit le k-ième polynôme symétrique, noté σ_k, comme : $\sigma_k(x_1, x_2, ..., x_n) = \sum_{I \in \mathcal{P}_k(\{1,...,n\})} \prod_{i \in I} x_i$. Si les $(r_i)_{1 \leq i \leq n}$ sont les racines de $P(x)$, éventuellement multiples, on a : $\sigma_k(r_1, r_2, \cdots, r_n) = (-1)^k \frac{a_{n-k}}{a_n}$.

Remarque 4. *Si $\rho(B) < 1$ (donc la méthode est convergente) alors $\|B^k\|^{\frac{1}{k}} < 1$ a partir d'un certain rang et donc $R_k(B) > 0$ à partir d'un certain rang.*

Le but étant d'avoir une estimation de la vitesse de convergence, il faut pouvoir calculer $\|B^k\|^{\frac{1}{k}}$ pour connaître $R_k(B)$. Or ce calcul peut être coûteux (même si on peut avoir une majoration de ce nombre si B est diagonalisable). On introduit alors la notion de vitesse de convergence asymptotique, plus facile à calculer :

Définition 16. *On appelle vitesse de convergence asymptotique de la matrice de l'itération B le nombre*

$$\mathcal{R}(B) = -\ln(\rho(B)).$$

Cette définition vient du fait que $\|B^k\|^{\frac{1}{k}}$ converge vers $\rho(B)$ quand k tend vers $+\infty$.

4.5 Les matrices hermitiennes définies positives

Théorème 25. *Si A est hermitienne inversible et si $A = M - N$ avec M inversible, on pose $B = Id - M^{-1}A$. On suppose que $M + M^* - A$ (qui est hermitienne) est définie positive. Alors $\rho(B) < 1$ équivaut à A définie positive.*

Démonstration. On montre que pour $y = Bx$, on a

$$(x|Ax) - (y|Ay) = ((x-y)|(M + M^* - A)(x-y)).$$

En effet,

$$(y|Ay) = (x - M^{-1}Ax|A(x - M^{-1}Ax)).$$

On utilise alors aussi $x - y = M^{-1}Ax$. Supposons A définie positive. Soit x vecteur propre de B de valeur propre λ. On obtient

$$(1 - |\lambda|^2)(x|Ax) = |1-\lambda|^2(x|(M + M^* - A)x).$$

Si $\lambda = 1$, on obtient $x = Bx$ et donc $M^{-1}Ax = 0$, ce qui est impossible. On a donc $|\lambda| < 1$. Supposons maintenant que $\rho(B) < 1$. Si A n'était pas définie positive, on aurait existence de $x_0 \neq 0$ tel que $\alpha_0 = (x_0|Ax_0) \leq 0$. Soit $x_n = B^n x_0$. Donc $\alpha_n = (x_n|Ax_n)$ tend vers 0. Or

$$\alpha_{n-1} - \alpha_n = ((x_{n-1} - x_n)|(M + M^* - A)(x_{n-1} - x_n)) > 0,$$

si $x_{n-1} - x_n \neq 0$ (si $x_{n-1} = x_n$, on en déduit que 1 est valeur propre de B, ce qui est impossible). Donc α_n est strictement décroissante ; son premier terme étant négatif ; elle ne peut pas converger vers 0. □

Corollaire 7. *Avec les notations et hypothèses du théorème précédent, si $M + M^* - A$ et A sont définies positives, alors la méthode itérative est convergente.*

Corollaire 8. *Si A est hermitienne, $A = D - E - F$ la décomposition classique, avec D définie positive (ses éléments diagonaux sont tous > 0), $\omega \in]0;2[$, alors la méthode SOR est convergente si et seulement si A est définie positive.*

Démonstration. On a $D = D^*$ et $F = E^*$, $M = D/\omega - E$, puis

$$M^* + M - A = \frac{2-\omega}{\omega} D,$$

qui est hermitienne. Pour $0 < \omega < 2$, puisque D est définie positive, on a $M + M^* - A$ définie positive ; donc la méthode converge ssi A est définie positive. □

Corollaire 9. *Si A est hermitienne définie positive, alors la méthode SOR converge si et seulement si $\omega \in]0; 2[$.*

Démonstration. Il suffit de montrer que D est définie positive. □

Chapitre 5

Résolution numérique d'équations différentielles ordinaires

Une équation différentielle est une équation qui implique une ou plusieurs dérivées d'une fonction inconnue. Si toutes les dérivées ne sont que par rapport à une seul variable indépendante, on a une *équation différentielle ordinaire* (EDO), tandis que l'on a une *équation aux dérivées partielles* (EDP), quand sont présentes des dérivées par rapport à plusieurs variables indépendantes. L'équation différentielle (ordinaire ou aux dérivées partielles) est d'*ordre* p, si p est l'ordre maximal des dérivées présentes.
On s'intéresse dans ce chapitre aux équations différentielles ordinaires (EDO) d'ordre 1. En anglais, on dit ODE (Ordinary Differential equations) et PDE (Partial Differential Equations).

5.1 Quelques problèmes

Les équations différentielles ordinaires permettent de décrire de nombreux phénomènes dans des domaines très variés.

5.1.1 Thermodynamique

Considérons un corps de température interne T dans un milieu ambiant de température constante T_e. Alors le transfert de chaleur entre le corps et le milieu ambiant peut être décrit par la loi de Stefan-Boltzmann

$$v(t) = \epsilon\gamma S(T^4(t) - T_e^4),$$

où t est la variable temporelle, ϵ est la constante de Boltzmann (égal à $5.6 \cdot 10^{-8} J/m^2K^4s$, où J est pour Joule, K pour Kelvin et bien sûr m pour mètre et s pour seconde), γ est la constante d'émissivité du corps, S la superficie du corps et v le taux de transfert de chaleur. Le taux de variation de l'énergie $E(t) = mCT(t)$ (où C est la chaleur spécifique au matériau constituant le corps) est égal en valeur absolue à ce taux. Donc, en écrivant $T(0) = T_0$, le calcul de $T(t)$ nécessite la solution de l'équation différentielle ordinaire

$$T'(t) = -\frac{v(t)}{mC}. \tag{5.1}$$

5.1.2 Dynamique des populations

Considérons une population de bactéries dans un environnement confiné dans lequel il ne peut y avoir plus que B éléments. Supposons qu'à l'instant initial, le nombre de bactéries est égal à $y_0 << B$ et que le taux d'accroissement est une constante $C > 0$. Dans ce cas, le taux de changement de la population est proportionnel au nombre de bactéries existantes, sous la restriction que ce nombre ne peut excéder B. Cela s'exprime par léquation différentielle

$$y'(t) = Cy(1 - \frac{y}{B}), \tag{5.2}$$

où la solution $y = y(t)$ désigne le nombre de bactéries à l'instant t.
Supposons que 2 populations y_1 et y_2 soient en compétition. A la place de (5.2), on aura

$$\begin{cases} y_1'(t) &= C_1 y_1(t)(1 - b_1 y_1 - d_2 y_2) \\ y_2'(t) &= C_2 y_1(t)(1 - b_2 y_2 - d_1 y_1) \end{cases}, \tag{5.3}$$

où $C1, C_2$ représentent les taux de croissance des deux populations, les coefficients d_1, d_2 définissent les interactions entre les deux espèces et b_1, b_2 sont liés à la quantité de nourriture disponible.

5.1.3 Le problème du pendule

Le mouvement d'un pendule de masse m, suspendu à un point O par un fil non pesant de longueur ℓ, en rotation d'angle $\theta(t)$ autour de O est gouverné par l'équation

$$\theta"(t) = -g \sin(\theta(t))/\ell.$$

L'angle $\theta(t)$ est mesuré par rapport à une verticale passant par O. On s'intéresse au mouvement entre l'instant 0 et l'instant $T > 0$.
On se donne des conditions initiales :

$$\theta(0) = \pi/3, \quad \theta'(0) = 0.$$

5.2 Le problème de Cauchy

On considère des équations différentielles d'ordre 1 (une équation d'ordre $p > 1$ peut toujours être réduite en un système de p équations d'ordre 1). Le cas de systèmes du premier ordre sera traité plus tard.
Une équation différentielle ordinaire admet en général une infinité de solutions : on connait le comportement de la dérivée, soit la fonction... à une constante près. Afin de définir cette constante, il faut imposer une condition supplémentaire qui décrit la valeur en un point donné de l'intervalle.
Par exemple, l'équation (5.2) admet une famille de solutions

$$y(t) = B\frac{\psi(t)}{1 + \psi(t)},$$

avec $\psi(t) = \exp(Ct + K)$, où K est une constante arbitraire. Si on impose $y(0) = 1$, on prend l'unique solution correspondant à la valeur $K = \ln\left(\frac{1}{B-1}\right)$.

5.2.1 Définition

On considère donc le *problème de Cauchy* : trouver $y \in C^1(I)$ [1] satisfaisant

$$y'(t) = f(t, y(t)), \quad y(t_0) = y_0. \tag{5.4}$$

On rappelle alors un résultat classique d'analyse :

Proposition 16. *Supposons que la fonction $f(t, y)$ soit :*
1. *continue par rapport à ses 2 arguments*
2. *Lipschitz-continue par rapport à son second argument, c'est-à-dire qu'il existe une constante $L > 0$ telle que*

$$|f(t, y_1) - f(t, y_2)| \leq L|y_1 - y_2|, \ \forall t \in I, \ \forall (y_1, y_2) \in \mathbb{R}^{\not\vDash}.$$

Alors la solution y du problème de Cauchy existe et est unique. Elle sastifait de plus

$$y(t) - y_0 = \int_{t_0}^{t} f(\tau, y(\tau)) d\tau. \tag{5.5}$$

Si f ne dépend pas de t, on dit que l'équation est *autonome*.

5.2.2 Problématique des solutions numériques

Malheureusement, des solutions explicites ne sont disponibles que pour certains types très spéciaux d'équations différentielles. Dans certains autres cas, la solution n'est disponible que sous forme implicite. C'est par exemple le cas pour l'équation $y' = (y - t)/(y + t)$, dont les solutions satisfont la relation implicite

$$\frac{1}{2} \ln(t^2 + y^2) + \text{arctg}(\frac{y}{t}) = C.$$

Dans d'autres circonstances, la solution n'est même pas représentable sous forme implicite, comme c'est le cas pour l'équation $y' = \exp(-t^2)$, dont la solution générale ne peut qu'être exprimée sous forme de série.
Pour ces raisons, on cherche à approcher la solution de *toute* famille d'équation différentielle ordinaire pour laquelle il existe des solutions.
La stratégie commune de toutes ces méthodes consiste à subdiviser l'intervalle $I = [t_0, T]$, avec $T < \infty$, en N_h intervalles de longueur $h = (T - t_0)/N_h$; h est appelé le *pas de discrétisation*. Ensuite, en chaque *noeud* $t_n (0 \leq n \leq N_h)$, on cherche une valeur inconnue u_n qui approche $y_n = y(t_n)$. L'ensemble des valeurs $\{u_0 = y_0, u_1, \ldots, u_{N_h}\}$ est notre *solution numérique*.

5.2.3 Solution locale

Il se peut que les conditions de régularité données dans la Proposition 16 ne soient vérifiées que dans un voisinage d'un point considéré. Si $f(t, y)$ est localement Lipschitz en (t_0, y_0) par rapport à y : il existe $L > 0$

$$|f(t, y_1) - f(t, y_2)| \leq L|y_1 - y_2|, \forall t \in J \subset I, \ \forall (y_1, y_2) \in \Sigma^2,$$

1. Continue et Dérivable

avec Σ un voisinage de y_0 de rayon r_Σ et J un voisinage de t_0 de rayon r_J, alors le problème de Cauchy admet une unique solution dans un voisinage de t_0, de rayon r_0 satisfaisant

$$0 < r_0 < \min(r_J, \frac{r_\Sigma}{M}, \frac{1}{L}),$$
$$M = \sup_{J\times\Sigma} |f(t,y)|.$$

Cette solution est appelée *solution locale*.
Notons que si f a une dérivée continue par rapport à y, cette condition est satisfaite, il suffit de prendre

$$L = \max_{J\times\Sigma} |\partial_y f(t,y)|.$$

Le problème de Cauchy admet une unique solution globale, si l'on peut prendre $J = I$ et $\Sigma = \mathbb{R}$, c'est-à-dire, si f est *uniformément Lipschitz* par rapport à y (et on retrouve le cadre de la Proposition 16).

5.3 Premiers exemples : les méthodes d'Euler

Une méthode classique, la *méthode d'Euler progressive* génère la solution numérique ainsi :

$$u_{n+1} = u_n + hf_n,\ n = 0, \dots, N_h - 1,$$

où l'on a utilisé la notation $f_n = f(t_n, u_n)$. Cette méthode s'interprète en prenant l'approximation $y'(t_n) \approx \frac{y(t_{n+1})-y(t_n)}{h}$.
On définit de manière similaire la *méthode d'Euler rétrograde* par

$$u_{n+1} = u_n + hf_{n+1},\ n = 0, \dots, N_h - 1,$$

en prenant cette fois-ci l'approximation $y'(t_{n+1}) \approx \frac{y(t_{n+1})-y(t_n)}{h}$.
Ces deux méthodes sont des exemples de *méthodes à un pas* : pour calculer la solution au point t_{n+1}, on n'a besoin de connaître l'information qu'au temps t_n. Plus précisément dans le cas de la méthode d'Euler progressive, on n'a besoin que de u_n pour calculer u_{n+1} ; dans le cas de la méthode d'Euler rétrograde, u_{n+1} dépend aussi de lui-même par l'intermédiaire de f_{n+1}. Ainsi, on appelle la première méthode, *méthode d'Euler explicite*, tandis que la deuxième s'appelle *méthode d'Euler implicite*.
Par exemple, la discrétisation de (5.2) par la méthode d'Euler explicite donne

$$u_{n+1} = u_n + hCu_n(1 - u_n/B),$$

tandis que pour la méthode d'Euler implicite on doit résoudre l'équation non linéaire

$$u_{n+1} = u_n + hCu_{n+1}(1 - u_{n+1}/B).$$

Ainsi, les méthodes implicites sont plus coûteuses, mais nous verrons qu'elles peuvent avoir de meilleures propriétés de stabilité.

5.4 Intégration numérique

En intégrant l'équation différentielle (5.4), on obtient

$$y(t_{n+1}) - y(t_n) = \int_{t_n}^{t_{n+1}} f(t, y(t))dt.$$

Supposons que f ne dépende pas de la deuxième variable. Dans ce cas on se ramène au calcul d'une intégrale :

$$y(t) = \int_{t_0}^{t} f(s)ds + y(t_0).$$

On va donc présenter ici quelques concepts relatifs à l'intégration numérique. Soient $a < b$, 2 réels et $f : [a, b] \to \mathbb{R}$, une fonction. On cherche à trouver une approximation numérique de $\int_a^b f(x)\mathrm{d}x$.

5.4.1 Principe

Le calcul de $\int_a^b f(x)\mathrm{d}x$ ne peut parfois pas se faire de manière exacte. En effet, d'une part f n'admet pas forcément de primitive connue ; d'autre part, il est possible que f ne soit connue qu'en certains points, comme cela peut arriver dans le cas de mesures physiques ou de schémas numériques. L'idée est alors de remplacer f par un polynôme qui l'approche "bien" et de calculer l'intégrale de ce polynôme.

5.4.2 Formule de quadrature

On fait l'approximation

$$\int_a^b f(x)\mathrm{dx} \simeq \sum_{i=0}^{n} \omega_i f(x_i). \tag{5.6}$$

La formule de droite est appelée formule de quadrature, qui est donnée par les poids ω_i et les points x_i pour $i = 0, \ldots, n$.

5.4.3 Intervalle

Connaissant une formule de quadrature sur l'intervalle $[-1, 1]$, on en déduit une formule sur un intervalle $[a, b]$ quelconque : si on fait le changement de variable $y = a + (x + 1)\frac{b-a}{2}$

$$\int_a^b f(y)\mathrm{d}y = \frac{b-a}{2} \int_{-1}^{1} f(a + (x+1)\frac{b-a}{2})\mathrm{d}x \simeq \frac{b-a}{2} \sum_{i=0}^{n} \omega_i f(a + (x_i + 1)\frac{b-a}{2}).$$

Pour la suite, on ne cherchera donc que des formules de quadrature sur l'intervalle $[-1, 1]$.

5.4.4 Formules composites

Lorsque l'on souhaite calculer une intégrale sur un intervalle donné $[a, b]$, on n'utilise en général pas la formule directement sur l'intervalle $[a, b]$, mais on le découpe en N sous intervalles (a priori de même taille, mais ce n'est pas forcé) et on applique la formule sur chacun des sous intervalles :

$$\int_a^b f(x)\mathrm{d}x = \sum_{i=0}^{N-1} \int_{a+i(b-a)/N}^{a+(i+1)(b-a)/N} f(x)\mathrm{d}x.$$

5.4.5 Formules de Newton-Cotes

La formule de Newton-Cotes à $n+1$ points correspond à utiliser la formule de quadrature avec les points équidistants $-1+2i/n$, $i=0,\ldots,n$, en choisissant les poids ω_i de telle sorte que la formule (5.6) soit exacte pour tous les polynômes de degré $\leq n$. On obtient alors les résultats suivants pour les valeurs de n allant de zéro à 9.

$$f(-1)$$
$$f(1)+f(-1)$$
$$\tfrac{1}{3}f(1)+\tfrac{4}{3}f(0)+\tfrac{1}{3}f(-1)$$
$$\tfrac{1}{4}f(-1)+\tfrac{3}{4}f\left(\tfrac{1}{3}\right)+\tfrac{3}{4}f\left(\tfrac{-1}{3}\right)+\tfrac{1}{4}f(1)$$
$$\tfrac{7}{45}f(1)+\tfrac{32}{45}f\left(\tfrac{1}{2}\right)+\tfrac{4}{15}f(0)+\tfrac{7}{45}f(-1)+\tfrac{32}{45}f\left(\tfrac{-1}{2}\right)$$
$$\tfrac{19}{144}f(-1)+\tfrac{25}{72}f\left(\tfrac{1}{5}\right)+\tfrac{25}{48}f\left(\tfrac{-3}{5}\right)+\tfrac{19}{144}f(1)+\tfrac{25}{48}f\left(\tfrac{3}{5}\right)+\tfrac{25}{72}f\left(\tfrac{-1}{5}\right)$$
$$\tfrac{41}{420}f(-1)+\tfrac{41}{420}f(1)+\tfrac{68}{105}f(0)+\tfrac{9}{140}f\left(\tfrac{-1}{3}\right)+\tfrac{9}{140}f\left(\tfrac{1}{3}\right)+\tfrac{18}{35}f\left(\tfrac{-2}{3}\right)+\tfrac{18}{35}f\left(\tfrac{2}{3}\right)$$
$$\tfrac{751}{8640}f(-1)+\tfrac{751}{8640}f(1)+\tfrac{3577}{8640}f\left(-\tfrac{5}{7}\right)+\tfrac{49}{320}f(-3/7)+\tfrac{2989}{8640}f(-1/7)+\tfrac{2989}{8640}f(1/7)+\tfrac{49}{320}f(3/7)+\tfrac{3577}{8640}f\left(\tfrac{5}{7}\right)$$
$$\tfrac{989}{14175}f(-1)+\tfrac{989}{14175}f(1)-\tfrac{908}{2835}f(0)-\tfrac{928}{14175}f\left(\tfrac{-1}{2}\right)-\tfrac{928}{14175}f\left(\tfrac{1}{2}\right)+\tfrac{5888}{14175}f\left(\tfrac{-3}{4}\right)+\tfrac{10496}{14175}f\left(\tfrac{-1}{4}\right)+\tfrac{10496}{14175}f\left(\tfrac{1}{4}\right)+\tfrac{5888}{14175}f\left(\tfrac{3}{4}\right)$$
$$\tfrac{2857}{44800}f(-1)+\tfrac{2857}{44800}f(1)+\tfrac{1209}{2800}f\left(\tfrac{-1}{3}\right)+\tfrac{1209}{2800}f\left(\tfrac{1}{3}\right)+\tfrac{2889}{22400}f\left(\tfrac{1}{9}\right)+\tfrac{27}{1120}f\left(\tfrac{5}{9}\right)+\tfrac{27}{1120}f\left(\tfrac{-5}{9}\right)$$
$$+\tfrac{2889}{22400}f\left(\tfrac{-1}{9}\right)+\tfrac{15741}{44800}f\left(\tfrac{-7}{9}\right)+\tfrac{15741}{44800}f\left(\tfrac{7}{9}\right)$$

Les premières formules s'appellent formules des rectangles (gauche), des trapèzes, de Simpson, Simpson 3/8, Boole. On se rend compte que pour $n=8$ et $n\geq 10$, il y a des poids négatifs, et la valeur absolue de ses poids devient de plus en plus grande, ce qui rend les formules inutilisables à cause des erreurs d'arrondi.

5.4.6 Degré de précision

Définition 17. *On dit qu'une formule de quadrature est d'ordre p, si la formule est exacte pour tous les polynômes de degré $\leq p$ et n'est plus vraie pour un polynôme de degré $p+1$.*

Sous certaines conditions, l'erreur de quadrature s'écrit alors sous la forme

$$\int_{-1}^{1} f(x)dx - \sum_{i=0}^{p} \omega_i f(x_i) = C_p f^{(p+1)}(\xi),\ C_p \neq 0. \tag{5.7}$$

pour une formule de quadrature d'ordre p.

5.4.7 Formules de Gauss

Etant donnée une formule de quadrature à $n+1$ points, $x_0,\ldots,x_n$, on cherche à avoir une formule de d'ordre le plus grand possible. Comme on a $\int_{-1}^{1}\prod_{i=0}^{n}(x-x_i)^2\mathrm{d}x \neq 0$, on en déduit que la formule ne pourra pas être d'ordre $> 2n+1$. D'où la question, peut-on avoir une formule de quadrature d'ordre $2n+1$?

•**Le cas** $n=0$**.** On doit avoir $\int_{-1}^{1}(x-x_0)\mathrm{d}x=0$ et donc on en déduit que $x_0=0$.

•**Le cas** $n=1$**.** On doit avoir

$$\int_{-1}^{1}(x-x_0)(x-x_1)\mathrm{d}x=0 \text{ et } \int_{-1}^{1}x(x-x_0)(x-x_1)\mathrm{d}x=0.$$

De la deuxième équation, on obtient $x_1=-x_0$ et on a donc $\int_0^1 x^2\mathrm{d}x = x_0^2$, ce qui donne $x_0=-1/\sqrt{3}$ et $x_1=1/\sqrt{3}$.

•**Le cas** $n=2$**.** On doit avoir

$$\int_{-1}^{1}x^i(x-x_0)(x-x_1)(x-x_2)\mathrm{d}x=0 \quad i=0,1,2.$$

En posant $X = x_0x_1x_2$ et $Y = x_0 + x_1 + x_2$, on tire du cas $i = 0$ et $i = 2$:

$$(1/3)Y + X = 0,\ (1/5)Y + (1/3)X = 0$$

et donc $X = Y = 0$, on prend alors $x_0 = -\alpha, x_1 = 0$ et $x_2 = \alpha$, l'équation pour $i = 1$ donne alors $1/5 - (1/3)\alpha^2 = 0$ et donc $\alpha = \sqrt{3/5}$.
•**Le cas** $n = k$**.** On cherche un polynôme $P_k = \prod_{i=0}^{k}(x - x_i)$ de degré $k + 1$ ayant toutes ses racines dans l'intervalle $[-1, 1]$ et tel que

$$\int_{-1}^{1} t^i P_k(t)\mathrm{d}t = 0, \quad i = 0, \ldots, k.$$

Pour cela, on prend $P_k(t) = (t - a_k)P_{k-1}(t) - b_kP_{k-2}(t)$ et il suffit de trouver a_k et b_k tels que

$$\int_{-1}^{1} P_k(t)P_{k-1}(t)\mathrm{d}t = 0 \text{ et } \int_{-1}^{1} P_k(t)P_{k-2}(t)\mathrm{d}t = 0.$$

On obtient alors

$$a_k = \frac{\int_{-1}^{1} tP_{k-1}(t)^2\mathrm{d}t}{\int_{-1}^{1} P_{k-1}(t)^2\mathrm{d}t} \text{ et } b_k = \frac{\int_{-1}^{1} tP_{k-1}(t)P_{k-2}(t)\mathrm{d}t}{\int_{-1}^{1} P_{k-2}(t)^2\mathrm{d}t}.$$

Il reste alors à voir que toutes les racines sont distinctes et bien dans l'intervalle $[-1, 1]$. Pour cela on considère les racines de multiplicité impaire $a_0, \ldots, a_j$. On a alors $P_k = Q_k \prod_{i=0}^{j}(t - a_i)$ et Q_k est de signe constant sur l'intervalle $[-1, 1]$. Supposons maintenant que $j \leq k$, on a alors $\int_{-1}^{1} Q_k \prod_{i=0}^{j}(t - a_i)^2\mathrm{d}t = 0$, ce qui est impossible. Donc on a bien $k+1$ racines distinctes de P_k dans l'intervalle $[-1, 1]$.
Notons aussi que les poids sont cette fois-ci positifs. En effet, en prenant

$$\ell_j(t) = \prod_{\substack{i=0 \\ i\neq j}}^{k} \frac{t - x_i}{x_j - x_i},$$

on a $0 < \int_{-1}^{1} \ell_j(t)^2\mathrm{d}t = \omega_j$.

5.5 Schémas à un pas explicites

On considère ici des schémas à un pas explicites.

Définition 18. *On dit qu'une méthode numérique est à un pas, si pour tout* $n \in \mathbb{N}$, u_{n+1} *ne dépend que de* u_n*. Autrement, on dit que le schéma est une méthode multi-pas (ou à pas multiples).*

Définition 19. *Une méthode est dite explicite si la valeur* u_{n+1} *peut être calculée directement à l'aide des valeurs précédentes* u_k, $k \leq n$ *(ou d'une partie d'entre elles). Une méthode est dite implicite, si* u_{n+1} *n'est défini que par une relation implicite faisant intervenir la fonction* f*.*

Le schéma d'Euler explicite
On écrit

$$y(t_{n+1}) - y(t_n) = \int_{t_n}^{t_{n+1}} f(s, y(s))\mathrm{d}s.$$

On utilise alors la formule des rectangles gauche pour approcher l'intégrale :

$$\int_{t_n}^{t_{n+1}} f(s, y(s))\mathrm{d}s \simeq hf(t_n, y(t_n)).$$

On retrouve alors le schéma d'Euler explicite précédemment introduit.
Schéma de Runge explicite
Une autre manière de faire est d'utiliser la formule du point milieu. On a alors

$$\int_{t_n}^{t_{n+1}} f(s, y(s))\mathrm{d}s \simeq hf(t_n + h/2, y(t_n + h/2)).$$

On remarque alors que l'on a besoin d'une approximation de $y(t_n+h/2)$ que l'on peut obtenir justement avec le schéma d'Euler explicite précédent :

$$y(t_n + h/2) \simeq y(t_n) + h/2f(t_n, y(t_n)).$$

Le schéma s'écrit alors

$$k_1 = f(t_n, u_n), \quad k_2 = f(t_n + h/2, u_n + (h/2)k_1), \quad u_{n+1} = u_n + hk_2.$$

On peut écrire

$$u_{n+1} = u_n + h\phi(t_n, u_n, h), \quad \phi(t_n, u_n, h) = f(t_n + h/2, u_n + (h/2)f(t_n, u_n)).$$

Dans le cas où l'on utilise la formule des trapèzes à la place du point milieu, on appelle cette méthode la méthode de Heun.

Schémas de Runge-Kutta à s étages
Une méthode de Runge-Kutta à s étages est donnée par

$k_1 = f(t_n, u_n), k_2 = f(t_n + c_2h, u_n + ha_{2,1}k_1), \ldots, k_s = f(t_n + c_sh, u_n + h(a_{s,1}k_1 + \ldots a_{s,s-1}k_s))$,
$u_{n+1} = u_n + h(b_1k_1 + \ldots b_sk_s)$.

On représente habituellement la méthode par l'écriture matricielle

$$\begin{array}{l} 0| \\ c_2|a_{2,1} \\ c_s|a_{s,1}, \ldots, a_{s,s-1} \\ \quad |b_1, \ldots, b_s. \end{array}$$

Par la suite, on supposera toujours que $c_i = \sum_{j=1}^{i-1} a_{i,j}$ pour $i = 2, \ldots, s$.
Cela signifie que l'on a

$$k_i = f(t_n + c_ih, u(t_n + c_ih)) + O(h^2).$$

La méthode la plus célèbre est la méthode de Runge-Kutta d'ordre 4 (RK4) donnée par l'écriture matricielle

$$\begin{array}{c|cccc} 0 & & & & \\ 1/2 & 1/2 & & & \\ 1/2 & 0 & 1/2 & & \\ 1 & 0 & 0 & 1 & \\ \hline & 1/6 & 1/3 & 1/3 & 1/6. \end{array}$$

Ecriture générale
De manière générale, on écrit un schéma à un pas sous la forme

$$u_0 = y(t_0),\ u_{n+1} = u_n + h\phi(t_n, u_n, h). \tag{5.8}$$

ϕ est une fonction de $\mathbb{R}^+ \times \mathbb{R}^d \times \mathbb{R}^+$ à valeurs dans $\mathbb{R}^d$ et est appelée *fonction d'incrément*. L'algorithme sera alors satisfaisant lorsque l'erreur $e_n = |u_n - y(t_n)|$ converge vers 0 pour $n = 0, \ldots, N$ lorsque le pas de temps h tend vers 0. Pour cela, nous allons introduire plusieurs notions : la **consistance** et la **stabilité** qui conduiront ensuite à la **convergence** de l'approximation vers la solution exacte.

5.6 Consistance, stabilité et convergence

Pour simplifier les notations, on prendra $t_0 = 0$. On utilisera aussi la notation $\Delta t = h = T/N$. On définit l'erreur

$$e_n(\Delta t) = y(t_n) - u_n,\ n = 0, \ldots, N,\ \Delta t = T/N.$$

Définition 20. *On dit que le schéma est convergent sur l'intervalle* $[0, T]$*, si l'on a*

$$\lim_{N\to\infty} \max_{n=0,\ldots,N} \|e_n(\Delta t)\| = 0.$$

Pour $p \in \mathbb{N}$*, on dit que que le schéma est convergent d'ordre* p *s'il existe une constante* C *ne dépendant que de* f, T *et* u_0 *tel que*

$$\max_{n=0,\ldots,N} \|e_n(\Delta t)\| \leq C\Delta t^p.$$

Erreur de consistance
En posant $y_n = y(t_n)$, on a

$$y_{n+1} = y_n + \Delta t\phi(t_n, y_n, \Delta t) + \epsilon_{n+1},\ n = 0, \ldots, N-1,$$

où ϵ_{n+1} est le résidu obtenu au temps t_{n+1}, lorsque l'on insère la solution exacte au temps t_n dans le schéma numérique.
Ecrivons le résidu sous la forme

$$\epsilon_{n+1} = \Delta t\tau_{n+1}(\Delta t).$$

La quantité $\tau_{n+1}(\Delta t)$ est appelée *erreur de troncature locale* (ETL) au noeud t_{n+1}. L'*erreur de troncature globale* est alors donnée par

$$\tau(\Delta t) = \max_{0\leq n\leq N-1} \|\tau_{n+1}(\Delta t)\|.$$

Définition 21. *On dit que le schéma est consistant, si l'on a*

$$\lim_{\Delta t \to 0} \tau(\Delta t) = 0,$$

pour tout $t \geq 0$ et toute solution u. Pour $p \in \mathbb{N}$, on dit que que le schéma est consistant d'ordre p s'il existe une constante C ne dépendant que de f, T et u_0 tel que

$$\tau(\Delta t) \leq C\Delta t^p,$$

pour tout $t \geq 0$ et toute solution u.

Proposition 17. (Caractérisation de la consistance) *Si la fonction $\phi \in C(\mathbb{R}^+ \times \mathbb{R}^d \times \mathbb{R}^+, \mathbb{R}^d)$ et si*

$$\phi(t, u, 0) = f(t, u),\ t \in [0, T],\ u \in \mathbb{R}^d,$$

alors le schéma est consistant.

Démonstration. Soit $y \in C^1([0, T], \mathbb{R}^d)$ la solution exacte de (5.4). On a alors

$$\tau_{n+1}(\Delta t) = \frac{1}{\Delta t} \int_{t_n}^{t_{n+1}} (\phi(s, y(s), 0) - \phi(t_n, y(t_n), \Delta t))\mathrm{d}s,$$

et donc $\tau(\Delta t)$ tend vers 0 lorsque Δt tend vers 0. □

Théorème 26. (Stabilité+consistance ⇒ Convergence)
(consistance) *Supposons que le schéma est consistant d'ordre p : il existe une constante $C > 0$ dépendant de f, T et y_0 telle que*

$$\tau(\Delta t) \leq C\Delta t^p.$$

(stabilité) *Supposons aussi que ϕ est continu et Lipschitzienne par rapport à la variable $u \in \mathbb{R}^d$.*

$$\|\phi(t, u, \Delta t) - \phi(t, v, \Delta t)\| \leq \Gamma\|u - v\|, \text{pour tout } t \geq 0,\ \Delta t \geq 0.$$

Alors la solution est convergente et on a l'estimation

$$\|e_n(\Delta t)\| \leq \frac{C}{\Gamma}(\exp(\Gamma t_n) - 1)\Delta t^p + \|e_0(\Delta t)\| \exp(\Gamma t_n),\ n = 0, \ldots, N.$$

Démonstration. On a

$$e_{n+1}(\Delta t) = e_n(\Delta t) + \Delta t(\phi(t_n, y(t_n), \Delta t) - \phi(t_n, u_n, \Delta t) + \tau_{n+1}(\Delta t)),$$

et donc

$$\|e_{n+1}(\Delta t)\| \leq \|e_n(\Delta t)\|(1 + \Gamma\Delta t) + C\Delta t^{p+1}.$$

On a alors

$$\|e^{n+1}(\Delta t)\| + \frac{C\Delta t^p}{\Gamma} \leq (1 + \Gamma\Delta t)(\|e_n(\Delta t)\| + \frac{C\Delta t^p}{\Gamma}).$$

□

On va maintenant voir à quelles conditions une méthode à un pas est d'ordre p. Une méthode à un pas étant définie par la formule

$$x_{n+1} = x_n + h\Phi(t_n, x_n, h),$$

l'erreur locale pour une méthode à un pas est

$$e_{n+1} = x(t_{n+1}, x_n, h) - x(t_n, x_n, h) - h\Phi(t_n, x_n, h).$$

Si on suppose que Φ et f sont de classe C^p, alors comme x vérifie $x' = f(t, x)$ x est de classe C^{p+1} et en écrivant un développement de Taylor de $t \mapsto x(t, x_n, h)$ et de $h \mapsto \Phi(t_n, x_n, h)$, l'erreur devient

$$\begin{aligned} e_{n+1} = h\frac{dx}{dt}(t_n, x_n, h) + \cdots + \frac{h^p}{p!}\frac{d^p x}{dt^p}(t_n, x_n, h) \\ - h(\Phi(t_n, x_n, 0) + \cdots + \frac{h^{p-1}}{(p-1)!}\frac{\partial^{p-1}\Phi}{\partial h^{p-1}}(t_n, x_n, 0)) + O(h^{p+1}). \end{aligned}$$

D'autre part, comme $\frac{dx}{dt}(t) = f(t, x(t))$, on a $\frac{d^p x}{dt^p}(t) = f^{[p-1]}(t, x(t))$, où l'on note $f^{[j]}(t, x(t)) = \frac{d^j}{dt^j} f(t, x(t))$. Avec cette notation, en identifiant les termes de même puissance en h_n, on en déduit la proposition suivante.

Proposition 18. *Une méthode à un pas caractérisée par la fonction Φ est d'ordre au moins égal à p si et seulement si*

$$\frac{\partial^{j-1}\Phi}{\partial h^{j-1}} = \frac{1}{j} f^{[j-1]}(t_n, x_n), \quad \textit{pour } 1 \leq j \leq p.$$

Corollaire 10. *Une méthode à un pas caractérisée par la fonction Φ est d'ordre au moins égal à 1 si et seulement si*

$$\Phi(t_n, x_n, 0) = f(t_n, x_n).$$

Démonstration. On applique la proposition précédente avec $p = 1$. □

On en déduit immédiatement que la méthode d'Euler est d'ordre 1 et que la méthode d'Euler est la seule méthode à un pas et à un étage d'ordre 1.

Corollaire 11. *Une méthode à un pas caractérisée par la fonction Φ est d'ordre au moins égal à 2 si et seulement si*

$$\Phi(t_n, x_n, 0) = f(t_n, x_n), \ \textit{et} \ \frac{\partial\Phi}{\partial h}(t_n, x_n, 0) = \frac{1}{2}\Big[\frac{\partial f}{\partial t}(t_n, x_n) + d_x f(t_n, x_n)(f(t_n, x_n)\Big].$$

Démonstration. On applique la proposition précédente avec $p = 2$. Dans ce cas

$$f^{[1]}(t, x) = \frac{d}{dt} f(t, x(t)) = \frac{\partial f}{\partial t}(t, x(t)) + d_x f(t, x(t))(\frac{dx}{dt}) = \frac{\partial f}{\partial t}(t, x(t)) + d_x f(t, x(t))(f(t, x(t)),$$

où $d_x f(t, x)(y) = \sum_{j=1}^n y_j \frac{\partial f}{\partial x_j}(t, x)$. □

Constructions de méthode de Runge-Kutta d'ordre 2. La fonction Φ pour une méthode à s étages est définie par

$$\begin{aligned}\Phi(t_n, x_n, h) &= b_1 k_1 + \cdots + b_s k_s \\ &= b_1 f(t_n, x_n) + \cdots + b_s f(t_n + c_s h, x_n + h(a_{s,1} f(t_n, x_n) + \cdots + a_{s,s-1} f(t_n, x_n))).\end{aligned}$$

Donc

$$\Phi(t_n, x_n, 0) = (b_1 + \cdots + b_s) f(t_n, x_n).$$

La condition d'ordre 1 entraîne donc que $b_1 + \cdots + b_s = 1$. Pour l'ordre 2, il faut également calculer $\frac{\partial \Phi}{\partial h}(t_n, x_n, 0)$. Or on a

$$\begin{aligned}\frac{\partial \Phi}{\partial h}(t_n, x_n, h) &= \sum_{i=1}^{s} b_i \frac{\partial k_i}{\partial h}(t_n, x_n, h), \\ &= \sum_{i=1}^{s} b_i \Big[c_i \frac{\partial f}{\partial t}(t_n + c_i h, x_n + h(a_{i,1} k_1 + \cdots + a_{i,i-1} k_{i-1})) \\ &\quad + d_x f(t_n + c_i h, x_n + h(a_{i,1} k_1 + \cdots + a_{i,i-1} k_{i-1}))(a_{i,1} k_1 + \ldots a_{i,i-1} k_{i-1}) \Big].\end{aligned}$$

Or, on a pour tout i, $k_i(t_n, x_n, 0) = f(t_n, x_n)$ et $\sum_{j=1}^{i-1} a_{i,j} = c_i$. Il en résulte que

$$\frac{\partial \Phi}{\partial h}(t_n, x_n, 0) = \sum_{i=1}^{s} b_i c_i \Big[\frac{\partial f}{\partial t}(t_n, x_n) + d_x f(t_n, x_n)(f(t_n, x_n)) \Big].$$

Et donc en utilisant la condition d'ordre 2

$$\frac{\partial \Phi}{\partial h}(t_n, x_n, 0) = \frac{1}{2} \Big[\frac{\partial f}{\partial t}(t_n, x_n) + d_x f(t_n, x_n)(f(t_n, x_n)) \Big],$$

il vient $\sum_{i=1}^{s} b_i c_i = \frac{1}{2}$.

On peut vérifier que les conditions d'ordre 2 sont vérifiées pour les méthodes de Runge et d'Euler amélioriée. Pour la première méthode, on a $c_2 = \frac{1}{2}$, $b_1 = 0$ et $b_2 = 1$. On a donc bien $b_1 + b_2 = 1$ et $b_1 c_1 + b_2 c_2 = \frac{1}{2}$. Pour la deuxième méthode, on a $c_2 = 1$, $b_1 = \frac{1}{2}$ et $b_2 = \frac{1}{2}$. On a donc également $b_1 + b_2 = 1$ et $b_1 c_1 + b_2 c_2 = \frac{1}{2}$.

Notons que si on se restreint aux méthodes à deux étages d'ordre 2. Le seul terme de A non nul est a_{21} qui est forcément égal à c_2 et c_1 est toujours nul. Les seules inconnues de la méthode de Runge-Kutta sont donc b_1, b_2 et c_1 qui vérifient les relations

$$\begin{aligned} b_1 + b_2 &= 1 \\ b_2 c_2 &= \frac{1}{2} \end{aligned}$$

Ainsi pour avoir une méthode de Runge-Kutta à deux étages d'ordre 2, on peut choisir $c_2 = \theta$ non nul quelconque, puis il en découle $b_2 = \frac{1}{2\theta}$ et $b_1 = 1 - \frac{1}{2\theta}$. Le tableau de Butcher pour une méthode d'ordre 2 à 2 étages générale s'écrit donc

$$\begin{array}{c|cc} 0 & & \\ \theta & \theta & \\ \hline & 1 - \frac{1}{2\theta} & \frac{1}{2\theta} \end{array}$$

Remarque 5. *On peut faire les mêmes calculs pour obtenir des ordres plus élévés, mais il deviennent de plus en plus compliqués. En ajoutant plus d'étages, on obtient plus de degrés de liberté pour atteindre un ordre donné. Mais plus il y a d'étages, plus il faut faire d'évaluations de fonctions et donc plus la méthode est coûteuse. Le nombre minimal d'étages pour atteindre l'ordre p est p et il existe des méthodes de Runge-Kutta d'ordre p à p étages seulement pour $p \leq 4$. C'est la raison pour laquelle la méthode à un pas la plus utilisée en pratique est la méthode d'ordre 4 qui est souvent appelée méthode de Runge-Kutta (ou RK4) et qui a le tableau de Butcher suivant.*

0				
$\frac{1}{2}$	$\frac{1}{2}$			
$\frac{1}{2}$	0	$\frac{1}{2}$		
1	0	0	1	
	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{6}$

5.7 Schémas implicites

Le schéma d'Euler implicite
On écrit à nouveau

$$y(t_{n+1}) - y(t_n) = \int_{t_n}^{t_{n+1}} f(s, y(s))\mathrm{d}s.$$

On utilise cette fois-ci la formule des rectangles à droite pour approcher l'intégrale :

$$\int_{t_n}^{t_{n+1}} f(s, y(s))\mathrm{d}s \simeq \Delta t f(t_{n+1}, y(t_{n+1})).$$

On retrouve alors le schéma d'Euler implicite

$$u_0 = y(0), \quad u_{n+1} = u_n + \Delta t f(t_{n+1}, u_{n+1}), \quad n = 0, \ldots, N-1.$$

Lemme 7. *Soit $g : \mathbb{R}^d \to \mathbb{R}^d$ une fonction satisfaisant*

$$\|g(x) - g(y)\| \leq L\|x - y\|, \textit{pout tout } (x, y) \in \left(\mathbb{R}^d\right)^2,$$

avec un nombre $0 < L < 1$, alors g admet un unique point fixe, i.e., il existe un unique $\ell \in \mathbb{R}^d$ tel que $g(\ell) = \ell$.

Théorème 27. *Si $f : \mathbb{R}^+ \times \mathbb{R}^d \to \mathbb{R}^d$ est une fonction continue et Lipschitzienne en $x \in \mathbb{R}^d$, i.e., il existe $L > 0$ tel que pour tout $(x, y) \in \left(\mathbb{R}^d\right)^2$*

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

alors il existe une unique solution u_{n+1} satisfaisant

$$u_{n+1} = u_n + \Delta t f(t_{n+1}, u_{n+1}),$$

dès lors que le pas de temps vérifie $\Delta t < 1/L$.

Le schéma de Crank-Nicolson
On utilise cette fois-ci la méthode des trapèzes

$$u_{n+1} = u_n + \frac{\Delta t}{2}(f(t_n, u_n) + f(t_{n+1}, u_{n+1})).$$

Méthodes de Runge Kutta implicites
On reprend la notation matricielle, mais cette fois-ci, on admet d'avoir plus de coefficients

$$\begin{array}{c|c} c_1 & a_{1,1}, \dots, a_{1,s} \\ c_2 & a_{2,1}, \dots, a_{2,s} \\ c_s & a_{s,1}, \dots, a_{s,s} \\ \hline & b_1, \dots, b_s. \end{array}$$

Méthodes de Runge Kutta semi-implicites
Un compromis est de n'avoir que s équations non linéaires indépendantes à résoudre en considérant

$$\begin{array}{c|c} c_1 & a_{1,1} \\ c_2 & a_{2,1}, a_{2,2} \\ c_s & a_{s,1}, \dots, a_{s,s} \\ \hline & b_1, \dots, b_s. \end{array}$$

5.8 Stabilité absolue

On peut se demander quel est l'intérêt d'utiliser des méthodes implicites, sachant que leur résolution est plus difficile. On considère ici le problème test

$$y'(t) = \lambda y(t), \quad y(0) = 1, \tag{5.9}$$

dont la solution explicite est donnée par $y(t) = \exp(\lambda t)$.

Définition 22. *Une méthode approchant le problème (5.9) est absolument stable si*

$$\|u_n\| \to 0, \quad \textit{lorsque} \quad t_n \to \infty. \tag{5.10}$$

La région de stabilité est alors définie par

$$\mathcal{A} = \{z = h\lambda : \textit{(5.10) est vérifiée}\}.$$

Enfin une méthode est dite A-stable ou inconditionnellement stable si pour tout λ tel que $\Re(\lambda) < 0$ on a (5.10).

On remarque que les méthodes implicites d'Euler et de Crank Nicholson sont A-stables, alors que les méthodes explicites ne sont pas A-stables. De manière générale, les méthodes explicites ne sont pas A-stables ; cela justifie donc l'utilisation de méthodes implicites. Par contre, certaines méthodes implicites peuvent être instables ou seulement *conditionnellement stables* (i.e. stables mais non A-stables).

5.9 Méthodes multi-pas

Définition 23. *Une méthode à k pas est une méthode qui s'écrit sous la forme*

$$\alpha_k u_{j+k} + \alpha_{k-1} u_{j+k-1} + \dots \alpha_0 u_j = h(\beta_k f_{j+k} + \dots + \beta_0 f_j), \quad \alpha_k \neq 0,$$

où l'on a posé $f_j = f(t_j, u_j)$. Si $\beta_k = 0$, la méthode est explicite et implicite sinon.

Pour définir les méthodes à pas multiples, on utilise des formules d'intégration numérique, de différentiation ou d'interpolation (cf Exercices).

Schéma saute-mouton (leap-frog)

$$u_{n+1} = u_{n-1} + 2\Delta t f(t_n, u_n).$$

Méthode de Simpson

$$u_{n+1} = u_{n-1} + \frac{\Delta t}{3}(f(t_{n-1}, u_{n-1}) + 4f(t_n, u_n) + f(t_{n+1}, u_{n+1})).$$

Schéma d'ordre 2 d'Adams-Bashforth

$$u_{n+1} = u_{n-1} + \frac{\Delta t}{2}(3f(t_n, u_n) - f(t_{n-1}, u_{n-1})).$$

Schémas prédicteur-correcteur

Supposons u_0 et u_1 donnés. On peut alors prédire par une formule explicite une première approximation $u_{n,0}$, $\quad n = 1, \ldots, N-1$,

$$u_{n+1,0} = u_n + \frac{\Delta t}{2}(3f(t_n, u_n) - f(t_{n-1}, u_{n-1})).$$

On effectue ensuite K étapes de correction

$$u_{n+1,k+1} = u_n + \frac{\Delta t}{2}(f(t_n, u_n) + f(t_{n+1}, u_{n+1,k})),$$

avec $k = 0, \ldots, K$, le nombre d'itérations K est choisi suffisament grand pour que la suite $u_{n+1,k}$ ne varie plus beaucoup par rapport à k.

5.10 Méthodes multi-pas (complément)

On va dans cette section construire des méthodes numériques qui calculent l'approximation x_{n+1} en fonction des approximations en plusieurs temps différents et pas seulement de x_n. Ces méthodes sont plus anciennes que les méthodes de Runge-Kutta. Elles ont été introduites par Adams dans les années 1850.

5.10.1 Méthode d'Adams explicite

On se donne une subdivision $t_0 < t_1 < \cdots < t_N = t_f$ de l'intervalle de temps sur lequel on veut résoudre l'équation différentielle. Comme pour les méthodes à un pas, on commence par intégrer l'équation entre t_n et t_{n+1}, ce qui donne

$$x(t_{n+1}) = x(t_n) + \int_{t_n}^{t_{n+1}} f(t, x(t))\, dt.$$

Pour une méthode à k pas, on introduit alors $p_{k-1}(t)$ le polynôme d'interpolation de Lagrange de degré $k-1$ aux points $(t_{n-k+1}, f(t_{n-k+1}, x_{n-k+1})), \ldots, (t_n, f(t_n, x_n))$ et on remplace $f(t, x(t))$ dans l'intégrale par ce polynôme d'interpolation $p_{k-1}(t)$. La méthode d'Adams explicite consiste donc à définir

$$x_{n+1} = x_n + \int_{t_n}^{t_{n+1}} p_{k-1}(t)\, dt.$$

Exemples

1. Méthode d'Adams explicite à un pas.
   Dans ce cas $p_0(t)$ est le polynôme d'interpolation de degré 0 de (t_n, x_n) qui vaut $p_0(t) = f(t_n, x_n)$. La méthode d'Adams explicite à un pas s'écrit donc

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

   ce qui correspond à la méthode d'Euler explicite.
2. Méthode d'Adams explicite à deux pas. On définit dans ce cas le polynôme de degré 1 $p_1(t)$ tel que $p_1(t_{n-1}) = f(t_{n-1}, x_{n-1})$ et $p_1(t_n) = f(t_n, x_n)$. On obtient alors

   $$p_1(t) = \frac{t_n - t}{t_n - t_{n-1}} f(t_{n-1}, x_{n-1}) + \frac{t - t_{n-1}}{t_n - t_{n-1}} f(t_n, x_n).$$

   En intégrant on obtient la méthode d'Adams explicite à deux pas

   $$x_{n+1} = x_n - \frac{(t_{n+1} - t_n)^2}{2(t_n - t_{n-1})} f(t_{n-1}, x_{n-1}) + \big(\frac{(t_{n+1} - t_{n-1})^2}{2(t_n - t_{n-1})} - \frac{1}{2}(t_n - t_{n-1})\big) f(t_n, x_n).$$

   Dans le cas d'un pas constant h cette méthode devient

   $$x_{n+1} = x_n + \frac{h}{2}(3f(t_n, x_n) - f(t_{n-1}, x_{n-1})).$$

3. Méthode d'Adams explicite à trois pas. Dans le cas d'un pas h constant cette méthode s'écrit

   $$x_{n+1} = x_n + \frac{h}{12}(23f(t_n, x_n) - 16f(t_{n-1}, x_{n-1}) + 5f(t_{n-2}, x_{n-2})).$$

Notons qu'il faut une technique différente pour initialiser une méthode multi-pas, qui est une récurrence portant sur plusieurs termes alors qu'on n'a qu'une condition initiale.

5.10.2 Méthode d'Adams implicite

Dans la méthode d'Adams explicite on utilise le polynôme d'interpolation hors de l'intervale $[t_{n-k+1}, t_n]$ sur lequel il est construit, ce qui peut entraîner des erreurs importantes. D'où l'idée pour améliorer la méthode de définir le polynôme d'interpolation p_{k-1} également à partir de la valeur encore inconnue x_{n+1}. Ainsi on définit $p_{k-1}(t)$ comme étant le polynôme d'interpolation de degré k aux points $(t_{n-k+1}, f(t_{n-k+1}, x_{n-k+1})), \ldots, (t_{n+1}, f(t_{n+1}, x_{n+1}))$. On obtient dans ce cas pour un pas h constant

1. Méthode d'Adams implicite à un pas

   $$x_{n+1} = x_n + hf(t_{n+1}, x_{n+1}).$$

   C'est la méthode d'Euler implicite.
2. Méthode d'Adams implicite à deux pas

   $$x_{n+1} = x_n + \frac{h}{2}(f(t_{n+1}, x_{n+1}) + f(t_n, x_n)).$$

3. Méthode d'Adams implicite à trois pas

$$x_{n+1} = x_n + \frac{h}{12}(5f(t_{n+1}, x_{n+1}) + 8f(t_n, x_n) - f(t_{n-1}, x_{n-1})).$$

L'inconvénient de la méthode d'Adams implicite (et des méthodes implicites en général) est que la formule donnant x_{n+1} dépend de $f(t_{n+1}, x_{n+1})$. Elle s'écrit sous la forme

$$x_{n+1} = \lambda_n + h\beta f(t_{n+1}, x_{n+1}). \tag{5.11}$$

L'inconnue x_{n+1} est donc la solution d'une équation fonctionnelle non linéaire dans le cas général qu'on doit résoudre par une méthode de type Newton.

5.10.3 Méthode prédicteur correcteur

Plutôt que de résoudre directement l'équation fonctionnelle (5.11) à laquelle on aboutit dans la méthode d'Adams implicite, on préfère souvent utiliser des méthodes du type prédicteur-correcteur qui consistent à prédire une valeur de x_{n+1} par une méthode explicite, puis à corriger une ou plusieurs fois cette valeur en utilisant la formule (5.11) où l'on met la valeur prédite dans le membre de droite.

Bibliographie

[1] Lascaux-Théodor, *Analyse Numérique matricielle appliquée à l'art de l'ingénieur I et II*, Dunod, 2000.

[2] Quarteroni, Sacco, Saleri, *Méthodes numériques, algorithmes, analyse et applications.*