CENTRALE MARSEILLE

MATHEMATIQUES 1A-ANALYSE NUMÉRIQUE

G. Chiavassa, J. Liandrat, M. Tournus

Remerciements

Un GRAND MERCI à Chantal Esmenard qui a tapé la première version de ce manuscrit sous Latex. Nous remercions chaleureusement Francois MAUGER (Centrale Marseille promotion 2009), Réda JABRAZKO (Centrale Marseille promotion 2011), Rodolfo Andrés NÙNEZ URIBE (Centrale Marseille promotion 2013), Anne Laure BAILLY (Centrale Marseille promotion 2016) et Romain Maviel (Centrale Marseille promo-entrante 2016) pour avoir porté à notre attention de nombreuses erreurs.
Nous remercions aussi par avance les lecteurs qui voudront bien nous faire part (par exemple par courrier électronique à jacques.liandrat@centrale-marseille.fr ou guillaume.chiavassa@centrale-marseille.fr ou magali.tournus@centrale-marseille.fr) de leurs suggestions ou corrections pour l'amélioration de ce texte.

Notations

v désigne un vecteur de $\mathbb{R}^n$, $v = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$,
$A \in L(\mathbb{R}^n)$ désigne une matrice carrée à coefficients réels, $A = (a_{ij})_{1 \leq i,j \leq n}$,
f désigne une fonction

Normes vectorielles

- pour $1 \leq p < \infty, ||v||_p = (\sum_{i=1}^{n} |v_i|^p)^{1/p}$,
- $||v||_\infty = sup_{1 \leq i \leq n} |v_i|$,
- Si $||A||$ est une matrice symétrique définie positive, $||x||_A = (x^T Ax)^{1/2}$.

Espaces

- $\forall n \in \mathbb{N}, P_n(\mathbb{R})$ désigne l'espace des polynômes à une variable à coefficients réels de degré inférieur ou égal à n

Normes matricielles

- En plus des propriétés de définition d'une norme, une norme matricielle vérifie $\forall A, B, ||AB|| \leq ||A||.||B||$,
- Norme subordonnée : $||A||_{E,F} = \sup_{x \neq 0} \frac{||Ax||_F}{||x||_E}$,
- $||A||_\infty = \sup_i \sum_j |a_{ij}|$,
- $||A||_2 = \sqrt{\rho(AA^*)}$,
- $||A||_1 = sup_j \sum_i |a_{ij}|$,

Normes fonctionnelles

- $\forall 1 \leq p < +\infty, \forall f \in L^p(\Omega), ||f||_p = (\int_\Omega |f|^p)^{1/p}$,
- $\forall f \in L^\infty(\Omega), ||f||_\infty = \sup_{x \in \Omega} |f(x)|$

Table des matières

Chapitre 1

Approximation Polynomiale

1.1 Méthode d'interpolation de Lagrange

Soit $f : [a, b] \to I\!R, f \in C^0[a, b]$ (fonction continue sur $[a, b]$).
On se donne $(n + 1)$ points $x_0, x_1, ...x_n$ dans $[a, b]$, 2 à 2 distincts.
Existe t-il un polynôme $p_n \in P_n$ (ensemble des polynômes de degré inférieur ou égal à n) tel que $p_n(x_i) = f(x_i)$ pour tout i de 0 à n ?

Pour $i \in [o, n]$ on définit : $l_i(x) = \frac{\Pi_{j\neq i}(x-x_j)}{\Pi_{j\neq i}(x_i-x_j)}$

On a clairement $l_i(x) \in P_n$ et $l_i(x_j) = \delta_{ij}$. Le polynôme défini par $p_n(x) = \sum_{i=0}^{n} f(x_i)l_i(x)$ répond au problème posé. On a le théorème suivant :

Théorème 1.1.1. Interpolation de Lagrange en dimension 1
$\forall n \in I\!N, (x_0, x_1, ...x_n) \in [a, b], f \in C^0([a, b])$, *avec* $x_i \neq x_j, i \neq j$, *le problème d'interpolation :* " **Trouver** $p \in P_n$ **tel que** $p_n(x_i) = f(x_i)$ **pour tout** $0 \leq i \leq n$" *admet une solution unique que l'on appelle polynôme d'interpolation de Lagrange de* f *aux points* $x_0, x_1...xn$. *Il est donné par :*

$$p_n(x) = \sum_{i=0}^{n} f(x_i)l_i(x). \tag{1.1}$$

Preuve

L'existence est déjà prouvée. L'unicité se prouve de la façon suivante :

Soit q_n un autre polynôme de P_n interpolant f aux points x_i. On a $q_n(x_i) = f(x_i) = q_n(x_i)$ donc x_i est racine de $q_n - p_n$ et $\Pi_{n+1}(x) = \prod_{j=0}^{n}(x - x_j)$ divise $q_n - p_n$. Comme $d^o\Pi_{n+1} = (n + 1)$ et $d^o(q_n - p_n) \leq n$ on a $p_n = q_n$

■

Remarque 1.1.2. *Les polynômes* $l_i(x)$ *sont appelés les polynômes interpolateurs élémentaires de Lagrange. Plus généralement étant donné un espace* V *et une suite de points* $(x_0, ...xn)$ *on appelle fonctions interpolatrices élémentaires de Lagrange (si elles existent et si elles sont uniques) les*

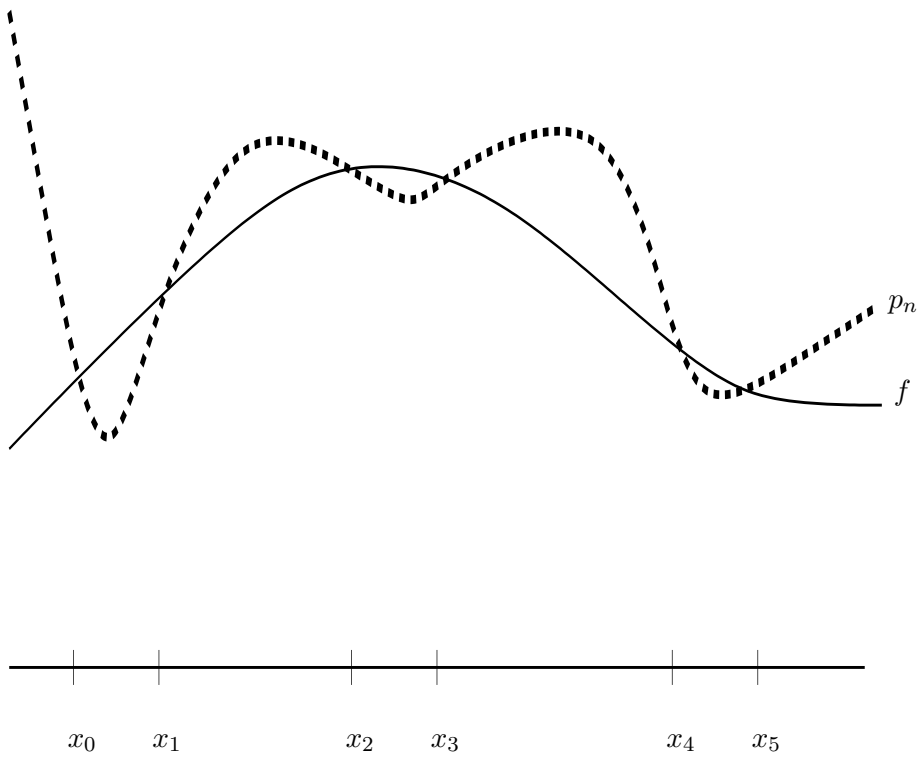


FIGURE 1.1 – Interpolation de Lagrange

fonctions l_i de V vérifiant : $l_i(x_j) = \delta_{i,j}$. Elles permettent de définir l'interpolation dans l'espace V également appelée la projection collocation de C^0 dans V : $\forall f \in C^0 \mapsto Pc(f)(x) = \sum f(x_i)li(x) \in V$.

1.1.1 Formule d'erreur

Théorème 1.1.3. Formule d'erreur de l'interpolation de Lagrange

On suppose que f est $n+1$ fois continuement dérivable sur $[a,b]$. Alors, pour tout $x \in [a,b]$, il existe un point $\xi_x \in]min\{x, x_i\}, max\{x, x_i\}[$ tel que :

$f(x) - p_n(x) = \frac{1}{(n+1)!} \prod_{n+1}(x) f^{(n+1)}(\xi_x)$.

Preuve

Si $x = x_i$ comme $\Pi_{n+1}(x) = \prod_{j=0}^{n}(x - x_j)$, on a $\Pi_{n+1}(x_i) = 0$ et tout point ξ_x convient.
Si $x \neq x_i$. On introduit p_{n+1} le polynôme d'interpolation de f de degré $n+1$ relativement aux points $x, x_0, ...x_n$. On a $p_{n+1} \in P_{n+1}$; $f(x) - p_n(x) = p_{n+1}(x) - p_n(x)$. Or $d^o(p_{n+1} - p_n) \leq n+1$ et $p_{n+1} - p_n$ s'annule en $(n+1)$ points $(x_0, ...x_n)$ donc :

$$p_{n+1}^{(t)} - p_n^{(t)} = c\Pi_{n+1}^{(t)}$$

Soit $g_{(t)} = f(t) - p_{n+1}(t)$ on a $g(t) = f(t) - p_n(t) - c\Pi_{n+1}(t)$. De plus, $g(t) = 0$ pour $t \in \{x, x_0, ...x_{n+1}\}$ donc (voir lemme 1.1.4) $\exists \xi_x \in]\min\{x, x_i\}, \max\{x, x_i\}[$ tel que $g^{(n+1)}(\xi_x) = 0$. Or $p_n^{(n+1)} = 0$ et $\Pi_{n+1}^{(n+1)} = (n+1)!$. Il s'en suit que $f^{(n+1)}(\xi_x) - p_n^{(n+1)}(\xi_x) - c\Pi_{(n+1)}^{n+1} = 0$. Finalement $c = \frac{f^{(n+1)}(\xi_x)}{(n+1)!}$ et $f(x) = p_n(x) + c\Pi_{n+1} = p_n(x) + \frac{f^{(n+1)}(\xi_x)}{(n+1)!}\Pi_{n+1}(x)$.

■

On a utilisé le lemme suivant qui est une conséquence directe du théorème de Rolle :

Lemme 1.1.4. *Soit g une fonction p fois continuement dérivable sur $[a, b]$. On suppose qu'il existe $p+1$ points $c_0 < c_1 < ...c_p$ de $[a, b]$ tels que $g(c_i) = 0$. Alors, $\xi \in]c_0, c_p[$ tel que $g^{(p)}(\xi) = 0$.*

On a $|f(x) - p_n(x)| \leq \frac{1}{(n+1)!}|f^{(n+1)}(\xi_x)||\Pi_{n+1}(x)|$ ce qui montre que l'erreur d'interpolation est liée d'une part à $||f^{(n+1)}||_\infty$ qui est d'autant plus grande que f oscille et d'autre part à $||\Pi_{n+1}||_\infty$ qui est liée à la répartition des points d'interpolation.

Remarque 1.1.5. *Pour une segmentation régulière de pas $x_{i+1} - x_i = h$ on a : $|f(x) - p_n(x)| \leq C.h^{n+1}, \forall x \in [x_0, x_n]$, la constante C dépendant de f.*

1.1.2 Méthode des différences divisées

Il s'agit d'une méthode simple et efficace fournissant le polynôme d'interpolation de Lagrange de f.

Soit p_k le polynôme d'interpolation de Lagrange de f aux points $x_0, x_1, ...x_k$. On note $f_{[}x_0, ...x_k]$ le coefficient de x^k dans $p_k(x)$ (coefficient directeur). En fait, la connaissance de tous les coefficients $f_{[}x_0, ...x_l], 0 \leq l \leq k$ suffit pour construire p_k. En effet :
$p_k - p_{k-1}$ et un polynôme de $d^o \leq k$, s'annulant en $x_0...x_{k-1}$ et admettant $f_{[}x_0, ...x_k]$ comme coefficient directeur donc :

$$p_k - p_{k-1} = f_{[}x_0, ...x_k](x - x_0)...(x - x_{k-1}).$$

Sachant que $p_0(x) = f(x_o)$ on obtient alors facilement

$$p_n(x) = f(x_0) + \sum_{k=1}^{n} f_{[x_0,...xk]}(x - x_0)...(x - x_{k-1}). \tag{1.2}$$

Lemme 1.1.6. *Le calcul de $f_{[}x_0, ...x_{k-1}]$ s'effectue d'autre part grâce à la formule suivante :*

$$\forall k \geq 1, f_{[x_0,x_1,...x_k]} = \frac{f_{[x_1...x_k]} - f_{[x_0,...x_{k-1}]}}{x_k - x_0} \tag{1.3}$$

Preuve

Soit q_{k-1} le polynôme de P_{k-1} qui interpole f aux points $(x_1, x_2...x_k)$. On pose

$$\tilde{p}_{k(x)} = \frac{(x-x_0)q_{k-1}(x)-(x-x_k)p_{k-1}}{x_k-x_0}$$

On a $\tilde{p}_k \in P_k$ et

$$\begin{cases} \tilde{p}_k(x_0) &= p_{k-1}(x_0) = f(x_0) \\ \tilde{p}_k(x_k) &= q_{k-1}(x_k) = f(x_k) \\ \tilde{p}_k(x_i) &= \frac{(x_i-x_0)f(x_i)-(x_i-x_k)f(x_i)}{x_k-x_0} = f(x_i), \forall i, 0 < i < k. \end{cases}$$

Donc $\tilde{p}_k = p_k$
En égalant les coefficients de x^kdans la définition de $\tilde{p}_k = p_k$ on obtient exactement la récurrence.

■

L'algorithme de calcul fonctionne alors de la façon suivante :

$$\begin{array}{ccccccccc} f(x_n) & \to & f_{[x_{n-1}x_n]} & \to & f_{[x_{n-2}x_{n-1}x_n]} & \cdots & \to & f_{[x_0 \ldots x_n]} \\ & \nearrow & & & & & & \\ f(x_{n-1})) & \to & f_{[x_{n-2}x_{n-1}]} & \nearrow & & & & \\ \vdots & \nearrow & & & & & & \\ \vdots & & & & & & & \\ f(x_0) & & & & & & & \end{array} \tag{1.4}$$

Pour évaluer numériquement le polynôme avec un minimum d erreur il est recommandé de l'écrire sous la forme :

$$p_n(x) = f[x_0] + (x - x_0)\Big(f[x_0, x_1] + (x - x_1)\Big(f[x_0, x_1, x_2] + ... + (x - x_n - 1)f[x_0 ... x_n]\Big)\Big)$$

Remarque 1.1.7. *Le cas où les points d'interpolation sont équidistants est particulièrement simple. On introduit l'opérateur*
aux différences finies $\Delta : (f_0, f_1 ... f_n) \to (\Delta f_0, \Delta f_1 ... \Delta f_{n-1})$

$$avec\ \Delta f_k = f_{k+1} - f_k.$$

On vérifie alors que $f[x_i, ...x_{i+k}] = \frac{\Delta^k f_i}{k! h^k}$

1.1.3 Stabilité et convergence de l'interpolation Lagrange

Soit L_n l'opérateur d'interpolation de Lagrange défini par :

Définition 1.1.8.

$$L_n : \begin{cases} C[a,b] & \to & P_n \\ f & \mapsto & p_n \end{cases}$$

Il est très important de connaître l'influence sur p_n, des erreurs sur les valeurs $f(x_i)$. En effet, dans la pratique on ne connaît que $\tilde{f}_i$, une approximation de $f(x_i)$. Il s'agit bien d'un problème de stabilité au sens ou l'on étudie l'influence d'une perturbation $(f(x_i) - (\tilde{f}_i))$ sur le résultat $(L_n(f) - \sum_{i=0}^{n} \tilde{f}_i l_i)$.
On a :

$$p_n(x) = \sum_{i=0}^{n} f(x_i) l_i(x) \text{ et } \tilde{p}_n(x) = \sum_{i=0}^{n} \tilde{f}_i l_i(x).$$

Donc

$$p_n(x) - \tilde{p}_n(x) = \sum_{i=0}^{n} \big(f(x_i) - \tilde{f}_i\big) l_i(x)$$

et

$$|p_n(x) - \tilde{p}_n(x)| \leq \sum_{i=0}^{n} |l_i(x)| \max_{i \in [0,n]} \left| f(x_i) - \tilde{f}_i \right|,$$

soit

$$|p_n(x) - \tilde{p_n}(x)| \leq \Lambda_n max \left| f(x_i) - \tilde{f}_i \right|,$$

avec

$$\Lambda_n = ||(\sum_{i=0}^{n} |l_i(x)|)||_\infty.$$

Finalement, $||p_n - \tilde{p_n}|| \leq \Lambda_n \max \left| f(x_i) - \tilde{f}_i \right|$.
Λ_n s'interprète donc comme le coefficient d'amplification des erreurs dans le procédé d'interpolation de Lagrange. Λ_n est la constante de Lebesgue associé aux points $\{x_i\}_{i=0}^n$ et à l'intervalle $[a, b]$. On démontre facilement que $\Lambda_n = \|L_n\|_\infty$ ou $\|\|_\infty$ est induite par la norme $\|\|_\infty$ dans $C^0[a, b]$.
En fait Λ_n est également lié aux problèmes d'erreur d'interpolation : Le théorème suivant compare l'erreur d'interpolation $\|f - p_n\|_\infty$ à l'erreur minimale correspondant à la meilleure approximation dans P_n.

Théorème 1.1.9. Interpolation/meilleure approximation
$\forall_f \in C^0[a, b]$ *on a* $\|f - p_n\|_\infty \leq (1 + \Lambda_n)\epsilon_n(f)$,

où $\epsilon_n(f) = d(f, P_n) = inf\{||f - q||_\infty, q \in P_n\}$ *est appelée degrés d'approximation de la fonction* f *par les polynômes de* P_n*, au sens de la convergence uniforme.*

Preuve

$\forall f \in C^0, \forall q \in P_n$, $f - p_n = f - q + q - p_n$ et $L_n(q) = q$. Donc, $||f - p_n||_\infty \leq \|f - q\|_\infty (1 + \Lambda_n)$ et en prenant la borne inférieure de $\|f - q\|_\infty$ pour $q \in P_n$ on obtient le résultat annoncé.

■

Remarque 1.1.10. *Pour minimiser* Λ_n, n *fixé, on est conduit à modifier le choix des points* $(x_0, ...x_n)$*. Si on appelle* $\bar{\Lambda}_n$ *la borne inférieure pour tous les choix possibles on peut démontrer que* $\bar{\Lambda}_n \sim_{n \to +\infty} \frac{2}{\pi} Logn$*. Les points d'interpolation associés à* $\bar{\Lambda}$ *ont été caractérisés. Il ne sont pas calculables facilement. Pour des points équidistants on a* $\Lambda_n \sim \frac{2^{n+1}}{e^{nLogn}}$ *ce qui est loin d'être optimal. Pour les points d'interpolation de Chebychev,* $\Lambda_n \sim \frac{2}{\pi} Log(n)$ *de sorte que les points de Chebychev sont quasiment optimaux.*

Le comportement de $\epsilon_n(f)$ est contrôlé par le théorème suivant où $w(f,h)$, le module de continuité de f est défini par :

$$w(f,h) = \max_{\substack{t,t' \in [a,b] \\ |t-t'| \leq h}} |f(t) - f(t')|.$$

Théorème 1.1.11. *(Jackson)*
Il existe un réel M, indépendant de n, a et b tel que pour tout $f \in C^0[a,b]$,

$$\epsilon_n(f) \leq M w(f, \frac{b-a}{n}). \tag{1.5}$$

$$\tag{1.6}$$

Une conséquence est le théorème de Weierstrass :

Théorème 1.1.12. *(Weierstrass) :*
Si $[a,b]$ est un intervalle borné de $\mathbb{R}$; l'ensemble des fonctions polynomiales est dense dans $C^0[a,b]$ pour la topologie de la convergence uniforme.

1.2 Méthode d'interpolation de Hermite

Au lieu de faire coïncider f et p_n aux points x_i de $[a,b]$ on cherche maintenant à faire coïncider f et p_n ainsi que leurs dérivées d'ordre inférieur ou égal à α_i aux points x_i.

Théorème 1.2.1. Interpolation de Hermite
Soient $(k+1)$ points $(x_0, x_1, ...x_k)$ et $(k+1)$ entiers (α_i). on pose $n = k + \alpha_0 + \alpha_1 + ...\alpha_k$.
Alors, étant donné une fonction f définie sur $[a,b]$ admettant des dérivées continues d'ordre α_i aux points x_i, il existe un polynôme $p_n \in P_n$ et un seul tel que $\forall (i,l), 0 \leq_i \leq k, 0 \leq l \leq \alpha_i$, $p_n^{(l)}(x_i) = f_n^{(l)}(x_i)$ où $g^{(l)}(x_i)$ désigne la dérivée d'ordre l de g au point x_i.
p_n est appelé polynôme d'interpolation de Hermite de la fonction f relativement aux points $x_0, ...x_k$ et aux entiers $\alpha_0, ...\alpha_k$.

Preuve

Les équations ci-dessus fournissent un système linéaire à $(n+1)$ inconnues (les coefficients de p_n) et $(n+1)$ équations $(k+1+\alpha_0+..., \alpha_k = n+1)$.

Il suffit donc, pour prouver existence et unicité de montrer que le problème homogène associé n'admet que la solution nulle.
(x_i) est racine d'ordre $(\alpha_i + 1)$ de p_n et donc $p_n(x) = \prod_{i=0}^{k} (x - x_i)^{\alpha i+1} q(x)$

or

$$\sum_{i=0}^{r} \alpha_i + 1 = n + 1$$

donc $q = 0$ et $p_n = 0$, ce qui conclut la preuve.

■

On s'intéresse en particulier aux polynômes $p_{il}, 0 \leq i \leq k, 0 \leq l \leq \alpha_i$ définis par :

$\forall (n,m), 0 \leq n \leq k, 0 \leq m \leq \alpha_i$, $p_{il}^m(x_n) = \delta_{(i,l),(n,m)}$.
D'après le théorème précédent ces polynômes existent et sont uniques. On a alors :

$$\forall f \text{ vérifiant les hypothèses du théorème précédent : } {p_n}^{(x)} = \sum f^{(l)}(x_i)\ pi_l(x).$$

Exercice

Si on pose

$$q_i(x) = \prod_{\substack{j=0,k \\ j \neq i}} \left(\frac{x - x_j}{x_i - x_j} \right)^{\alpha_j + 1}$$

Vérifier que $p_{i\alpha_i}(x) = \frac{(x-x_i)^{\alpha_i}}{\alpha_i!} q_i(x)$ et que
pour $l = \alpha_{i-1}, ...0$

$$p_{il}(x) = \frac{(x - x_i)^l}{l!} q_i(x) - \sum_{j=l+1}^{\alpha_i} {C_j}^l {q_i}^{(j-l)}(x_i) p_{ij}(x).$$

□

Remarque 1.2.2. *L'interpolation de Lagrange est un cas particulier de l'interpolation de Hermite correspondant au choix $\alpha_0 = \alpha_1 = ...\alpha_k = 0$, $k = n$. L'approximation par le polynôme de Taylor est un autre cas particulier correspondant à $k = 0$ et $\alpha_0 = n$*

Erreur d'interpolation :

On a le théorème suivant :

Théorème 1.2.3. Erreur dans l'interpolation de Hermite
On suppose $f \in C^{n+1}[a,b]$. Alors, pour tout $x \in [a,b]$, il existe ξ_x appartenant au plus petit intervalle fermé contenant $x, x_0, x_1 ... x_k$ tel que :

$$f(x) - p_n(x) = \frac{1}{(n+1)!} \Pi_{n+1}(x) f^{(n+1)}(\xi_x) \ \textit{Avec}\ \Pi_{n+1}(x) = \Pi_{i=0}^k (x - x_i)^{\alpha_i + 1}$$

1.3 Meilleure approximation

Nous avons vu que l'interpolation de Lagrange pouvait donner de mauvais résultats comme procédé d'approximation d'une fonction continue sur un intervalle puisque la constante de Lebesgue Λ_n tend vers l'infini quand n tend vers l'infini(voir dans [3] l'analyse du phénomène de Runge).

On s'intéresse donc à d'autres approximations.
La question de la meilleure approximation correspond au problème suivant : étant donnée une distance d et $f \in C^0[a,b]$ existe -t-il un polynôme $p_n \in P_n$ qui minimise la distance de f à P_n, c'est à dire qui réalise le minimum de $\{d(f,p), p \in P_n\}$ et ce polynôme est-il unique ?
On a le premier résultat suivant :
Soit un espace vectoriel de fonctions E, muni d'une norme $\|.\|_E$ tel que $P_n \subset E$. Soit d la distance associée à cette norme ($d(f,g) = ||f-g||_E$) alors :

Théorème 1.3.1. Existence d'une meilleure approximation
Pour toute fonction f de E il existe au moins un polynôme $p_n \in P_n$ tel que

$$d(f,p_n) = \begin{array}{l} inf\{d(f'q)\} \\ q \in P_n \end{array} \tag{1.7}$$

Un tel polynôme est appelé polynôme de meilleure approximation de f dans E par un polynôme de degré $\leq n$.

Preuve

Pour $p = 0$ on a $||f-p||_E = ||f||_E$ donc $d(f,P_n) \leq ||f||_E$. Soit K l'ensemble des polynômes de P_n tels que $d(f,P_n) \leq ||f||_E$. K est une partie fermée et bornée de P_n, donc compacte puisque P_n est de dimension finie. Donc la fonction $p \to \|f-p\|_E$ qui est continue atteint sa borne inférieure en un point de K.

■

Un choix classique pour E est $E = L^p(a,b)$, $1 \leq p \leq +\infty$ (muni de la norme $\|f\|_p = (\int_a^b |f|^p\,dx)^{1/p}$), avec $1 \leq p < +\infty$ et $\|f\|_\infty = sup(|f(x)|)$. Le polynôme p_n est alors appelé polynôme de meilleure approximation L^p de f. Pour $p = 2$ on parle de la meilleure approximation au sens des moindres carrés.

Attention, il n'y a pas unicité en général comme le montrent les exemples suivants :
Exemple

Exemple 1 $E = L^1(-1,1); f(x) = 1$ si $x > 0, f(x) = -1$ si $x \leq 0$

Pour tout $\alpha \in [-1,1], p_0(x) = \alpha$ réalise la meilleure approximation L^1 de f par un polynôme constant et $\|f-p_0\|_{L^1} = 2$.

Exemple 2 $E = L^\infty(-1,1); f(x) = 1$ si $x > 0, f(x) = -1$ si $x \leq 0$

$\forall \alpha \in [0,2], P_1(x) = \alpha x$ réalise la meilleure approximation de f par un polynôme de degré 1 et $\|f-p_1\|_{L^\infty(-1,1)} = 1$.

◇

Dans 2 situations classiques il est possible de démontrer l'unicité du polynôme de meilleure approximation ; il s'agit de la meilleure approximation uniforme sur un intervalle borné et de la meilleure approximation quadratique.

1.3.1 Meilleure approximation uniforme sur un intervalle borné

Soit $[a, b]$ un intervalle borné, on a le théorème suivant :

Théorème 1.3.2. Meilleure approximation uniforme sur un intervalle borné
Pour tout $f \in C^0[a, b]$ et $n \in \mathbb{N}$, il existe un unique polynôme $p_n \in P_n$ qui réalise le minimum de $||f - q||_\infty$ quand q parcourt P_n.
Ce polynôme s'appelle le polynôme de meilleure approximation uniforme de f dans P_n.

On a alors $||f - p_n||_\infty = \epsilon_n(f)$.
D'après le théorème de Jackson (1.1.11) on sait que $||f - p_n||_\infty \to 0$ quand $n \to \infty$.

1.3.2 Meilleure approximation quadratique

On s'intéresse maintenant au cas où E est un espace de Hilbert (c'est à dire un espace vectoriel muni d'un produit scalaire, complet pour la norme induite).
Soit $]a, b[$ un intervalle ouvert borné ou non de $\mathbb{R}$. on se donne une fonction $w(x)$, définie et continue, strictement positive sur $]a, b[$ telle que $\int_b^a |x|^n w(x)dx < +\infty$ pour tout n (on appelle une telle fonction une fonction poids).
On considère $E = L^2_w(a, b) = \left\{f / f\sqrt{w} \in L^2(a, b)\right\}$. D'après les hypothèses faites sur w, on a, pour tout n, $P_n \subset E$.
D'autre part, la norme $\|.\|_{2,w}$ est définie à partir du produit scalaire :

$< f, g >= \int_b^a f(x)g(x)w(x)dx$.

On a tout d'abord le théorème suivant :

Théorème 1.3.3. Polynômes orthogonaux
Il existe une suite de polynômes p_n et une seule vérifiant :
i) $d^o p_n = n$
ii) Le coefficient de plus haut degré de p_n vaut 1.
iii) $\forall q \in P_{n-1}, < p_n, q >= 0$

On a de plus la relation de récurrence : $p_n(x) = (x - \lambda_n)p_{n-1} - \mu_n p_{n-2}$ avec : $\mu_n = \frac{\|p_{n-1}\|^2_{2,w}}{\|p_{n-2}\|^2_{2,w}}$ et $\lambda_n = \frac{<xp_{n-1}, p_{n-1}>}{\|p_{n-1}\|^2_{2,w}}$.

Preuve

On construit une suite p_n en utilisant le procédé d'orthogonalisation de Graam-Schmidt sur la base canonique $1, x, x^2, ..x^n$ dans cet ordre.
On a $p_0(x) = 1, p_1(x) = x - \frac{<x,1>1}{\|1\|^2_{2,w}}$ et $p_n(x) = x^n - \sum_{1=0}^{n-1} \lambda_{i,n} p_i(x)$, avec $\lambda_{i,n} = \frac{<x^n, p_i>}{<p_i, p_i>}$.
Les p_n sont bien deux à deux orthogonaux et vérifient ii). L'unicité de cette suite découle directement de la condition d'orthogonalité.

■

Exercice

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

□

La suite $\frac{p_n}{\|p_n\|_{2,w}}$ constitue une base orthonormée de p_n.

Exemple

Les exemples classiques sont :

1) $[(a,b] = [-1,1]; w(x) = (1-x)^\alpha(1+x)^\beta$ avec $\alpha > -1, \beta > -1$. Les polynômes (p_n) sont les polynômes de Jacobi.

   $\alpha = \beta = 0$ fournit les polynômes de Legendre (on a alors $w = 1$)

   $\alpha = \beta = -\frac{1}{2}$ fournit les polynômes de Chebychev de première espèce

   $\alpha = \beta = 1/2$ fournit les polynômes de Chebychev de deuxième espèce

2) $[a,b[= [0,+\infty[, w(x) = e^{-x}$ Les polynômes (p_n) sont les polynômes de Laguerre

3) $]a,b[=]-\infty,+\infty[, w(x) = e^{-x^2}$ Les polynômes (p_n) sont les polynômes de Hermite

◇

Exercice

Déduire des exemples précédents des polynômes orthogonaux sur n'importe quel intervalle (a,b) de $\mathbb{R}$.

□

Le problème de la meilleure approximation quadratique de f dans P_n est résolu par le théorème suivant :

Théorème 1.3.4. Meilleure approximation quadratique
Il existe un unique polynôme $q_n \in P_n$ qui vérifie

$$\|f - q_n\|_{2,w} = \inf_{q \in p_n} \|f - q\|_{2,w} \tag{1.8}$$

Ce polynôme est donné par $q_n(x) = \sum_{i=0}^{n} \frac{<f,p_i>}{<p_i,p_i>} p_i(x)$ et c'est l'unique polynôme $q_n \in P_n$ tel que $< f, q >=< q_n, q > \forall q \in P_n$.

Preuve

La dernière relation définit q_n comme la projection orthogonale de f dans P_n. Il y a existence et unicité de ce polynôme. Il minimise la distance de f à p_n, en effet :

$$\begin{aligned} \forall q \in p_n, \|f-q\|^2 &= \|f - q_n + q_n - q\|^2 \\ &= \|f-q_n\|^2 + 2\,(f-q_n, q_n - q) + \|q_n - q\|^2 \\ &= \|f-q_n\|^2 + \|q_n - q\|^2 \end{aligned}$$

et $q_n = \sum \frac{<f,p_i>}{<pi,pi>} p_i(x)$.

■

Nous allons étudier maintenant le problème de la convergence de q_n vers f quand n tend vers $+\infty$.

Théorème 1.3.5. Convergence de la meilleure approximation quadratique
Si]a,b[est un intervalle borné alors

$$\lim_{n\to+\infty} \|f - q_n\|_{2,w} = 0$$

pour tout $f \in C^0_{]a,b[}$

Attention, la condition $]a,b[$ *borné est essentielle*

Preuve

Si $f \in C^0_{[a,b]}$ on introduit r_n son polynôme de meilleure approximation uniforme. On a

$$\|f - q_n\|_2 \leq \|f - r_n\|_2 \leq C_w \|f - r_n\|_\infty$$

où $C_w = (\int_a^b |w(x)|\, d_x)^{1/2}$.
Or on sait que pour $f \in C^0[a,b]$,

$$\lim_{n\to+\infty} \|f - r_n\|_\infty = 0, \text{donc} \lim_{n\to+\infty} \|f - q_n\|_2 = 0.$$

Si $f \notin C^0[a,b]$ mais uniquement à $C^0]a,b[$ on se ramène au cas précédent à l'aide d'une fonction appropriée.

■

Les polynômes orthogonaux possèdent la propriété générale :

Théorème 1.3.6. Racines des polynômes orthogonaux
Pour tout poids w sur $]a,b[$ *le polynôme* p_n *possède n racines distinctes dans l'intervalle* $]a,b[$.

Preuve

Soient $x_1, ...x_k$ les zéros distincts de p_n contenus dans $]a,b[$ et $m_1, ...m_k$ leur multiplicité. On a $m_1 + ...m_k \leq n$.
Si m_i est pair on pose $\xi_i = 0$, si m_i est impair on pose $\xi_i = 1$.
Soit $q(x) = \Pi_{i=1}^k (x - x_i)^{\xi_i}$. On a $d^o(q) \leq k \leq n$ et le polynôme $p_n.q$ admet dans $]a,b[$ les zéros (x_i) avec multiplicité paire $(m_i + \xi_i)$. Donc $p_n q$ est de signe constant dans $]a,b[/\{x_1...x_k\}$ et

$$< p_n, q >= \int_a^b p_n(x)q(x)w(x)d_x \neq 0.$$

Comme p_n est orthogonal à P_{n-1}, on a nécessairement $\deg(q) \geq n$ et $m_1 = m_2 = ... = m_k = 1$. Finalement le polynôme p_n possède n racines distinctes dans $]a,b[$.

■

Chapitre 2

Intégration numérique, formules de quadratures

Étant donné un intervalle $[a, b]$ et $w(x) > 0$ une fonction poids définie sur $]a, b[$, l'objet de ce chapitre est d'approcher pour toute fonction f continue et intégrable sur $[a, b]$, l'intégrale $\int_\alpha^\beta f(x)w(x)dx$ par une 'formule du type :

$$\sum_{i=0}^{k} \lambda_i f(x_i), \tag{2.1}$$

où $\lambda_i \in I\!R$ et $x_i \in [a, b]$.
Une telle formule s'appelle une formule de quadrature. Il faut bien noter que **les poids λ_i et les points x_i qui la définissent entièrement sont indépendants de f**. Nous allons étudier comment construire de telles formules et comment contrôler l'erreur d'évaluation de l'intégrale, $E(f) = |\int_\alpha^\beta f(x)w(x)dx - \sum_{i=0}^{k} \lambda_i f(x_i)|$.

Il existe classiquement deux types de constructions aboutissant à des formules du type (2.1) :

- Les méthodes composées, lorsque $[a, b]$ est borné et $w(x) = 1$; on introduit une segmentation $(\alpha_i)_{i=0}^k$ avec $a = \alpha < \alpha_1 < \cdots < \alpha_i < \alpha_{i+1} < \cdots < \alpha_k = b$ et on approche chaque intégrale $\int_{\alpha_i}^{\alpha_{i+1}} f(x)(d_x)$ par une formule de quadrature élémentaire obtenue en remplaçant f par un polynôme d'interpolation sur $[\alpha_i, \alpha_i + 1]$.
- Les méthodes de Gauss, pour des poids $w(x)$ et des intervalles (a, b) particuliers où on approche f par un polynôme d'interpolation sur l'intervalle (a, b) en des points choisis de manière à obtenir une formule de quadrature exacte ($\int_a^b f(x)w(x)d_x = \sum_{i=0}^{k} \lambda_i f(x_i)$) quand f est un polynôme de degré le plus élevé possible. Il se trouve que les x_i sont les racines des polynômes orthogonaux (voir théorème 1.3.3) associés au poids $w(x)$ sur $[a, b]$.

2.1 Méthodes composées

Exemple

Somme de Riemann :On choisi $\xi_i \in [\alpha_i, \alpha_{i+1}]$ et, sur chaque intervalle $[\alpha_i, \alpha_{i+1}]$, on remplace f par le polynôme de degré 0 qui interpole f au point ξ_i. On a alors $po(x) = f(\xi_i)$. L'approximation de l'intégrale sur $[\alpha_i, \alpha_{i+1}]$ est alors

$$\int_{\alpha_i}^{\alpha_{i+1}} f(x)dx \simeq (\alpha_{i+1} - \alpha_i) f(\xi_i)$$

et la formule de quadrature sur tout l'intervalle $[\alpha, \beta]$ s'écrit :

$$\int_{\alpha}^{\beta} f(x)dx \simeq \sum_{i=0}^{k-1} (\alpha_{i+1} - \alpha_i) f(\xi_i)$$

- pour $\xi_i = \alpha_i$ on obtient la formule des rectangles à gauche

$$\int_{\alpha}^{\beta} f(x)dx \simeq \sum_{i=0}^{k-1} (\alpha_{i+1} - \alpha_i) f(\alpha_i).$$

- pour $\xi_i = \alpha_{i+1}$ on obtient la formule des rectangles à droite et pour $\xi_= \frac{\alpha_i + \alpha_{i+1}}{2}$ la formule du point milieu.

Formule des trapèzes : Ici, on remplace sur $[\alpha_i, \alpha_{i+1}]$ la fonction f par p_1 le polynôme de degré 1 qui interpole f aux points α_i et α_{i+1}. On obtient alors :

$$\int_{\alpha_i}^{\alpha_{i+1}} f(x)dx \simeq \int_{\alpha_i}^{\alpha_{i+1}} p_1(x)dx = \frac{\alpha_{i+1} - \alpha_i}{2} \left(f(\alpha_i) + f(\alpha_{i+1}) \right)$$

et

$$\int_{\alpha}^{\beta} f(x)dx \simeq \sum_{i=0}^{k-1} \frac{\alpha_{i+1} - \alpha_i}{2} (f(\alpha_i) + f(\alpha_{i+1}))$$

$\diamond$

Définition 2.1.1. *Ordre d'une formule de quadrature*
Une formule de quadrature est dite d'ordre p si elle est exacte (erreur $E(f)$ nulle) pour tout polynôme de P_p et inexacte pour au moins un polynôme de P_{p+1}.

Exercice

Calculer l'ordre des formule de Riemann et des trapèzes (pour la formule de Riemann on discutera en fonction du choix du point ξ_i).

□

D'une façon générale, dans une méthode composée, pour estimer l'intégrale $\int_{\alpha}^{\beta} f(x)dx$ on commence par écrire :

$$\int_{\alpha}^{\beta} f(x)dx = \sum_{i=0}^{k-1} \int_{\alpha_i}^{\alpha_{i+1}} f(x)dx. \tag{2.2}$$

Ensuite, on effectue pour chaque intégrale le changement de variable $x = \alpha_{i+\frac{1}{2}} + \frac{sh_i}{2}$ avec $\alpha_{i+\frac{1}{2}} = \frac{\alpha_i + \alpha_{i+1}}{2}$ et $h_i = \alpha_{i+1} - \alpha_i$. Alors, si on définit $\varphi_i(s) = f(\alpha_{i+\frac{1}{2}} + \frac{sh_i}{2})$ on a

$$\int_{\alpha_i}^{\alpha_{i+1}} f(x)dx = \frac{hi}{2}\int_{-1}^{1}\varphi_i(s)ds,$$

et on est donc ramené à approcher k intégrales sur l'intervalle fixe $[-1,1]$.

L'approche générale consiste alors à choisir une formule de quadrature élémentaire sur l'intervalle $[-1,1]$ et à l'utiliser pour approcher chacune des intégrales de la somme (2.2).

Pour construire cette formule de quadrature élémentaire, on se donne $(l+1)$ point$\tau_0, \tau_1, ..., \tau_l$ et, pour approcher $\int_{-1}^{1}\varphi(s)ds$ on remplace φ par son polynôme d'interpolation de Lagrange $p_l \in P_l$ relativement aux point τ_j.
On a :

$$p_l(s) = \sum_{j=0}^{l}\varphi(\tau_j)l_j(s)$$

avec

$$l_j(s) = \frac{\Pi_{i\neq j}(s-\tau_i)}{\Pi_{i\neq j}(\tau_j-\tau_i)}.$$

Alors :

$$\int_{-1}^{1}\varphi(s)ds \simeq 2\sum_{j=0}^{l}\varphi(\tau_j)\omega_j$$

où

$$\omega_j = \frac{1}{2}\int_{-1}^{1} lj(s)ds$$

pour $j = 0, 1, \cdots, l$.

Finalement si on pose $\alpha_{ij} = \frac{\alpha_i+\alpha_{i+1}}{2} + \frac{h_i}{2}\tau_j$ on obtient sur l'intervalle $[\alpha_i, \alpha_{i+1}]$:

$$\int_{+\alpha_i}^{\alpha_{i+1}} f(x)dx \simeq hi\sum_{j=0}^{l} f(\alpha_{ij})\omega_j.$$

Par sommation on obtient alors la formule composée :

$$\int_{\alpha}^{\beta} f(x)dx \simeq \sum_{i=0}^{k-1}\left(hi\sum_{j=0}^{l} f(\alpha_{ij})\omega_j\right) = T_{kl}(f) \tag{2.3}$$

Exercice

Donner une borne inférieure de l'ordre de la formule composée (2.3).

□

Exemple
Formules de Newton Cotes :

Elles sont obtenues en choisissant les points τ_j équidistants suivants $\tau_j = -1 + \frac{2j}{l}(0 \leq j \leq l)$. Pour $l \geq 8$ les coefficients ω_j deviennent "'grands"' et de signes mélangés ce qui rend ces formules très sensibles aux erreurs d'arrondis. On ne les utilise donc que pour $l \leq 7$.

◇

Exercice

E1 Expliciter les formules de Newton Cotes pour $l = 1, 2, 4$

$l = 1$ Formule des trapèzes : $\int_{-1}^{1} \varphi(x)dx \simeq \varphi(-1) + \varphi(1) \qquad (\omega_0 = \omega_1 = 1/2)$

$l = 2$ Formule de Simpson : $\int_{-1}^{1} \delta(x)dx \simeq \frac{1}{3}(\delta(-1) + 4\delta(0) + \delta(1)) \qquad \left(\omega_0 = \omega_2 = \frac{1}{6}, \omega_1 = \frac{2}{3}\right)$

$l = 4$ Formule de Boole-Villarceau : $\omega_0 = \omega_4 = \frac{7}{90}; \omega_1 = \omega_3 = \frac{16}{45}; \omega_2 = \frac{2}{15}$

E2 Démontrer que l'ordre des formules de Newton Cotes est l si l est impair et $l + 1$ si l est pair.

□

Le comportement de l'erreur associé à une formule de quadrature composé est donné par le premier théorème suivant :

Théorème 2.1.2. Convergence des formules de quadrature composées
Pour une formule élémentaire fixée et si $\forall_j \in [0, l], \tau_j \in [-1, 1]$ *alors :*

Pour toute fonction f *intégrable au sens de Riemann sur* $[\alpha, \beta], T_{kl}(f)$ *tend vers* $\int_\alpha^\beta f(x)dx$ *lorsque le diamètre*

$$\begin{array}{l} \delta_k = max(h_i) \\ 0 \leq i \leq k-1 \end{array}$$

de la subdivision $(\alpha_0, \cdots \alpha_k)$ *de l'intervalle* $[\alpha, \beta]$ *tend vers* 0.

Preuve
En commutant l'ordre des sommations on a

$$T_{kl} = \sum_{j=0}^{l} \omega_j \sum_{i=0}^{k-1} f(\alpha_{ij})hi = \sum_{j=0}^{l} \omega_j I_{jk}$$

avec $I_{jk} = \sum_{i=0}^{k-1} h_i f(\alpha_{ij}), \alpha_{ij} \in [\alpha_i, \alpha_{i+1}]$.
I_{jk} est donc une somme de Riemann et $\lim_{\delta_k \to 0} I_{jk} = \int_\alpha^\beta f(x)dx$.
Donc

$$\begin{array}{ll} \lim & T_{kl} = \left(\sum_{j=0}^{l} \omega_j\right) \int_\alpha^\beta f(x)dx. \\ \delta_k \to 0 & \end{array}$$

Il suffit alors de vérifier que $\sum_{i=0}^{l} \omega_j = 1$ pour conclure.

■

Exercice

Démontrer, en utilisant que $L_n(P_k) = P_k$ si $k \leq n$, que $\sum_{i=0}^{l} \omega_j = 1$.

□

On peut se poser la question de la convergence de T_{kl} vers $\int_\alpha^\beta f(x)dx$ quand $l \to +\infty$. En général, il n'y a pas convergence même en supposant $f \in \subset^\infty [a, b]$).

2.1.1 Évaluation de l'erreur dans les méthodes d'intégration composées

:
Nous nous intéressons ici au contrôle de l'erreur $E(f) = |\int_\alpha^\beta f(x)w(x)dx - \sum_{i=0}^k \lambda_i f(x_i)|$. On a le théorème suivant :
Théorème : Étant donné une formule de quadrature sur $[\alpha, \beta]$ déduite d'une formule élémentaire d'ordre p sur $[-1, 1]$. Soit $h = \max|\alpha_i + 1 - \alpha_i|$ le pas de la segmentation correspondante.
Alors, il existe une constant C_p telle que pour tout $f \in C^{p+1}[\alpha, \beta]$:

$$|E(f)| \leq C_p \frac{(\beta - \alpha)}{p! 2^{p+2}} ||f^{(p+1)}||_\infty h^{p+1}. \tag{2.4}$$

L'erreur, pour une fonction C^{p+1} converge donc avec h vers 0, comme h^{p+1}, comme ceci est visible sur la figure 2.1.

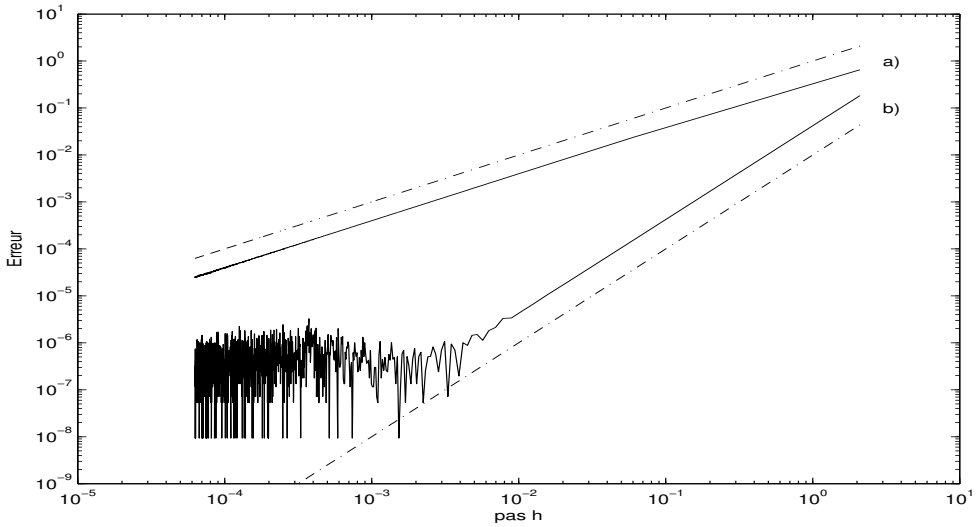


FIGURE 2.1 – *Erreur $E(f)$ en fonction de h ; $f(x) = x^4$; $[\alpha, \beta] = [0, 2\pi]$. Le pas local h_i est constant et égal à $h = \frac{2\pi}{N}$. Les deux axes sont en échelle logarithmique. a) : en trait plein erreur mesurée pour la formule des rectangles à gauche et en pointillé la courbe $y = h$. b) : en trait plein erreur mesurée pour la formule des trapèzes et en pointillé la courbe $y = h^2/100$.*

2.2 Formules de quadrature de Gauss

Ici, on appelle $(x_0, x_1, \cdots x_{l-1})$ les points de la formule de quadrature intérieurs à $]\alpha, \beta[$ et $x_l, \cdots x_k$ les points extérieurs à $]\alpha, \beta[$. On considère donc la formule de quadrature à $(k+1)$ points (l points intérieurs, $(k+1-l)$ points extérieurs) :

$$\int_\alpha^\beta f(x)w(x)dx \simeq \sum_{i=0}^k \lambda_i f(x_i)$$

Le lemme suivant répond à la question de la construction d'une formule de quadrature d'ordre maximum :

Lemme 2.2.1. *Étant donnés 2 entiers k et l vérifiant $k+1 \geq l \geq 0$ et $k+1-l$ réels fixés $x_l, x_{l+1}, \cdots, x_k$ extérieurs à $]\alpha,\beta[$, il existe une formule et une seule du type décrit plus haut qui soit d'ordre $k+l$. Cet ordre est maximum.*

Cette formule de quadrature maximise l'ordre compte tenu de la répartition (points extérieurs/points intérieurs) et du nombre de points $(k+1)$.

Preuve

Unicité : On suppose que les points et les poids d'une formule de quadrature satisfaisant le lemme existent.
Ne s'annulant pas sur $]\alpha,\beta[$, le polynôme $\prod_{i=l}^{k}(x-x_i)$ garde un signe constant sur $]\alpha,\beta[$.Il existe donc un entier ξ, valant soit $+1$ soit -1 tel que la fonction $\theta(x) = \xi w(x)\prod_{i=l}^{k}(x-x_i)$ soit strictement positive sur $]\alpha,\beta[$. C'est alors une fonction poids sur $]\alpha,\beta[$.

Soit $p_l(x) = \prod_{i=0}^{l-1}(x-x_i)$. Pour tout q avec $d^0q \leq l-1$ on a

$$\int_\alpha^\beta p_l(x)q(x)\theta(x)dx = \xi\int_\alpha^\beta \prod_{i=0}^{k}(x-x_i)q(x)w(x)dx = 0$$

car la formule de quadrature, d'ordre $k+l$, est exacte pour $p(x) = q(x)\prod_{i=0}^{k}(x-x_i)$ et que ce polynôme s'annule sur tous les points de la formule de quadrature. D'après le théorème 1.3.3 le polynôme p_l est le polynôme orthogonal de degré $k+1$ pour le poids $\theta(x)$ sur $]\alpha,\beta[$. Les points $x_0, x_1, \cdots x_l$ qui sont ses racines sont donc bien définies de manière unique. De plus, si on considère

$$l_i(x) = \frac{\prod_{j=0,j\neq i}^{k}(x-x_j)}{\prod_{j=0,j\neq i}^{k}(x_i-x_j)}$$

on a $d^0l_i = k \leq k+l$ et la formule de quadrature qui par hypothèse est exacte dans ce cas fournit $\lambda_i = \int_\alpha^\beta li(x)w(x)dx$.
Les (x_i,λ_i) sont donc uniques.

Existence : Nous allons montrer que l'unique formule de quadrature définit plus haut convient.
Soit p un polynôme de degré inférieur ou égal à k. Comme p s'écrit $p(x) = \sum_{i=0}^{k}p(x_i)l_i(x)$, on obtient

$$\int_\alpha^\beta p(x)w(x)dx = \sum_{i=0}^{k}\lambda_i p(x_i),$$

et la formule de quadrature est exacte pour tout polynôme de $d^o \leq k$.

Soit maintenant p un polynôme de $d^o \leq k+l$.On effectue sa division euclidienne par le polynôme $\Pi_{i=0}^{k}(x-x_i)$. Il existes alors 2 polynômes r et q avec $d^or \leq k$ et $d^0q \leq l-1$
tels que $p(x) = \prod_{i=0}^{k}(x-x_i)q(x) + r(x)$. Alors :

$$\begin{array}{rcl}
\int_\alpha^\beta p(x)w(x)dx & = & \xi(\int_\alpha^\beta q(x)\prod_{i=0}^{l-1}(x-x_i)\theta(x)dx) + \int_\alpha^\beta r(x)w(x)dx \\
& = & \xi(\int_\alpha^\beta q(x)p_l(x)\theta(x)dx) + \int_\alpha^\beta r(x)w(x)dx \\
& = & 0 + \sum_{i=0}^{k}\lambda_i r(xi)
\end{array}$$

car d'une part p_l est le $(l+1)$ ième polynôme orthogonal et $d_q^o \leq l-1$ et, d'autre part $d^o r \leq k$ et la formule de quadrature est exacte pour tout polynôme de degré inférieur ou égal à k.
Or, d'après la définition de q et r, $p(x_i) = r(x_i)$ donc la formule définie plus haut est exacte pour les polynômes de degré $\leq k+l$. On admettra qu'il existe un polynôme de degré $k+l+1$ pour lequel la formule est fausse ce qui permet de conclure la démonstration.

■

Remarque 2.2.2. *Il faut bien noter que le poids servant à définir les polynômes orthogonaux dont les racines fournissent les points de la formule de quadrature* **dépend des points extérieurs donnés à priori**.

Nous donnons maintenant 2 applications classiques de ce théorème.

1) Pour $k = l-1$, tous les points sont intérieurs. On a alors le résultat suivant :

Théorème 2.2.3. Formules de quadrature de Gauss *Il existe une formule d'intégration et une seule à $(k+1)$ points intérieurs d'ordre $2k+1$. Elle est obtenue en prenant pour points (x_i) les racines du $(k+2)^{ieme}$ polynôme orthogonal pour le poids $w(x)$ sur $]\alpha, \beta[$ et pour poids λ_i,*

$$\lambda_i = \int_\alpha^\beta li(x) w(x) dx.$$

De plus pour tout $i, \lambda_i > 0$

Preuve

Il s'agit d'une application directe du lemme 2.2.1. La seule chose à montrer est que $\lambda_i > 0$. Il suffit de remarquer que $l_i^2 \in P_{2k+1}$. La formule de quadrature donne alors :

$$\lambda_i = \int_\alpha^\beta li^2(x) w(x) dx > 0.$$

■

2) Pour le $k = l+1$, c'est à dire pour deux points extérieurs, avec $x_{k-1} = \alpha, x_k = \beta$ on a le résultat suivant :

Théorème 2.2.4. Formules de quadrature de Gauss Lobatto
Il existe une formule d'intégration et une seule à $(k+1)$ points dont deux extérieurs à $]\alpha, \beta[(x_{k-1} = \alpha, x_k = \beta)$ d'ordre $2k-1$. Elle est obtenue en prenant points $x_i, 0 \leq i \leq k-2$ les racines du k^{ieme} polynôme orthogonal pour le poids $\theta(x) = w(x)(x-\alpha)(\beta - x)$ et pour poids λ_i,

$$\lambda_i = \int_\alpha^\beta li(x) w(x) dx > 0.$$

Cas particuliers d'usage classique :

Les méthodes de Gauss réalisent un ordre maximal pour un nombre fixé $(k+1)$ de points d'interpolation. Cependant, la complexité des calculs des polynômes orthogonaux fait que les méthodes de Gauss sont surtout utilisées pour les polynômes de Legendre et de Chebyshev :

Voici les premières formules dites de Gauss Legendre :

$w(x) = 1, [\alpha, \beta] = [-1, 1]$

k	$\prod_{k+2}$	$x_0, x_1, \cdots, x_k$	$\lambda_0, \cdots \lambda_R$	N
0	x	0	2	1
1	$x^2 - 1/3$	$-\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}$	$1, 1$	3
2	$x^3 - 3/5x$	$-\sqrt{\frac{3}{5}}, 0, +\sqrt{\frac{3}{5}}$	$\frac{5}{9}, \frac{8}{9}, \frac{5}{9}$	5
3	$x^4 - \frac{6}{7}x^2 + 3/35$	$\pm\sqrt{\frac{3}{7} \pm \frac{2}{7}\sqrt{\frac{6}{5}}}$	$\frac{1}{2} - \frac{1}{6}\sqrt{\frac{5}{6}}, \frac{1}{2} + \frac{1}{6}\sqrt{\frac{5}{6}}$	7

Exercice Obtenir par application du théorème 2.2.3 les points et poids ci dessus et vérifier l'ordre des formules obtenues.

□

Lorsqu'on prend $k = l$ et $x_k = \beta$ (*resp.* $x_k = \alpha$) on obtient la méthode de Gauss Radau à droite (resp. à gauche) d'ordre $2k$. Les coefficients λ_i sont positifs.

Exercice

Expliciter les premières formules de Gauss Radau à droite ou à gauche.

□

Chapitre 3

Résolution numérique de systèmes linéaires

3.1 Quelques rappels d'algèbre linéaire

L'outil le plus important de l'analyse numérique est l'algèbre linéaire et la théorie matricielle. C'est bien sur l'outil essentiel pour les problèmes de calcul de solutions d'équations linéaires, de calcul de valeurs ou de vecteurs propres. Mais c'est également un outil essentiel pour résoudre de nombreux autres problèmes : équations non linéaires, équations différentielles, théorie de l'approximation puisque tôt ou tard, on finit par se ramener à un problème linéaire (ou plus souvent à une suite de problèmes linéaires).

3.1.1 Inversibilité, valeurs propres et formes canoniques

Pour tout vecteur $x \in \mathbb{C}^n$, x^T désigne le vecteur transposé $x^T = (x_1, \cdots x_n)$(vecteur ligne) et x^* désigne le vecteur adjoint, $x^* = (\bar{x}_1, \bar{x}_2, \cdots \bar{x}_n)$.
L'application : $\begin{cases} \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R} \\ (u,v) \longmapsto v^T u = u^T v = \sum_{i=1}^n u_i v_i \end{cases}$ est appelée produit scalaire euclidien.

(Le produit scalaire hermitien est défini à partir de $u.v = \overline{v.u} = \sum u_i \bar{v}_i$)

Pour toute matrice $A \in L(\mathbb{R}^m, \mathbb{R}^n)$ de terme général $[a_{ij}]$ on note A^T la matrice $[a_{ji}]$ et on a $\forall u \in \mathbb{R}^m, \forall v \in \mathbb{R}^n (Au, v) =< u, A^T v >$.
Le rang d'une matrice A est le nombre de ses vecteurs colonnes indépendants. On note pour toute matrice carrée ($m = n$), $det(A)$ le déterminant de A et A^{-1} l'inverse de A si elle existe. A est alors dite non singulière ou inversible.
On a le théorème suivant :

Théorème 3.1.1. Inversibilité des matrices
Pour $A \in L(\mathbb{C}^n)$ les propositions suivantes sont équivalentes :
(i) A est non singulière (ou inversible),
(ii) det $A \neq 0$,
(iii) le système linéaire $Ax = 0$ admet la solution $x = 0$ comme unique solution,
(iv)$\forall b \in \mathbb{C}^n), Ax = b$ admet une solution unique,
(v) Les vecteurs colonnes de A sont linéairement indépendants,

(*vi*) *la matrice est de rang n.*

La matrice A est dite :

- symétrique si A est réelle et $A^T = A$,
- hermitienne si $A^* = A$,
- orthogonale si $AA^T = A^T A = I$,
- unitaire si $AA^* = A^* A = I$,
- normale si $AA^* = A^* A$.

La trace d'une matrice A est définie par $tr(A) = \sum_{i=1}^{n} a_{ii}$.
Si $A \in L(\mathbb{C}^n)$, le scalaire λ est une valeur propres de A si il existe $x \neq 0$ tel que $Ax = \lambda x$. Le vecteur x est appelé vecteur propre de A. De même, λ est une valeur propre de A si $\det(A - \lambda I) = 0$. On appelle polynôme caractéristique de A le polynôme $P_A(z) = \det(A - zI)$.

On note $S(A)$ le spectre de A, c'est dire l'ensemble des valeurs propres de A et $\rho(A) = \max_{\lambda \in S(A)} |\lambda|$ le rayon spectral de A.

Exemple
Les matrices triangulaires, c 'est dire de la forme :
$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ 0 & \ddots & a_{nn} \end{bmatrix}$ ou $A = \begin{bmatrix} a_{11} & & \\ \vdots & \ddots & 0 \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}$
ont pour spectre l 'ensemble des éléments diagonaux.

$\diamond$

On a les résultats classiques suivant :

$$tr(A) = \sum_{\lambda_i \in S(A)} \lambda_i(A) \tag{3.1}$$

$$\det(A) = \prod_{\lambda_i \in S(A)} \lambda_i(A) \tag{3.2}$$

$$tr(A+B) = tr(A) + tr(B) \tag{3.3}$$

$$\det(AB) = \det(A)\det(B). \tag{3.4}$$

Exercice Soit P une matrice inversible et A une matrice quelconque. Calculer $\det(P^{-1})$,$\det(P^{-1}AP)$ et $S(P^{-1}AP)$.

□

Une matrice A est dite diagonalisable si il existe P inversible telle que $P^{-1}AP$ est diagonale. Les éléments diagonaux de $P^{-1}AP$ sont alors les valeurs propres de A ; la $j$ieme colonne de P est formée des composantes d'un vecteur propre associé à λ_j ; Il existe une base formée de vecteurs propres de A.

Une matrice A est positive semi définie si $x^T Ax \geq 0$, $\forall x \in \mathbb{R}^n$ et définie positive si $xAx > 0 \forall x \in \mathbb{R}^n, x \neq 0$

Deux matrices A et B sont semblables si il existe une matrice P inversible telle que $B = P^{-1}AP$.

Théorème 3.1.2. Spectre de matrices semblables
Si $(A, B) \in L^2(\mathbb{R}^n)$ sont semblables alors A et B ont les mêmes valeurs propres.

Théorème 3.1.3. Réduction des matrices symétriques
Si $A \in L(\mathbb{R}^n)$ est symétrique alors A est semblable à une matrice diagonale,i.e il existe P inversible telle que $P^{-1}AP$ est diagonale. De plus P est orthogonale.

Théorème 3.1.4. Réduction des matrices normales
Si A est normale, A est semblable à une matrice diagonale i.e il existe P inversible telle que $P^{-1}AP$ est diagonale. De plus, P est unitaire.

Théorème 3.1.5. Réduction des matrices complexes *(Shur)*
Pour toute matrice $A \in L(\mathbb{C}^n)$, il existe une matrice inversible $P \in L(\mathbb{C}^n)$ telle que $P^{-1}AP$ est une matrice triangulaire. De plus P est unitaire.

Théorème 3.1.6. Forme canonique de Jordan
Toute matrice $A \in L(\mathbb{C}^n)$ est semblable à la matrice diagonale par blocs :

$$J = \begin{bmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_m \end{bmatrix}$$

où chaque bloc J_i est soit une matrice égale à $\lambda_i I$ soit une matrice de la forme :

$$J_i = \begin{bmatrix} 1 & \lambda_i & \cdots & 0 \\ 0 & 1 & \ddots & 0 \\ 0 & 0 & \ddots & \lambda_i \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

où λ_i parcourt le spectre de A.

3.1.2 Normes vectorielles et normes matricielles

Soit V un espace vectoriel sur $\mathbb{R}$

Une norme est une application de V dans $\mathbb{R}$ qui vérifie

$.\|u\| = 0 \Leftrightarrow u = 0, \|u\| \geq 0 \forall u \in v$

$.\|\alpha u\| = |\alpha| \|u\| \forall \alpha \in \mathbb{R}, \forall u \in V$

$.\|u + v\| \leq \|u\| + \|v\|, \forall (u, v) \in V^2$

2 normes $\|.\|$ et $\|.\|'$ sont équivalentes si et seulement si il existe 2 constantes $C > 0$ et $C' < +\infty$ telles que :

$$\forall u \in V, C\,\|u\| \leq \|u\|' \leq C'\,\|u\|$$

Une norme matricielle est une norme dans $\mathbb{R}^{n^2}$ qui vérifie en plus :$\|AB\| \leq \|A\|\,.\,\|B\|$. Il s'agit d'une propriété de compatibilité avec le produit des matrices.

Norme matricielle subordonnée : Étant donnée $\|.\|$ et $\|.\|'$ 2 normes vectorielles sur $\mathbb{C}^n$

$$\|A\| = \sup_{x \neq 0} \frac{\|A_x\|'}{\|x\|}$$

est une norme matricielle appelé norme subordonnée aux normes vectorielles $\|.\|$ et $\|.\|'$. C'est bien sur un cas particulier de la définition usuelle de la norme d'une application linéaire.
Une norme subordonnée vérifie :$\|I\| = 1$

Attention : Il existe des normes matricielles non subordonnées à des normes matricielles.

Exercice

Vérifier que $\|A\| = \left(\sum_{i,j} |a_{ij}|^2\right)^{1/2}$ n'est pas une norme subordonnée.

□

Formule de Neumann : $\forall A \in L(\mathbb{C}^n)$ telle que $\rho(A) < 1$ alors $(I-A)^{-1}$ existe et $(I-A)^{-1} = \sum_{i=0}^{\infty} A^i$

$\forall A \in L(\mathbb{C}^n)$, pour toute norme $\|\|$, subordonnée ou non $\rho(A) \leq \|A\|$

$\forall A \in L(\mathbb{C}^n), \forall \varepsilon > 0 \qquad \exists\, \|.\|$ une norme sur $\mathbb{C}^n$ telle que :

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

.

3.1.3 Généralités sur l'analyse numérique matricielle

Les deux problèmes fondamentaux de l'analyse numérique matricielle sont la résolution de systèmes linéaires et le calcul de vecteurs propres et de valeurs propres.

(1) Étant donné A une matrice inversible et un vecteur b on cherche u tel que $Au = b$

(2) Étant donné A une matrice carrée, on cherche les valeurs propres de A (ou certaines d'entre elles) et éventuellement les vecteurs associés *ie* : $\lambda, p \neq 0$ tels que $(A - \lambda I)p = 0$

Il est bien clair que, la matrice A étant inversible, les **formules de Cramer** donnent explicitement la solution du problème $Au = b$. **MAIS**, comme nous le verrons dans ce qui suit ces formules ne sont pas utilisées dans la pratique car elle sont prohibitivement **coûteuses** en nombre d'opérations et donc en temps de calcul. La recherche de valeurs propres qui est équivalente au calcul des racines de polynômes est aussi coûteuse en nombre d'opérations.
Une des premières caractéristiques à considérer quand on est confronté à un problème matriciel est le nombre et la répartition des éléments nuls de A.

Une matrice creuse est une matrice possédant "beaucoup" de zéros et une matrice pleine possède pas ou "peu" de zéros.

La répartition des zéros de A donne lieu à la classification suivante :

- Une matrice bande est une matrice telle que il existe $b < n$ avec $a_{ij} = 0$ dès que $|i - j| \geq b$

Pour $b = 2$ la matrice est dite tridiagonale

$$\begin{pmatrix} . & . & & & & & \\ . & . & . & & & & \\ & . & . & . & 0 & & \\ & & . & . & . & & \\ & & & . & . & . & \\ & & 0 & . & . & . & \\ & & & & . & . & \end{pmatrix} \tag{3.5}$$

- Matrice triangulaire supérieure

$$\begin{pmatrix} . & . & . & . & . & . & . \\ & . & . & . & . & . & . \\ & & . & . & . & . & . \\ & & & . & . & . & . \\ & & 0 & & . & . & . \\ & & & & & . & . \\ & & & & & & . \end{pmatrix} \tag{3.6}$$

- Matrice de Hessenberg

$$\begin{pmatrix} . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ & . & . & . & . & . & . \\ & & . & . & . & . & . \\ & & & . & . & . & . \\ & & 0 & & . & . & . \\ & & & & & . & . \end{pmatrix} \tag{3.7}$$

- Matrice diagonale par blocs

$$\begin{pmatrix} & \cdots & & & & & & & \\ \vdots & A_{11} & \vdots & 0 & & & & & \\ & \cdots & & \cdots & & & & & \\ & 0 & \vdots & A_{22} & \vdots & 0 & & & \\ & & & \cdots & & & & & \\ & & & & & \ddots & & 0 & \\ & & & & & & & \cdots & \\ & & & & & 0 & \vdots & A_{nn} & \vdots \\ & & & & & & & \cdots & \end{pmatrix} \tag{3.8}$$

- Matrice tridiagonale par blocs

$$\begin{pmatrix} & \cdots & & \cdots & & & & & \\ \vdots & A_{11} & \vdots & A_{12} & \vdots & & & & \\ & \cdots & & \cdots & & & & & \\ \vdots & A_{21} & \vdots & A_{22} & \vdots & \ddots & & & \\ & \cdots & & \cdots & & & & \cdots & \\ & & & \ddots & & \ddots & \vdots & A_{n-1,n} & \vdots \\ & & & & & \cdots & & \cdots & \\ & & & & \vdots & A_{n,n-1} & \vdots & A_{nn} & \vdots \\ & & & & & \cdots & & \cdots & \end{pmatrix} \tag{3.9}$$

3.2 Conditionnement d'un système linéaire à matrice inversible

Exemple

Soit le système linéaire :

$$\begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \begin{pmatrix} U_1 \\ U_2 \\ U_3 \\ U_4 \end{pmatrix} = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix}.$$

Sa solution est $\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$. Si on considère le système perturbé :

$$\begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \begin{pmatrix} U'_1 \\ U'_2 \\ U'_3 \\ U'_4 \end{pmatrix} = \begin{pmatrix} 32,1 \\ 22,9 \\ 33,1 \\ 30,9 \end{pmatrix}$$

Sa solution est $\begin{pmatrix} 9,2 \\ -12,6 \\ 4,5 \\ 1,1 \end{pmatrix}$.

Une perturbation de 10^{-1} sur chaque composante du second membre entraîne des perturbations de l'ordre de la dizaine sur la solution. Si on raisonne en perturbation relative, une perturbation relative de l'ordre de $\frac{1}{200}$ème sur le second membre conduit à une perturbation relative de l'ordre de 10 sur la solution, soit une amplification des perturbations relatives de l'ordre de 2000.
Pourtant A est "sympathique" : elle est symétrique, $\det(A) = 1$, ses composantes ont toutes le même ordre de grandeur et il en est de même pour l'inverse de A qui est :

$$\begin{pmatrix} 25 & -41 & 10 & -6 \\ -41 & 68 & -17 & 10 \\ 10 & -17 & 5 & -3 \\ -6 & 10 & -3 & 2 \end{pmatrix}.$$

$\diamond$

On s'intéresse ici aux effets sur la solution u, de perturbations sur les données du problème qui sont A et b. On appelle δA et δb ces perturbations.
On considère ici que $\delta A = 0$.
Si $Ax = b$ alors il existe δx la perturbation induite par les perturbations δA et δb telle que $A(x + \delta x) = b + \delta_b$. Soit $||.||$ une norme vectorielle quelconque et $||.||$ la norme matricielle subordonnée. On a $\delta x = A^{-1}\delta_b$ d'où :

$$\left.\begin{array}{c} ||b|| \leq ||A||.||x|| \\ ||\delta x|| \leq ||A^{-1}||.||\delta_b|| \end{array}\right\} \Rightarrow \frac{||\delta x||}{||x||} \leq ||A^{-1}||.||A||.\frac{||\delta_b||}{||b||}.$$

Exercice

Obtenir une majoration du même type pour $\delta A \neq 0$, $\delta b = 0$.

□

On voit ici que la perturbation relative induite sur la solution $\frac{||\delta x||}{||x||}$ est contrôlée par la perturbation relative sur les données via un facteur multiplicatif $||A^{-1}||.||A||$ que l'on appelle conditionnement de A relativement à la norme matricielle $||.||$.

Théorème 3.2.1. Conditionnement d'une matrice inversible
Pour toute matrice A inversible on appelle conditionnement relativement à la norme $||.||$ le réel $cond(A) = ||A^{-1}||.||A||$. Il vérifie les propriétés suivantes :

1)cond $(A) \geq 1$, cond $(A) =$ cond (A^{-1}), cond $(\alpha A) =$ cond (A) $\quad ,\forall \alpha \neq 0$.
2)Si on appelle cond $_2(A)$ le conditionnement de A relativement à la norme 2 alors cond $_2(A) = \frac{\mu_n(A)}{\mu_1(A)}$ où $\mu_1 > 0$, $\quad \mu_n > 0$ désignent les plus petites et plus grandes valeurs singulières de A. (on appelle valeur singulière d'une matrice carrée A, les racines positives des valeurs propres de la matrice hermitienne A^*A))
3)Si A est normale $cond_2(A) = \frac{\lambda n(A)}{\lambda_1(A)}$ où $\lambda_1(A)$ et $\lambda_n(A)$ sont les valeurs propres de A de module minimum et maximum.
4)Le conditionnement $cond_2(A)$ d'une matrice unitaire ou orthogonale vaut 1.
5)Si U est unitaire alors $cond_2(A) = cond_2(UA) =$ cond $_2(U^*AU)$

Exercice

Vérifier numériquement que pour la matrice A de l'exemple précédent $\text{cond}_2(A) = \frac{\lambda_4}{\lambda_1} \simeq 2984$

□

3.2.1 Conditionnement d'un problème de valeurs propres :

Exemple

Soit la matrice

$$A_\varepsilon = \begin{pmatrix} 0 & 0 & \cdots & 0 & \varepsilon \\ 1 & 0 & \cdots & & 0 \\ 0 & 1 & 0 & & 0 \\ 0 & 0 & \ddots & & 0 \\ 0 & \cdots & 0 & 1 & 0 \end{pmatrix}.$$

Pour $\varepsilon = 0$, toutes les valeurs propres de A valent 0 ; pour $\varepsilon \neq 0$ le polynôme caractéristique de A vaut $(-1)^n(z^n - \varepsilon)$ et les valeurs propres de A sont donc les racines n^{iemes} de ε. Par exemple pour $n = 40$ et $\varepsilon = 10^{-40}$ on obtient $\forall_i|\lambda_i| = 10^{-1}$.Ainsi une modification de 10^{-40} sur la matrice induit une modification de 10^{-1} sur les valeurs propres, soit une amplification de 10^{39} !.
◇

On s'intéresse ici aux effets, sur le spectre de A, de perturbations sur la donnée du problème, c'est à dire la matrice A. Il n'y a pas de raison pour que le conditionnement du problème de valeur propre soit relié au conditionnement de A (ici A n'est pas obligatoirement inversible). On a en particulier le théorème suivant :

Théorème 3.2.2. Conditionnement d'un problème de valeurs propres
Soit A une matrice diagonalisable et P une matrice inversible telle que $P^{-1}AP = diag(\lambda_i)$. Soit $||.||$ une norme matricielle telle que
$||diag(\lambda_i)|| = \max_i |\lambda_i|$ *où* $S(A) = \{\lambda_i, 1 \leq i \leq m\}$.
Alors pour toute matrice δA, on a $S(A + \delta A) \subset \bigcup_{i=1}^m D_i$
où $D_i = \{z \in \mathbb{C} / |z - \lambda_i| \leq cond\ (P)||\delta A||\}$.

Exercice
Démontrer ce théorème
□

Ainsi, l'amplification des perturbations, c'est a dire le conditionnement d'un problème de valeurs propres ne dépend pas du conditionnement de la matrice du problème mais de celui des matrices de passage qui diagonalisaient la matrice du problème. On définit $\mathrm{condV}(A) = \inf\{\mathrm{cond}(P)$ telle que $P^{-1}AP = \mathrm{diag}(\lambda$ comme le conditionnement du problème du calcul des valeurs propres de la matrice A.

Exercice
Démontrer que pour toute matrice normale A, $\mathrm{condV}(A) = 1$.
□

3.3 Résolution de systèmes linéaires

Nous allons aborder successivement dans ce chapitre les trois grandes familles de résolution de systèmes linéaires qui sont les méthodes directes, les méthodes itératives et les méthodes de descente. Avant de commencer l'étude de ces méthodes il convient de se convaincre que les formule de Cramer ne sont pas utilisable dans la pratique pour estimer numériquement la solution d'un grand système linéaire.

Un petit retour sur les formules de Cramer :

On rappelle que les formules de Cramer donnent la solution du système $Ax = b$ sous la forme :

$$\forall i, 1 \leq i \leq n, x_i = \frac{\det\,(B_i)}{\det\,A} \text{ où } B_i = \begin{pmatrix} a_{11} & a_{1,i-1} & b_1 & a_{1,i+1} & a_{1n} \\ & & | & & \\ & & | & & \\ a_{n1} & & b_n & & a_{nn} \end{pmatrix}. \tag{3.10}$$

Sachant que calcul d'un déterminant entraine $n!-1$ additions et $(n-1)n!$ multiplications, le nombre d'opérations nécessaire au calcul de la solution par les formules de Cramer est :

$$\begin{cases} (n+1)! \text{ additions} \\ (n+2)! \text{ multiplications} \\ n \text{ divisions.} \end{cases}$$

La complexité des formules de Cramer est donc $O((n+1)!)$.
Exemple : pour $n = 10$ il faut donc effectuer 400.000.000 opérations pour obtenir la solution par les formules de Cramer. Pour $n = 100$ il en faut environ 10^{162} (utiliser la formule de Stirling). On se convaincra dans l'exercice suivant que ce nombre d'opération est prohibitif même pour les plus puissants ordinateurs alors que la taille considérée ($n = 100$) est de plusieurs ordres de grandeur inférieure aux besoins standards actuels.

Exercice

Évaluer le temps de calcul (CPU) nécessaire a un ordinateur ayant une capacité ce calcul de 100 Megaflops. On conseille de réfléchir à l'unité de temps convenable pour l'évaluation de ce temps.

□

3.4 Les méthodes directes de résolution de systèmes linéaires

Le principe des méthodes directes étudiées ici, revient à la détermination explicite ou pas d'une matrice M inversible et telle que la matrice MA soit triangulaire supérieure puis à la résolution directe de ce système triangulaire. L'archétype des méthodes directes est la méthode dite du pivot de Gauss.

Exercice

Calculer la complexité de l'algorithme dit de remontée pour expliciter la solution d'un système triangulaire supérieur.

□

3.4.1 Méthode du pivot de GAUSS

Elle comporte 2 étapes :
- Élimination successive des variables et modification du second membre

- Résolution par la méthode de remontée

Définition de la méthode

Étape d'élimination :

1iere étape :
- Comme $\det(A) \neq 0$, l'un au moins des éléments de la première colonne de A est diffèrent de 0. On choisit l'un de ces coefficients a_{p1} et on l'appelle le premier pivot.

- On échange la ligne du pivot (la p^{ieme} d'après nos notations) avec la première ligne.

Exercice

Vérifier que cette opération revient à multiplier à gauche l'équation $Ax = b$ par une matrice P_{1p} que l'on calculera.

□

-On annule, par des combinaisons linéaires entre lignes tous les éléments de la première colonne situés sous la diagonale.

Exercice

Vérifier que cette opération revient à multiplier à gauche l'équation $Ax = b$ par une matrice d'élimination E_1 que l'on calculera.

□

On obtient à cette étape le nouveau système $E_1P_1Ax = E_1P_1b$ et la matrice $A_1 = E_1P_1A$ a la forme :

$$A_1 = \begin{pmatrix} \alpha_{11} & \alpha_{12} & - & - & \alpha_{1n} \\ 0 & b_{22} & & & b_{2n} \\ | & | & & & \\ | & | & & & \\ | & | & & & \\ | & | & & & \\ 0 & b_{n2} & - & - & b_{nn} \end{pmatrix}. \tag{3.11}$$

(par convention on pose $A_0 = A$)
Comme $\det P_{1p} = \det E_1 = 1$ la matrice A_1 est inversible et donc l'un au moins des éléments $(A_1)_{i2}, (2 \leq i \leq n)$ est diffèrent de 0 et peut donc servir de pivot pour la deuxième étape.
2ieme étape et suivantes : on recommence les mêmes opérations en ne considérant que la sous matrice $((A_1)_{ij})2 \leq i, j \leq n$ et on effectue $(n-1)$ étapes au total.

Après la $(n-1)$ieme étape, la matrice A_{n-1} obtenue est triangulaire supérieure.
Étape de remontée
On calcule alors successivement les coordonnées $x_i, n \geq i \geq 1$ de la solution de ce système triangulaire qui coïncide avec la solution du problème $Ax = b$ initial.

Remarque 3.4.1. *Il est désavantageux numériquement d'utiliser des pivots "petits"(la matrice d'élimination E est alors mal conditionnée).*

Exercice

Établir la forme générale de la matrice d'élimination à l'étape p et calculer son conditionnement

□

Pratiquement, on utilise les stratégies suivantes :

- Pivot partiel : on choisit à la kieme étape le plus grand élément de la kieme colonne de la sous matrice A_k,
- Pivot total : On choisit à la kieme étape le plus grand élément de la sous matrice, il faut donc utiliser des permutations de colonnes pour le ramener en kieme colonne, puis, classiquement des permutations de lignes pour le ramener en kieme ligne.

Complexité de la méthode de Gauss

En utilisant les formules classiques :

$$\begin{array}{l}\sum_{k=1}^{n-1} k^2 = \frac{n(n-1)(2n-1)}{6} \\ \sum_{k=1}^{n-1} k = \frac{n(n-1)}{2}\end{array} \tag{3.12}$$

on obtient l'évaluation du nombre d'opérations relatives aux deux étapes précédentes suivante :

$$\left\{\begin{array}{l} O(\frac{n}{3}^3)\text{additions} \\ 0(\frac{n}{3}^3)\text{multiplications} \\ 0(\frac{n}{2}^2)\text{divisions}\end{array}\right. \tag{3.13}$$

On obtient de façon précise que la complexité de la méthode de Gauss est équivalente, quand $n \to +\infty$ à $\frac{2n^3}{3}$. La méthode de Gauss demande donc environ 7.10^5 opérations pour $n = 100$. Il est instructif de comparer ce nombre a celui nécessaire pour la mise en oeuvre des formules de Cramer.

3.4.2 factorisation L.U d'une matrice

Dans le cas où l'on n'utilise pas de stratégie de pivot (c'est à dire que l'on utilise systématiquement $(A_l)_{l+1,l+1}$ comme pivot alors la matrice de passage entre l'étape l et l'étape $l+1$ est la matrice triangulaires inférieure d'élimination E_{l+1} c'est à dire que l'on a $A_{l+1} = E_{l+1}A_l$. La matrice E_{l+1} est inversible et E_{l+1}^{-1} est également triangulaire inférieure.

Exercice

Expliciter la matrice E_{l+1}^{-1}.

□

Comme la matrice A_{n-1} est triangulaire supérieure on peut écrire :
$A = E_1^{-1} \quad E_1 A = E_1^{-1}E_2^{-1}A_2 = ... = (E_{n-1}...E_1)^{-1}A_{n-1} = LU$ avec $L = (\Pi_{i=1}^{n} E_{n-i})^{-1}$ et $U = A_{n-1}$. Par construction L est triangulaire inférieure et U est triangulaire supérieure.
Pour pouvoir factoriser une matrice sous la forme LU avec L une matrice triangulaire inférieure et Uune matrice triangulaire supérieure, il suffit donc que l'on puisse appliquer la méthode de Gauss sans avoir a permuter de ligne pour avoir un pivot non nul. Le théorème suivant donne des conditions suffisantes pour que ce soit le cas. Il donne aussi des conditions particulières sur L pour assurer son unicité.

Théorème 3.4.2. Factorisation L.U d'une matrice *Soit $A = (aij)$ une matrice carrée d'ordre*

n telle que les n sous matrices

$$\Delta_k = \begin{pmatrix} a_{11} & . & . & . & a_{1k} \\ . & & & & \\ . & & & & \\ . & & & & \\ a_{k1} & . & . & . & a_{kk} \end{pmatrix} \qquad 1 \leq k \leq n$$

sont inversibles. Alors il existe une unique matrice triangulaire inférieure $L = (l_{ij})$ *avec* $l_{ii} = 1, 1 \leq i \leq n$ *et une unique matrice triangulaire supérieure* U *telles que* $A = LU$.

Remarque 3.4.3. *En fait si la condition suffisante de ce théorème n'est pas vérifiée on peut toujours s'y ramener après des permutations préalables de lignes.*

L'intérêt majeur de l'existence de la factorisation $L.U$ apparaît quand on doit résoudre plusieurs systèmes linéaires associés à la même matrice A. Il suffit en effet de calculer une fois pour toutes L et U puis de résoudre ensuite à chaque fois 2 systèmes triangulaires ($Lv = b$ puis $Ux = v$), ce qui est bien moins coûteux que d'appliquer à chaque fois la méthode de Gauss.

Une application particulièrement utilisée concerne le cas des matrices tridiagonales. Un algorithme classique de factorisation LU et de résolution pour de telles matrices est l'algorithme de Thomas ([1]).

3.5 Factorisation et méthode de Cholesky :

Les matrices symétriques définies positives vérifient les hypothèses du théorème 3.4.2 ; en effet, les n sous matrices considérées dans le théorème 3.4.2 sont elles aussi symétriques définies positives et donc inversibles. Les matrices symétriques définies positives admettent donc une factorisation LU. En fait la symétrie conduit à une factorisation analogue mais ne faisant intervenir qu'une seule matrice et sa transposée. On a le théorème suivant :

Théorème 3.5.1. Factorisation de Cholesky
Si A *est une matrice symétrique définie positive, il existe au moins une matrice réelle triangulaire inférieure* B *telle que* $A = BB^T$. *Si on impose que tous les éléments diagonaux de* B *soient strictement positifs alors la factorisation correspondante est unique.*

Preuve

Il est facile de calculer les coefficients de la matrice B. On pose à priori :

$$B = \begin{pmatrix} b_{11} & & & & \\ b_{21} & b_{22} & & 0 & \\ . & & . & & \\ . & & & . & \\ . & & & & . \\ b_{n1} & & . & . & . & b_{nn} \end{pmatrix} \tag{3.14}$$

$$A = BB^T \Rightarrow aij = (BB^T)ij = \sum_{k=1}^{n} \quad b_{ik} \quad b_{jk} = \sum_{k=1}^{min(i,j)} b_{ik} \quad b_{jk} \tag{3.15}$$

Comme A est symétrique il suffit de vérifier cette relation pour $i \leq j$ $\quad ,i.e$:

$$a_{ij} = \sum_{k=1}^{i} \quad b_{ik} \quad b_{jk} \qquad 1 \leq i \leq j \leq n \tag{3.16}$$

On détermine alors b_{ij} colonne par colonne en commençant par l'élément diagonal :

- colonne 1 :

$$\begin{array}{ccccc} j=1 & a_{11} = b_{11}^2 & \longrightarrow & b_{11} = \sqrt{a_{11}} \\ j=2 & a_{12} = b_{11} \quad b_{21} & \longrightarrow & b_{21} = a_{12}/b_{11} \\ \vdots & \vdots & & \\ j=n & a_{1n} = b_1 \quad b_{n-1} & \longrightarrow & b_{n1} = a_{1n}/b_{11} \end{array} \tag{3.17}$$

- Pour la ième colonne de B. (après avoir déterminé les $(i-1)$ premières)

$$\begin{array}{ccccc} j=i: & a_{ii} = \sum_{k=1}^{i} b_{ik} b_{jk} & \longrightarrow & b_{ii} = \sqrt{a_{ii} - \sum_{i=1}^{k-1} (b_{ik})^2} \\ j=i+1 & a_{i,i+1} = \sum_{k=1}^{i} b_{ik} b_{i+1,k} & \longrightarrow & b_{i+1,i} = \frac{a_{i,i+1} - \sum_{k=1}^{i-1} b_{ik} \quad b_{i+1,k}}{b_{ii}} \\ \vdots & & & \vdots \\ j=n & a_{in} = \sum_{k=1}^{i} b_{ik} b_{nk} & \longrightarrow & b_{n,}i = \frac{a_{in} - \sum_{k=1}^{i-1} b_{ik} b_{nk}}{b_{ii}} \end{array} \tag{3.18}$$

On obtient ainsi, existence et unicité pour la matrice B.

■

La méthode de Cholesky pour résoudre $Au = b$ consiste donc à calculer la matrice B telle que $BB^T = A$ puis, comme après la factorisationLU à résoudre successivement les deux systèmes linéaires à matrice triangulaire.

Estimation du nombre d'opérations :

Le calcul de B s'effectue on $\frac{n^3}{3} + \frac{n^2}{2}$ opérations et la résolution des systèmes triangulaires demande $2n^2$ opérations, ce qui se compare favorablement à l'estimation $\frac{2n^3}{3}$ pour la méthode de Gauss.

Remarque 3.5.2. *Une fois factorisée sous la forme BB^T la matrice A a son déterminant facilement calculable sous la forme*

$$det\ A = (\prod_{i=1}^{n} b_{ii})^2$$

3.6 Méthodes itératives de résolution de systèmes linéaires

3.6.1 Généralités

La base des méthodes itératives décrites dans ce chapitre est la réécriture de l'équation $Ax = b$ sous la forme $x = Bx + c$. En effet cette forme présente la solution du système linéaire initial comme un point fixe. On pense alors immédiatement aux itérations de point fixe (théorème de Banach) qui s'écrivent :

$$\begin{cases} x_0 \text{ donné} \\ x_{k+1} = B \quad x_k + c \quad , k \geq 0 \end{cases} \tag{3.19}$$

On dit que la méthode itérative est convergente si la suite x_k converge dans $\mathbb{R}^n$ pour tout vecteur initial x_0.

Pour ce type de méthodes il y a donc deux difficultés : construire une matrice B et un vecteur c et ensuite s'assurer de la convergence des itérations.
On a tout d'abord le théorème suivant :

Théorème 3.6.1. *Les propositions suivantes sont équivalentes :*
1) La méthode itérative (3.19) est convergente
2) $\rho(B) < 1$
3) $||B|| < 1$ *pour au moins une norme matricielle*

Preuve

On introduit l'erreur à l'itération k définie par : $e_k = u - u_k$; on a alors, $e_{k+1} = Be_k$ et la convergence de la méthode équivaut à la convergence de e_k vers 0.

- 1) $\Rightarrow$ 2) On raisonne par l'absurde : si $\rho(B) \geq 1$ alors il existe une valeur propre de B, $\lambda \geq 1$ et un vecteur propre associé v. On choisit alors $u_0 = u - v$. Alors $e_k = \lambda^k v$ et e_k ne peut pas tendre vers 0.
- 2) $\Rightarrow$ 3) : On sait que pour tout $\varepsilon > 0$ il existe au moins une norme matricielle telle que $||B|| \leq \rho(B) + \varepsilon$. Pour ε assez petit on a donc $||B|| < 1$.
- 3) $\Rightarrow$ 1) : Pour la norme exhibée en 3), on a $||e_k|| \leq ||B||^k ||e_0||$ et donc la convergence.

■

Remarque 3.6.2. *La méthode itérative définie plus haut n'est autre que la méthode de Picard (méthodes des approximations successives) pour trouver un point fixe de l'application* $f : v \longmapsto Bv + c$ *qui est une contraction si la propriété 3) du théorème précédent est vérifiée.*

Remarque 3.6.3. *Pour une matrice normale,* $||B||_2 = \rho(B)$ *et la vitesse de convergence* L_2 *(décroissance de* $||e_k||_2$*)est liée à* $(\rho(B))^k$*. Plus généralement le théorème suivant renseigne sur la vitesse de convergence de la méthode itérative :*

Théorème 3.6.4. *- Soit* $||.||$ *une norme vectorielle quelconque et soit* u *tel que* $u = Bu + c$*. On étudie la méthode itérative définie par :* $u_{k+1} = Bu_k + c, k \geq 0$*. Alors :*

$$\lim_{k \to +\infty} \sup_{||u_0 - u|| = 1} ||u_k - u||^{1/k} = \rho(B) \tag{3.20}$$

- Si $u = Bu + c = \tilde{B}u + \tilde{c}$ *on peut alors comparer les deux méthodes itératives :*

$$u_{k+1} = Bu_k + c \text{ et } \tilde{u}_{k+1} = \tilde{B}\tilde{u}_k + \tilde{c} \tag{3.21}$$

On suppose que $\rho(B) < \rho(\tilde{B})$ *et que* $u_0 = \tilde{u}_0$
Alors, quelque soit $\varepsilon > 0$*, il existe* $l(\varepsilon)$ *tel que :*

$$k \geq l \Rightarrow \sup ||U_0 - U|| = 1 \left\{ \frac{||\tilde{U}_k - U||}{||U_k - U||} \right\}^{1/k} \geq \frac{\rho(\tilde{B})}{\rho(B) + \varepsilon} \tag{3.22}$$

L'étude des méthodes itératives consiste à répondre aux deux problèmes suivant :

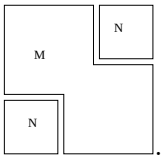
1) Est ce que la méthode itérative de matrice B converge ?
2) Si on dispose de plusieurs méthodes itératives, laquelle choisir ?
Les théorèmes 3.6.1 et 3.6.4 donnent des réponses théoriques mais ils sont souvent difficiles à appliquer dans la pratique.

3.7 Méthodes de Jacobi, Gauss Seidel et de relaxation

Ces trois méthodes sont des cas particuliers de l'approche suivante : Étant donné le système $Au = b$, supposons que A se décompose sous la forme $A = M - N$ où M est inversible et "facile" à inverser. On a $Au = b \Leftrightarrow u = M^{-1}Nu + M^{-1}b$ qui est de la forme $u = Bu + c$ et on introduit donc la méthode itérative $u_{k+1} = Bu_k + c$ avec $B = M^{-1}N$ et $C = M^{-1}b$. On est plus précisemment conduit à inverser une cascade de systèmes linéaires du type $Mu_{k+1} = Nu_k + b \qquad k \geq 0$.

Les méthodes de Jacobi et de Gauss Seidel correspondent à des matrices M et N "disjointes" ($M_{ij} = 0$ si $N_{ij} \neq 0$) ce qui n'est pas le cas de la méthode de relaxation.

Exemple de décomposition en matrices "disjointes" :

$A =$ 

.

3.7.1 Méthode de Jacobi :

Soit A une matrice inversible telle que que $a_{ii} \neq 0 \qquad 1 \leq i \leq n$. On écrit A sous la forme : A=D-E-F (décomposition par points de la matrice A)s avec :

$D_{ij} = a_{ij}\delta_{ij}$
$E_{ij} = -a_{ij}$ si $i > j$,0 sinon
$F_{ij} = -a_{ij}$ si $i < j$,0 sinon

$$A = \begin{pmatrix} a_{11} & & & \\ & & -F & \\ & D & & \\ -E & & & \\ & & & a_{nn} \end{pmatrix} \tag{3.23}$$

D est inversible par hypothèse et on construits alors la suite :

$$Du_{k+1} = (E + F)u_k + b \tag{3.24}$$

La matrice d'itération de cette méthode vaut $D^{-1}(E + F) = D^{-1}(D - A) = I - D^{-1}A$ et est dite matrice de Jacobi. Les calculs effectifs d'une itération de la méthode de Jacobi se présentent sous la forme :

$$\begin{pmatrix} a_{11}u_1^{(k+1)} \\ a_{22}u_2^{(k+1)} \\ \vdots \\ a_{nn}u_n^{(k+1)} \end{pmatrix} = \begin{pmatrix} 0 & -a_{12}u_2^{(k)} & \cdots & \cdots & \cdots & -a_{1n}u_n^{(k)} \\ -a_{21}u_1^{(k)} & 0 & -a_{23}u_3^{(k)} & \cdots & \cdots & -a_{2n}u_n^{(k)} \\ & & & & & \\ -a_{n1}u_1^{(k)} & -a_{n2}u_2^{(k)} & \cdots & \cdots & -a_{n,n-1}u_{n-1}^{(k)} & 0 \end{pmatrix} + b. \tag{3.25}$$

3.7.2 Méthode de Gauss-Seidel

Dans la méthode précédente il semble que l'on pourrait améliorer la méthode en utilisant "mieux" les quantités déjà calculées à l'étape $(k+1)$. Par exemple pour calculer $u_2^{(k+1)}$ il semblerait judicieux d'utiliser $u_1^{(k+1)}$ au lieu de $u_1^{(k)}$, de même pour $u_3^{(k+1)}$, ... On remplace alors le système précédent par :

$$\begin{matrix} a_{11}u_1^{(k+1)} \\ a_{22}u_2^{(k+1)} \\ \vdots \\ a_{nn}u_n^{(k+1)} \end{matrix} = \begin{pmatrix} 0 & -a_{12}u_2^{(k)} & \cdots & \cdots & \cdots & -a_{1n}u_n^{(k)} \\ -a_{21}u_1^{(k+1)} & 0 & -a_{23}u_3^{(k)} & \cdots & \cdots & -a_{2n}u_n^{(k)} \\ & & & & & \\ -a_{n1}u_1^{(k+1)} & \cdots & \cdots & \cdots & -a_{n,n-1}u_{n-1}^{(k+1)} & 0 \end{pmatrix} + b, \tag{3.26}$$

ce qui revient à écrire :

$$Du_{k+1} = Eu_{k+1} + Fu_k + b \text{ soit, } u_{k+1} = (D-E)^{-1}Fu_k + (D-E)^{-1}b \tag{3.27}$$

Cette méthode est la méthode itérative de Gauss-Seidel. $(D-E)^{-1}F$ est la matrice de la méthode itérative de Gauss-Seidel.

Les besoins en mémoire de la méthode de Gauss-Seidel sont nettement plus faibles que ceux de la méthode de Jacobi (avantage déterminant pour les "grands systèmes").

3.7.3 Méthode de relaxation

Une itération de la méthode de Gauss-Seidel s'écrit :

$$(D-E)\,u_{k+1} = Fu_k + b \tag{3.28}$$

On souhaiterait perturber un petit peu cette méthode afin de réduire, si possible le rayon spectral de la matrice d'itération. On choisit $\left(\frac{D}{\omega} - E\right)$ comme matrice d'itération $(\omega \neq 0)$ et on écrit alors :

$$A = D - E - F = \left(\frac{D}{\omega} - E\right) - \left(\frac{1-\omega}{\omega}\right)D - F \tag{3.29}$$

On est conduit à la récurrence :

$$\left(\frac{D}{\omega} - E\right)u_{k+1} = \left(\frac{1-\omega}{\omega}D + F\right)u_k + b \qquad k \geq 0 \tag{3.30}$$

C'est la méthode de relaxation.

La matrice d'itération de la méthode de relaxation s'écrit :

$$\mathcal{L}_\omega = \left(\frac{D}{\omega} - E\right)^{-1}\left(\frac{1-\omega}{\omega}D + F\right) = (D - \omega E)^{-1}\left((1-\omega)D + \omega F\right) \tag{3.31}$$

Pour $\omega > 1$ on dit que la méthode est une méthode de sur-relaxation. Elle est dite de sous-relaxation pour $\omega < 1$.

Remarque 3.7.1. *Une décomposition par bloc du type $A = D - E - F$ est également possible (par opposition à une décomposition par points déjà utilisée). Elle conduit aux méthodes de Jacobi, Gauss-Seidel et de relaxation par blocs.*

Exemple de décomposition par blocs, A_{ii}, E et F sont des matrices "disjointes".

$$A = \begin{pmatrix} A_{11} & & & -F \\ & A_{22} & & \\ & & A_{33} & \\ -E & & & A_{44} \end{pmatrix}. \tag{3.32}$$

3.7.4 Convergence des méthodes de Jacobi, de Gauss-Seidel et de relaxation

Plusieurs résultats sont classiques et concernent le rayon spectral de la matrice d'itération de ces méthodes, donc la convergence des méthodes itératives. On cite un seul résultat que l'on ne démontre pas :

Théorème 3.7.2. *Condition générale suffisante de convergence*
Soit A une matrice hermitienne définie positive décomposée sous la forme $A = M - N$ avec M inversible. Si $(M^* + N)$ est définie positive alors $\rho(M^{-1}N) < 1$

En particulier la méthode de Jacobi converge pour une matrice hermitienne définie positive si $2D - A$ est définie positive. Par application à la méthode de relaxation on a :

Théorème 3.7.3. *(Condition suffisante de convergence de la méthode de relaxation)*
Si A est hermitienne définie positive, la méthode de relaxation par points (ou par bloc) converge pour $0 < \omega < 2$.

On a le théorème suivant sur le choix optimal du facteur de relaxation :

Théorème 3.7.4. *Soit A une matrice hermitienne, définie positive et tridiagonale par blocs. Alors les méthodes de Jacobi, Gauss-Seidel et Relaxation convergent pour $0 < \omega < 2$. (On a donc $\rho(J) < 1$ si J est la matrice d'itération de Jacobi). Il existe un unique paramètre de relaxation optimal :*

$$\omega_0 = \frac{2}{1 + \sqrt{1 - (\rho(J))^2}} \tag{3.33}$$

3.8 Méthodes de descente, méthodes du gradient et du gradient conjugué

La présentation faite ici suppose que A est une matrice symétrique définie positive (S.D.P.) mais de nombreuses autres versions de l'algorithme sont disponibles dans la littérature. (Voir par exemple le livre de Théodor et Lascaux [4]).

3.8.1 Introduction

Soient $A \in I\!R^n \times I\!R^n$ une matrice S.D.P et $b \in I\!R^n$. On considère le problème suivant :

$$(P_1)\text{Trouver } x \in I\!R^n \text{ tel que } Ax = b.$$

Si on introduit $J(u) = \frac{1}{2}(Au, u) - (b, u)$, on a le résultat classique suivant :

Théorème 3.8.1. *1) J est strictement convexe sur $I\!R^n$, c'est à dire que pour tous $u, v \in I\!R^n, 0 \leq \theta \leq 1$,*

$$J(\theta u + (1-\theta)v) < \theta J(u) + (1-\theta)J(v) \tag{3.34}$$

2) J admet un minimum unique $u \in I\!R^n$,

3) $J'(u) = 0$, où J' désigne le gradient de J au point u,

4) u est la solution unique du problème (P_1).

Le problème (P_1) est donc équivalent à celui de la recherche du minimum de J

$$(P_2)\begin{cases} \text{trouver } x \in I\!R^n \text{ tel que} \\ J(x) < J(y), \forall y \in I\!R^n, y \neq x \end{cases} \tag{3.35}$$

3.8.2 Méthodes de descente

Pour minimiser la fonctionnelle J on emploie une méthode itérative qui s'écrit :

$$x_{k+1} = x_k + \alpha_k p_k \tag{3.36}$$

où $\begin{cases} p_k \text{ est un vecteur (la direction de descente à l'étape } k) \\ \alpha_k \text{ est un scalaire (le pas de la descente à l'étape } k). \end{cases}$
La difficulté est de choisir α_k et p_k à chaque itération (ou pas de descente) afin de diminuer à chaque pas la valeur de J.
On note $r_k = b - Ax_k = -\nabla J(x_k)$ le résidu au point x_k.

Méthodes à pas optimal

On appelle méthode à pas optimal une méthode de descente où, à chaque pas on choisit la valeur de α_k qui minimise J dans la direction p_k.Il existe une expression explicite pour cette valeur dite pas optimal :

On vérifie tout d'abord que minimiser $J(x) = \frac{1}{2} < Ax, x > - < b, x >$ revient à minimiser $E(x)$ avec $E(x) = ||x-u||_A^2 =< x-u, A(x-u) >$ où u est la solution de $Au = b$.

En effet

$$\begin{aligned} E(x) &= < x, Ax > - < x, Au > - < u, Ax > + < u, Au > \\ &= 2J(x) + 2 < b, x > - < x, b > - < b, x > + < u, Au > \\ &= 2J(x) + ||u||_A^2. \end{aligned}$$

Le pas optimal α_k dans la direction p_k est l'argument qui minimise $J(x_k + \alpha p_k)$ ou, de façon équivalente $E(x_k + \alpha p_k)$.

Or

$$\begin{aligned} E(x_k + \alpha p_k) &= < x_k + \alpha_k p_k - u, A(x_k + \alpha_k p_k - u) > \\ &= < (x_k - u) + \alpha_k p_k, A((x_k - u) + \alpha_k p_k) > \\ &= < x_k - u, A(x_k - u) > + 2\alpha_k (< p_k, A(x_k - u) >) + \alpha_k^2 < Ap_k, p_k > \\ &= E(x_k) - 2\alpha_k < p_k, r_k > + \alpha_k^2 < Ap_k, p_k > . \end{aligned}$$

Comme $< Ap_k, p_k > \neq 0$, le minimum est atteint pour :

$$\alpha_k = \frac{< r_k, p_k >}{< Ap_k, p_k >} \tag{3.37}$$

Exercice

Vérifier que $< r_{k+1}, p_k >= 0$.

□

La convergence des méthodes de descente à pas optimal est assurée par le théorème suivant :

Théorème 3.8.2. *Pour toute suite x_k avec α_k optimal et toute direction p_k telle que :*

$$\exists \mu / \forall_k \left\langle \frac{r_k}{||r_k||}, \frac{p_k}{||p_k||} \right\rangle \geq \mu > 0, \tag{3.38}$$

la suite x_k converge vers la solution u.

Méthodes du gradient

On appelle méthode du gradient une méthode de descente où la direction de descente p_k est le gradient r_k. Suivant le choix de α_k une telle méthode peut être à pas optimal ou pas.

- La méthode du gradient à pas optimal s'écrit :

$$x_{k+1} = x_k + \alpha_k r_k \text{ et } \alpha_k = \frac{||r_k||^2}{< Ar_k, r_k >} \tag{3.39}$$

D'après le théorème 3.8.2, la méthode de gradient à pas optimal converge. On démontre que la vitesse de convergence de la méthode du gradient à pas optimal est $\left(\frac{K(A)-1}{K(A)+1}\right)$ c'est à dire que $E(x_k) \leq E(x_0) \left(\frac{K(A)-1}{K(A)+1}\right)^{2k}$.

On rappelle que $K(A)$ est le conditionnement 2 de A (ici $K(A) = \frac{\lambda_{max}}{\lambda_{min}}$ car A est symétrique) et que $E(x) = (A(x-u), x-u)$.

Remarque 3.8.3. *Si $K(A) = 1$, toutes les valeurs propres de A sont égales. On a $A = \lambda I$ et $E(x) = \lambda ||x - \tilde{x}||_A^2$. L'équation $E(x) = cste$ est l'équation d'une sphère et la méthode du gradient à pas optimal converge en une itération.*

- La méthode du gradient à pas constant (dite aussi méthode de Richardson) s'écrit :

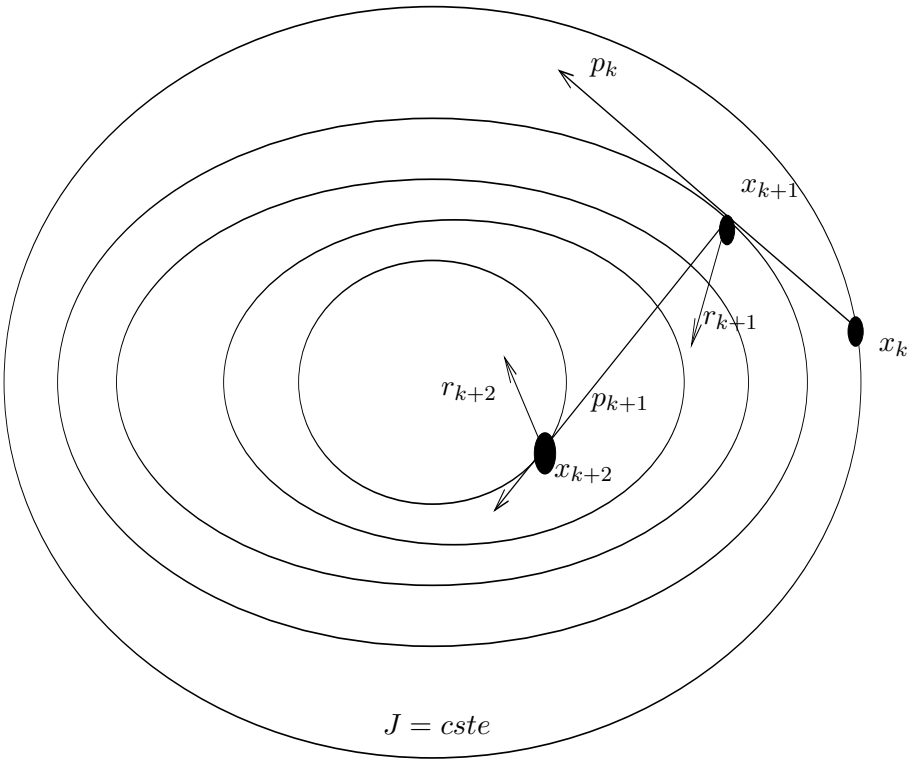


FIGURE 3.1 – Méthode de descente à pas optimal

$x_{k+1} = x_k + \alpha(b - Ax_k) = x_k + \alpha A(\tilde{x} - x_k)$

et $e_k = x_k - \tilde{x}$ vérifie $e_{k+1} = e_k - \alpha A e_k = (I - \alpha A)e_k$

Soit $e_k = (I - \alpha A)^k e_0$ et il y a convergence si $\rho(I - \alpha A) < 1$.

Le choix optimal de α, qui minimise $\rho(I - \alpha A)$ est $\alpha = \frac{2}{\lambda_{min} + \lambda_{max}}$

Méthodes du gradient conjugué

On souhaite utiliser de nouvelles directions de descente p_k avec α_k minimum local.
Comme on a alors $< p_{k-1}, r_k >= 0$, On choisit p_k dans le plan formé par les deux directions orthogonales p_{k-1} et r_k.

On écrit donc $p_k = r_k + \beta_k p_{k-1}$ et on détermine β_k pour que le facteur de réduction soit maximum, étant donné que le coefficient α_k est toujours optimal.
On obtient :

$$\beta_k = -\frac{< Ap_{k-1}, r_k >}{< Ap_{k-1}, p_{k-1} >} \tag{3.40}$$

On a alors $< Ap_{k-1}, p_k >=< Ap_{k-1}, r_k + \beta_k p_{k-1} >=< Ap_{k-1}, r_k > - < Ap_{k-1}, r_k >= 0$

On dit que p_k et p_{k-1} sont A conjugués, *i.e* ils sont orthogonaux pour le produit scalaire défini par la matrice définie positive A.

On obtient(exercice) : $< r_{k+1}, r_k >= 0$ et $\beta_k = \frac{||r_k||^2}{||r_{k-1}||}$.

L'algorithme est alors :
$$\begin{cases} \alpha_k = \frac{||r_k||^2}{<Ap_k, p_k>} \\ x_{k+1} = x_k + \alpha_k p_k \\ r_{k+1} = r_k - \alpha_k A p_k \\ \beta_{k+1} = \frac{||r_{k+1}||^2}{||r_k||} \\ p_{k+1} = r_{k+1} + \beta_{k+1} P_k \end{cases}$$

On démontre par récurrence que :

$$\begin{array}{l} < p_k, Ap_l >= 0 \qquad \forall \quad k \neq l \\ < r_k, r_l >= 0 \qquad \forall \quad k \neq l \\ < r_k, p_l >= 0 \qquad 0 \leq l < k \\ < r_k, p_k >=< r_k, r_k > \\ \alpha_k = \frac{<r_k, r_k>}{<Ap_k, p_k>} \\ \beta_{k+1} = \frac{<r_{k+1}, r_{k+1}>}{<r_k, r_k>} \end{array} \tag{3.41}$$

On peut alors voir l'algorithme de gradient conjugué comme un algorithme de minimisation qui fonctionne de la façon suivante :

A partir de $x_0 \in I\!R^n$ on construit de manière itérative une base $p_0, p_1, ...p_{n-1}$ de $I\!R^n$ orthogonale pour le produit scalaire induit par A.

On réalise à chaque étape le minimum de J sur le sous espace $(p_0, p_1, ...p_{k-1})$ et non uniquement à l'itération correspondant à l'étape dans un plan particulier.
Initialisation :

$$\begin{cases} x_0 \in I\!R^n \text{ quelconque} \\ r_0 = b - Ax_0 \\ p_0 = r_0 \end{cases}$$

Itérations :

$k = 0, 1, ...$

1) Minimisation de J sur $x_0, p_0, ...p_k$

Calcul de la nouvelle direction de descente...

$p_{k+1} = r_{k+1} + \beta_{k+1}...$

avec$\beta_{k+1} = \frac{-<r_{k+1}, Ap_k>}{<Ap_k, p_k>}$

On a alors le théorème :

Théorème 3.8.4. *L'algorithme du gradient conjugué converge dans $I\!R^n$ en au plus n itérations vers la solution.*

Preuve
A chaque itération x_k réalise le minimum de J sur le sous espace E_k avec $dim E_k = k$ et pour

$k = n, E_k = I\!R^n$ ∎

Attention : Cette estimation est théorique. Pratiquement à cause des erreurs d'arrondis la convergence a lieu en plus n itérations, d'autant plus nombreuses que $K(A)$ est grand.

Évaluation du coût de calcul :
A chaque itération le coût est de $2nn_z$ où n_z est le nombre de coefficients non réels de la matrice A.

Vitesse de convergence :
La vitesse de convergence dépend de $\left(\frac{\sqrt{K(A)}-1}{\sqrt{K(A)}+1}\right)$.
En effet,

$$E(x_k) \leq 4. \left(\frac{\sqrt{K(A)}-1}{\sqrt{K(A)}+1}\right)^{2k} E(x_0)$$

où $K(A) = \text{cond}_2(A)$

Pour l'algorithme de gradient conjugué comme pour les autres algorithmes itératifs on a intérêt à utiliser des matrices ayant un conditionnement proche de 1. Pour toute matrice C régulière, $C^{-1}Ax = C^{-1}b$ est un système linéaire algébriquement équivalent au système $Ax = b$ mais un choix convenable de C peut conduire à un conditionnement de $(C^{-1}A)$ plus petit que celui de A. Si on souhaite utiliser la méthode de gradient il faut que $C^{-1}A$ soit S.D.P. Comment choisir C ?

On a en fait des méthodes itératives classiques qui à partir de $A = M - N$ proposent de calculer $x_{k+1} = M^{-1}Nx_k + M^{-1}b$. Pour la méthode $SSOR$ par exemple avec $\omega = 1$ on a :
$M = (D + L) \quad D^{-1}(D + L)^t$. C'est une matrice factorisée sous la forme de Cholesky (LDL^t) et cette factorisation est une factorisation approchée de la matrice A.

On choisit donc comme matrice C, soit une des matrices M des méthodes itératives, soit des factorisations incomplètes de A de type Cholesky. On parle alors de l'algorithme de gradient conjugué pré-conditionné.

(On parle de factorisation incomplète de Cholesky car si A est multibandes, L à des coefficients dont le module diminue quand on s'éloigne des bandes principales de A).

Exemple de pré-conditionnement :

On part de $Ax = b$ et on pré-conditionne A par $C = D$, la diagonale de A.
On applique ensuite la méthode de Richardson (méthode de gradient à pas fixe $\alpha_k = 1$)
$x_{k+1} = x_k + r_k \quad$ avec $r_k = D^{-1}b - D^{-1}Ax_k$

Soit $Dx_{k+1} = Dx_k + (b - Ax_k)$
C'est à dire $Dx_{k+1} = (D - A)x_k + b \qquad$ si on écrit $A = D - E - F$ on obtient :
$Dx_{k+1} = (E + F)x_k + b$, c'est la méthode de Jacobi.

(La méthode $x_{k+1} = x_k + \alpha r_k$ avec pré-conditionnement D, donne la méthode de Jacobi relaxée).

Remarque 3.8.5. *: - La méthode du gradient conjugué a de multiples extensions au cas non symétrique (méthode de l'équation normale,du résidu minimal etc...) - Les méthodes de factorisation incomplètes se généralisent pour le pré-conditionnement de problèmes non symétriques et*

sont basés sur les factorisation LU (au lieu de LL^t).

Exemples : - Méthode de l'équation normale. $Ax = b$ remplacée par $A^tAx = A^tb$ et (A^tA) est symétrique définie positive. (Son principal inconvénient vient du fait que $K_2(A^tA) = K_2^2(A)$ et comme $K_2(A) \geq 1$ la vitesse de convergence est plus faible qui si on utilisait A directement ; en plus on est obligé de faire 2 multiplications matrice vecteur au lieu d'une).
- Méthode orthomin : Le principe est de minimiser $\|r_k\|^2$ ou r_k est le résidu à chaque étape.

Chapitre 4

Méthodes de différences finies pour les équations différentielles

La première partie de ce chapitre présente les principaux résultats théoriques concernant les équations différentielles d'ordre 1, à savoir les théorèmes d'existence et d'unicité, de dépendance continue et de régularité. L'obtention d'une solution exacte analytique étant souvent impossible, on a recours en pratique à des méthodes numériques permettant d'approcher cette solution. L'objet de la seconde partie est la présentation et l'étude de ces méthodes.

4.1 Quelques résultats théoriques

On s'intéresse dans toute cette partie au problème suivant, dit *problème de Cauchy :*
Étant donné f une fonction de $\mathbb{R} \times \mathbb{R}^n$ dans $\mathbb{R}^n, t_0 \in \mathbb{R}$ et $y_0 \in \mathbb{R}$

Trouver y telle que : $\left\{ \begin{array}{l} y(t_0) = y_0 \\ y'(t) = f(t,y), \quad t > t_0 \end{array} \right.$

Nous commençons par donner les principaux théorèmes d'existence.

Théorème 4.1.1. *théorème de Cauchy Péano :*
Si f est continue au voisinage de $(t_0, y_0) \in I_0 \times \mathbb{R}^n$, alors il existe un voisinage J_0 de t_0 et une fonction $y \in C^1(J_0)$ tels que :

$$\left\{ \begin{array}{ll} \forall t \in J_0 & y'(t) = f(t,y) \\ y(t_0) = y_0 & \end{array} \right. \tag{4.1}$$

On appelle :
- *Solution locale* : la donnée du couple (I, y) où I est un voisinage de t_0 dans I_0 et où $y \in C^1(I)$ telle que

$$\left\{ \begin{array}{ll} y(t_0) = y_0 & \\ \forall t \in & y'(t) = f(t,y) \end{array} \right. \tag{4.2}$$

-*Prolongement* : La solution locale (J, z) prolonge (I, y) si $I \subset J$ et $\forall t \in I \quad y(t) = z(t)$. Si $I \neq J$ le prolongement est dit strict.

- *Solution maximale* : (I, y) est maximale si il n'existe pas de prolongement strict.

- *Solution globale* : $I = I_0$

Exemples :

$$\cdot \begin{cases} y'(t) = -2ty^2 & t \in I\!R \\ y(0) = 1 \end{cases} \quad \text{a une solution globale } \left\{ I\!R, \quad y(t) = \frac{1}{1+t^2} \right\} \tag{4.3}$$

$$\cdot \begin{cases} y'(t) = 2ty^2 & t \in I\!R \\ y(0) = 1 \end{cases} \quad \text{a une solution maximale } \left\{]-1, 1[, \quad y(t) = \frac{1}{1-t^2} \right\} \tag{4.4}$$

et il n'y a pas de solution globale .

Les résultats suivants concernent l'unicité de la solution.

Définition 4.1.2. *On dira que le problème*

$$\begin{cases} y(0) = y_0 \\ \forall t \in I_0, \quad y'(t) = f(t, y) \end{cases}$$

admet une solution et une seule s'il admet une solution globale sur $I = I_0$ et si toute solution locale est la restriction de cette solution globale.

Théorème 4.1.3. *On suppose $I_0 = [t_0, t_0 + h]$ et f sur $I_0 \times I\!R^n$.*
S'il existe $l \in \mathcal{L}^1(I_0)$ (fonction intégrable sur I_0) telle que :

$$\forall t \in I_0, \quad \forall y, z \in I\!R^n \qquad < f(t, y) - f(t, z), y - z > \leq l(t)|y - z|^2 \tag{4.5}$$

Alors le problème $\begin{cases} y' = f(t, y) \forall t \in I_0 \\ y(0) = y_0 \end{cases}$
admet une solution et une seule.

Preuve

- Nous admettrons que l'existence locale (Théorème 1) donne une solution globale sur I_0
- Pour démontrer l'unicité on prend (I, z) une solution locale et (I_0, y) une solution globale. On pose

$$L(t) = \int_{t0}^{t} l(s)ds \text{ et } \varphi(t) = e^{-2L(t)} \quad |y(t) - z(t)|^2 \qquad \forall t \in I \tag{4.6}$$

Alors

$$\forall t \in I \quad \begin{array}{ll} \varphi'(t) & = e^{-2L(t)} \left(\frac{d}{dt}|y(t) - z(t)|^2 - |y(t) - z(t)|^2 2l(t) \right) \\ & = 2e^{-2L(t)} \left(< y'(t) - z'(t), y(t) - z(t) > -l(t)|y(t) - z(t)|^2 \right) \end{array} \tag{4.7}$$

et

$$\varphi'(t) = e^{-2L(t)} \left\{ < f(t, y(t)) - f(t, z(t)), y(t) - z(t) > -l(t)|y(t) - z(t)|^2 \right\} \tag{4.8}$$

et d'après l'hypothèse sur f
$\varphi'(t) \leq 0$ donc $\forall t \quad 0 \leq \varphi(t) \leq \varphi(t_0) = 0$ et $\varphi(t) = 0$ d'où $y(t) = z(t)$ sur I. ∎

Nous pouvons obtenir le corollaire suivant :

Théorème 4.1.4. *Théorème de Cauchy Lipschitz :*
On suppose que la fonction f est continue sur $I_0 \times \mathbb{R}^n$ et qu'il existe L tel que :

$$\forall (t,y),(t,z) \in I_0 \times \mathbb{R}^n \qquad |f(t,y) - f(t,z)| \leq L|y-z| \tag{4.9}$$

alors le problème

$$\begin{cases} y' = f(t,y) & t \in I_0 \\ y(0) = y_0 \end{cases} \tag{4.10}$$

admet une solution unique.

Preuve
C'est une conséquence immédiate du théorème précédent :

$$|< f(t,y) - f(t,z), y - z >| \leq L|y-z|^2 \text{ et } l(t) = L \text{ et } l \in \mathcal{L}^1_{I_0} \text{ pour } I_0 \text{ borné .} \tag{4.11}$$

■

On peut donner une autre version intéressante :

Théorème 4.1.5. *Théorème de Cauchy Lipschitz :*
Soit le problème de Cauchy

$$\begin{cases} y' = f(t,y) & t \in [t_1,t_2] \\ y(t_0) = y_0 & y(t) \in \mathbb{R}^n \end{cases} f \begin{cases} \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n \\ (t,y) \longmapsto f(t,y) \end{cases} \tag{4.12}$$

Si f est continue de $[t_1,t_2] \times \mathbb{R}^n$ dans $\mathbb{R}^n$, et si f est uniformément Lipschitz par rapport à u :

$$\forall t \in [t_1,t_2], \quad \forall (u,v) \in \mathbb{R}^n, \qquad |f(t,u) - f(t,v)| \leq L|u-v| \tag{4.13}$$

($|.|$ désigne une norme qq de $\mathbb{R}^n$)
Alors

$$\forall t_0 \in [t_1,t_2], \quad \forall y_0 \in \mathbb{R}^n, \exists! y \in \mathcal{C}\left([t_1,t_2];\mathbb{R}^n\right) \tag{4.14}$$

qui satisfait le problème de Cauchy.

4.1.1 Exemple des systèmes linéaires à coefficients constants

Soit A une matrice $A \in M_n(\mathbb{K})$ avec $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$.
On étudie $\begin{cases} U'(t) = AU(t) \\ U(t_0) = u_0 \end{cases}$
On a $f(t,U) = AU(t)$ et$|f(t,U)| \leq ||A||.|U(t)|$.
$L = ||A||$ convient ($||A||$ est la norme de A
subordonnée à la norme $|.|$ de $\mathbb{K}^n$). Nous allons maintenant expliciter la solution.

Pour cela on considère l'équation différentielle $\begin{cases} M' = AM \text{ où } M(t) \in M_n(\mathbb{K}) \\ M(0) = I \end{cases}$

On a $f(t,M) = AM$ et comme $||A.M|| \leq ||A||.||M||$ (propriété spécifique des normes matricielles), on a encore existence et unicité de la solution M.

Étudions $U(t) = M(t-t_0)U_0$. On a :

$$\begin{cases} U'(t) = AM(t-t_0)U_0 = AU(t) \\ U(t_0) = U_0 \end{cases} \tag{4.15}$$

Donc $M(t-t_0)U_0$ est la solution du problème initial.
Donnons quelques propriétés de $M(t)$:

1) $M(t)$ commute avec A.
En effet soit $B(t) = AM(t) - M(t)A$.
On a :$\begin{cases} U'(t) = AM(t-t_0)U_0 = AU(t) \\ U(t_0) = U_O \end{cases}$ Comme $B(t) = 0$ est solution et que celle ci est unique on $B(t) = 0$.

2)$M(t+s) = M(t)M(s)$
En effet soit $S(t) = M(t+s) - M(t)M(s)$
$\begin{cases} S'(t) = AM(t+s) - AM(t)M(s) = AS \\ S(0) = 0 \end{cases}$
donc $S = 0$.

3)$M(t) = \sum_{n=0}^{\infty} \frac{A^n t^n}{n!}$
En effet $\frac{dM^{(p)}}{dt^{(p)}} = A^p.M$, et comme $||A^p|| \leq ||A||^p$ et que la série de terme général $\frac{||A||^p}{p!}t^p$ converge, on a bien l'égalité.
On note $M(t) = e^{At}$.
On vient en particulier de montrer que $e^{A(t+s)} = e^{At}e^{As}$ et $e^{At}A = Ae^{At}$.
Quelques autres propriétés :

4) Si A et B commutent $e^{(A+B)t} = e^{At}e^{Bt} = e^{Bt}e^{At}$

$e^{(A+B)t} = \sum \frac{(A+B)^p}{p!}t^p$ on applique la formule du binôme puisque A et B commutent.

$$\begin{aligned} e^{(A+B)t} &= \sum_p \frac{1}{p!} \sum_n C_p^n A^n B^{p-n} t^p = \sum_{n,p} \frac{A^n}{n!}\frac{B^{p-n}}{(p-n)!}t^p \\ &= \left(\sum \frac{A^n t^n}{n!}\right)\left(\sum \frac{B^n t^n}{n!}\right) \\ &= e^{At}.e^{Bt} \end{aligned}$$

5) Si A est hermitienne, e^{At} est hermitienne définie positive.
- En effet on a $M' = AM = MA$, soit en passant à l'adjoint $\begin{cases} M'^* = AM^* \\ M^*(0) = I \end{cases}$

Donc M^* est solution et $M^* = M$

- Soit x un vecteur propre de A correspondant à la valeur propre λ. On a

$$\begin{cases} \frac{d}{dt}(e^{At}x) = A(e^{At}x) = e^{At}\lambda x = \lambda(e^{At}x) \\ e^{A.0x} = x \qquad \text{donc } e^{At}x = e^{\lambda t}x \end{cases}$$

Comme on peut décomposer A sur une base de vecteurs propres on voit que A et e^{At} ont même vecteurs propres et que les valeurs propres de e^{At} sont $e^{\lambda t}$. e^{At} est donc définie positive.

6) $\det(e^{At}) = e^{t(\text{trace(A)})}$
On a $e^{At} = \sum \frac{A^n t^n}{n!}$
On peut rendre A triangulaire : $\exists P/ \quad P^{-1}AP = B$ avec B triangulaire.

$$e^{At} = \sum_n (PBP^{-1})^n \frac{t^n}{n!} = P \sum_n \frac{B^n t^n}{n!} P^{-1} = P e^{Bt} P^{-1}$$

$$\det(e^{At}) = \det(e^{Bt}) \text{ or si } B_{ii} = b_i, B_{ii}^p = b_i^p \text{ donc } (e^{Bt})_{ii} = e^{b_i t}$$

$$\text{et } \det(e^{Bt}) = \prod_i e^{b_i t} = e^{t(\text{trace B})} = e^{t(\text{trace A})}$$

Formule de Duhamel

Soit $g \in \mathcal{C}\left(([0,T]); I\!R^n\right)$ et soit $A \in M_n(I\!R)$

$$\text{Le système } \begin{cases} U'(t) = AU + g \\ U(0) = U_0 \end{cases} \tag{4.16}$$

possède une solution unique donnée par la formule de Duhamel

$$U(t) = e^{At}U_0 + \int_0^t e^{A(t-s)} g(s) ds \tag{4.17}$$

Preuve
$f(t,U) = AU + g$ vérifie $|f(t,U) - f(t,V)| \leq ||A|||U - V|$ donc d'après le théorème de Cauchy-Lipschitz il y existence et unicité de la solution U.
Soit $V(t) = e^{-At}U(t)$, on a

$$\begin{aligned} V'(t) &= -Ae^{-At}U + e^{-At}(AU + g) \\ &= e^{-At}g(t) \end{aligned}$$

et comme $V(0) = U_0$

$$V(t) = U_0 + \int_0^t e^{-As} g(s) ds$$

soit

$$U(t) = e^{At}V = e^{At}U_0 + \int_0^t e^{A(t-s)} g(s) ds \tag{4.18}$$

■

4.1.2 Exemple des systèmes linéaires à coefficients variables

Soient $A(t)$ une application continue de $[0,T]$ dans $M_n(I\!R)$ et g une application continue de $[0,T]$ dans $I\!R$. On pose $f(t,U) = A(t)U + g(t)$. On considère le problème de Cauchy

$$\begin{cases} U'(t) = f(t,U) \\ U(0) = U_0 \end{cases} \tag{4.19}$$

Comme $[0,T]$ est compact $||A(t)||$ atteint son maximum sur $[0,T]$.
$L = \max_{t\in[0,T]} ||A(t)||$ convient pour appliquer le théorème de Cauchy-Lipschitz.

1)Cas scalaire : $U'(t) = a(t)U + g(t)$
Si $g = 0$, on a

$$U(t) = U_0 \exp\,(\int_0^t a(s)ds) \tag{4.20}$$

Si $g \neq 0$ on applique la méthode de la variation de la constante qui donne :

$$U'(t) = U_0' \quad \exp\,(\int_0^t a(s)ds) + U_0 a(t) \exp\,(\int_0^t a(s)ds) = a(t)U_0 \exp\,(\int_0^t a(s)ds) + g(t)$$

soit

$$U_0'(t) = \exp\,(-\int_0^t a(s)ds)g(t) \text{ et en posant } a_1(t) = \int_0^t a(s)ds$$

$$U_0(t) = \int_0^t e^{-a_1(s)}g(s)ds + U_0$$

et finalement

$$U(t) = U_0 \exp\,(a_1(t)) + \int_0^t \exp\,(a_1(t) - a_1(s))g(s)ds.$$

Cas de la dimension de $n \geq 2$
Dans ce cas la solution $U = A(t)U$ ne s'exprime pas simplement en fonction d'une exponentielle de matrice parce que, dans le cas général des conditions nécessaires de commutations entre matrices ne sont pas vérifiées.

Pourtant, il existe un formalisme général qui rappelle l'exponentielle : on considère le système différentiel matriciel

$$\left\{ \begin{array}{l} \frac{\partial G}{\partial t}(t,s) = A(t)G(t,s) \\ G(s,s) = I \qquad G(t,s) \in M_n(I\!R) \end{array} \right. \tag{4.21}$$

Ce système vérifie les hypothèses de Cauchy Lipschitz et admet donc une solution unique. De plus si on pose $U(t) = G(t,0)U_0$ on a $\left\{ \begin{array}{l} U'(t) = A(t)U_0 \\ U(0) = U_0 \end{array} \right.$ donc $G(t,s)$ est l'opérateur qui associe à une condition initiale U_0 au temps s la solution du problème de Cauchy au temps t.
$G(t,s)$ est l'opérateur résolvant du système différentiel.
Quand on considère le problème de Cauchy avec $g \neq 0$, l'unique solution vérifiant $\left\{ \begin{array}{l} U'(t) = A(t) + g(t) \\ U(0) = U_0 \end{array} \right.$
est donné par

$$U(t) = G(t,0)U_0 + \int_0^t G(t,s)g(s)ds \tag{4.22}$$

Preuve

On sait que la solution est unique et dans $\mathcal{C}\,([t_0, t_0+T], I\!R^n)$. Si on dérive la formule ci dessus on a :

$$\begin{array}{rcl} U'(t) & = & G'(t,t_0)U_0 + G(t,t)g(t) \\ & = & A(t)U(t) + g(t) \end{array}$$

avec en plus $U(t) = U_0$ et U est donc la solution unique au problème de Cauchy avec $g \neq 0$ ∎

$G(t,s)$ vérifie la propriété suivante qui rappelle l'exponentielle :

$$G(\tau,t).G(t,s) = G(\tau,s)$$

En effet si on considère $B(\tau) = G(\tau,t)G(t,s) - G(\tau,s)$ on a

$$\begin{aligned} B'(\tau) &= A(\tau)G(\tau,t)G(t,s) - A(\tau)G(\tau,s) \\ &= A(\tau)\left(B(\tau)\right) \end{aligned}$$

De plus $B(t) = 0$. Comme ce problème de Cauchy admet une solution unique alors $B = 0$.

Lemme 4.1.6. *Lemme de Gronwall*
Soit u une fonction continuement différentiable de $[t_0,T]$ dans $\mathbb{R}^n$ Soient ϕ et ψ deux fonctions intégrables sur $[t_0,T]$ et presque partout positives ou nulles. Si

$$|u'(t)| \leq \phi(t) + \psi(t)|u(t)| \text{ sur } [t_0,T] \text{ alors, si } \psi(t) = \int_{t_0}^{t} \psi(s)ds$$

alors

$$\forall t \in [t_0,T], |u(t)| \leq |U_0|e^{\psi(t-t_0)} + \int_{t_0}^{t} \phi(s)e^{\psi(t)-\psi(s)}ds \tag{4.23}$$

Preuve
Soit

$$h(t,\varepsilon) = e^{\psi(t)}\left(|U_0| + \varepsilon\right) + \int_{t_0}^{t} e^{(\psi(t)-\psi(s))}\phi(s)ds$$

On a

$$h(t_0,\varepsilon) = e^{\psi(t)}\left(|U_0| + \varepsilon\right) > |U(t_0)|$$

d'autre part $h'(t,\varepsilon) = \psi(t)h(t,\varepsilon) + \phi(t)$ et h est $\mathcal{C}^1$

Par continuité, $\exists\tau$ tel que $\forall t \in [t_0,\tau]$ $\quad |U(t)| \leq h(t)\varepsilon$.

On va montrer que $\tau = T$ et ensuite on passera à la limite $\varepsilon \to 0$.

Si $\tau < T$ comme l'intervalle $[t_0,\tau]$ est maximum $|U(t)| = h(\tau,\varepsilon)$.
D'autre part si $t \leq \tau$

$$\begin{aligned} |u(t)| &\leq |u_0| + \textstyle\int_{t_0}^{t} |u'(s)|ds \\ &\leq |u_0| + \textstyle\int_{t_0}^{t} \left(\phi(s) + \psi(s)|u(s)|\right) ds \\ &\leq |u_0| + \textstyle\int_{t_0}^{t} \left(\phi(s) + \psi(s)h(s,\varepsilon)\right) ds \end{aligned}$$

On a $h'(t,\varepsilon) = \phi(t) + h(t,\varepsilon)\psi(t)$
Donc

$$\begin{aligned} \textstyle\int_{t_0}^{t} \psi(s)h(s,\varepsilon)ds &= \textstyle\int_{t_0}^{t}(h' - \phi)ds \\ &= h(t,\varepsilon) - |u_0| - |\varepsilon| - \textstyle\int_{t_0}^{t} \phi(s)ds \end{aligned}$$

Donc

$$|u(t)| \leq h(t,\varepsilon) - \varepsilon$$

Ceci n'est pas possible pour $t = \tau$!.
Finalement on fait tendre ε vers 0.
■

Application du Lemme de Gronwall
Le lemme précédent permet d'étudier la dépendance de la solution par rapport aux données

Lemme 4.1.7. *On note C_L l'ensemble des fonctions f, continues de $[t_1, t_2] \times \mathbb{R}^n$ qui satisfont*

$$\forall t \in [t_1, t_2], \forall (u, v) \in \mathbb{R}^n, \qquad |f(t, u) - f(t, v)| \leq L|u - v| \tag{4.24}$$

Alors l'application qui au couple $(f, u_0) \in C_L \times \mathbb{R}^n$ associe la solution du problème de Cauchy $\begin{cases} y' = f(t, y) & t \in [t_1, t_2] \\ y(t_0) = u_0 \end{cases}$ *est continue*

De plus si $g \in C_L$ et si $\begin{cases} \dot{z} = g(s, v) \\ z(t_0) = v_0 \end{cases}$ *alors*

$$|y(t) - z(t)| \leq e^{L(t-t_0)}|u_0 - v_0| + \int_{t_0}^{t} |g(s, v) - f(s, v)|\, e^{L|t-s|} ds \tag{4.25}$$

Preuve
$u'(t) = f(t, u)$
$v'(t) = g(t, u)$
avec $w(t) = u(t) - w(t)$ on a

$$\begin{array}{ll} |\dot{w}(t)| & \leq |f(t, u) - f(t, v)| + |f(t, v) - g(t, v)| \\ & \leq L|u(t) - v(t)| + \phi(t) \\ & \leq L|w(t)| + \phi(t) \\ \text{avec} & \phi(t) = |f(t, v) - g(t, v)| \end{array}$$

L'application du Lemme de Gronwall donne :

$$|w(t)| \leq e^{L|t-t_0|} \left(|u(t_0) - v(t_0)|\right) + \int_{t_0}^{t} e^{L(t-s)} \left|f(t, v) - g(t, v)\right| ds$$

■

Régularité de la solution du problème de Cauchy
Soit y la solution du problème de Cauchy. Si f est p fois différentiable alors y est $(p + 1)$ fois différentiable et $y^{(p+1)}(t) = f^{[p]}(t, y)$
avec $f^{[p]}$ définie par : $\begin{cases} f^{[0]} & = f(t, y) \\ f^{[l+1]}(t, y) & = \frac{\partial}{\partial t} f^{[l]}(t, y) + D_y f^{[l]}(t, y).f(t, y) \end{cases}$
(Dérivée totale par rapport à t)

4.1.3 Équations différentielles d'ordre supérieur à un

On appelle système différentiel d'ordre p une équation de la forme :

$$y^{(p)} = f(t, y, y', \cdots y^{(p-1)}) \tag{4.26}$$

Où $f : U \to I\!R^n$ est une application continue sur $U \subset I\!R^n \times (I\!R^m)^{(p)}$

Une solution est une application de $I_0 \to I\!R^m, y, p$ fois dérivable telle que :

$$\begin{cases} \forall t \in I_0, \qquad \big(t, y(t), y(t), \cdots y^{(p-1)}(t)\big) \in U \subset I\!R \times (I\!R^m)^p \\ y^{(p)}(t) = f\left(t, y, y' \cdots y^{(p-1)}\right) \end{cases} \tag{4.27}$$

Un système différentiel d'ordre p est équivalent à :

$$\begin{cases} \frac{dy_0}{dt} & = y_1 \\ \vdots & \\ \frac{dy_{p-2}}{dt} & = y_{p-1} \\ \frac{dy_{p-1}}{dt} & = f(t, y_0, y_1 \cdots y_{p-1}) \end{cases} \tag{4.28}$$

L'inconnue est $(y_0, y_1 \cdots y_{p-1})^t$ et appartient à $(I\!R^m)^p$

Théorème 4.1.8. *Le problème de Cauchy s'écrit :*

$$\begin{cases} Y^{(p)} = f\left(t, y, y', \cdots, y^{(p-1)}\right) \qquad t \in I. \\ y^{(0)}(0) = y_0 \\ y^{(1)}(0) = y_1 \\ \vdots \\ y^{(p-1)}(0) = y_{p-1} \end{cases} \tag{4.29}$$

Si f est localement lipschitzienne en $(y_0, y_1, \cdots y_{p-1})$ sur U alors le problème de Cauchy ci dessus admet une solution et une seule sur I_0.

4.1.4 Problèmes bien posés, bien conditionnés, problèmes raides

Définition 4.1.9. *On dit qu'un problème de Cauchy est bien posé mathématiquement si la solution est unique et dépend continuement de la donnée initiale.*

Exemple 1 :

$$\begin{cases} y' = 2\sqrt{|y|} \\ y(0) = 0 \end{cases} \in [0, +\infty[$$

Ce problème admet les solutions

$$\begin{cases} y(t) = 0 \qquad t \in [0, a] \\ y(t) = (t-a)^2 \qquad t \in [a, +\infty[\qquad \forall a \in [0, +\infty[\end{cases}$$

$$\left(\begin{array}{l} y' = 2\sqrt{|y|} \\ y(0) = \alpha \qquad t \in [0, +\infty[\text{ admet la solution } y(t) = (t + \sqrt{\alpha})^2 \end{array} \right)$$

On utilise la méthode d'Euler :$y_{n+1} = y_n + 2h_n\sqrt{|y_n|}$

- Si $y_0 = 0, \qquad y_n = 0 \qquad \forall_n \geq 0$
- Si $y_0 = \epsilon, \qquad y_{n+1} = y_n + 2h_n\sqrt{y_n} \longrightarrow y(t) \simeq (t + \sqrt{\varepsilon})^2$ quand $h \to 0$

Il n'y a ni unicité (a est qq) ni continuité de la solution par rapport à la condition initiale.

La fonction $f(t,y)$ vaut $2\sqrt{|y|}$ et n'est pas lipschitzienne en 0!.

Les résultats généraux sur les équations différentielles montrent que le problème de Cauchy $y' = f(t,y)$ est mathématiquement bien posé dès que $f(t,y)$est localement lipschitzienne en y.

Définition 4.1.10. *On dit qu'un problème de Cauchy est numériquement bien posé si la continuité de la solution par rapport à la donnée initiale est "assez bonne" pour que la solution ne soit pas perturbée par une erreur initiale ou des erreurs d'arrondi faibles.*
"Assez bonne" veut dire que la constante de continuité est petite par rapport à la précision des calculs. Cette définition ne fait pas référence à la méthode de calcul utilisée.

Exemple 2 :

$\begin{cases} y' = 3y - 1 \qquad t \in]0,10] \\ y(0) = \frac{1}{3} \end{cases}$ a pour solution $y(t) = \frac{1}{3}$
Si on prend $\tilde{y}(0) = \frac{1}{3} + \varepsilon$ on obtient $\tilde{y}(t) = \frac{1}{3} + \varepsilon e^{3t}$ et $y(10) - \tilde{y}(10) = \varepsilon e^{30} \simeq 10^{13}\varepsilon$
Ce problème est mathématiquement bien posé mais numériquement mal posé si la précision des calculs est de l'ordre de 10^{-10}.

Exemple 3 :

Un problème numériquement bien posé peut aussi soulever des difficultés pour un schéma numérique particulier.
Nous verrons en exercice de TD que le problème

$$\begin{cases} y' = -150y + 30 \qquad t \in [0,1] \\ y(0) = \frac{1}{5} \end{cases} \tag{4.30}$$

est mathématiquement et numériquement bien posé, mais que l'obtention d'une solution numérique acceptable avec le schéma d'Euler nécessite d'imposer une restriction sur le pas de discrétisation.

Définition 4.1.11. *On dit qu'un problème est bien conditionné si les méthodes numériques usuelles peuvent en donner la solution en un temps de calcul raisonnable. Le problème de conditionnement est lié à la valeur de la constante de stabilité M($M < 10^{\alpha}$ si la précision est $10^{-\alpha}$). Sinon on dit que le problème est raide.*

4.2 Analyse numérique de la méthode d'Euler

Soit le problème différentiel suivant :

Trouver y tel que
$\forall t \in [t_0, t_0+T] \begin{cases} y'(t) = f(t, y(t)) \\ y(t_0) = \eta \end{cases}$
Nous supposons dans ce chapitre que f est continue sur $[t_0, t_0+T] \times I\!R^n$ et vérifie une hypothèse de Lipschitz :

$$\exists L \quad / \quad \begin{array}{l} \forall t \in [t_0, t_0+T] \\ \forall (y,z) \in I\!R^n \times I\!R^m \end{array} \qquad |f(t,y) - f(t,z)| \leq L|y-z| \tag{4.31}$$

Nous savons alors (voir les théorèmes précédents) que le problème de Cauchy admet une solution et une seule. Nous allons l'approcher de façon discrète de la façon suivante :

- On se donne une subdivision $\{t_0 < t_1 < \cdots < t_N = (T+t_0)\}$ de $[t_0, t_0+T]$ et on pose $h_n = t_{n+1} - t_n$ pour $n = 0, N-1$.
h_n est le pas de la discrétisation et on note $h = max(h_n)$

En utilisant le problème de Cauchy on a :

$$\begin{aligned} y(t_{n+1}) &= y(t_n) + \int_{t_n}^{t_{n+1}} y'(t)dt \\ &= y(t_n) + \int_{t_n}^{t_{n+1}} f(t, y(t))dt \end{aligned} \tag{4.32}$$

Un schéma numérique est un algorithme qui permet de calculer une approximation de $y(t_n)$ pour $n = 1, N-1$, sachant que l'on connaît y_0, une approximation $y(t_0)$. On note y_n l'approximation de $y(t_n)$.

En utilisant une approximation de l'intégrale de type Riemann, (formule de quadrature élémentaire = rectangle à gauche), on obtient le schéma d'Euler explicite.

Précisément, le schéma d'Euler explicite s'écrit :

$$\begin{cases} y_0 = \eta \\ y_{n+1} = y_n + h_n f(t_n, y_n) \text{ pour } n = 0, \text{ à } N-1 \end{cases} \tag{4.33}$$

Un schéma d'Euler implicite s'écrirait :

$$\begin{cases} y_0 = \eta \\ y_{n+1} = y_n + h_n f(t_{n+1}, y_{n+1}) \text{ pour } n = 0, \text{ à } N-1 \end{cases} \tag{4.34}$$

Des schémas un peu plus complexes font intervenir les valeurs de f aux pas $(n-i, n-(i-1), n)$. Par exemple le schéma Adams Basforth s'écrit :

$$\begin{cases} y_0 = \eta \\ y_1 = y_0 + h_0 f(t_0, y_0) \\ y_{n+1} = y_n + h_n \left(\frac{3}{2} f(t_n, y_n) - \frac{1}{2} f(t_{n-1}, y_{n-1})\right) \text{ pour } n = 1 \text{ à } N-1 \end{cases}$$

Quand on utilise un schéma numérique on s'intéresse à l'erreur sur la solution calculée c'est à dire $e_n = y(t_n) - y_n$.
En fait, quand on calcule vraiment la solution ce n'est pas exactement le schéma d'Euler présente plus haut que l'on utilise mais un schéma perturbé par une erreur d'arrondis (de la machine ou de l'individu ou des deux) à chaque étape. Ce schéma "réel" s'écrit :

$$\begin{cases} y_0^* = \eta^* \\ y_{n+1}^* = y_n^* + h_n \left(f(t_n, y_n^*) + \mu_n\right) + \rho_n \end{cases} \tag{4.35}$$

où μ_n désigne l'erreur effectuée sur l'estimation de f et ρ_n l'erreur d'arrondi effectué sur le calcul de y_n.
On souhaite alors :

1) que si les perturbations (ρ_n, μ_n) sont petites alors y_n^* reste proche de y_n. Il s'agit là d'une notion de stabilité qui correspond très souvent à des problèmes de conditionnement des problèmes. La stabilité du schéma est définie par le comportement de (y_{n+1}, z_{n+1}) vis à vis de (y_n, z_n) où

$$\begin{cases} y_{n+1} = y_n + h_n f(t_n, y_n) \\ z_{n+1} = z_n + h_n f(t_n, y_n) + \alpha_n \end{cases} \tag{4.36}$$

2) que le schéma utilisé approche convenablement l'équation considérée. Il s'agit là d'une notion de consistance. On définit plus précieusement l'erreur de consistance :

$$\varepsilon_n = y(t_{n+1}) - y(t_n) - h_n f(t_n, y(t_n)) \tag{4.37}$$

qui représente l'erreur introduite quand on applique le schéma à la solution du problème de Cauchy au temps t_n et que l'on compare le résultat obtenu à la solution au temps t_{n+1}.

Nous allons maintenant donner une estimation de l'erreur dans la méthode d'Euler.

4.2.1 Majoration de l'erreur dans la méthode d'Euler

Nous allons d'abord majorer l'erreur de consistance ε_n puis nous allons établir une relation entre l'erreur de la méthode e_n et l'erreur de consistance ε_n.
On a :

$$\varepsilon_n = y(t_{n+1}) - y(t_n) - h_n f(t_n, y(t_n)) \tag{4.38}$$
$$= \int_{t_n}^{t_{n+1}} (y'(t) - y'(t_n))\, dt \tag{4.39}$$

soit

$$|\varepsilon_n| \leq \omega(y'; h) h_n \tag{4.40}$$

où $\omega(y'; h)$ est le module de continuité de y' :

$$\omega(y'; h) = \sup_{|t-a| \leq h} \left| y'(t) - y'(a) \right|$$

D'après les hypothèses faites sur f, (continue), y' est continue et donc $\lim_{h \to 0} \omega(y'; h) = 0$ et par conséquent $\lim_{h \to 0} \varepsilon_n = 0$.
- Si de plus $y \in C^2[t_0, t_0 + T]$ par application de la formule de Taylor avec reste intégral on a :

$$\varepsilon_n = y(t_{n+1}) - y(t_n) - h_n y'(t_n) = \int_{t_n}^{t_{n+1}} (t_{n+1} - s) y''(s) ds \tag{4.41}$$

(intégration par parties) soit

$$|\varepsilon_n| \leq h_n \int_{t_n}^{t_{n+1}} \left| y''(s) \right| ds \tag{4.42}$$

puis

$$\sum_0^N |\varepsilon_n| \leq h \int_0^t |y''(s)| ds \leq ch \tag{4.43}$$

En conclusion, quand f est de classe C^1 l'ordre du schéma d'Euler est 1.

Revenons à l'erreur globale de la méthode : (c'est cette erreur là qui doit tendre vers 0 si la méthode converge)

$$\begin{aligned} e_{n+1} &= y(t_{n+1}) - y_{n+1} = y(t_{n+1}) - y(t_n) + y(t_n) - y_n + y_n - y_{n+1} \\ &= y(t_{n+1}) - y(t_n) + y_n - y_{n+1} + e_n \\ &= e_n + \varepsilon_n + h_n f\left(t_n, y(t_n)\right) - h_n f(t_n, y_n) \end{aligned}$$

d'où

$$e_{n+1} = e_n + h_n \left(f(t_n, y(t_n)) - f(t_n, y_n)\right) + \varepsilon_n$$

f étant Lipschitzienne, $\underline{|e_{n+1}| \leq (1 + h_n L)|e_n| + |\varepsilon_n|}$

Nous avons finalement les deux estimations suivantes :

$$\left\{ \begin{array}{l} |\varepsilon_n| \leq h_n \omega(y'; h) \\ |e_{n+1}| \leq (1 + h_n L)|e_n| + |\varepsilon_n| \end{array} \right. \tag{4.44}$$

Nous avons besoin du Lemme suivant pour conclure :

Lemme 4.2.1. *Lemme de Gronwall discret*
Soient θ_n et α_n, deux suites de réels positifs ou nuls vérifiant : $\forall n \geq 0 \quad \theta_{n+1} \leq (1 + h_n L)\theta_n + \alpha_n$ où $h_n = t_{n+1} - t_n$
alors

$$\forall n \geq 0 \quad \theta_n \leq e^{L(t_n - t_0)}\theta_0 + \sum_{i=0}^{n-1} e^{L(t_n - t_{i+1})}\alpha_i \tag{4.45}$$

Preuve
Celle-ci se fait par récurrence.
Pour $n = 0$: $\theta_0 \leq \theta_0$.
Pour $n = 1$:

$$\begin{aligned} \theta_1 &\leq e^{L(t_1 - t_0)}\theta_0 + e^{L(t_1 - t_0)}\alpha_0 \\ &\leq e^{L(t_1 - t_0)}\theta_0 + \alpha_0 \end{aligned}$$

Pour $n = 0$, on a $\theta_1 \leq (1 + h_0 L)\theta_0 + \alpha_0 = (1 + L(t_1 - t_0))\theta_0 + \alpha_0$.
or $(1 + x) \leq e^x \qquad \forall x \in R$, et l'assertion est vérifiée aux premiers rangs.

Supposons qu'au rang n :

$$\theta_n \leq e^{L(t_n - t_0)}\theta_0 + \sum_{i=0}^{n-1} e^{L(t_n - t_{i+i})}\alpha_i$$

alors

$$\begin{aligned} \theta_{n+1} &\leq \left(1 + (t_{n+1} - t_n)L\right)\theta_n + \alpha_n \\ &\leq \left(1 + (t_{n+1} - t_n)L\right) e^{L(t_n - t_0)}\theta_0 \\ &+ \left(1 + (t_{n+1} - t_n)L\right) \sum_{i=0}^{n-1} e^{L(t_n - t_{i+1})}\alpha_i \end{aligned}$$

En utilisant toujours la majoration $1 + (t_{n+1} - t_n)L \leq e^{L(t_{n+1} - t_n)}$, on a :

$$\begin{aligned} \theta_{n+1} &\leq e^{L(t_{n+1} - t_0)} + \sum_{i=0}^{n-1} e^{L(t_{n+1} - t_{i+1})}\alpha_i + \alpha_n \\ &\leq e^{L(t_{n+1} - t_0)} + \sum_{i=0}^{n} e^{L(t_{n+1} - t_{i+1})}\alpha_i \end{aligned}$$

■

Appliquons maintenant ce Lemme à e_n et α_n. On obtient :

$$\begin{aligned}
|e_n| &\leq e^{L(t_n-t_0)}|e_0| + \sum_{i=0}^{n-1} e^{L(t_n-t_{i+1})}\omega(y';h)h_i \\
&\leq e^{L(t_n-t_0)}|e_0| + \omega(y';h)\sum_{i=0}^{n-1}\int_{t_i}^{t_{i+1}} ds\ e^{L(t_n-t_{n+i})} \\
&\leq e^{L(t_n-t_0)}|e_0| + \omega(y';h)\sum_{i=0}^{n-1}\int_{t_i}^{t_{i+1}} e^{L(t_n-s)}ds \\
&\leq e^{L(t_n-t_0)}|e_0| + \omega(y';h)\int_{t_0}^{t_n} e^{L(t_n-s)}ds \\
&\leq e^{L(t_n-t_0)}|e_0| + \omega(y';h)\left(\frac{e^{L(t_n-t_0)}-1}{L}\right) \text{ pour } L \neq 0 \\
&\leq |e_0| + \omega(y';h)(t_n-t_0) \text{ pour } L = 0
\end{aligned}$$

En revenant à la définition de $e_n = y(t_n) - y_n$

$$|y(t_n) - y_n| \leq e^{L(t_n-t_0)}|\eta - y_0| + \omega(y',h)\left(\frac{e^{L(t_n-t_0)}-1}{L}\right) \qquad (L \neq 0) \tag{4.46}$$

Si $\lim_{h\to 0} y_0 = \eta$ et $\lim_{h\to 0}\omega(y';h) = 0$ alors

$$\lim_{h\to 0}\max_{0\leq n\leq N}|y(t_n) - y_n| = 0 \tag{4.47}$$

et la méthode d'Euler est convergente

Si on suppose $y \in C^2(t_0, t_{0+T})$ on a également

$$|y(t_n) - y_n| \leq e^{L(t_n-t_0)}|\eta - y_0| + h\int_{t_0}^{t_n} e^{L(t_n-s)}y''(s)ds$$

.
Il existe d'autres versions des hypothèses sur f, plus réalistes, qui conduisent elles aussi à des estimations du même genre de l'erreur de méthode e_n et donc, pour lesquelles le schéma d'Euler est convergent.

4.2.2 Effets des erreurs arrondis

Comme nous l'avons dit plus haut la solution effectivement calculée est celle vérifiant le schéma perturbé :

$$y^*_{n+1} = y^*_n + h_n f(t_n, y^*_n) + h_n\mu_n + \rho_n \tag{4.48}$$

où μ_n désigne l'erreur avec laquelle est estimée la fonction f et ρ_n l'erreur due aux arrondis. On peut supposer raisonnablement que $|\rho_n| \leq \rho$ et $|\mu_n| \leq \mu$.

On note $e_n^* = y(t_n) - y_n^*$ l'erreur du schéma perturbé. On a

$$e_{n+1}^* = y(t_{n+1}) - y_{n+1}^* = \underbrace{y(t_{n+1}) - y(t_n) - h_n f(t_n, y(t_n))}_{\varepsilon_n} + \underbrace{y(t_n) - y_n^*}_{e_n^*}$$

$$+ \quad h_n f(t_n, y(t_n)) - h_n f(t_n, y_n^*) - h_n \mu_n - \rho_n$$

C'est à dire :

$$e_{n+1}^* = e_n^* + h_n \left(f(t_n, y(t_n)) - f(t_n, y_n^*) \right) + \varepsilon_n - h_n \mu_n - \rho_n$$

d'où

$$|e_{n+1}^*| \leq |e_n^*|(1 + h_n L) + \varepsilon_n + h_n \mu + \rho$$

Si on note $\alpha_n = \varepsilon_n + h_n \mu + \rho$ on a :

$$|e_{n+1}^*| \leq (1 + h_n L)|e_n^*| + \alpha_n \text{ avec } |\alpha_n| \leq h \int_{t_n}^{t_{n+1}} |y''(s)| ds + h_n \mu + \rho$$

($y \in C^2[t_0, t_{0+T}]$). On peut appliquer à nouveau le lemme de Gronwall :

$$|e_n^*| \leq e^{L(t_n - t_o)} |e_0^*| + h \int_{t_o}^{t_n} e^{L(t_n - s)} |y''(s)| ds + \mu \sum_{i=0}^{n-1} e^{L(t_n - t_{i+1})} h_n + \rho \sum_{i=0}^{n-1} e^{L(t_n - t_{i+1})}$$

On choisit $h = h_n, \quad \forall_n$ (pas constant) et :

$$|y(t_n) - y_n^*| \leq \underbrace{e^{L(t_n - t_o)} |y(t_0) - y_0^*|}_{\text{erreur due à l'erreur à l'instant initial}}$$

$$+ \quad \underbrace{h \int_{t_o}^{t_n} e^{L(t_n - s)} |y''(s)| ds}_{\text{erreur de consistance}}$$

$$+ \quad \underbrace{\frac{\mu e^{L(t_n - t_0)} - 1}{L}}_{\text{erreur due à l'imprécision sur } f}$$

$$+ \quad \underbrace{\rho e^{L(t_n - t_0)}}_{\text{erreur d'arrondis}}$$

On se fixe à l'instant final $T = t_0 + N\ h$ en faisant tendre h vers 0 (N tend donc vers l'infini).
On a :

- L'erreur due à l'erreur à l'instant initial est indépendante de h
- L'erreur de consistance tend vers 0 quand $h \to 0$ comme h (schéma d'ordre 1)
- L'erreur due à l'imprécision sur f est indépendante de h
- L'erreur d'arrondis tend vers l'infini avec N et $\left(N = \frac{T - t_0}{h}\right)$

On a $|y(t_N) - y_N^*| \leq A + B\ h + \frac{\rho C}{h}$ si on suppose que le pas h_n est fixé et égal à $\frac{T}{N}$.

Soit $\varphi(h) = A + B\ h + \frac{\rho C}{h}$ alors $\varphi'(h) = B - \frac{\rho C}{h^2}$ et φ est minimum pour $h = h^* = \sqrt{\frac{\rho C}{B}}$

Pour $h = h^* \qquad Bh^* = \sqrt{\rho C B} = \frac{\rho C}{h^*}$ et les deux erreurs (de consistance et arrondis) sont égales.
D'un point de vue pratique on évitera de prendre des valeurs de h pour lesquelles l'erreur de la méthode est d'ordre inférieur à l'erreur d'arrondi.

4.3 Étude générale des méthodes à un pas

Soit le problème de Cauchy :

$$\begin{cases} y'(t) = f(t, y(t)) & t \in [t_0, t_0 + T] \\ y(t_0) = \eta \text{ donné} \end{cases} \tag{4.49}$$

ou f est une fonction continue de $[t_0, t_{0+T}] \times I\!R^2$ telle que
$|f(t, y) - f(t, z)| < L|y - z|$.

On se donne une subdivision $t_0 < 1 < \cdots < t_N = t_{0+T}$ de $[t_0, t_{0+T}]$ et on pose $h_n = t_{n+1} - t_n$ et $h = \max\limits_{0 \le n \le N} h_n$.

Une méthode à un pas s'écrit

$$\begin{cases} y_{n+1} = y_n + h_n \phi(t_n, y_n, h_n) & n \ge 0 \\ y_0 = \eta_h \end{cases} \tag{4.50}$$

On supposera que ϕ est continue de $[t_0, t_{0+T}] \times I\!R \times [0, h^*]$ dans $I\!R$, $(h^* > 0)$ et ne dépend que de f.

(Pour la méthode d'Euler : $\phi(t, y, h)) = f(t, y))$

4.3.1 Quelques définitions :

Consistance :
La méthode ci dessus est consistante avec l'équation différentielle initiale si, pour toute solution y de l'équation $y' = f(y, t)$, la somme des erreurs de consistance ε_n :

$$\sum_{n=0}^{N-1} |\varepsilon_n| = \sum_{n=0}^{N-1} |y(t_{n+1}) - y(t_n) - h_n \phi(t_n, y_n, h_n)| \tag{4.51}$$

tend vers 0 lorsque h tend vers 0.
L'erreur de consistance au temps t_n, ε_n, représente l'erreur que l'on fait au nième pas en remplaçant l'équation différentielle par le schéma numérique.

Stabilité :
On dit que la méthode ci dessus est stable si il existe une constante M indépendante de h telle que, pour toutes suites y_n, z_n, ρ_n vérifiant :

$$\begin{aligned} y_{n+1} &= y_n + h_n \phi(t_n, y_n; h_n) \\ z_{n+1} &= z_n + h_n \phi(t_n, z_n; h_n) + \rho_n \end{aligned} \tag{4.52}$$

on ait :

$$\max_{0 \le n \le N} |z_n - y_n| \le M \left(|y_0 - z_0| + \sum_{n<N} |\rho_N| \right) \tag{4.53}$$

Il s'agit en fait de la notion de continuité de la solution du schéma vis à vis des données ϕ (perturbées par ρ_n) et y_0(perturbée par $\eta = |y_0 - z_0|$), M ne dépend pas de h !

Convergence :
La méthode ci dessus est convergente si les conditions $h_n \to 0$ et $\eta_h \to \eta$ donnent :

$$\max_{0\leq n\leq N} |y(t_n) - y_n| \to 0 \tag{4.54}$$

L'erreur de consistance au temps t_n, ε_n, représente l'erreur que l'on fait au nième pas en remplaçant l'équation différentielle par le schéma numérique.

Théorème 4.3.1. *Si la méthode à un pas est stable et consistante alors elle est convergente*

Preuve
y_n vérifie $y_{n+1} = y_n + h_n\phi(t_n, y_n, h_n)$
Soit z_n tel que $z_n = y(t_n)$ alors :
$z_{n+1} = z_n + h_n\phi(t_n, z_n, h_n) + \{y(t_{n+1}) - y(t_n) - h_n\phi(t_n, y(t_n), h_n)\}$
Soit, avec $z_{n+1} = z_n + h_n\phi(t_n, z_n, h_n) + \varepsilon_n$.
La méthode étant consistante on a :

$$\lim_{h\to 0} \sum_{n=0}^{N-1} |\varepsilon_n| = 0 \tag{4.55}$$

Or grâce à la stabilité on a :

$$\max_{0\leq n\leq N} |z_n - y_n| \leq M \left(|\eta_h - \eta| + \sum_n |\varepsilon_n| \right) \tag{4.56}$$

avec $\lim_{h\to 0} |\eta_h - \eta| = 0$, on obtient

$$\lim_{h\to 0} z_n - y_n = 0 \qquad \text{soit} \qquad \lim_{h\to 0} y_n = y(t_n) \tag{4.57}$$

c'est à dire que la méthode est convergente. ∎

Le point essentiel est que l'erreur de consistance ε_n est justement la perturbation qui fait passer de la suite $y(t_n)$ à la suite y_n à chaque étape, i.e, la suite $y(t_n)$ est la suite y_n perturbée pour l'erreur de constitance ε_n.

4.3.2 Convergence des méthodes à un pas :

Une condition *nécessaire et suffisante* pour que la méthode à un pas définie ci dessus soit consistante est que

$$\phi(t, y, 0) = f(t, y), \qquad \forall t \in [t_0, t_{0+T}], \quad \forall y \in I\!R$$

(On démontre facilement que c'est suffisant. On ne démontrera pas la condition nécessaire).

Une condition *suffisante de stabilité* est qu'il existe une constante Λ telle que :

$$\forall t \in [t_0, t_{0+T}], \quad \forall (y, z) \in I\!R, \quad \forall h \in [0, h^*], \quad |\phi(t, y, h) - \phi(t, z, h)| \leq \Lambda |y - z| \tag{4.58}$$

On a alors :

$$|y_n - z_n| \leq e^{\Lambda T} \left(|y_0 - z_0| + \sum_{i=1}^{n-1} |\varepsilon_i| \right) \tag{4.59}$$

(et $M = e^{\Lambda T}$)

Preuve

$$\begin{cases} y_{n+1} = y_n + h_n\phi(t_n, y_n; h_n) \\ z_{n+1} = z_n + h_n\phi(t_n, z_n; h_n) + \varepsilon_n \end{cases} \tag{4.60}$$

Soit

$$|y_{n+1} - z_{n+1}| \leq (1 + h_n\Lambda)|y_n - z_n| + |\varepsilon_n|$$

D'après le lemme de Gronwall :

$$|y_n - z_n| \leq e^{\Lambda(t_n - t_0)}|y_0 - z_0| + \sum_{i=0}^{n-1} e^{\Lambda(t_n - t_{i+1})}|\varepsilon_i|$$

■

Théorème 4.3.2. *On suppose que la fonction ϕ vérifie $\phi(t, y; 0) = f(t, y)$ et la condition de Lipschitz précédente. Alors la méthode à un pas est convergente.*

Preuve
Ce théorème découle directement des conditions ci dessus et du théorème de la page suivante. ■

4.3.3 Ordre d'une méthode à un pas

On précise la notion de consistance en introduisant la notion d'ordre.

Définition 4.3.3. *La méthode à un pas définie plus haut est d'ordre p ($p > 0$) si il existe un réel K indépendant de y et de ϕ tel que :*

$$\sum_{n=0}^{n-1} |\varepsilon_n| = \sum_{n=0}^{n-1} |y(t_{n+1}) - y(t_n) - h_n\phi(t_n, y(t_n); h_n)| \leq Kh^p \tag{4.61}$$

Théorème 4.3.4. *Pour une méthode stable avec constante de stabilité M on a alors, si elle est d'ordre p,*

$$\max |y_n - y(t_n)| \leq M\left(|y_0 - y(t_0)| + \sum |\varepsilon_n|\right) \leq M\left(|y_0 - y(t_0)| + Kh^p\right) \tag{4.62}$$

Si la constante M n'est pas trop grande ($\leq 10^2$), une méthode d'ordre 3 avec un pas de $h = 10^{-2}$ permet d'obtenir une précision de 10^{-4}.

Preuve
Ce résultat est une application directe des notions de stabilité et d'ordre d'une méthode. ■

Théorème 4.3.5. *On suppose que f est p fois différentiable continuement dans $[t_0, t_0 + T] \times I\!R$ et que les fonctions $\phi, \frac{\partial\phi}{\partial h}, \cdots \frac{\partial^p\phi}{\partial h^p}$ existent et sont continues dans $[t_0, t_0 + T] \times I\!R \times [0, h^*]$.*
Alors une condition nécessaire est suffisante pour que la méthode soit d'ordre au moins p s'écrit :

$$\forall (t, y) \in [t_0, t_0 + T] \times I\!R$$

$$\left.\begin{array}{l}\phi(t,y;0)=f(t,y)\\ \frac{\partial\phi}{\partial h}(t,y;0)=\frac{1}{2}f^{(1)}(t,y)\\ \vdots\\ \frac{\partial^{p-1}\phi}{\partial h^{p-1}}(t,y;0)=\frac{1}{p}f^{(p-1)}(t,y)\end{array}\right\}\ \frac{\partial^l\phi}{\partial h^l}(t,y;0)=\frac{1}{l+1}f^{(l)}(t,y)\quad 0\le l\le p-1 \tag{4.63}$$

où $f^{(l)}$ désigne la dérivée totale de f par rapport à t :

$$f^{(0)}=f(t,y);\ f^{(k)}(t,y)=\frac{\partial}{\partial t}f^{(k-1)}(t,y)+D_yf^{(k-1)}(t,y).f(t,y) \tag{4.64}$$

Preuve
On prend ϕ de classe C^p :

$$\phi(t_n,y_n,h_n)=\sum_{l=0}^{p}\frac{h_n^l}{l!}\phi^{(l)}(t_n,y_n,0)+\mathcal{O}(h_n^p)$$

Si f est de classe C^p par rapport à y et t, y est de classe C^{p+1} d'où :

$$\begin{aligned}y(t_{n+1})-y(t_n) &= y(t_n+h_n)-y(t_n)\\ &= \sum_{k=1}^{p+1}\frac{h_n^k}{k!}\ y^{(k)}(t_n)+\mathcal{O}(h_n^{p+1})\\ &= \sum_{l=0}^{p}\frac{h_n^{(l+1)}}{(l+1)!}\ f^{(l)}(t_n,y_n)+\mathcal{O}(h_n^{p+1})\end{aligned}$$

et

$$\begin{aligned}\varepsilon_n &= |y(t_{(n+1)}-y(t_n)-h_n\phi(t_n,y_n,h_n)|\\ &= \sum_{l=0}^{p}\frac{h_n^{l+1}}{l!}\left(\frac{1}{l+1}f^{(l)}(t_n,y_n)-\phi^{(l)}(t_n,y_n,0)+\mathcal{O}(h_n^{p+1})\right)\end{aligned}$$

Donc $\varepsilon_n\le K\ h_n^{p+1}$ si et seulement si $\frac{1}{l+1}f^{(l)}(t_n,y_n)=\phi^{(l)}(t_n,y_n,0)$ pour $0\le l\le p-1$.

Si $\varepsilon_n\le K\ h_n^{p+1}$ alors $\sum|\varepsilon_n|\le K'h^p$ où $K'=TK$ et la condition est donc suffisante.
La condition nécessaire se prouve par l'absurde. ∎

4.3.4 Influence des erreurs d'arrondi

L'erreur globale $\max_{0\le n\le N}|y_n-y(t_n)|\le M\left(|y_0-y(t_0)|+K\ h^p\right)$ est une erreur théorique qui ne tient pas compte des erreurs d'arrondi.
Dans la réalité l'ordinateur a calculé, non pas la suite y_n mais une valeur $\tilde{y}_n$ dans laquelle interviendront

- Une erreur d'arrondi ρ_n sur $\phi(t_n,\tilde{y}_n,h_n)$
- Une erreur d'arrondi σ_n sur $\tilde{y}_{n+1}$

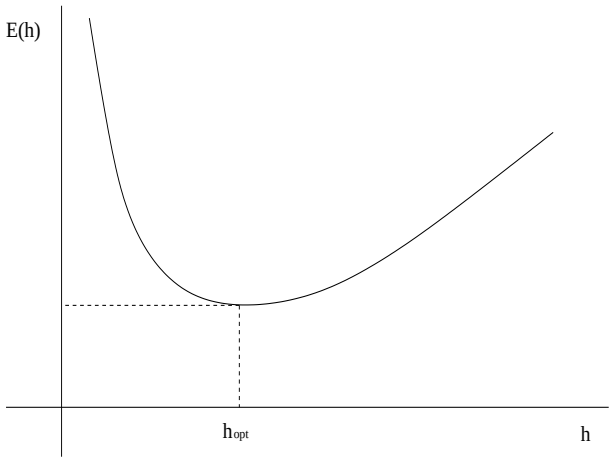


FIGURE 4.1 – influence des erreurs d'arrondi sur l'erreur globale

Soit $\tilde{y}_{n+1} = \tilde{y}_n + h_n\phi(t_n, \tilde{y}_n + h_n) + h_n\rho_n + \sigma_n$.
On pose $\tilde{y}_0 = y_0 + \varepsilon_0$ et on suppose que $\forall_n, \quad |\rho_n| < \rho$ et $|\sigma_n| < \sigma$
On a

$$\begin{aligned} \max_{0\leq n\leq N} |\tilde{y}_n - y_n| &\leq M(|\varepsilon_0| + \sum_{0\leq n\leq N} h_n\ |\rho_n| + |\sigma_n|) \text{ (stabilité)} && (4.65)\\ &\leq M\left(|\varepsilon_0| + T\rho + N\sigma\right) && (4.66)\end{aligned}$$

On sait aussi que $\max|y_n - y(t_n)| \leq M(|y_0 - y(t_0)| + Kh^p)$
d'où $\max|\tilde{y}_n - y(t_n)| \leq M(|\varepsilon_0| + |y_0 - y(t_0)| + T\rho + N\sigma + K\ h^p)$.
Si $h_n = h$ (pas constant) alors $T = Nh$ et

$$E(h) \leq M\left(|\varepsilon_0| + |y_0 - y(t_0)| + T\rho\right) + MT\left(\frac{\sigma}{h} + \frac{K}{T}h^p\right) \tag{4.67}$$

avec $E(h) = \max|\tilde{y}_n - y(t_n)|$.

L'erreur $E(h)$ passe par un *minimum* pour $h_{opt} = \left(\frac{\sigma T}{\rho K}\right)^{\frac{1}{p+1}}$ (voir figure 4.1)

En particulier si on diminue le pas l'erreur augmente !
Ceci est dû au fait que le nombre de pas $N = \frac{T}{h}$ augmente et
avec lui les erreurs d'arrondi. Les erreurs d'arrondi en $\frac{\sigma}{h}$ l'emportent sur l'erreur théorique en h^p!.

4.4 Exemples de méthodes à un pas

4.4.1 Méthode du développement de Taylor

On prend

$$\phi(t, y; h) = f(t, y) + \frac{h}{2}f^{(1)}(t, y) + \cdots + \frac{h^{p-1}}{p!}f^{(p-1)}(t, y) \tag{4.68}$$

Si on écrit

$$f(t,y) = f(t_n,y) + (t-t_n)f^{(1)}(t_n,y) + \cdots + \frac{(t-t_n)^m}{m!}f^{(m)}(t_n,y) + \cdots \tag{4.69}$$

puis

$$\begin{aligned} y(t_{n+1}) &= y(t_n) + \int_{t_n}^{t_{n+1}} f(t,y)dt \\ &\simeq y(t_n) + h_n f(t_n,y) + \frac{h_n^2}{2}f^{(1)}(t_n,y) + \cdots + \frac{(h_n)^{m+1}}{(m+1)!}f^{(m)}(t_n,y) + \cdots \end{aligned}$$

cela revient à tronquer la série à l'ordre pr.

- Pour $p=1$ on retrouve la méthode d'Euler.
- Les hypothèses pour pouvoir caractériser l'ordre d'une méthode sont trivialement vérifiées et la méthode du développement de Taylor est d'ordre p.
- Si f et $f^{(k)}$ pour $0 \le k \le p-1$ sont lipschitziennes ie :

  $$\exists L_k \ / \ \forall t \in [t_0, t_0+T], \quad \forall (y,z) \in I\!R \qquad |f^{(k)}(t,y) - f^{(k)}(t,z)| \le L_k |y-z|$$

  alors ϕ vérifie aussi la condition de Lipschitz

  $$|\phi(t,y;h) - \phi(t,z;h)| \le \Lambda |y-z|$$

  avec $\Lambda = L_0 + \frac{h}{2}L_1 + \cdots + \frac{h^{p-1}}{(p)!}L_{p-1}$ et pour $h \le h_{max}$,

  $$\Lambda \le \Lambda_{max} = L_0 + \frac{h_{max}}{2}L_1 + \cdots + \frac{h_{max}^{p-1}}{p!}L_{p-1}$$

  C'est une condition suffisante de stabilité et la méthode du développement de Taylor est donc stable. (constante de stabilité $e^{\Lambda T_{max}}$).

- La méthode du développement de Taylor est alors convergente(théorème)
- Du point de vue pratique, les méthodes du développement de Taylor présentent l'inconvénient d'utiliser $f, f^{(1)}, \cdots f^{(p-1)}$ ce qui mobilise beaucoup de mémoire. On évite généralement ces méthodes sauf dans des cas particuliers où il est facile de calculer les dérivées totales de f par rapport à t.

4.4.2 Méthodes de Runge-Kutta

On se donne q réels $c_1, c_2, \cdots c_q$; $\quad c_i \in [0,1]$ distincts ou non et on leur associe des formules de quadrature numériques :

$$\int_0^{c_i} \psi(t)dt \simeq \sum_{j=1}^{q} a_{ij}\psi(c_j) \qquad i = 1, \cdots, q \tag{4.70}$$

et

$$\int_0^1 \psi(t)dt \simeq \sum_{j=1}^{q} \quad b_j\psi(c_j) \tag{4.71}$$

On considère les instants intermédiaires $t_{n,i} = t_n + c_i h_n$.
Si y est solution de $y' = f(t,y)$ alors

$$y(t_{n,i}) = y(t_n) + \int_{t_n}^{t_{n,i}} f(t,y)dt \simeq y(t_n) + h_n \sum_{j=1}^{q} a_{ij} f(t_{n,j}, y(t_{n,j})) \tag{4.72}$$

et

$$y(t_{n+1}) \simeq y(t_n) + \sum_{j=1}^{q} b_j \quad f(t_{n,j}, y(t_{n,j})). \tag{4.73}$$

On appelle schéma de Runge-Kutta à q pas intermédiaires le schéma :

$$1) \quad y_{n,i} = y_n + h_n \sum_{j=1}^{q} a_{ij} f(t_{n,j}, y_{n,j}) \qquad i = 1, ..., q \tag{4.74}$$

$$2) \quad y_{n+1} = y_n + h_n \sum_{j=1}^{q} b_j f(t_{n,j}, y_{n,j}) \tag{4.75}$$

$$3) \quad t_{n,i} = t_n + c_i h_n \qquad i = 1, \cdots, q \tag{4.76}$$

La première ligne est en fait un système de q équations à q inconnues $y_{n,j} (j = 1, q)$. Une fois résolu il permet de calculer y_{n+1} par la ligne 2).

Si on écrit $\phi(t, y; h) = \sum_{j=1}^{q} b_j f(t_n + c_j h_n, y_j)$ où
$y_k = y_n + h_n \sum_{j=1}^{q} a_{kj} f(t_n + c_j h_n, y_j)$ alors la méthode de Runge Kutta s'écrit :

$$y_{n+1} = y_n + h_n \phi(t_n, y_n; h_n) \tag{4.77}$$

et c'est bien une méthode à un pas.

Une méthode de Runge Kutta est entièrement définie quand on connaît q et les coefficients a_{ij}, b_j et les réels c_j.

La coutume est de représenter une méthode de Runge Kutta sous la forme du tableau suivant :

$$\begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \cdots & a_{1q} \\ c_2 & a_{21} & a_{22} & & a_{2q} \\ \vdots & \vdots & \vdots & & \vdots \\ c_q & a_{q1} & a_{q2} & & a_{qq} \\ \hline & b_1 & b_2 & \cdots & b_q \end{array}$$

Exemple 1 :

• $q = 1$

$$\begin{array}{c|c} 0 & 0 \\ \hline & \\ & 1 \end{array}$$

$y_{n,1} = y_n + h_n.0$

$y_{n+1} = y_n + h_n f(t_n, y_n)$
Il s'agit de la méthode d'Euler explicite

Exemple 2 :

- $q = 2$

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \alpha & \alpha & 0 \\ \hline & 1-\frac{1}{2\alpha} & \frac{1}{2\alpha} \end{array} \qquad (\alpha \neq 0)$$

Donne

$$\begin{aligned} y_{n,1} &= y_n + 0 \\ y_{n,2} &= y_n + h_n \alpha f(t_{n,1}, y_{n,1}) = y_n + h_n \alpha f(t_n, y_n) \end{aligned} \tag{4.78}$$

et

$$y_{n+1} = y_n + h_n \left(1 - \frac{1}{2\alpha}\right) f(t_n, y_n) + h_n \frac{1}{2\alpha} f(t_n + \alpha h_n, y_n + h_n \alpha f(t_n, y_n)) \tag{4.79}$$

Pour $\alpha = \frac{1}{2}$ on a :

$$y_{n+1} = y_n + h_n f(t_n + \frac{h_n}{2}, y_n + \frac{h_n}{2} f(t_n, y_n)) \tag{4.80}$$

C'est la méthode d'Euler modifiée ou méthode du point milieu.

Exemple 3 : Méthode de Runge Kutta d'ordre 4 classique
Elle correspond à $q = 4$ et au tableau suivant :

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

On en déduit l'algorithme suivant :

$$\begin{array}{ll} t_{n,1} = t_n; & y_{n,1} = y_n \\ t_{n,2} = t_n + \frac{h_n}{2}; & y_{n,2} = y_n + \frac{h_n}{2} f(t_n, y_n) \\ t_{n,3} = t_{n,2}; & y_{n,3} = y_n + \frac{h_n}{2} (t_{n,2}, y_{n,2}) \\ t_{n,4} = t_{n+1}; & y_{n,4} = y_n + h_n f(t_{n,3}, y_{n,3}) \end{array} \tag{4.81}$$

et pour terminer :

$$y_{n+1} = y_n + h_n \left(\frac{1}{6} f(t_n, y_n) + \frac{2}{6} f(t_{n,2}, y_{n,2}) + \frac{2}{6} f(t_{n,3}, y_{n,3}) + \frac{1}{6} f(t_{n,4}, y_{n,4})\right) \tag{4.82}$$

Les méthodes d'intégration de chaque étape sont :

$$\begin{aligned}
M_2 &: \int_0^{1/2} g(t)dt \simeq \frac{1}{2}g(0) \quad \text{rectangle à gauche} \\
M_3 &: \int_0^{1/2} g(t)dt \simeq \frac{1}{2}g(1/2) \quad \text{rectangle à droite} \\
M_4 &: \int_0^{1} g(t)dt = g(1/2) \quad \text{point milieu} \\
M &: \int_0^{1} g(t)dt = \frac{1}{6}g(0) + \frac{2}{3}g(1/2) + \frac{1}{6}g(1) \quad \text{Simpson}
\end{aligned}$$

Cette méthode est d'ordre 4

4.4.3 Stabilité et ordre des méthodes de Runge-Kutta :

On introduit les matrices et vecteurs suivants :

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1q} \\ \vdots & & & \\ \vdots & & & \\ a_{q1} & \cdots & \cdots & a_{qq} \end{pmatrix} \qquad C = \begin{pmatrix} c_1 & & & & \\ & \ddots & & & \\ & & c_2 & & 0 \\ & & & \ddots & \\ 0 & & & & c_q \end{pmatrix} \tag{4.83}$$

$$b = \begin{pmatrix} b_1 \\ \vdots \\ \vdots \\ b_q \end{pmatrix} \quad e = \begin{pmatrix} 1 \\ \vdots \\ \vdots \\ 1 \end{pmatrix} \tag{4.84}$$

- Si A est strictement triangulaire inférieure ($a_{ij} = 0$ dès que $i \leq j$) la résolution du système donnant les $y_{n,i}$ est immédiate ; $y_{n,1} = y_n$ et on obtient $y_{n,2}$ à partir de $y_{n,1}$. On dit que la méthode RK est explicite.

- Si A est triangulaire inférieure ($a_{ij} = 0$ dès que $i < j$) le calcul de $y_{n,i}$ se fait par résolution successive de q équations à une inconnue. RK est dite semi-implicite.

- Autrement RK est dite implicite et le calcul de $y_{n,i}$ se fait par résolution globale d'un système de q équations à q inconnues.

Stabilité :
Sous l'hypothèse $h_n L\rho(|A|) < 1$ où $|A|$ désigne la matrice $|a_{ij}|$, L la constante de Lipschitz de f et $\rho(|A|)$ le rayon spectral de $|A|$,
le schéma de Runge-Kutta admet une solution unique.
De plus si $h^* L\rho(|A|) < 1$, le schéma est stable pour tout $h,\ 0 < h \leq h^*$.

Ordre :
Une condition nécessaire et suffisante pour qu'une méthode RK soit d'ordre p pour toute fonction f suffisamment régulière est qu'elle vérifie les conditions $A(k)$ suivantes pour $1 \leq k \leq p$

p	$A(p)$
1	$b^t e = 1$
2	$b^t Ce = b^t Ae = 1/2$
3	$b^t C^2 e = b^t CAe = b^t (Ae)^2 = 1/3$ $b^t ACe = b^t A^2 e = \frac{1}{6}$

e est le vecteur de $(1, ..., 1)$ de $\mathbb{R}^q$ où q est la taille de la matrice de Runge-Kutta.

On applique, pour trouver ces résultats, le critère défini par le théorème 4.3.5 qui consiste à évaluer les dérivées $\frac{\partial^k \phi}{\partial h^k}(t, y, 0)$ sachant que ϕ est définie par :

$$\phi(t, y, h) = \sum_{j=1}^{q} b_j f(t + c_j h, y_j) \tag{4.85}$$

4.5 Méthodes à pas multiples

On s'intéresse comme au chapitre précédent au problème de Cauchy

$$\begin{cases} y' = f(t, y) \quad (t, y) \in [t_0, t_0 + T] \times \mathbb{R} \\ y(t_0) = \eta \end{cases} \tag{4.86}$$

On appelle méthode numérique à $r + 1$ pas, toute relation du type :

$$y_{n+1} = \psi(t_n, y_n, h_n; t_{n-1}, y_{n-1}, h_{n-1}; \cdots; t_{n-r}, y_{n-r}, h_{n-r}) \quad -1 \leq r,\ r \in \mathbb{Z} \tag{4.87}$$

Leur intérêt vient du fait que l'on peut obtenir un ordre élevé pour des complexités de calcul nettement inférieures à la méthode de Runge Kutta.
Nous allons étudier deux méthodes :

1) Les méthodes d'Adams Bashforth
2) Les méthodes d'Adams Moulton

4.5.1 Erreur de consistance et ordre

Soit y une solution exacte du problème de Cauchy. L'erreur de consistance ε_n est $\varepsilon_n = y(t_{n+1}) - y_{n+1}$,
où y_{n+1} est calculée à partir des $r+1$ valeurs $y(t_n) \cdots y(t_{n-r})$ par $y_{n+1} = \psi(t_n, y(t_n), h_n; \cdots; t_{n-r}, y(t_{n-r}), h_{n-r})$.

La méthode est dite d'ordre p si

$$\sum_{n=0}^{N-1} |\varepsilon_n| \leq C\ h^p \tag{4.88}$$

4.5.2 Stabilité

Définition 4.5.1. *On dit qu'une méthode à $r + 1$ pas est stable si pour toutes suites :*

$$\tilde{y}_{n+1} = \psi(t_{n-i}, \tilde{y}_{n-i}, h_{n-i}) + \varepsilon'_n \quad 0 \leq i \leq r,\ r \leq n < N \tag{4.89}$$

alors

$$\max_{0\leq n\leq N}|\tilde{y}_n - y_n| \leq M(\underbrace{\max_{0\leq n\leq r}|\tilde{y}_n - y_n|}_{\text{erreur sur les } r \text{ premier pas}} + \sum_{r\leq n<N}|\varepsilon'_n|) \tag{4.90}$$

4.5.3 Les méthodes d'Adams Bashforth :

h_n étant un pas variable et y étant la solution du problème de Cauchy, on écrit :

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

Si on suppose que $y(t_{n-i})$ et$f_{n-i} = f(t_{n-i}, y(t_{n-i}))$ sont déjà calculés pour $0 \leq i \leq r$, l'idée de la méthode est de remplacer $f(t, y(t))$ sur l'intervalle $[t_n, t_{n+1}]$ par son polynôme d'interpolation aux points $t_n, \cdots, t_{n-r}$.

Soit $p_{n,r}(t)$ le polynôme de degré r qui interpole $((t_{n-i}, f_{n-i}))_{i=0}^{r}$, alors

$$p_{n,r}(t) = \sum_{i=0}^{r} f_{n-i}\ L_{n,i,r}(t)$$

où

$$L_{n,i,r}(t) = \prod_{0\leq j\leq r, j\neq i} \frac{t - t_{n-j}}{t_{n-i} - t_{n-j}}$$

est le polynôme de Lagrange de degré r associé au point t_{n-i}.
On écrit donc :

$$\begin{aligned} y(t_{n+1}) &\simeq y(t_n) + \int_{t_n}^{t_{n+1}} p_{n,r}(t)dt \\ &= y(t_n) + h_n \sum_{i=0}^{r} b_{n,i,r} f_{n-i} \end{aligned}$$

avec

$$b_{n,i,r} = \frac{1}{h_n}\int_{t_n}^{t_{n+1}} L_{n,i,r}(t)dt$$

Le schéma AB_{r+1} s'écrit alors :

$$\begin{aligned} y_{n+1} &= y_n + h_n \sum_{i=0}^{r} b_{n,i,r} f_{n-i} \\ t_{n+1} &= t_n + h_n \\ f_{n+1} &= f(t_{n+1}, y_{n+1}) \end{aligned} \tag{4.91}$$

Son intérêt provient du fait qu'une seule estimation de f est nécessaire à chaque pas (contrairement à une méthode de Runge Kutta). Ceci est intéressant car l'estimation de f peut être très coûteuse.

Exemples :
• $AB_1 : r = 0$

$L_{n,0} = 1$ et donc $p_{n,0} = f_n$. Par suite
$b_{n,0,0} = 1$ et l'on retrouve le schéma d'Euler explicite

$$y_{n+1} = y_n + h_n f_n$$

• $AB_2 : r = 1$
On obtient immédiatement

$$p_{n,1}(t) = f_n + \frac{f_n - f_{n-1}}{t_n - t_{n-1}}(t - t_n) \tag{4.92}$$

et

$$\begin{aligned} \int_{t_n}^{t_{n+1}} p_{n,1}(t)dt &= f_n h_n + \frac{f_n - fn - 1}{t_n - t_{n-1}} \left[\frac{1}{2}(t - t_n)^2\right]_{t_n}^{t_{n+1}} \\ &= h_n(f_n + \frac{h_n}{2h_{n-1}}(f_n - f_{n-1})) \end{aligned} \tag{4.93}$$

D'où le schéma :

$$\begin{aligned} y_{n+1} &= y_n + h_n(f_n + \frac{h_n}{2h_{n-1}}(f_n - f_{n-1})) \\ t_{n+1} &= t_n + h_n \\ f_{n+1} &= f(t_{n+1}, y_{n+1}) \end{aligned} \tag{4.94}$$

Pour un pas constant $h_n = h$, AB_2 s'écrit :

$$\begin{cases} y_{n+1} = y_n + h(\frac{3}{2}f_n - \frac{1}{2}f_{n-1}) \\ t_{n+1} = t_n + h \\ f_{n+1} = f(t_{n+1}, y_{n+1}) \end{cases} \tag{4.95}$$

• Dans le cas ou $h_n = h$ (pas constant)
$b_{n,i,r}$ ne dépend plus de n (on le note $b_{i,r}$ et on obtient les quelques valeurs classiques suivantes :

r	$b_{0,r}$	$b_{1,r}$	$b_{2,r}$	$b_{3,r}$	$\beta_r = \sum \lvert b_{i,r}\rvert$
0	1				1
1	$\frac{3}{2}$	$\frac{-1}{2}$			2
2	$\frac{23}{12}$	$-\frac{16}{12}$	$\frac{5}{12}$		$3,66\cdots$
3	$\frac{55}{24}$	$-\frac{59}{24}$	$\frac{37}{24}$	$-\frac{9}{24}$	$6,6\cdots$

En appliquant ces schémas pour $f = 1$, on montre que nécessairement

$$\sum_{0 \le i \le r} b_{i,r} = 1. \tag{4.96}$$

4.5.4 Erreur de consistance et ordre de la méthode AB_{r+1}

Consistance :
D'après la définition générale on écrit :

$$\varepsilon_n = y(t_{n+1}) - y_{n+1} \tag{4.97}$$

où y_{n+1} est calculé à partir de $\left\{ \begin{array}{l} y(t_n) \cdots y(t_{n-r}) \\ h_n, \cdots\cdots h_{n-r} \end{array} \right.$
Soit

$$\varepsilon_n = y(t_{n+1}) - \left(y(t_n) + \int_{t_n}^{t_{n+1}} p_{n,r}(t)dt \right) \tag{4.98}$$

$$\varepsilon_n = \int_{t_n}^{t_{n+1}} \left(y'(t) - p_{n,r}(t) \right) dt$$

Le théorème de la moyenne donne : $\exists\ \theta \in]t_n, t_{n+1}[$ tel que
$\varepsilon_n = h_n \left(y'(\theta) - p_{n,r}(\theta) \right)$

L'erreur de consistance est donc une erreur d'interpolation (y' par $p_{n,r}$) qui est de l'ordre h_n^{r+1} puisque l'interpolation est d'ordre r.

En appliquant les résultats connus sur l'erreur d'interpolation on obtient :

$$|\varepsilon_n| \leq C h_n\ h^{r+1} \quad \text{et } \sum |\varepsilon_n| \leq CT\ h^{r+1} \tag{4.99}$$

La méthode AB_{r+1} est d'ordre $r+1$.

Stabilité :

Théorème 4.5.2. *On suppose que $f(t,y)$ est k Lipschitzienne en y et que les sommes $\sum_{0 \leq i \leq r} |b_n, i, r|$ sont majorées indépendamment de n par une constante β_r.*

$$\beta_r = \max_n \sum_{0 \leq i \leq r} |b_{n,i,r}| \tag{4.100}$$

alors la méthode AB_{r+1} est d'ordre $r+1$ est stable avec une constante de stabilité $S = \exp(\beta_r KT)$.

Preuve
On compare la suite récurrente non perturbée :

$$\begin{aligned} y_{n+1} &= y_n + h_n \sum_{i=0}^{r} b_{n,i,r} f_{n-i} \\ f_{n-i} &= f(t_{n-i}, y_{n-i}) \end{aligned}$$

avec la suite perturbée par ε_n

$$\begin{aligned} \tilde{y}_{n+1} &= \tilde{y}_n + h_n \sum_{i=0}^{r} b_{n,i,r} \tilde{f}_{n-i} + \varepsilon_n \\ \tilde{f}_{n-i} &= f(t_{n-i}, \tilde{y}_{n-i}) \end{aligned}$$

Soit

$$\sigma_n = \max_{0\leq i\leq n} |y_i - \tilde{y}_i|$$

on a

$$\left|\tilde{f}_{n-i} - f_{n-i}\right| \leq k\,|\tilde{y}_{n-i} - y_{n-i}| \leq k\ \sigma_n \qquad (f \text{ est lipschitzienne})$$

d'où

$$\begin{aligned} |\tilde{y}_{n+1} - y_{n+1}| &\leq \sigma_n + h_n \sum_{i=0}^{r} |b_{n,i,r}|\, k\sigma_n + |\varepsilon_n| \\ &\leq (1 + kh_n\beta_r)\sigma_n + |\varepsilon_n| \end{aligned}$$

Or

$$\begin{aligned} \sigma_{n+1} &= \max\left(|y_{n+1} - \tilde{y}_{n+1}|, \sigma_n\right) \text{ d'où :} \\ \sigma_{n+1} &\leq (1 + kh_n\beta_r)\sigma_n + |\varepsilon_n| \end{aligned}$$

On applique le lemme de Gronwall qui donne :

$$\begin{aligned} \sigma_n &\leq e^{\beta_r k(t_N - t_r)} \left(\sigma_r + \sum_{r\leq n\leq N} |\varepsilon_n|\right) \\ &\leq e^{\beta_r kT} \left(\sigma_r + \sum_{r\leq n\leq N} |\varepsilon_n|\right) \end{aligned}$$

Si on pose $M = e^{\beta_r kT}$, M est bien indépendante de h et on a bien

$$\max_{0\leq n\leq N} |y_n - \tilde{y}_n| \leq M \left(\max_{0\leq n\leq r} |y_n - \tilde{y}_n| + \sum_{r\leq n\leq N} |\varepsilon_n|\right) \tag{4.101}$$

qui est la définition de la stabilité du schéma à $r+1$ pas. ∎

Le tableau 4.5.3 montre que β_r augmente assez rapidement avec r ce qui implique que la stabilité devient de moins en moins "bonne" quand le nombre de pas augmente. Cette mauvaise stabilité est un des inconvénients majeurs des méthodes d'Adams Bashforth lorsque r est grand. On se limite en pratique à $r = 1$ ou $r = 2$.

4.5.5 Méthodes d'Adams-Moulton

L'idée est du même type que pour la méthode d'Adams Basforth mais on approxime ici $f(t, y(t))$ par son polynôme d'interpolation aux points t_{n+1},$t_n, \cdots t_n - r$.
On considère donc le polynôme $p^*_{n,r}(t)$ de degré $(r+1)$ qui interpole les points (t_{n-i}, f_{n-i}); $\quad -1 \leq i \leq r$.

On a alors

$$p^*_{n,r}(t) = \sum_{-1\leq i\leq r} f_{n-i} L^*_{n,i,r}(t)$$

avec

$$L^*_{n,i,r}(t) = \prod_{-1\leq j\leq r; j\neq i} \frac{t-t_{n-j}}{t_{n-i}-t_{n-j}}$$

On a :

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

et l'interpolation donne

$$y_{(t_{n+1})} \simeq y(t_n) + h_n \sum_{-1\leq i\leq r} b^*_{n,i,r}\ f_{n-i} \tag{4.102}$$

avec

$$b^*_{n,i,r} = \frac{1}{h_n}\int_{t_n}^{t_{n+1}} L^*_{n,i,r}(t)\ dt \tag{4.103}$$

Le schéma AM_{r+1} s'écrit donc :

$$y_{n+1} - h_n b^*_{n,-1,r} f(t_{n+1}, y_{n+1}) = y_n + h_n \sum_{0\leq i\leq r} b^*_{n,i,r} f_{n-i} \tag{4.104}$$

Contrairement aux schémas de type AB, y_{n+1} n'est pas donné ici par une expression explicite. Pour cette raison on dit que la méthode AM est implicite. On a recours à une méthode itérative pour le calcul de y_{n+1}.

Si on pose $u_n = y_n + h_n \sum_{0\leq i\leq r} b^*_{n,i,r} f_{n-i}$ le point y_{n+1} est solution de

$$x = u_n + h_n b^*_{n,-1,r} f(t_{n+1}, x) \tag{4.105}$$

soit à résoudre l'équation $x = \varphi(x)$ où $\varphi(x) = u_n + h_n b^*_{n,-1,r} f(t_{n+1}, x)$.
Si f est dérivable, on a $\varphi'(x) = h_n b^*_{n,-1,r} f'_y(t_{n+1}, x)$. φ est donc Lipschitzienne de rapport $h_n b^*_{n,-1,r} L$.
Si on suppose que $h_n b^*_{n,-1,r} L < 1$ pour h_n assez petit, φ est contractante (cte plus petite que 1) et l'algorithme de point fixe

$$x_p = \varphi(x_p) \tag{4.106}$$

converge vers la solution y_{n+1}.

Exemples :

• AM_0 : $r = 0$
$p^*_{n,0}$ est le polynôme de degré 1 qui interpole f entre (t_n, f_n) et (t_{n+1}, f_{n+1}) :

$$p^*_{n,0}(t) = f_n + \frac{f_{n+1}-f_n}{t_{n+1}-t_n}(t-t_n)$$

$$\begin{aligned}
\int_{t_n}^{t_{n+1}} p^*_{n,0}(t) &= f_n h_n + \frac{f_{n+1}-f_n}{h_n}\left[\frac{1}{2}(t-t_n)^2\right]_{t_n}^{t_{n+1}} \\
&= h_n\left(f_n + \frac{1}{2}(f_{n+1}-f_n)\right) \\
&= h_n\left(\frac{1}{2}f_{n+1} + \frac{1}{2}f_n\right)
\end{aligned}$$

donc $b^*_{n,-1,0} = b^*_{n,0,0} = \frac{1}{2}$ et

$$y_{n+1} = y_n + \frac{h_n}{2}\left(f(t_{n+1}, y_{n+1}) + f(t_n, y_n)\right) \tag{4.107}$$

Cette méthode s'appelle la méthode de Crank Nicolson.

• AM_{-1} : $r = -1$
On obtient le schéma $y_{n+1} = y_n + h_n f(t_{n+1}, y_{n+1})$ qui est la méthode d'Euler implicite (dite aussi Euler rétrograde).

Remarque : On peut démontrer que, sous des hypothèses peu restrictives sur le pas il est possible de contrôler

$$\sum_{i=-1}^{r} |b^*_{n,i,r}| \tag{4.108}$$

On a toujours $\frac{1}{r+2} \leq b^*_{n,-1,r} \leq 1/2$

Voici quelques valeurs classiques pour des pas constants ($b^*_{n,i,r}$ est alors indépendant de n).

r	$b^*_{-1,r}$	$b^*_{0,r}$	$b^*_{1,r}$	$b^*_{2,r}$	$b^*_{3,r}$	$\beta^*_r = \sum \lvert b^*_{i,r}\rvert$	β_{r+1}
0	$\frac{1}{2}$	$\frac{1}{2}$				1	2
1	$\frac{5}{12}$	$\frac{8}{12}$	$-\frac{1}{12}$			1,16	3,66
2	$\frac{9}{24}$	$\frac{19}{24}$	$-\frac{5}{24}$	$\frac{1}{24}$		1,41	6,66
3	$\frac{251}{720}$	$\frac{646}{720}$	$-\frac{264}{720}$	$\frac{106}{720}$	$-\frac{19}{720}$	1,78	12,64

Erreur de consistance et ordre des méthodes d'Adams Moulton

La méthode AM_{r+1} est d'ordre $r+2$ dès que f est k lipschitzienne par rapport à y.
Si on note $\beta^*_r = \max_n \sum_{i=1}^{r} |b^*_{n,i,r}|$ et $\gamma^*_n = \max_n |b^*_{n,-1,r}|$, alors dés que $h_{max} < \frac{1}{k\gamma^*_r}$ (condition pour que la méthode itérative qui permet de calculer y_{n+1} converge) alors AM_{r+1} est stable et la constante de stabilité vaut $S = \exp\left(\frac{b^*_r kT}{1-\gamma^*_r k h_{max}}\right)$.
Si $h_{max} << \frac{1}{k\gamma^*_r}$, alors $S = \exp\left(b^*_r kT\right)$.
Le tableau précédent montre que la méthode AM_{r+1} est "beaucoup plus stable" que la méthode AB_{r+2}.

Cependant malgré cet avantage le caractère implicite de la méthode AM limite son emploi.

Chapitre 5

Approximation par différences finies des équations aux dérivées partielles

5.1 Exemples d'équations aux dérivées partielles et classification

5.1.1 Un peu d'histoire

— [**1642-1727**] Newton, [**1647-1716**] Leibnitz, théorie du mouvement et mouvement des planètes, équations différentielles ordinaires et [**1685-1731**] Taylor, différences finies, développement de...

— [**1717-1783**] D'Alembert, [**1768-1830**] Fourier et [**1707-1783**] Euler, Equations des ondes, de la chaleur, principes de la mécanique des fluides, équations aux dérivés partielles.

— [**1826-1866**] Rieman, [**1785-1836**] Navier et [**1819,1903**] Stokes proposent les EDP qui portent leur nom ; L'équation de Rieman admet des SOLUTIONS DISCONTINUES...

— [**1906-1998**] Leray, [**1915-2002**] Schwartz, [**1928-2001**] Lions, solutions faibles, distributions, analyse numérique des EDP

Passage des équations différentielles ordinaires (EDO, ODE)
aux équations aux dérivées partielles (EDP, PDE) :
Quand le nombre d'inconnues tend vers $+\infty$ comme par exemple :

— le nombre de particules fluides : Mécanique des fluides,

— le nombre de pixels : EDP pour le traitement d'images,

— Le nombre de véhicules : EDP pour la modélisation du traffic routier,

— le nombre de fluidités : EDP de la finance.

Notions de dérivée :
Besoin de définir une dérivée faible valable pour des fonctions non continues car certaines EDP admettent des solutions discontinues

5.1.2 Exemples d'équations aux dérivées partielles du second ordre

Ω désigne un ouvert de $\mathbb{R}^m$ de frontière Γ.
Equation de Laplace : EQUATION ELLIPTIQUE (équation stationnaire)

$$\left\{ \begin{array}{rcl} -\Delta u & = & f \text{ sur } \Omega \subset \mathbb{R}^m \\ u|_\Gamma & = & u_0 \end{array} \right. \tag{5.1}$$

Propriétés :

— la donnée aux bords influe sur la solution sur tout le domaine

— la solution est plus régulière que f sur l'intérieur de Ω

— le principe du maximum contraint les extrema de u

Equation de la chaleur : EQUATION PARABOLIQUE (équation d'évolution)

$$\begin{cases} \frac{\partial u}{\partial t} - \Delta u &= f \text{ sur } \Omega \times [0,T] \\ u|_\Gamma &= g \\ u(.,t=0) &= u_0 \text{ sur } \Omega, \end{cases} \tag{5.2}$$

Propriétés :

— solution pour tout temps, indéfiniement continuement dérivable

— EDO (temps), EDP elliptique (espace)

Equation des ondes : EQUATION HYPERBOLIQUE
(équation d'évolution)

$$\begin{cases} \frac{\partial^2 u}{\partial t^2} - c^2 \Delta u = f \text{ sur } \Omega \times [0,T] \\ u|_\Gamma = u_0 \\ \frac{\partial u}{\partial n}(.,t=0) = u_1 \text{ sur } \Gamma, \end{cases} \tag{5.3}$$

Propriétés :

— Solutions du type $u(x,t) = a(x-ct) + b(x+ct)$ quand $f = 0$

— Notion de propagation, de courbes caractéristiques

— Solutions discontinues possibles

5.2 Méthodes de différences finies pour les EDP elliptiques et paraboliques

Dans cette partie on suppose que les resultats d'existence et d'unicité de la solution sont acquis.

5.2.1 Méthodes de différences finies pour les EDP elliptiques

Soit Ω un ouvert du plan de frontière Γ rectangulaire, soit g une fonction continue définie sur Ω et soit l'équation de Laplace :

$$\begin{cases} -u_{xx} - u_{yy} = f \\ \text{Conditions aux limites sur } \Gamma. \end{cases} \tag{5.4}$$

Objectif

On introduit une grille de pas h de points $\{X_\lambda, \lambda \in \Lambda\}$, sur Ω. On cherche une méthode pour calculer un vecteur $U_h = \{U_\lambda, \lambda \in \Lambda\}^t \in I\!R^{\text{card}(\Lambda)}$ tel que :

$$||U_h - u(X_\lambda)|| \to 0 \quad \text{quand} \quad \begin{cases} h & \to & 0 \\ \text{erreur sur } f & \to & 0 \\ \text{erreur sur les conditions aux limites} & \to & 0 \end{cases}$$

C'est une propriété de **convergence**. Elle se définit par rapport à une norme dans $I\!R^{\text{card}(\Lambda)}$, généralement la norme $||.||_\infty$ ou la norme $||.||_2$. Une difficulté vient du fait que quand $h \to 0$, $\text{card}(\Lambda) \to +\infty$.

Base de la méthode :

On remplace les opérateurs différentiels par des opérateurs aux différences finies que l'on construit en utilisant les développements de Taylor (hypothèse de régularité).

On obtient **l'équation discrétisée** qui correspond à un système linéaire de la forme :

$$A_h U_h = F_h$$

— Stabilité :

Le schéma aux différences finies est stable si et seulement si les quantités $||A_h^{-1}||$ sont bornées indépendamment de h.

— Consistance : Si $\tilde{f}$ correspond à une évaluation, sans autre erreur que l'erreur de consistance, du membre de droite (incluant les conditions aux limites)) alors, le schéma aux différences finies est consistant si et seulement si pour $\mathcal{U}_h = (u(X_\lambda))^t$ et $\mathcal{F}_h = (\tilde{f}(X_\lambda))^t$ on a :

$$||A_h \mathcal{U}_h - \mathcal{F}_h|| \to 0.$$
$$h \to 0$$

Si

$$||A_h \mathcal{U}_h - \mathcal{F}_h|| \sim h^p$$
$$h \to 0$$

le schéma aux différences finies est dit d'ordre p.

Attention : La consistance de l'approximation complète tient compte du schéma utilisé pour approcher l' EDP ET de la prise en compte des conditions aux limites.

— Convergence :

Théorème de Lax

Si le schéma est stable et consistant alors il est convergent

Analyse du problème de Dirichlet en dimension 1 :

$$\begin{cases} -u_{xx} = f, x \in]0,1[\\ u(0) = \alpha, u(1) = \beta \end{cases}$$

On choisit $x_\lambda = \frac{\lambda}{N}, \lambda = 1, ...N$. On pose $h = \frac{1}{N}$. On obtient :

$$A_h = -\frac{1}{h^2}\begin{pmatrix} -2 & 1 & 0 & ... & ... & 0 \\ 1 & -2 & 1 & 0 & .. & 0 \\ 0 & 1 & -2 & 1 & ... & 0 \\ ... & ... & ... & ... & ... & ... \\ 0 & ... & 0 & 1 & -2 & 1 \\ 0 & ... & 0 & 0 & 1 & -2 \end{pmatrix}$$

$$F_h = \begin{pmatrix} f(\frac{1}{N}) + \frac{\alpha}{h^2} \\ f(\frac{2}{N}). \\ . \\ . \\ f(\frac{N-2}{N}) \\ f(\frac{N-1}{N}) + \frac{\beta}{h^2} \end{pmatrix}$$

et

$$||A_h\mathcal{U}_h - \mathcal{F}_h||_\infty \leq \frac{h^2}{12}||u^{(4)}||_\infty$$

De plus :

$$||A_h^{-1}||_\infty \leq \frac{1}{8}$$

Il y a consistance à l'ordre 2 et stabilité, donc convergence.

5.2.2 Méthodes de différences finies pour les EDP paraboliques

3.2.2.1 Cas général

0n s'intéresse au problème suivant, Ω étant un ouvert polyhédrique de $\mathbb{R}^m$ de frontière Γ :

$$\begin{cases} \frac{\partial u}{\partial t} + Au &= f \text{ sur } \Omega \times [0,T] \\ u|_\Gamma &= g \\ u(.,t=0) &= u_0 \text{ sur } \Omega. \end{cases}$$

Discrétisation en espace

On remplace l'opérateur différentiel en espace A par un opérateur aux différences finies A_h après introduction, comme dans le cas elliptique d'une grille de card(Λ) points et en tenant compte des conditions aux limites. Le système semi-discrétisé en espace s'écrit :

$$\begin{cases} \frac{dU_h}{dt} + A_hU_h &= f_h \text{ sur } [0,T] \\ U_h(t=0) &= (u_0)_h \end{cases}$$

c'est un système différentiel dont la solution est la fonction de t : $U_h(t) = \{U_\lambda(t)\}^t$.

Discrétisation en temps

On choisit une segmentation de $[0,T]$ et on utilise un schéma aux différences finies pour EDO (voir cours d'analyse numérique de 1A). Par exemple pour un schéma à un pas et une segmentation à pas constant Δt, on obtient à chaque pas de temps n un système linéaire qui s'écrit :

$$\left\{ \begin{array}{lcl} U_{h,n+1} & = & C_h U_{h,n} + F_{h,n} \\ U_{h,0} & = & (u_0)_h \end{array} \right.$$

— Stabilité : (Cas où C_h est indépendant de n)
Le schéma aux différences finies est stable si et seulement si les quantités $||(C_h)^n||$ sont bornées indépendamment de h dès que $n\Delta t \leq T$

— Consistance :
Le schéma aux différences finies est consistant si et seulement si quand on prend $\mathcal{U}_{h,n} = u(n\Delta t, x_\Lambda)$ alors si on appelle $\mathcal{V}_{h,n+1}$ le vecteur que fournit le schéma et si on suppose que u est régulière, on a :

$$\left\{ \begin{array}{cc} \Sigma_n ||\mathcal{V}_{h,n+1} - u((n+1)\Delta t, x_\Lambda)|| & \rightarrow 0 \\ h & \rightarrow 0 \\ \Delta t & \rightarrow 0 \\ \text{Erreurs (bord, C. init., f)} & \rightarrow 0 \end{array} \right.$$

Si

$$\left\{ \begin{array}{cc} \Sigma_n ||\mathcal{V}_{h,n+1} - u((n+1)\Delta t, x_\Lambda)|| & \sim h^p + (\Delta t)^q \\ h & \rightarrow 0 \\ \Delta t & \rightarrow 0 \\ \text{Erreurs (bord, C. init., f)} & \rightarrow 0 \end{array} \right.$$

le schéma est dit d'ordre p en espace et d'ordre q en temps.

— Convergence :
On dispose encore une fois d'un théorème de Lax :
Théorème de Lax
Si le schéma est stable et consistant alors il est convergent.

3.2.2.2 Schémas à pas constant (h) pour des EDP à coefficients contants posées sur $\mathbb{R}^m$. Etude de la stabilité pour la norme L^2.
On se place ici dans le cas scalaire, $m = 1$.
La discrétisation totale s'écrit $A_{h,\Delta t} U^{n+1} = B_{h,\Delta t} U^n + C^n$, où $A_{h,\Delta t}$ et $B_{h,\Delta t}$ sont des matrices telles que $[A_{h,\Delta t}]_{ij} = \alpha(j-i)$ et $[B_{h,\Delta t}]_{ij} = \beta(j-i)$.
On définit : $S(\omega) = \frac{\sum_{k\in\mathbb{Z}} \beta(k) e^{-ik\omega}}{\sum_{k\in\mathbb{Z}} \alpha(k) e^{-ik\omega}}$.
S s'appelle le **symbole** de l'opérateur $\hat{C}$.

Condition de stabilité de Von Neumann
Le schéma aux différences finies est L^2 stable si et seulement si :

$$\exists K \text{ tel que } \sup_{\omega\in\mathbb{R}} |S(\omega)| \leq 1 + K\Delta t$$

où la constante M est indépendante de h et de Δt.

Il se peut que la condition de stabilité de Von Neumann ne soit vérifiée que sous certaines conditions liant h et Δt, on dit alors que le schéma est conditionnellement stable.

Remarques

— L'étude de la stabilité par le critère de Von Neumann peut également s'effectuer pour des schémas aux différences finies appliqués à des opérateurs non paraboliques, pourvu que les opérateurs soient à coefficients constants et le schéma aux différences finies soit à pas constants.

— Dans $\mathbb{R}^m, m > 1$ ou dans le cas de schémas à plusieurs pas, le symbole devient un opérateur linéaire et le critère de Von Neumann porte sur le module de ses valeurs propres. C' est alors une condition nécessaire.

Bibliographie

[1] CIARLET, P.G. :"Introduction à l'analyse matricielle et à l'optimisation", Masson (1989)

[2] CROUZEIX, M., MIGNOT,A.L. :"Analyse numérique des équations différentielles", Masson (1989)

[3] DEMAILLY, J.P. :"Analyse numérique des équations différentielles", PUG (1991)

[4] LASCAUX,P., THEODOR,R. :"Analyse numérique matricielle appliquée à l'art de l'ingénieur", Masson (1994)

[5] LIANDRAT,J. :"Analyse numérique élémentaire", Cours polycopié ESM2 (2002)

[6] LIANDRAT,J. :"Compléments d'analyse numérique", Cours polycopié ESM2 (2002)

[7] SCHATZMANN, M. :"Analyse numérique" Intereditions 1991)

[8] ALLAIRE, G. :"Analyse numérique et optimisation", Editions de l'Ecole Polytechnique, 2005, www.cmap.polytechnique.fr/∼allaire/livre2.html (2012)