Résolution numérique des Équations Différentielles Ordinaires

L3 Mapi^3

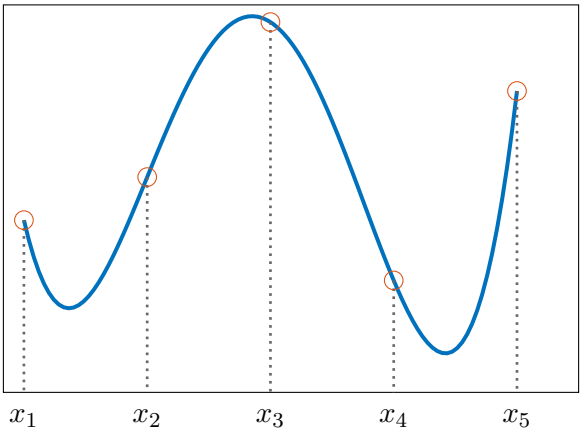
Christophe Besse

First printing, November 2016

Table des matières

1. Interpolation polynomiale

On dispose d'une série de couples de points (x_i, f_i), $i \in \{0, \cdots, n\}$. Le but de l'interpolation est de contruire un polynôme p qui prenne les valeurs f_i aux points x_i. Si on suppose que les valeurs f_i sont issues de l'évaluation d'une fonction f en x_i, nous tenterons de quantifier l'erreur $|f(t) - p(t)|$.



On ne présente ici que l'interpolation de Lagrange. Il en existe d'autres comme par exemple l'interpolation de Hermite qui outre les valeurs f_i s'interesse également aux valeurs de la dérivée en x_i.

Notations

- On note $\mathbb{P}_n$ l'ensemble des polynômes d'une variable (réelle ou complexe) de degré $\leq n$. $\mathbb{P}_n$ est un espace vectoriel de dimension $n+1$
- Soit $[a, b] \in \mathbb{R}$. On note $C^0([a, b])$ l'ensemble des fonctions continues sur $[a, b]$ et

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

- On note $C^m([a, b])$ l'ensemble des fonctions de classe C^m sur $[a, b]$.

1.1 Interpolation de Lagrange

On considère $n+1$ points distincts, pas nécessairement ordonnés $(x_0, x_1, \cdots, x_n)$ de $[a, b]$ et on considère une fonction $f \in C^0([a, b])$. Nous souhaitons répondre à la question

Existe-t-il un polynôme $p \in \mathbb{P}_n$ tel que $p(x_i) = f(x_i)$, $0 \leq i \leq n$?

Définition 1.1.1 — Polynômes de Lagrange. On définit les polynômes de Lagrange associés aux points $(x_0, \cdots, x_n)$ par

$$\begin{aligned} l_i(x) &= \frac{(x-x_0)\cdots(x-x_{i-1})(x-x_{i+1})\cdots(x-x_n)}{(x_i-x_0)\cdots(x_i-x_{i-1})(x_i-x_{i+1})\cdots(x_i-x_n)}, \quad 0 \le i \le n, \\ &= \prod_{0\le j\le n, j\neq i} \frac{(x-x_j)}{(x_i-x_j)}. \end{aligned} \tag{1.1}$$

On a $l_i(x_j) = \delta_{ij}$ $\forall (i,j) \in \{0, \cdots, n\}$ où δ_{ij} est le symbole de Kronecker, et degré$(l_i) = n$.

Proposition 1.1.1 $(l_0, \cdots, l_n)$ est une base de $\mathbb{P}_n$.

Preuve Cette famille est évidemment génératrice. Le point clé est de savoir si elle est libre. Soit $(a_0, \cdots, a_n) \in \mathbb{R}^{n+1}$ et $x \in \mathbb{R}$. Alors, si $\sum_{i=0}^n a_i l_i(x)$ pour tout x, on a pour tout j, $\sum_{i=0}^n a_i l_i(x_j) = 0$. Comme $l_i(x_j) = \delta_{ij}$, cela implique $a_j = 0$ pour tout j.

Ainsi, la famille est libre et génératrice et c'est donc une base. □

Théorème 1.1.2 Le problème : trouver $p \in \mathbb{P}_n$ tel que $p(x_i) = f(x_i)$, $\forall 0 \le i \le n$ admet une unique solution donnée par

$$p(x) = \sum_{i=0}^n f(x_i)\, l_i(x). \tag{1.2}$$

p s'appelle le polynôme d'interpolation de Lagrange, noté p_n.

Preuve

Existence : on vérifie aisément que le polynôme p donné par (1.2) répond à la question.
Unicité : soit $q \in \mathbb{P}_n$ tel que $q(x_i) = f(x_i)$, $\forall 0 \le i \le n$ et $r = p - q \in \mathbb{P}_n$. On a ainsi $r(x_i) = 0$ $\forall 0 \le i \le n$. Il existe donc un polynôme A tel que

$$\underbrace{r(x)}_{d^\circ\, n} = A(x) \underbrace{(x-x_0)(x-x_1)\cdots(x-x_n)}_{d^\circ\,(n+1)}.$$

Donc, si $A \neq 0$, r devrait être de degré $n+1$. La seule possibilité est que $A \equiv 0$ et donc $r = 0$. □

R On pose $\Pi_{n+1}(x) = (x-x_0)(x-x_1)\cdots(x-x_n)$. Alors

$$l_i(x) = \frac{\Pi_{n+1}(x)}{(x-x_i)\Pi'_{n+1}(x_i)}. \tag{1.3}$$

1.2 Étude de l'erreur d'interpolation et stabilité

En pratique, on commet systèmatiquement des erreurs car un ordinateur ne travaille qu'avec un nombre limité de chiffres significatifs. Il est donc important de connaître l'influence sur le résultat final des erreurs commises sur les données. On remplace (ici volontairement) les vraies valeurs $f(x_i)$ par des valeurs approchées f_i et on regarde l'incidence sur p_n. On note ce nouveau polynôme d'interpolation $\tilde{p}_n(x) = \sum_{i=0}^n f_i l_i(x)$. L'erreur commise est donc

$$\begin{aligned} |\tilde{p}_n(x) - p_n(x)| &= \left| \sum_{i=0}^n (f_i - f_i(x)) l_i(x) \right| \\ &\le \sum_{i=0}^n |f_i - f_i(x)|\ |l_i(x)| \\ &\le \max_i |f_i - f_i(x))| \sum_{i=0}^n |l_i(x)|. \end{aligned}$$

On note la constante de Lebesgue associée aux points $x_0, \cdots, x_n$

$$\Lambda_n = \max_{x\in[a,b]} \sum_{i=0}^{n} |l_i(x)|.$$

Ainsi,

$$\|\tilde{p}_n(x) - p_n(x)\|_\infty \leq \Lambda_n \max_i |f_i - f_i(x)|\,.$$

L'erreur commise sur les $f(x_i)$ est donc amplifiée (ou atténuée) par la constante de Lebesgue.

Proposition 1.2.1 On introduit l'application linéaire

$$\begin{array}{rccc} \mathscr{L}_n : & C^0([a,b]) & \to & \mathbb{P}_n \\ & f & \mapsto & p_n \end{array}$$

qui à $f \in C^0([a,b])$ associe son unique polynôme d'interpolation de Lagrange aux points $x_0, \cdots, x_n$. Alors, la norme de $\mathscr{L}_n$ est Λ_n, c'est à dire

$$|||\mathscr{L}_n||| := \sup_{\substack{f \in C^0([a,b]) \\ f \neq 0}} \frac{\|\mathscr{L}_n(f)\|}{\|f\|_\infty} = \Lambda_n.$$

Preuve on commence par montrer $|||\mathscr{L}_n||| \leq \Lambda_n$. On a

$$\begin{aligned} |\mathscr{L}_n(f)(x)| = |p_n(x)| &= \left|\sum_{i=0}^{n} f_i(x) l_i(x)\right| \\ &\leq \sum_{i=0}^{n} |f_i(x)|\ |l_i(x)| \\ &\leq \|f\|_\infty \sum_{i=0}^{n} |l_i(x)| \\ &\leq \Lambda_n \|f\|_\infty. \end{aligned}$$

Pour obtenir l'égalité, on se demande s'il existe une fonction $f \in C^0([a,b])$ telle que $\|\mathscr{L}_n(f)\|_\infty = \Lambda_n \|f\|_\infty$. Il n'est pas du tout sur qu'une telle fonction existe car le *sup* peut ne pas être atteint. Supposons que c'est le cas. Cela impliquerait

1. $\|f\|_\infty \sum_{i=0}^{n} |l_i(x)| = \Lambda_n \|f\|_\infty$ signifie que x est un point de maximum de la fonction $y \mapsto \sum_i |l_i(y)|$. Or, un tel point existe car cette fonction est continue sur $[a,b]$, intervalle fermé borné de $\mathbb{R}$.
2. $\sum_{i=0}^{n} \|f\|_\infty = \sum_{i=0}^{n} |f_i(x)|\ |l_i(x)|$ signifie que $|f(x_i)| = \|f\|_\infty$ pour tout i. On peut supposer que $\|f\|_\infty = 1$.
3. $\sum_{i=0}^{n} |f_i(x)|\ |l_i(x)| = \left|\sum_{i=0}^{n} f_i(x) l_i(x)\right|$ signifie que les $f_i(x)l_i(x)$ ont tous le même signe. On peut supposer que toutes ces quantités sont positives.

En combinant (2) et (3), on voit qu'on peut prendre $f(x_i) = 1$ si $l_i(x) \geq 0$ et $f(x_i) = -1$ si $l_i(x) < 0$. De plus, si on suppose les points ordonnés $x_0 < x_1 < \cdots < x_n$, on peut choisir f affine par morceaux (c'est à dire affine sur chaque intervalle $[x_i, x_{i+1}]$) et constante pour $x \leq x_0$ et $x \geq x_n$. Alors, on vérifie aisément que f satisfaite $\|\mathscr{L}_n(f)\|_\infty = \Lambda_n \|f\|_\infty$. □

Notre première estimation de l'erreur est donnée par

Théorème 1.2.2 Pour toute fonction $f : [a,b] \to \mathbb{R}$, on a

$$\|f - \mathscr{L}_n(f)\|_\infty \leq (1 + \Lambda_n) d(f, \mathbb{P}_n)$$

où $d(f,\mathbb{P}_n) = \inf_{q\in\mathbb{P}_n} \|f-q\|_\infty$.

Preuve : par unicité du polynôme d'interpolation, on a $\mathscr{L}_n(q) = q$, $\forall q \in \mathbb{P}_n$. On écrit alors

$$\begin{aligned}\|f - \mathscr{L}_n(f)\|_\infty &= \|f - q + \mathscr{L}_n q - \mathscr{L}_n(f)\|_\infty \\ &\leq \|f-q\|_\infty + \|\mathscr{L}_n(q-f)\|_\infty \\ &\leq \|f-q\|_\infty + \Lambda_n \|f-q\|_\infty.\end{aligned}$$

Le résultat en découle en prenant l'infimum. □

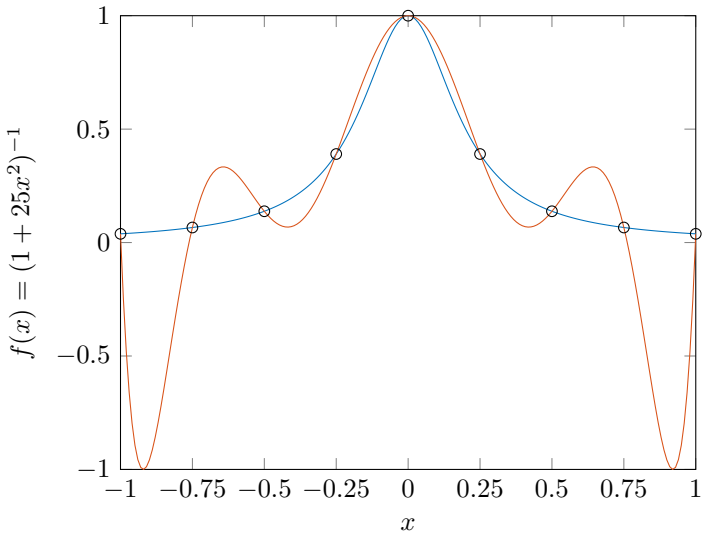
R Ce théorème est une forme indeterminée car on verra que $\lim_{n\to\infty}\Lambda_n = +\infty$ et que $\lim_{n\to\infty} d(f,\mathbb{P}_n) = 0$. En effet, d'après le théorème de Weierstrass, toute fonction continue sur $[a,b]$ est limite uniforme de polynômes.

■ **Exemple 1.1** — points équidistants : $x_i = a + i(b-a)/n$, $i = 0,1,\cdots,n$. Alors, si $x_i \in [-1,1]$, on a l'estimation $\Lambda_n \approx \dfrac{2^{n+1}}{e\, n \log(n)}$.

— points de Chebyshev : $x_i = \dfrac{a+b}{2} + \dfrac{b-a}{2}\cos\left(\dfrac{(2i+1)\pi}{2n+2}\right)$. Alors, $\Lambda_n \approx \dfrac{2}{\pi}\log(n)$.

■

■ **Exemple 1.2** *Effet de Runge*. Dans le cas des fonctions du type $g_a(x) = \frac{1}{1+a^2x^2}$ ou $h_a(x) = \frac{1}{a^2+x^2}$, on constate que l'interpolation via les points équidistants ne donne pas un résultat proche de la fonction interpolée. On constate des oscillations de grandes amplitudes près des bord (voir figure ci-dessous).



Fonction $f(x) = 1/(1+25x^2)$ et son interpolée.

■

Il est naturel de penser que l'erreur est meilleure dans le cas d'une fonction régulière (même si le phénomène de Runge existe aussi). C'est l'objet du résultat suivant

Théorème 1.2.3 Soit $f : [a,b] \to \mathbb{R}$ une fonction $(n+1)$ fois dérivable sur $]a,b[$ telle que f, f', $\cdots$, $f^{(n)}$ soient continues sur $[a,b]$. Alors, $\forall x \in [a,b]$, $\exists \xi_x \in]a,b[$ telle que

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

R — On voit que ce résultat dépend des valeurs des dérivées de f : c'est assez intuitif car plus une fonction oscille, plus elle est difficile à interpoler.

— $\frac{1}{(n+1)!}$ est de plus en plus petit quand $n \to \infty$.
— On peut interpréter la quantité Π_{n+1} comme une « mesure » de répartition des points $x_0, \cdots, x_n$ dans $[a,b]$. La meilleure interpolation est donc celle pour laquelle $\|\Pi_{n+1}\|_\infty$ est minimale.

Pour démontrer le théorème précédent, on a besoin du lemme de Rolle

Lemme 1.2.4 — (Rolle). Si $f \in C^1([a,b])$ et $f(a) = f(b) = 0$, $a \neq b$, alors $\exists \xi \in [a,b]$ telle que $f'(\xi) = 0$.

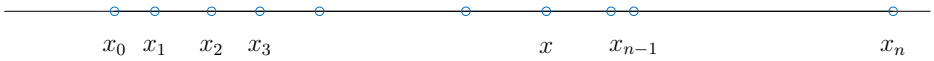
Preuve du théorème :
— Si x est l'un des x_i, les deux membres de l'égalité sont nuls et donc le résultat est acquis.
— Si $x \neq x_i$. La preuve se fait par récurrence sur n.
• $n = 0$ soit $x \neq x_0$. Le théorème fondamental de l'analyse nous dit que $\exists \xi \in]x_0, x[$ tel que

$$f(x) = \underbrace{f(x_0)}_{p_0(x)} + (x - x_0) f'(\xi).$$

Pour $n = 0$, $l_0(x) = 1$, $\Pi_1(x) = (x - x_0)$ et $p_0(x) = f(x_0)$. Ainsi

$$f(x) = p_0(x) = (x - x_0) f'(\xi) = \Pi_1(x) f'(\xi).$$

• $n > 0$ Soit p_n le polynôme de Lagrange associé à $(x_0, \cdots, x_n)$. On considère la fonction $\psi(y) = f(y) - p_n(y) - A\Pi_{n+1}(y)$ où A est une constante à déterminer. On a :
— $\psi(x_i) = 0$, $0 \leq i \leq n$ car $\Pi_{n+1}(x_i) = 0$ et $f(x_i) = p_n(x_i)$
— On choisit A telle que $\psi(x) = 0$. Ainsi, $A = \dfrac{f(x) - p_n(x)}{\Pi_{n+1}(x)}$. Cela est possible car comme $x \neq x_i$, $\Pi_{n+1}(x) \neq 0$.
Ainsi, ψ s'annule $(n+2)$ fois.



On a donc $\psi(y) = 0$ si $y = x$ ou $y = x_i$, $0 \leq i \leq n$. Si on applique le lemme de Rolle à l'intervalle $[x_0, x_1]$, alors il existe un point $x_0 < \xi_{01} < x_1$ tel que $\psi'(\xi_{01}) = 0$. En appliquant Rolle à chaque sous intervalle, on a que ψ' s'annule en $(n+1)$ points. En répétant ce processus, ψ'' s'annule en n points. Et ainsi de suite pour en conclure que $\psi^{(n+1)}$ s'annule en 1 point : $\exists \xi_x$ tel que $\psi^{(n+1)}(\xi_x) = 0$. Mais, on a

$$\psi^{(n+1)}(y) := f^{(n+1)}(y) - p_n^{(n+1)}(y) - A\Pi_{n+1}^{(n+1)}(y).$$

Comme $p_n \in \mathbb{P}_n$, $p_n^{(n+1)}(y) = 0$ et on a $\Pi_{n+1}^{(n+1)}(y) = (n+1)!$. En évaluant $\psi^{(n+1)}$ en ξ_x, on obtient

$$0 = f^{(n+1)}(\xi_x) - A(n+1)!$$

et donc

$$A = \frac{f^{(n+1)}(\xi_x)}{(n+1)!}.$$

En comparant les deux expressions de A, on a le résultat.

□

1.3 Calcul pratique du polynôme d'interpolation de Lagrange

Soit $p(x) = a_0 + a_1 x + \cdots + a_n x^n$. Lors de l'évaluation de ce polynôme, si on calcule chaque terme indépendemment, cela conduit à $(n-1)$ multiplications pour calculer successivement les x^k puis encore $(n-1)$ multiplications pour effectuer les produits avec les coefficients a_k et n additions ce qui au final

fait $2(n-1)$ multiplications et n additions. Une autre méthode consiste en l'utilisation de l'algorithme d'Horner

$$\begin{aligned} u_0 &= a_n x + a_{n-1} \\ u_1 &= x u_0 + a_{n-2} \\ u_2 &= x u_1 + a_{n-3} \\ &\vdots \end{aligned}$$

et $u_{n-1} = p_n(x)$. On fait donc n multiplications et n additions et on a donc un gain d'un facteur 2.

1.3.1 Différences divisées

On se donne une subdivision de $[a,b]$ par $x_0, \cdots, x_n$. On va calculer successivement $p_0, p_1, \cdots, p_n$ où $p_k \in \mathbb{P} - k$ est le polynôme associé à $x_0, \cdots, x_k$. On définit donc $p_0(x) = f(x_0)$.

On note $f[x_0, \cdots, x_k]$ le coefficient de plus haut degré de p_k. Comme $p_k - p_{k-1}$ est de degré k et s'annule en $x_0, \cdots, x_{k-1}$, on a

$$\begin{array}{ccccc} (p_k - p_{k-1})(x) & = & f[x_0, \cdots, x_k] & \Pi_k(x) \\ d^o \leq k & & \mathrm{C^{te}} & d^o(k). \end{array}$$

Ainsi $p_1(x) - p_0(x) = f[x_0, x_1]\Pi_1(x)$ et donc

$$p_1(x) = p_0(x) + f[x_0, x_1]\Pi_1(x) = f(x_0) + f[x_0, x_1]\Pi_1(x).$$

De même, $p_2(x) - p_1(x) = f[x_0, x_1, x_2]\Pi_2(x)$ et donc

$$p_2(x) = f(x_0) + f[x_0, x_1]\Pi_1(x) + f[x_0, x_1, x_2]\Pi_2(x).$$

Par récurrence, on obtient la **formule de Newton**

$$p_n(x) = f(x_0) + \sum_{k=1}^{n} f[x_0, \cdots, x_k]\Pi_k(x).$$

Il reste à calculer les $f[x_0, \cdots, x_k]$ de manière efficace.

Lemme 1.3.1 Pour tout $0 \leq i \leq n$, on a $f[x_i] = f(x_i)$ et pour $1 \leq k \leq n$,

$$f[x_0, \cdots, x_k] = \frac{f[x_1, \cdots, x_k] - f[x_0, \cdots, x_{k-1}]}{x_k - x_0}.$$

Ainsi, dès que l'on sait calculer des $f[\cdots]$ à k termes, on calcule facilement des $f[\cdots]$ à $k+1$ termes.
Preuve : Soit $q_{k-1} \in \mathbb{P}_{k-1}$ l'unique polynôme d'interpolation aux points $x_1, \cdots, x_k$. Par définition, le coefficient du terme dominant x^{k-1} est $f[x_1, \cdots, x_k]$. Soit

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

Mais, p_{k-1} et q_{k-1} appartiennent à $\mathbb{P}_{k-1}$. Ainsi, $\tilde{p}_k \in \mathbb{P}_k$ et

$$\tilde{p}_k(x_0) = -\frac{(x_0 - x_k)p_{k-1}(x_0)}{x_k - x_0} = f(x_0),$$

$$\tilde{p}_k(x_k) = \frac{(x_k - x_0)q_{k-1}(x_k)}{x_k - x_0} = f(x_k),$$

et pour $1 \leq j \leq k-1$,

$$\begin{aligned} \tilde{p}_k(x_j) &= \frac{(x_j - x_0)q_{k-1}(x_j) - (x_j - x_k)p_{k-1}(x_j)}{x_k - x_0}, \\ &= \frac{(x_j - x_0)f(x_j) - (x_j - x_k)f(x_j)}{x_k - x_0}, \\ &= f(x_j)\frac{(x_j - x_0) - (x_j - x_k)}{x_k - x_0} = f(x_j). \end{aligned}$$

Par unicité du polynôme d'interpolation, on a $\tilde{p}_k = p_k$. Mais

$$\begin{aligned} p_k(x) &= a_k x^k + \cdots \\ p_{k-1}(x) &= a_{k-1}x^{k-1} + \cdots \\ q_{k-1}(x) &= b_{k-1}x^{k-1} + \cdots \\ \tilde{p}_k(x) &= \frac{-a_{k-1} + b_{k-1}}{x_k - x_0}x^k + \cdots \end{aligned}$$

et donc $a_k = (b_{k-1} - a_{k-1})/(x_k - x_0)$ et on en conclut

$$f[x_0, \cdots, x_k] = \frac{f[x_1, \cdots, x_k] - f[x_0, \cdots, x_{k-1}]}{x_k - x_0}.$$

□

1.3.2 Algorithme de Horner

Pour calculer p_k en pratique, on calcule les $f[\cdots]$ puis on utilise la formule de Newton. On utilise le lemme pour déterminer la valeur de tous les $f[x_0, \cdots, x_k]$ qui sont obtenus en bout de lignes dans le tableau suivant. On les notes $T[k]$

$$\begin{array}{c|cccccc} T[0] & f(x_0) & & & & & \\ & & \searrow & & & & \\ T[1] & f(x_1) & \to & f[x_0,x_1] & & & \\ & & \searrow & & \searrow & & \\ T[2] & f(x_2) & \to & f[x_1,x_2] & \to & f[x_0,x_1,x_2] & \\ \vdots & \vdots & & & & & \\ \vdots & \vdots & \searrow & & \searrow & & \\ T[n-2] & f(x_{n-2}) & \to & f[x_{n-3},x_{n-2}] & \to & f[x_{n-4},x_{n-3},x_{n-2}] & \cdots \\ & & \searrow & & \searrow & & \\ T[n-1] & f(x_{n-1}) & \to & f[x_{n-2},x_{n-1}] & \to & f[x_{n-3},x_{n-2},x_{n-1}] & \cdots \\ & & \searrow & & \searrow & & \\ T[n] & f(x_n) & \to & f[x_{n-1},x_n] & \to & f[x_{n-2},x_{n-1},x_n] & \cdots \end{array}$$

et on reconstruit enfin le polynôme de Lagrange par la suite

$$\begin{aligned} U[n] &= T[n] \\ U[n-1] &= T[n-1] + (x - x_{n-1})U[n] \\ &\vdots \\ U[k] &= T[k] + (x - x_k)U[k+1] \\ &\vdots \\ U[0] &= p_n(x). \end{aligned}$$

■ **Exemple 1.3** On souhaite calculer le polynôme d'interpolation de la fonction $f(x) = x^2 + 1$ évaluée en $f(0) = 1$, $f(1) = 2$ et $f(2) = 5$. ■

Bien entendu, par unicité du polynôme d'interpolation de Lagrange, il faut que $p_2 = f$. Le tableau de calcul des $f[\cdots]$ est

$$\begin{array}{c|cccccc} T[0] & f(x_0) = 1 & & & & & \\ & & \searrow & & & & \\ T[1] & f(x_1) = 2 & \to & f[x_0,x_1] = \frac{2-1}{1} = 1 & & & \\ & & \searrow & & \searrow & & \\ T[2] & f(x_2) & \to & f[x_1,x_2] = \frac{5-2}{1} = 3 & \to & f[x_0,x_1,x_2] = \frac{3-1}{2} = 1. \end{array}$$

On a alors $U[2] = T[2] = 1$, puis $U[1] = T[1] + (x - x_1)U[2] = 1 + (x - 1)$ et enfin

$$U[0] = T[0] + (x - x_0)U[1] = 1 + x(1 + (x - 1)) = x^2 + 1 = p_2(x).$$

1.4 Exercices

Exercice 1.1 On note $l_i(x) = \prod_{j=0, j\neq i}^{n} \frac{x - x_j}{x_i - x_j}$, $0 \leq i \leq n$ les polynômes de Lagrange relatifs aux points $x_0, x_1, \cdots, x_n$ et $\pi_{k+1}(x) = \prod_{j=0}^{k}(x - x_j)$. Montre que $\forall\ 0 \leq i \leq n$, $\forall x \in \mathbb{R}$ et $x \neq x_i$

$$l_i(x) = \frac{\pi_{n+1}(x)}{(x - x_i)\pi'_{n+1}(x_i)}.$$

■

Exercice 1.2 **Interpolation de Lagrange**
Soit $f(x) = \cos(x)$.

1. Calculer le polynôme p_1 d'interpolation de Lagrange de f relativement aux points $x_0 = 0$ et $x_1 = \pi$.
2. Calculer le polynôme p_2 d'interpolation de Lagrange de f relativement aux points $x_0 = 0$, $x_2 = \frac{\pi}{2}$ et $x_1 = \pi$.
3. Montrer que $\forall\ x \in [x_0, x_1]$,
$$|f(x) - p_1(x)| \leq \frac{\pi^2}{8},$$
$$|f(x) - p_2(x)| \leq \frac{\sqrt{3}\pi^3}{216}.$$

■

Exercice 1.3 **Interpolation de Hermite**
L'objectif est de contruire un polynôme d'interpolation d'une fonction f en utilisant des valeurs de f et de sa dérivée f' aux points $(x_i)_{0\leq i\leq n}$ d'un intervalle $[a, b]$.

1. On définit les polynômes $H_i(x) = (1 - 2l'_i(x_i)(x - x_i))l_i^2(x)$ et $\tilde{H}_i(x) = (x - x_i)l_i^2(x)$. Montrer que $H_i(x_j) = \delta_{i,j}$, $H'_i(x_j) = 0$, et $\tilde{H}_i(x_j) = 0$,, $\tilde{H}'_i(x_j) = \delta_{i,j}$.
2. Soit p_n le polynôme de degré $2n + 1$ qui s'écrit
$$p_n(x) = \sum_{i=0}^{n} f(x_i)H_i(x) + \sum_{i=0}^{n} f'(x_i)\tilde{H}_i(x).$$
Montrer que p_n ainsi défini est l'unique polynôme d'interpolation (dit de Hermite) qui vérifie $p_n(x_i) = f(x_i)$ et $p'_n(x_i) = f'(x_i)$, $0 \leq i \leq n$.

■

Exercice 1.4 **Symétrie des différences divisées**
Soient $x_0, x_1, \cdots, x_n$ des points distincts d'un intervalle $[a, b]$.

1. Démontrer l'identité
$$f[x_0, \cdots, x_n] = \sum_{j=0}^{n} f(x_j) \prod_{k=0, k\neq j}^{n} \frac{1}{x_j - x_k}.$$
2. En déduire que la différence divisée $f[x_0, \cdots, x_n]$ est une fonction symétrique, c'est à dire que pour toute permutation σ
$$f[x_{\sigma(0)}, x_{\sigma(1)}, \cdots, x_{\sigma(n)}] = f[x_0, x_1, \cdots, x_n].$$

■

Exercice 1.5 **Convergence de l'interpolation de Lagrange**

Soient $\alpha > 1$, une fonciton f définie sur $[-1,1]$ par $f(x) = 1/(x-\alpha)$ et p_n le polynôme d'interpolation de Lagrange de f aux $n+1$ points distincts $x_i = -1+ih$, $i \in \{0,\cdots,n\}$ et $h = 2/n$.

1. Montrer que si $\alpha > 3$, alors $\lim_{n\to\infty} \|f - p_n\|_\infty = 0$.
2. Dans la pratique, nous préférons utiliser des polynômes de degré peu élevé sur chaque intervalle $[x_i, x_{i+1}]$, $i \in \{0,\cdots,n-1\}$. Notons f_n la fonction continue tel que $f_{n|[x_i,x_{i+1}]}$ est un polynômes de degré 1 et $f_n(x_i) = f(x_i)$ pour tout $0 \le i \le n$. On suppose ici que $\alpha \notin [-1,1]$.
   (a) Soit $i = 0,\cdots,n-1$. Écrire l'approximation de Lagrange de degré 1 $p_1^{(i)}$ de f sur l'intervalle $[x_i, x_{i+1}]$.
   (b) Montrer que $\|f - f_n\|_\infty \le C/n^2$ et donc que f_n converge uniformément vers f lorsque n tend vers l'infini.

■

2. Intégration numérique

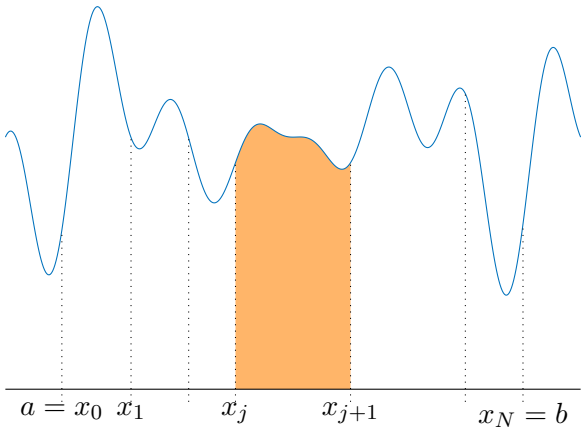
L'objectif de ce chapitre est connaissant une fonction $f \in C^0([a,b])$ réaliser un calcul approché de $\int_a^b f(x)dx$.

2.1 Formules de quadrature et leur ordre

On introduit une subdivision de $[a,b]$ en sous intervalles

$$a = x_0 < x_1 < x_2 < \cdots < x_n = b$$

et on utilise la relation de Chasles $\int_a^b f(x)dx = \sum_{j=0}^{N-1} \int_{x_j}^{x_{j+1}} f(x)dx$.



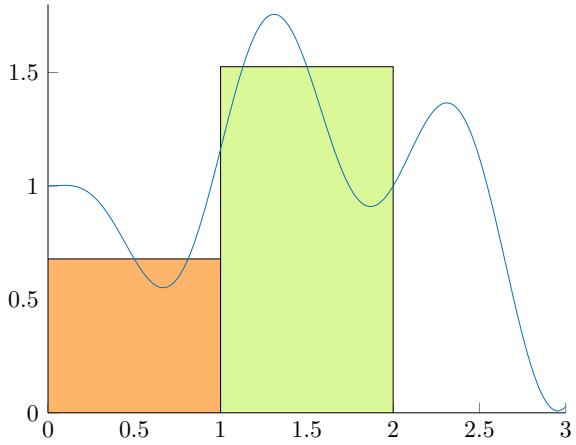
On est ramené au calcul d'intégrales sur des intervalles de longueur plus petite. On note $h_j = x_{j+1} - x_j$ la longueur de ces intervales et on fait le changement de variables $x = x_j + yh_j$

$$\int_{x_j}^{x_{j+1}} f(x)dx = h_j \int_0^1 f(x_j + yh_j)dy = h_j \int_0^1 g(y)dy$$

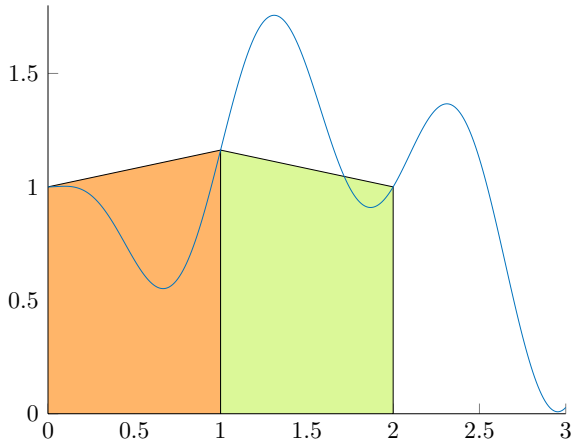
avec $g(y) = f(x_j + yh_j)$. On est finalement ramené au calcul approché de $\int_0^1 g(x)dx$.

■ **Exemple 2.1**

1 Formule du point milieu $\int_0^1 g(x)dx \approx g(\frac{1}{2})$.



2 Formule des trapèzes $\int_0^1 g(x)dx \approx \frac{1}{2}(g(0)+g(1))$.

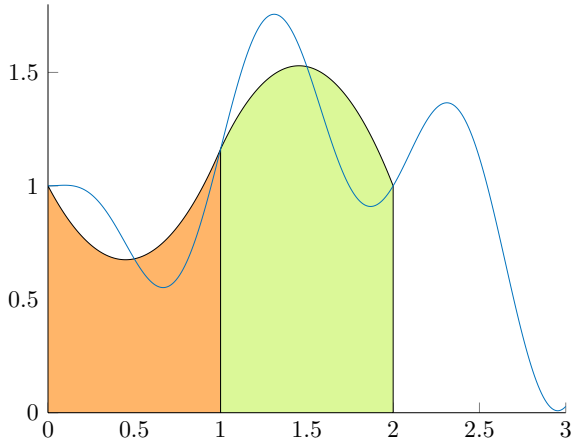


Si g est un polynôme de degré 1, c'est à dire $g(t) = \alpha + \beta t$, on a

$$\int_0^1 \alpha + \beta t dt = \alpha + \frac{\beta}{2}.$$

Mais $g(\frac{1}{2}) = \alpha + \frac{\beta}{2}$ et $\frac{1}{2}(g(0)+g(1)) = \alpha + \frac{\beta}{2}$. Ainsi, les formules du point milieu et des trapèzes sont exactes pour les polynômes de degré ≤ 1.

3 Formule de Simpson. On approche g par une parabole (polynôme de degré 2) sur $[0,1]$ qui passe par les points $(0, g(0))$, $(1/2, g(1/2))$ et $(1, g(1))$. C'est le polynôme d'interpolation de Lagrange. Alors $\int_0^1 g(x)dx \approx \frac{1}{6}\left(g(0) + 4g(\frac{1}{2}) + g(1)\right)$.



Si on remplace g par un polynôme de degré 2 : $g(t) = \alpha + \beta t + \gamma t^2$, on a $\int_0^1 g(t)dt = \alpha + \beta/2 + \gamma/3 = (g(0) + 4g(\frac{1}{2}) + g(1))/6$. Ainsi, la formule est exacte pour les polynôme de degré inférieur ou égal à 2. On aurait pu de douter de ce résultat car l'unique polynôme d'interpolation de Lagrange d'un polynôme est le polynôme lui même.

4 Généralisation On subdivise l'intervalle $[0, 1]$ avec s point équidistants et on remplace g par son polynôme d'interpolation de degré $(s-1)$ de Lagrange en ces points

$$\left(\frac{i}{s-1}, g\left(\frac{i}{s-1}\right)\right), \quad 0 \leq i \leq s-1.$$

Ce sont les **formules de Newton-Cotes**. ∎

Définition 2.1.1 Une formule de quadrature à s étages est donnée par

$$\int_0^1 g(t)dt \approx \sum_{i=1}^{s} b_i g(c_i) \tag{2.1}$$

Les c_i sont les nœuds de la formule de quadrature (point d'interpolation) et les b_i sont les poids.

R Au delà de $s = 10$, les poids "explosent", c'est à dire $b_i \gg 1$.

Définition 2.1.2 On dit que l'ordre de la formule de quadrature (2.1) est $p \in \mathbb{N}$ si la formule est exacte pour tous les polynômes de degré inférieur ou égale à $p-1$, c'est à dire

$$\int_0^1 g(t)dt = \sum_{i=1}^{s} b_i g(c_i), \quad \forall \deg(g) \leq p-1.$$

R

— on voit que les formules du point milieu et des trapèzes sont d'ordre 2. La formule de Newton-Cotes a pour ordre $p \geq s$.
— Une autre définition peut être donnée via l'erreur sur $[x_j, x_{j+1}]$

$$E_j(f) = \int_{x_j}^{x_{j+1}} f(x)dx - h_j \sum_{i=1}^{s} b_i f(x + j + h_j c_i).$$

Ainsi, (2.1) est d'ordre $p \in \mathbb{N}$ si et seulement si

$$\begin{cases} E_j(f) = 0, \ \forall f \in \mathbb{P}_{p-1} \\ \exists p \in \mathbb{P}_p, \ E_j(p) \neq 0 \end{cases}$$

Théorème 2.1.1 La formule de quadrature (2.1) a un order p si et seulement si

$$\sum_{i=1}^{s} b_i c_i^{q-1} = \frac{1}{q}, \quad \text{pour } q = 1, 2, \cdots, p.$$

Preuve

$\Rightarrow$ On a $\int_0^1 g(t)dt = \sum_{i=1}^{s} b_i g(c_i)$ pour tout polynôme g de degré inférieur ou égal à $p-1$. On prend donc $g(t) = t^{q-1}$. Ainsi, on a $\int_0^1 g(t)dt = \left[\frac{t^q}{q}\right]_0^1 = \frac{1}{q}$ et donc $\sum_{i=1}^{s} b_i c_i^{q-1} = \frac{1}{q}$.

$\Leftarrow$ Un polynôme de degré $p-1$ est une combinaison linéaire de $1, t, \cdots, t^{p-1}$. Or, les expressions $\int_0^1 g(t)dt$ et $\sum_{i=1}\hat{}sb_i g(c_i)$ sont linéaires en g. On remplace g par la combinaison linéaire d'où le résultat. □

R D'après ce théorème, pour qu'une méthode de quadrature de Newton-Cotes soit d'ordre positif ou nul, il faut et il suffit que $\sum_{i=1}^{s} b_i = 1$. Il suffit en effet de prendre $q = 1$ dans le théorème.

En fixant les nœuds $c_1, \cdots, c_s$ (distincts), la condition $\sum_{i=1}^{s} b_i c_i^{q-1} = \frac{1}{q}$ pour $q = 1, \cdots, p$ est un système linéaire pour les poids b_i

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ c_1 & c_2 & \cdots & c_s \\ \vdots & & & \\ c_1^{s-1} & c_2^{s-1} & \cdots c_1^{s-1} & \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_s \end{pmatrix} = \begin{pmatrix} 1 \\ 1/2 \\ \vdots \\ 1/s \end{pmatrix}. \tag{2.2}$$

La matrice de ce système linéaire est un matrice de Vandermonde dont on sait qu'elle est inversible. La résolution de ce système linéaire nous donne une formule de quadrature d'ordre $p \geq s$.

R Il existe une autre méthode pour trouver les poids. On connait les c_i, $1 \leq i \leq s$. Soit p_n le polynôme d'interpolation de g défini par $p_n(t) = \sum_{i=1}^{s} g(c_i) l_i(t)$. Alors,

$$\int_0^1 g(t)dt \approx \int_0^1 \sum_{i=1}^{s} g(c_i) l_i(t)dt = \sum_{i=1}^{s} g(c_i) \int_0^1 l_i(t)dt.$$

Ainsi, $b_i = \int_0^1 l_i(t)dt$.

Vérifions maintenant la formule $\sum_{i=1}^{s} b_i c_i^{q-1} = \frac{1}{q}$ pour la formule de Simpson. On a

$$b_1 = \frac{1}{6},\ b_2 = \frac{4}{6},\ b_3 = \frac{1}{6},\ c_1 = 0,\ c_2 = \frac{1}{2},\ c_3 = 1.$$

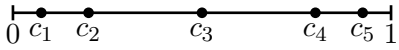
Il est évident que cette formule est vérifiée pour $q = 1, 2, 3$. Que se passe-t'il pour $q = 4$? On a

$$\frac{1}{6}0^3 + \frac{4}{6}\left(\frac{1}{2}\right)^3 + \frac{1}{6}1^3 = \frac{1}{4}.$$

Pour $q = 5$, ceci conduit à

$$\frac{1}{6}0^4 + \frac{4}{6}\left(\frac{1}{2}\right)^4 + \frac{1}{6}1^4 = \frac{5}{24} \neq \frac{1}{5}.$$

La méthode de Simpson, dont nous savions qu'elle était d'ordre 3, est en fait d'ordre 4 !. Elle est aussi exacte pour les polynômes de degré 3. Ceci est une propriété générale pour une formule symétrique.



Théorème 2.1.2 Une formule de quadrature symétrique avec s impair (c'est à dire $c_i = 1 - c_{s+1-i}$ et $b_i = b_{s+1-i}$ pour tout i) a toujours un ordre p pair. Ainsi, si elle est exacte pour les polynômes de degré inférieur ou égal à $2m-2$, elle est automatiquement exacte pour les polynômes de degré $2m-1$.

Preuve Soit g un polynôme de degré $2m-1$. On peut l'écrire

$$g(t) = C\left(t - \frac{1}{2}\right)^{2m-1} + \underbrace{g_1(t)}_{d^o \leq 2m-2}.$$

Il suffit donc de montrer que une formule symétrique est exacte pour $(t-1/2)^{2m-1}$. Or, c'est un polynôme symétrique par rapport à $t = 1/2$. Ainsi

$$\int_0^1 \left(t - \frac{1}{2}\right)^{2m-1} dt = 0.$$

Or, pour une formule symétrique, on a

$$b_i \left(c_i - \frac{1}{2}\right)^{2m-1} + b_{s+1-i}\left(c_{s+1-i} - \frac{1}{2}\right)^{2m-1} = 0.$$

L'approximation de $\int_0^1 \left(t - \frac{1}{2}\right)^{2m-1} dt = 0$ vaut donc aussi 0. □

2.2 Étude de l'erreur

On suppose pour simplifier qu'on a subdivisé $[a, b]$ en sous intervalles réguliers de longueur fixe h, soit $h = (b-a)/N$.

Définition 2.2.1 L'erreur globale d'une formule de quadrature est définie par

$$E(f) = \int_a^b f(x)dx - \sum_{j=0}^{N-1} h \sum_{i=1}^{s} b_i f(x_j + c_i h).$$

Commençons par étudier l'erreur commise sur un sous-intervalle de longueur h

$$\begin{aligned} E(f, x_0, h) &= \int_{x_0}^{x_0+h} f(x)dx - h\sum_{i=1}^{s} b_i f(x_0 + c_i h) \\ &= h\left(\int_0^1 f(x_0 + th)dt - \sum_{i=1}^{s} b_i f(x_0 + c_i h)\right). \end{aligned}$$

Rappel : soit f une fonction C^∞ au voisinage d'un point a, alors la série de Taylor de f est $\sum_0^n \frac{f^{(n)}(a)}{n!}(x-a)^n$. C'est une série entière qui a donc un rayon de convergence R.

■ **Exemple 2.2**
— $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$, $R = \infty$.
— $\ln(x) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} x^n$, $R = 1$.

■

Supposons que f soit suffisamment différentiable, on peut alors remplacer $f(x_0+th)$ et $f(x_0+c_ih)$ par leurs séries de Taylor (développées autour de x_0) sur leur rayon de convergence R_f. Alors

$$\begin{aligned} E(f,x_0,h) &= h\left(\int_0^1\left(\sum_{q\geq 0}\frac{f^{(q)}(x_0)}{q!}(th)^q\right)dt - \sum_{i=1}^{s} b_i\left(\sum_{q\geq 0}\frac{f^{(q)}(x_0)}{q!}(c_ih)^q\right)\right) \\ &= \sum_{q\geq 0}\frac{h^{q+1}}{q!}\left(\int_0^1 t^q dt - \sum_{i=1}^{s} b_i c_i^q\right) f^{(q)}(x_0). \end{aligned}$$

Comme la formule de quadrature est d'ordre p, on a

$$E(f,x_0,h) = \frac{h^{q+1}}{q!}\left(\frac{1}{p+1} - \sum_{i=1}^{s} b_i c_i^p\right) f^{(p)}(x_0) + O(h^{p+2}).$$

On appelle $C = \dfrac{1}{q!}\left(\dfrac{1}{p+1} - \displaystyle\sum_{i=1}^{s} b_i c_i^p\right)$ **constante d'erreur**.

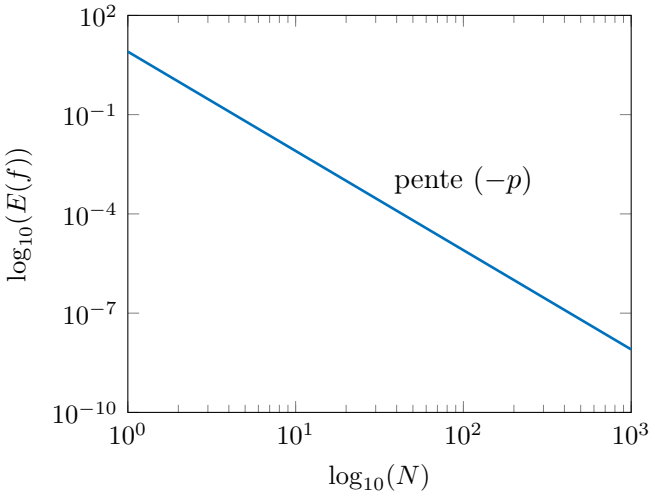
Si on suppose que h est petit de telle sorte que le terme $O(h^{p+2})$ soit négligeable par rapport au terme $Ch^{p+1}f^{(p)}(x_0)$, alors on a

$$\begin{aligned} E(f) &= \sum_{j=0}^{N-1} E(f,x_j,h) \\ &\approx Ch^p \sum_{j=0}^{N-1} hf^{(p)}(x_j) \\ &\approx Ch^p \int_a^b f^{(p)}(x)dx \\ &= Ch^p(f^{(p-1)}(b) - f^{(p-1)}(a)) \\ &= C^1 h^p \end{aligned}$$

Les formules précédentes permettent de voir que

$$\begin{aligned} \log_{10}(E(f)) &\approx \log_{10}(C^1) + p\log_{10}(h) \\ &\approx \log_{10}(C^1) + p\log_{10}(1/N) \\ &\approx \log_{10}(C^1) - p\log_{10}(N), \end{aligned}$$

où N est le nombre de points de discrétisation. On a donc une dépendance linéaire entre $\log_{10}(E(f))$ et $\log_{10}(N)$, avec une pente $-p$.



La formule d'erreur donnée ci-dessus est peu précise. On ne demande en effet que à f d'être suffisament dérivable. On donne donc le théorème suivant

Théorème 2.2.1 Considérons une formule de quadrature d'ordre p et $k \in \mathbb{N}$, $k \leq p$. Si $f : [x_0, x_0+h] \to \mathbb{R}$ est $k-$fois continuement différentiable $f \in C^k([x_0, x_0+h])$, on a

$$E(f, x_0, h) = h^{k+1} \int_0^1 N_k(\tau) f^{(k)}(x_0 + \tau h) d\tau$$

où $N_k(\tau)$, le noyau de Peano, est donné par

$$N_k(\tau) = \frac{(1-\tau)^k}{k!} - \sum_{i=1}^{s} b_i \frac{(c_i - \tau)_+^{k-1}}{(k-1)!}$$

où

$$(\sigma)_+^{k-1} = \begin{cases} \sigma^{k-1} \text{ si } \sigma > 0 \\ 0 \text{ si } \sigma \leq 0 \end{cases}$$

R Si on demande $f \in C^k([x_0, x_0+h])$, alors $N_{p+1}(0) = C$ la constante d'erreur.

Preuve On utilise la formule de Taylor avec reste intégral

$$f(x_0 + th) = \sum_{j=0}^{k-1} \frac{(th)^j}{j!} f^{(j)}(x_0) + h^k \int_0^t \frac{(t-\tau)^{k-1}}{(k-1)!} f^{(k)}(x_o + \tau h) d\tau.$$

On a

$$\begin{aligned} E(f, x_0, h) &= h \left[\int_0^1 f(x_0 + ht) dt - \sum_{i=1}^{s} b_i f(x_0 + c_i h) \right] \\ &= h \left[\underbrace{\sum_{j=0}^{k-1} \int_0^1 \frac{(th)^j}{j!} dt f^{(j)}(x_0)}_{A} + h^k \int_0^1 \int_0^t \frac{(t-\tau)^{k-1}}{(k-1)!} f^{(k)}(x_0 + \tau h) d\tau dt - \sum_{i=1}^{s} b_i f(x_0 + c_i h) \right] \end{aligned}$$

Or, pour $0 \leq t \leq 1$, on a $\int_0^t (t-\tau)^{(k-1)} g(\tau) d\tau = \int_0^1 (t-\tau)_+^{k-1} g(\tau) d\tau$ et

$$\sum_{i=1}^{s} b_i f(x_0 + c_i h) = \underbrace{\sum_{j=0}^{k-1} \sum_{i=1}^{s} b_i \frac{(c_i h)^j}{j!} f^{(j)}(x_0)}_{B} + \sum_{i=1}^{s} b_i h^k \int_0^{c_i} \frac{(c_i - \tau)^{k-1}}{(k-1)!} f^{(k)}(x_0 + \tau h) d\tau.$$

Comme la formule de quadrature est d'ordre $p \geq k$, on a $A - B = 0$. Il reste donc

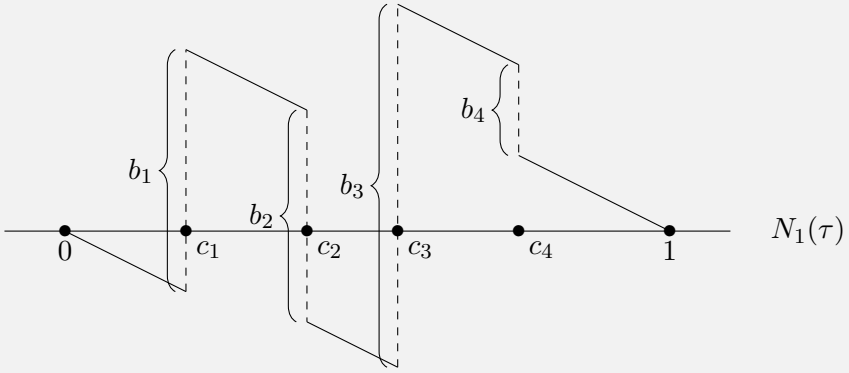
$$E(f, x_0, h) = h^{k+1} \int_0^1 \left(\underbrace{\int_0^1 \frac{(t-\tau)_+^{k-1}}{(k-1)!} dt}_{\frac{(1-\tau)^{k-1}}{k!}} - \sum_{i=1}^{s} b_i \frac{(c_i - \tau)_+^{k-1}}{(k-1)!} \right) f^{(k)}(x_0 + \tau h) d\tau$$

d'où le résultat □

Théorème 2.2.2 — Propriété du noyau de Peano. On considère une formule de quadrature d'ordre p et k un entier tel que $1 \leq k \leq p$. On a

1. $N_k'(\tau) = -N_{k-1}(\tau)$, $k \geq 2$ et $\tau \neq c_i$ quand $k = 2$
2. $N_k(1) = 0$, $k \geq 1$ si $c_i \leq 1$ $(1 \leq i \leq s)$
3. $N_k(0) = 0$, $k \geq 2$ si $c_i \geq 0$ $(1 \leq i \leq s)$

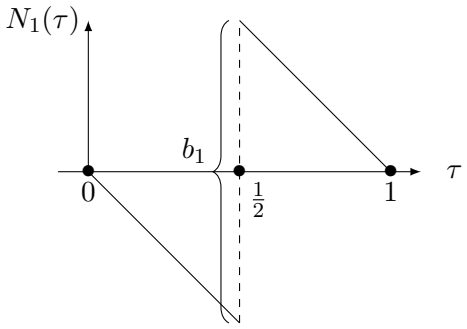
4. $\int_0^1 N_p(\tau)d\tau = \frac{1}{p!}(\frac{1}{p-1} - \sum_{i=1}^s b_i c_i^p) = C$, où C est la constante d'erreur
5. $N_1(\tau)$ est affine par morceaux, de pente -1, avec des sauts de hauteur b_i aux points c_i, $(1 \le i \le s)$



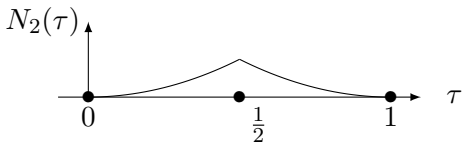
■ **Exemple 2.3**

Formule du point milieu $c_i = 1/2$, $b_1 = 1$. Alors,

$$N_1(\tau) = (1-\tau) - \underbrace{(1/2-\tau)_+^0}_{\begin{array}{ll}1 & \text{si } \tau < 1/2\\ 0 & \text{si } \tau \ge 1/2\end{array}} = \begin{cases} -\tau & \text{si } \tau < 1/2 \\ 1-\tau & \text{si } \tau \ge 1/2 \end{cases}$$



$$N_2(\tau) = \frac{(1-\tau)^2}{2} - \underbrace{(1/2-\tau)_+^1}_{\begin{array}{ll}1/2-\tau & \text{si } \tau < 1/2\\ 0 & \text{si } \tau \ge 1/2\end{array}} = \begin{cases} \tau^2/2 & \text{si } \tau < 1/2 \\ (1-\tau)^2/2 & \text{si } \tau \ge 1/2 \end{cases}$$



■

R On peut également définir le noyau de PEano comme

$$E\left(t \mapsto \frac{(t-\tau)_+^{k-1}}{(k-1)!}, 0, 1\right), \quad 0 \le \tau < 1.$$

On peut maintenant facilement estimer l'erreur globale

$$E(f) = \int_a^b f(x)dx - \sum_{j=0}^{N-1} h_j \sum_{i=1}^{s} b_i f(x_j + c_i h_j).$$

Théorème 2.2.3 Soit $f : [a,b] \to \mathbb{R}$, $f \in C^k([a,b])$ et une formule de quadrature égalem à p ($p \geq k$). Alors, on a

$$|E(f) \leq h^k(b-a) \int_0^1 |N_k(\tau)| d\tau \cdot \underbrace{\max_{x\in[a,b]} |f^{(k)}(x)|}_{\|f^{(k)}\|_{L^\infty([a,b])}}$$

où $h = \max_j h_j$.

Preuve On sait par définition que

$$E(f, x_0, h) = h^{k+1} \int_0^1 N_k(\tau) f^{(k)}(x_0 + \tau h) d\tau.$$

Ainsi

$$\begin{aligned} |E(f, x_0, h)| &\leq h^{k+1} \int_0^1 |N_k(\tau)||f^{(k)}(x_0 + \tau h)| d\tau \\ &\leq h^{k+1} \int_0^1 |N_k(\tau)| d\tau \ \|f^{(k)}\|_\infty. \end{aligned}$$

Si on somme par rapport à j, on a

$$|E(f)| \leq \sum_{j=0}^{N-1} |E(f, x_j, h_j)| \leq \sum_{j=0}^{N-1} \underbrace{h_j^{k+1}}_{\leq h^k h_j} \int_0^1 |N_k(\tau)| d\tau \underbrace{\max_{x\in[x_j,x_{j+1}]} |f^{(k)}(x)|}_{\leq \|f^{(k)}\|_{L^\infty([a,b])}}.$$

Utilisant le fait que $\sum_{j=0}^{N-1} h_j = b - a$ termine la démonstration. □

2.3 Formules d'ordre supérieur

On a vu que si on fixe les c_i, on obtient de manière unique les poids b_i pour une formule d'ordre $p \geq s$ (résolution du système de Vandermonde). La question naturelle est : peut-on optimiser les c_i pour que l'ordre soit augmenté ?

Théorème 2.3.1 Soit $(b_i, c_i)_{1\leq u\leq s}$ une formule de quadrature d'ordre p *geqs* et

$$\Pi_s(t) = (t - c_1) \cdots (t - c_s).$$

Alors, l'ordre est supérieur ou égal à $s + m$ si et seulement si

$$\int_0^1 \Pi_s(t) g(t) dt = 0, \quad \forall g \in \mathbb{P}_{m-1}.$$

Preuve Soit $f = in\mathbb{P}_{s+m-1}$, alors,

$$f(t) = \Pi_s(t) \underbrace{g(t)}_{d^o g \leq m-1} + \underbrace{r(t)}_{d^o r \leq s-1}.$$

Ainsi,

$$\int_0^1 f(t) dt = \int_0^1 \Pi_s(t) g(t) dt + \int_0^1 r(t) dt$$

et pour l'approximation

$$\sum_{i=1}^s b_i f(c_i) = \sum_{i=1}^s b_i \underbrace{\Pi_s(c_i)}_{=0} g(c_i) + \sum_{i=1}^s b_i r(c_i).$$

Comme la formule est exacte pour $r(t)$ (ordre supérieur ou égal à s par hypothèse), elle est exacte pour $f(t)$ si et seulement si $\int_0^1 \Pi_s(t) g(t) dt = 0$. □

■ **Exemple 2.4** — pour qu'une formule de quadrature à $s = 3$ étages ait un ordre supérieur ou égal à 4, il faut

$$0 = \int_0^1 (t - c_1)(t - c_2)(t - c_3)dt = \frac{1}{4} - (c_1 + c_2 + c_3)\frac{1}{3} + (c_1c_2 + c_1c_3 + c_2c_3)\frac{1}{2} - c_1c_2c_3$$

ce qui donne

$$c_3 = \frac{1/4 + (c_1 + c_2)/3 + c_1c_2/2}{1/3 - (c_1 + c_2)/2 + c_1c_2}.$$

— si $s = 3$, pour avoir $p = 6$, il faut vérifier trois conditions

$$\int_0^1 \Pi_s(t)g(t)dt = 0 \text{ pour } g(t) \in \Pi_{m-1}, \quad m = 1, 2, 3.$$

■

Théorème 2.3.2 Si p est l'ordre d'une formule de quadrature à s étages, alors $p \geq 2s$.

2.4 Polynômes orthogonaux de Legendre

Si on a $s = 3$ et qu'on veut $p = 6$ (c'est à dire $p = 2s$), on a vu qu'il faut vérifier les trois conditions

$$\int_0^1 \Pi_s(t) \begin{pmatrix} 1 \\ t \\ t^2 \end{pmatrix} dt = 0.$$

Cela conduit à des calculs longs et fastidieux. Pour rendre les calculs plus simples et symétriques, on fait le changement de variables $\tau = 2t - 1$. Comme $t \in [0, 1]$, on a $\tau \in [-1, 1]$. Le problème est donc : trouver $p_k(\tau) \in \mathbb{P}_k$, $\forall k \in \mathbb{N}^*$ tel que

$$\int_{-1}^1 p_k(\tau)g(\tau)d\tau \text{ où } deg(g) \leq k - 1.$$

On va utiliser les polynômes orthogonaux de Legendre. La relation d'orthogonalité entre polynômes est donnée par

$$< p_k, p_j >= \int_{-1}^1 p_k(\tau)p_j(\tau)d\tau = \delta_{kj}.$$

Les polynômes de Legendre sont donnés par

$$p_k(\tau) = \frac{1}{2^k k!}\frac{d^k}{d\tau^k}\left((\tau^2 - 1)^k\right).$$

Alors, le polynôme $p_s(2t - 1)$ joue le rôle de $\Pi_s(t)$ pour avoir l'ordre $p = 2s$. On a par exemple

— $p_0(\tau) = 1$, $p_1(\tau) = \tau$, $p_2(\tau) = \frac{3}{2}\tau^2 - \frac{1}{2}$.
— $p_k(\tau) = p_k(-\tau)$ si k est pair.
— $p_k(\tau) = -p_k(-\tau)$ si k est impair.

On a aussi la formule de récurrence

$$(k + 1)p_{k+1}(\tau) = (2k + 1)\tau p_k(\tau) - kp_{k-1}(\tau), \quad \forall k \geq 1.$$

Les racines de $p_k(\tau)$ sont réelles, simples et dans $]-1, 1[$.

2.5 Formule de quadrature de Gauss

On cherche des formules d'ordre $p = 2s$ avec $\Pi_s(t) = cp_s(2t - 1)$ où p_s est le polynôme de Legendre de degré s. On a

$$\int_0^1 p_s(2t - 1)g(2t - 1)dt = \frac{1}{2}\int_{-1}^1 p_s(\tau)g(\tau)d\tau = 0, \quad \deg(g) \geq s - 1.$$

Comme les racines de $p_s(2t - 1) \in]0, 1[$ et sont réelles, on a le théorème suivant.

Théorème 2.5.1 — Gauss 1814. $\forall s \in \mathbb{N}^*$, il existe une unique formule de quadrature à s étages d'ordre $p = 2s$. Elle est donnée par $c_1, \cdots, c_s$ racines de $p_s(2t-1)$ et les b_i sont calculés par le système linéaire de Vandermonde.

■ **Exemple 2.5** — $s = 1$: on obtient la formule du point milieu $\int_0^1 g(t)dt \approx g(1/2)$

— $s = 2$: on obtient la formule du point milieu $\displaystyle\int_0^1 g(t)dt \approx \frac{1}{2}g(\frac{1}{2} - \frac{\sqrt{3}}{6}) + \frac{1}{2}g(\frac{1}{2} + \frac{\sqrt{3}}{6})$.

■

2.6 Exercices

Exercice 2.1 Soit $(b_i, c_i)_{i=1}^s$ une formule de quadrature d'ordre $\geq s$. Montrer que

$$b_i = \int_0^1 l_i(x)\,dx, \quad \text{où } l_i(x) = \prod_{j=1, j\neq i}^{s} \frac{x - c_j}{c_i - c_j}.$$

■

Exercice 2.2 Montrer que si les nœuds d'une formule de quadrature à s étages satisfont $c_i = 1 - c_{s+1-i}$ (pour tous i) et si la formule a un ordre $p \geq s$, alors on a nécessairement $b_i = b_{s+1-i}$, c'est à dire la formule est symétrique. ■

Exercice 2.3 Calculer les formules de Newton-Cotes pour

$$(c_i) = (0, 1/3, 2/3, 1), \quad (c_i) = (0, 1/4, 2/4, 3/4, 1)$$

et déterminer l'ordre de ces formules de quadrature.
`Indication` : les calculs se simplifient en utilisant l'exercice 2. ■

Exercice 2.4 Formule de Radau
Déterminer c_2, b_1, b_2 dans la formule de quadrature

$$\int_0^1 g(t)dt \approx b_1 g(0) + b_2 g(c_2)$$

afin que son ordre soit maximal. ■

Exercice 2.5 Noyau de Péano

1. On considère la formule de quadrature des rectangles à gauche sur l'intervalle $[0, 1]$. Calculer le noyau de Péano de cette quadrature.
2. Calculer le noyau de Péano de la quadrature par la formule des trapèzes.
3. On veut estimer $I(f) = \int_0^1 f(t)dt$. On pose pour $f \in C^1([0, 1])$ la formule

$$J(f) = w_0 f(0) + w_1 f'(0) + w_2 f'(\xi), \quad \xi \in]0, 1[, \ w_i \in \mathbb{R}.$$

   (a) Déterminer ξ et les w_i pour $J(t) = I(t)$ tout polynôme de degré ≤ 3.
   (b) Soit $E(f) = I(f) - J(f)$. Calculer $E(x \to x^4)$ et en déduire l'ordre de la méthode.
   (c) Déterminer le noyau de Péano

■

Exercice 2.6 1. Donner les trois polynômes p_1, p_2 et p_3 de degré 2 tels que

$$\begin{cases} p_1(0) = 1, \\ p_1'(0) = 0, \\ p_1(1) = 0, \end{cases} \quad \begin{cases} p_2(0) = 0, \\ p_2'(0) = 1, \\ p_2(1) = 0, \end{cases} \quad \begin{cases} p_3(0) = 0, \\ p_3'(0) = 0, \\ p_3(1) = 1. \end{cases}$$

et montrer qu'ils forment une base de P_2.

2. Soit $f \in C^3([0,1])$. Montrer qu'il existe un unique polynôme $p \in P_2$ tel que

$$p(0) = f(0), \quad p'(0) = f'(0), \quad p(1) = f(1),$$

et l'exprimer dans la base (p_1, p_2, p_3).

3. Calculer $J = \int_0^1 p(x)dx$.

4. Soit $g(t) = f(t) - p(t) - \frac{f(x)-p(x)}{x^2(x-1)} t^2(t-1)$. Montrer qu'il existe $\xi \in]0,1[$ tel que $g'''(\xi) = 0$ et en déduire une majoration de $\|f - p\|_\infty$.

5. Soit $I = \int_0^1 f(x)dx$. Donner une estimation de l'erreur $I - J$.

■

Exercice 2.7 On considère la formule de quadrature

$$\int_0^\pi f(x)\sin(2x)\,dx = w_1 f(x_1) + w_2 f(x_2) + E(f).$$

1. Déterminer w_1, w_2, x_1 et x_2 de telle sorte que la formule soit exacte pour les polynômes de P_3.
2. Vérifier qu'en réalité cette formule est encore exacte pour les polynômes de degré 4.

■

Exercice 2.8 Soit $f : \mathbb{R} \to \mathbb{R}$, donné par

$$f(x) = a_0 + \sum_{k=1}^{m} (a_k \cos(kx) + b_k \sin(kx)), \quad a_k, b_k \in \mathbb{R}.$$

1. Quelle est la valeur exacte de $\int_0^{2\pi} f(x)dx$?
2. Appliquer la règle des trapèzes à $\int_0^{2\pi} f(x)dx$ avec $h = 2\pi/N$. À partir de quelle valeur de N le résultat est-il exact ?
3. Déterminer une formule de quadrature à $s = 3$ étages d'ordre 6.
4. Appliquer la formule précédente au calcul de $\int_0^{2\pi} f(x)dx$ et répondez à la même question que sous (2).
5. Quelle formule de quadrature proposez-vous pour l'intégration numérique d'une fonction périodique ?

■

3. EDO - Introduction

Lors de la découverte des équations différentielles ordinaires (que l'on notera dorénavant EDO) dans les études supérieures, elles peuvent souvent être résolues avec un papier et un crayon. Cependant, et c'est ce qui motive l'utilisation de méthodes numériques, la plupart des EDOs ne peuvent être résolues explicitement en terme de fonctions simples. De la même manière que l'on ne peut pas donner de formule explicite à $\int_a^b e^{-t^2} dt$, il est généralement impossible de résoudre une EDO explicitement.

On commence donc ce chapitre en présentant quelques modèles simples d'EDOs.

■ **Exemple 3.1** On considère l'EDO

$$x'(t) = \sin(t) - x(t). \tag{3.1}$$

La solution de l'équation homogène $x' = -x$ est donnée par $x_H(t) = Ae^{-t}$. L'application de la méthode de la variation de la constante conduit à $x = A(t)e^{-t}$ et donc $A'(t) = \sin(t)e^t$. En intégrant cette relation on a

$$A(t) = C + e^t(\sin(t) - \cos(t))$$

d'où la solution générale

$$x(t) = Ce^{-t} + \frac{1}{2}(\sin(t) - \cos(t))$$

où C est une constante arbitraire.

En revanche, si l'on considère

$$x'(t) = \sin(t) - 0.1x^3(t), \tag{3.2}$$

on n'a pas accès à une solution analytique. Pourtant, les solutions se ressemblemt de manière remarquable (voir les figures 3.1 et 3.2) ■

De manière générale, une EDO linéaire du type

$$x'(t) = \lambda(t)x(t) + f(t)$$

a pour solution

$$x(t) = Ag(t) + g(t) \int_0^1 \frac{f(s)}{g(s)} ds$$

avec

$$g(t) = \exp\left(\int_0^t \lambda(s) ds\right).$$

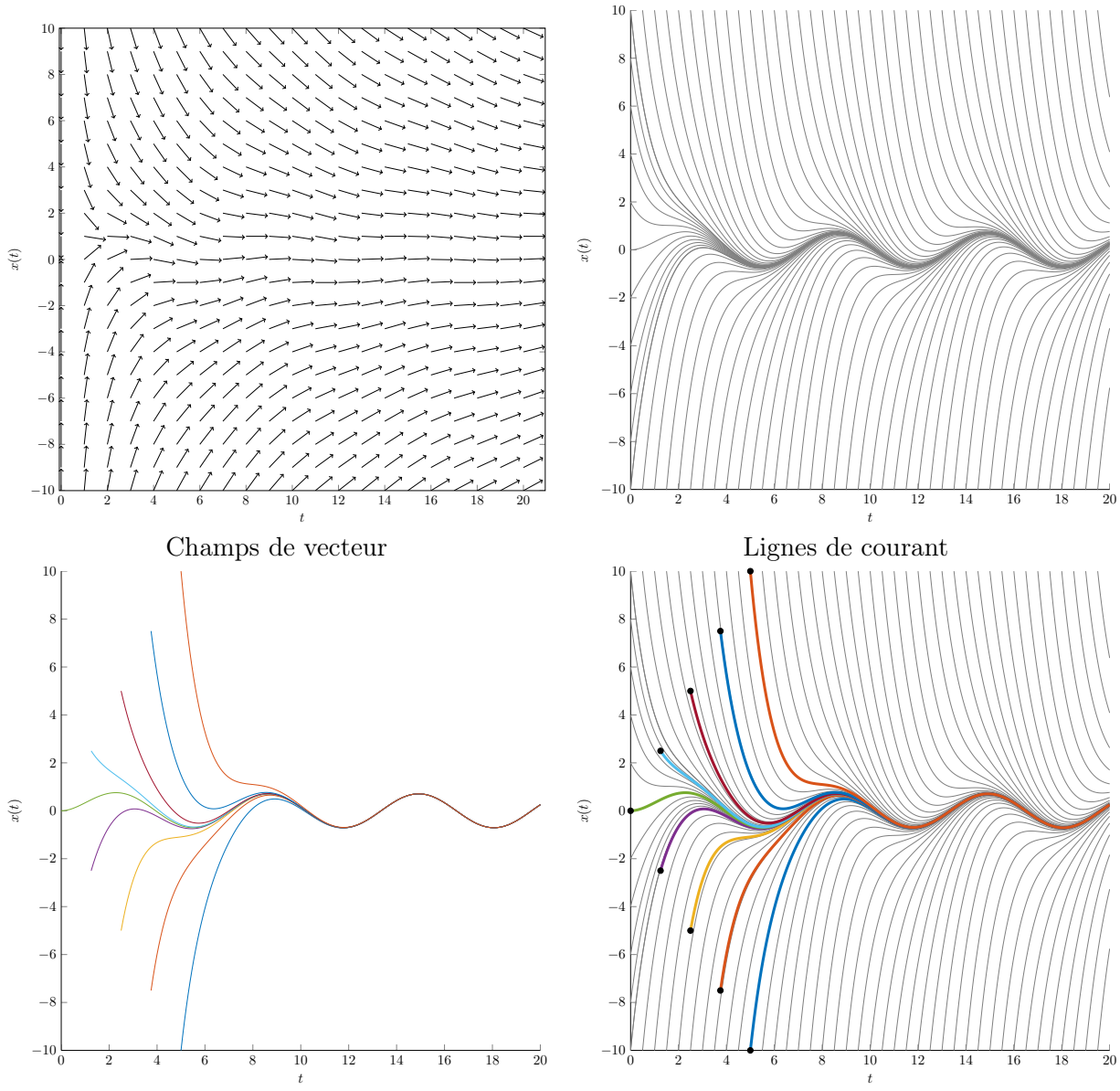


FIGURE 3.1 – Représentation des solutions de l'équation (3.1)

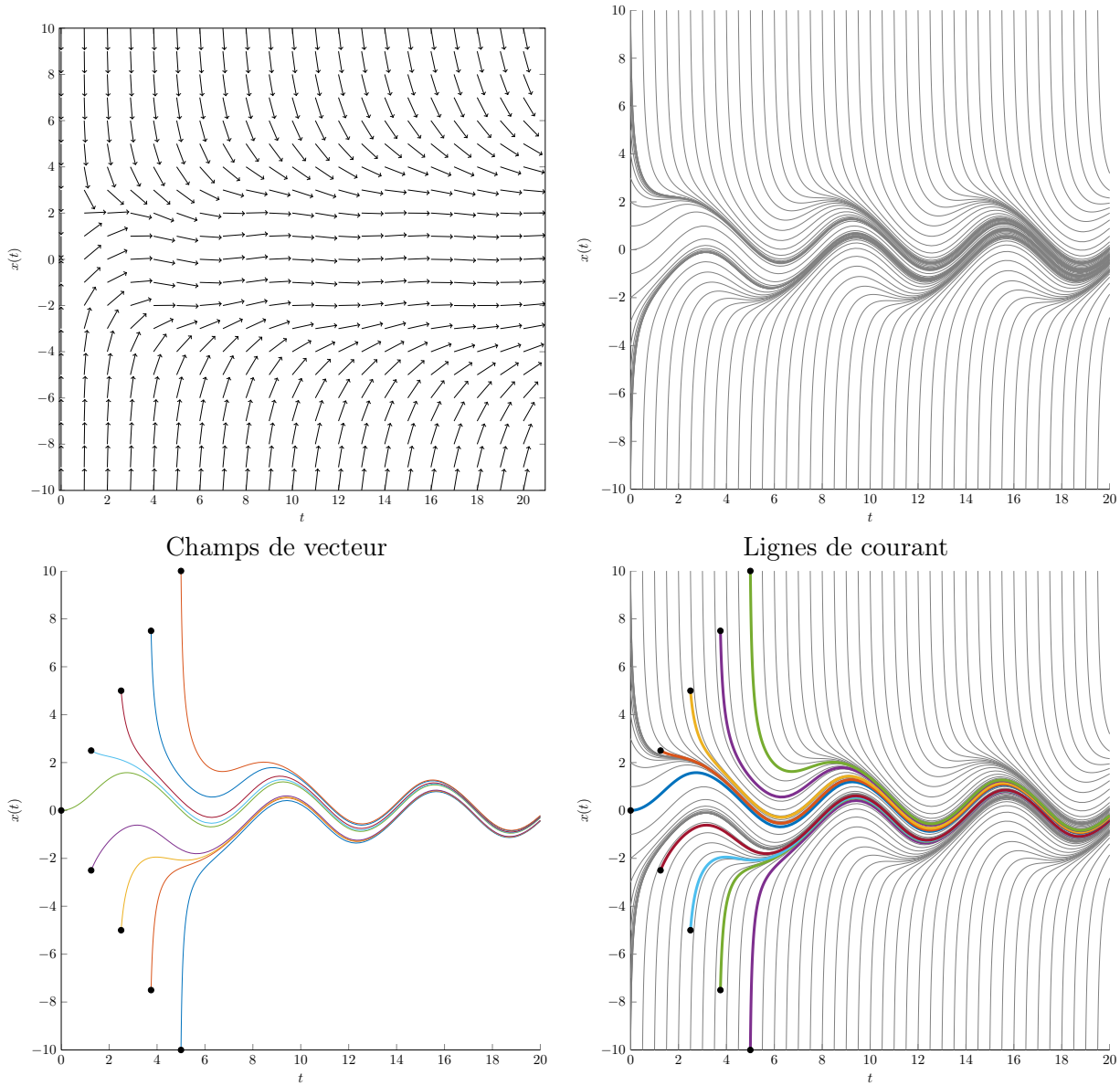


FIGURE 3.2 – Représentation des solutions de l'équation (3.2)

On s'interesse dans ce cours à la résolution d'EDO du premier ordre du type

$$\begin{cases} x'(t) = f(t, x(t)), & t \leq t_0 \\ x(t_0) = x_I \end{cases} \tag{3.3}$$

On supposera toujours que le problème (3.3) a une solution dans un intervalle $[t_0, t_f]$, $t_f > t_0$. La théorie est vue dans le cours d'analyse des équations différentielles ordinaires.

Bien que la forme (3.3) semble faire référence à une équation scalaire, elle peut également être appliquée aux systèmes d'EDOs. En effet, considérons

$$\begin{cases} u'(t) = p(t, u, v), & u(t_0) = \eta_0, \\ v'(t) = q(t, u, v), & v(t_0) = \eta_1. \end{cases}$$

Si on définit $\mathbf{x}(t) = (u, v)^T$, $\mathbf{f}(t, \mathbf{x}) = (p(t, u, v), q(t, u, v))^T$, $\boldsymbol{\eta} = (\eta_0, \eta_1)^T$, alors le système d'EDO s'écrit

$$\begin{cases} \mathbf{x}'(t) = \mathbf{f}(t, \mathbf{x}(t)), & t \leq t_0 \\ \mathbf{x}(t_0) = \boldsymbol{\eta} \end{cases}$$

Ainsi, $\mathbf{x}$, $\boldsymbol{\eta}$ sont des éléments de $\mathbb{R}^2$ et $\mathbf{f} : \mathbb{R} \times \mathbb{R}^2 \to \mathbb{R}^2$. Si le système contient m équations, alors, $\mathbf{x} \in \mathbb{R}^m$ et $\mathbf{f} : \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}^m$.

Les EDOs dans lesquelles le temps n'apparait pas explicitement sont dites autonomes. Par exemple, $x'(t) = x(t)(1 - x(t))$ est autonome alors que $x'(t) = (1 - 2t)x(t)$ ne l'est pas. On peut toujours réécrire une EDO non autonome en un système autonome. Par exemple, pour $x'(t) = (1-2t)x(t)$ avec $x(t_0) = \eta$, on pose $\mathbf{x}(t) = (t, x(t))^T$ et $\mathbf{f}(\mathbf{x}(t)) = (1, (1 - 2t)x(t))^T$ et on a

$$\begin{cases} \mathbf{x}'(t) = \mathbf{f}(\mathbf{x}(t)), & t \leq t_0 \\ \mathbf{x}(t_0) = (t_0, \eta_0)^T \end{cases}$$

■ **Exemple 3.2 — Équation de Lotka-Volterra.** Le modèle de Lotka-Volterra (1925) donne un modèle simple de conflit entre des populations de proies et de prédateurs. Supposons par exemple que les nombres de lapins et de renards dans une certaine région au temps t sont $u(t)$ et $v(t)$. Alors, si on a des populations initiales $u(0)$ et $v(0)$ au temps initial $t_0 = 0$, leurs nombres peuvent évoluer selon le système autonome

$$\begin{cases} u'(t) = 0,05u(t)(1 - 0,01v(t)), \\ v'(t) = 0,1u(t)(0,005u(t) - 2). \end{cases}$$

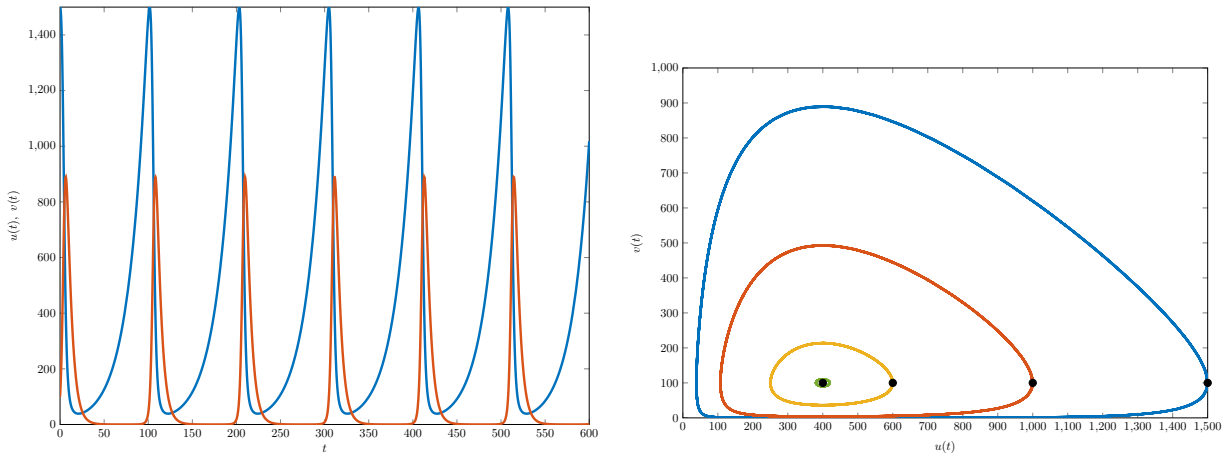
Ce système reproduit un certain nombre de comportements prévisibles :

- si le nombre de renard $v(t)$ augmente, alors plus de lapins sont mangés et donc on a décroissance du taux $u'(t)$ auquel ils se reproduisent (moins d'individus)
- si le nombre de lapins $u(t)$ augmente, alors un plus grand nombre de lapins sont susceptibles de se rencontrer et de se reproduire, le taux auquel les lapins se reproduisent augmente donc.

Si on se donne une population initiale de 1500 lapins et de 100 renards, on voit sur la figure 3.3-(a) l'évolution de $u(t)$ et $v(t)$ pour $t = 0$ à $t = 600$. Sur la figure 3.3-(b), on a trois courbes dans le plan (u, v) pour les données initiales respectives $[600, 100]$, $[1000, 100]$ et $[1500, 100]$ qui donnent un idée du portrait de phase. On constate un comportement périodique qui se déduit du fait que ces courbes sont fermées avec un centre de rotation en $[400, 100]$. ■

■ **Exemple 3.3 — Poursuite renard-lapin.** Les courbes de poursuite apparaissent naturellement dans des scénarios militaires ou de proies-prédateurs. Imaginons qu'un lapin suive un chemin prédéfini $(r(t), s(t))$ dans le but de perturber les intentions du renard. Supposons également que le renard se déplace à une vitesse qui est un facteur constant k fois la vitesse du lapin. On suppose de plus que le renard poursuit le lapin de telle sorte que à tout temps sa trajectoire soit tangente à celle du lapin. Alors, on peut montrer que la trajectoire du renard satisfait $(x(t), y(t))$ avec

$$\begin{cases} x'(t) = R(t)(r(t) - x(t)), \\ y'(t) = R(t)(s(t) - y(t)), \end{cases}$$



(a) Courbes solutions (b) Portrait de phase

FIGURE 3.3 – Solutions de l'équation de Lotka-Volterra

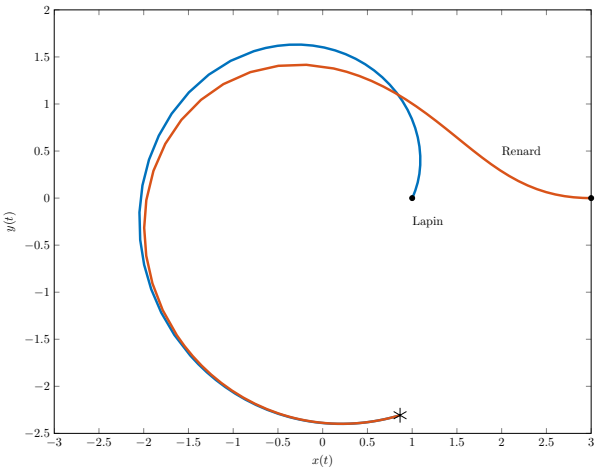


FIGURE 3.4 – Scénario de poursuite

où

$$R(t) = \frac{k\sqrt{r'(t)^2 + s'(t)^2}}{\sqrt{(r(t) - x(t))^2 + (s(t) - y(t))^2}}.$$

Si r et s sont connus pour tout temps, alors c'est un système d'EDO à $m = 2$ composantes. Si le lapin suit une trajectoire en spirale,

$$\begin{pmatrix} r(t) \\ s(t) \end{pmatrix} = \sqrt{1+t} \begin{pmatrix} \cos(t) \\ \sin(t) \end{pmatrix},$$

on a la solution donnée représentée sur la figure 3.4 où on résoud jusqu'à $T = 5,0710$. À ce temps précis, le renard rattrape le lapin et l'EDO devient mal posée (division par 0 dans la définition de $R(t)$). ■

■ **Exemple 3.4 — Propagation de zombies.** Les EDOs sont souvent utilisées en épidémiologie et en dynamique des populations pour décrire la dissémination d'une maladie. Imaginons ici une invasion de zombies (cela fonctionne aussi avec la propagation d'un virus). A chaque temps, on enregistre

— $H(t)$: la population d'humains,
— $Z(t)$: la population de zombies,
— $R(t)$: la population de zombies supprimés qui peuvent revenir après un certain temps des zombies.

Il est supposé qu'un zombie puisse convertir irrémédiablement un humain en zombie. D'un autre côté, les zombies ne peuvent pas être tués, mais un humain courageur peut temporairement envoyer un

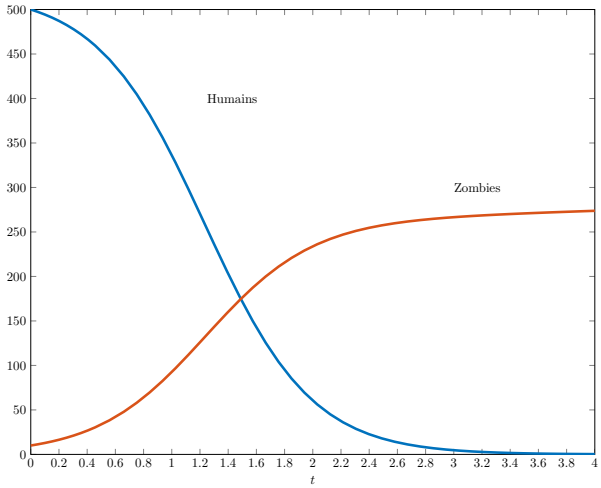


FIGURE 3.5 – Évolution de deux populations

zombie dans un état "supprimé". Le modèle le plus simple est

$$\begin{cases} H'(t) = -\beta H(t)Z(t), \\ Z'(t) = \beta H(t)Z(t) + \zeta R(t) - \alpha H(t)Z(t), \\ R'(t) = \alpha H(t)Z(t) - \zeta R(t), \end{cases}$$

où α, β et ζ sont trois constantes positives représentants
- α : une rencontre humain-zombie qui supprime des zombies
- β : une rencontre humain-zombie qui convertit un humain en zombie
- ζ : zombies supprimés qui retournent au statut zombie.

On représente sur la figure 3.5 la solution pour $\beta = 0,01$, $\alpha = 0,005$, $\zeta = 0,02$, $H(0) = 500$, $Z(0) = 10$ et $R(0) = 0$. ■

De nombreux modèles mettent en jeu des dérivées d'ordre plus élevé, par exemple les équations de Newton pour décrire le mouvement. On peut transformer ces équations en système du premier ordre.

■ Exemple 3.5 — Oscillateur de Van der Pol.

$$\begin{cases} x''(t) + 10(1 - x^2(t))x'(t) + x(t) = \sin(\pi t), \\ x(t_0) = \eta_0, \ x'(t_0) = \eta_1. \end{cases}$$

On pose $u = x$ et $v = x'$. Alors, l'équation du second ordre devient

$$\begin{cases} u'(t) = v(t), \\ v'(t) = -10(1 - u^2(t))v'(t) + u(t) + \sin(\pi t), \end{cases}$$

avec $u(t_0) = \eta_0$ et $v(t_0) = \eta_1$. Il sufft de poser $\mathbf{x}(t) = (u, v)^T$ et

$$\mathbf{f}(t, \mathbf{x}(t)) = \begin{pmatrix} v(t) \\ -10(1 - u^2(t))v'(t) + u(t) + \sin(\pi t) \end{pmatrix}$$

pour transformer l'équation du second ordre en un système d'EDOs du premier ordre. ■

Par extension, si on considère $x^{(m)}(t) = f(t, x(t), x'(t), \cdots, x^{(m-1)}(t))$, on pose

$$\begin{array}{ll} x_1(t) = x(t), & x_1'(t) = x_2(t), \\ x_2(t) = x'(t), & x_2'(t) = x_3(t), \\ \vdots \quad \text{et on a} & \vdots \\ x_m(t) = x^{(m-1)}(t) & x_{m-1}'(t) = x_m(t). \end{array}$$

Avec

$$\mathbf{x}(t) = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix}, \quad \mathbf{f}(t, \mathbf{x}(t)) = \begin{pmatrix} x_2(t) \\ x_3(t) \\ \vdots \\ f(t, x_1(t), x_2(t), \cdots, x_m(t)) \end{pmatrix}, \quad \boldsymbol{\eta} = \begin{pmatrix} x(0) \\ x'(0) \\ \vdots \\ x^{(m-1)}(0) \end{pmatrix}$$

l'équation d'ordre m est trasnsformée sous la forme du système $\mathbf{x}(t) = \mathbf{f}(t, \mathbf{x}(t))$.

4. La méthode d'Euler

On souhaite comme on l'a vu au chapitre précédent calculer les solutions des problèmes (3.3) de la forme

$$\begin{cases} x'(t) = f(t, x(t)), & t \geq t_0 \\ x(t_0) = \eta \end{cases}$$

qui possède une solution pour $t \in [t_0, t_f]$. On introduit une discrétisation de l'intervalle $[t_0, t_f]$ à l'aide du pas de maillage $h > 0$: $h = (t_f - t_0)/N$ et $t_n = t_0 + nh$, $0 \leq n \leq N$. On va chercher des approximations en ces temps discrets de la suite de nombres $x(t_0)$, $x(t_0 + h)$, $x(t_0 + 2h)$, ..., $x(t_0 + nh)$.

4.1 Exemples

Soit $h = 0,3$ et on considère l'EDO

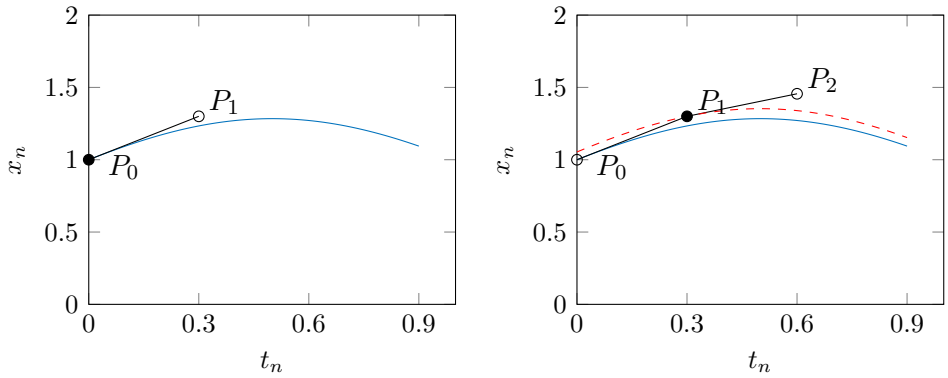
$$\begin{cases} x'(t) = (1 - 2t)x(t), & t \in [0, 0.9], \\ x(t_0) = 1. \end{cases}$$

La solution exacte est $x(t) = \exp(1/4 - (1/2 - t)^2)$. De manière générale, on a $x(t) = x(t_0)\exp(t_0^2 - t_0)\exp(t - t^2)$. Le fait de connaitre une solution exacte nous permet de juger du degré de précision de nos approximations. En $t = 0$, on a $x(0) = 1$ et donc $x'(0) = 1$. Cette information nous permet de construire la tangente à la courbe solution en $t = 0$ qui est ainsi d'équation $x = 1 + t$. On passe donc du point P_0 au point P_1 en évaluant x à l'extrémité de la tangente en $t = h$ (voir Figure 4.1-(a)). Donc, $x(t_0 + h) = x(0.3) \approx 1 + h = 1.3$.

On reproduit le même calcul avec comme donnée initiale $x(h) = 1.3$. La tangente en h a donc comme taux d'accroissement $x'(h) = (1 - 2h)x(h) = 0.52$ et son équation est $x = 1.3 + 0.52t$. On évalue la tangente en $t = 2h$ et on a $x(2h) \approx 1.3 + 0.52h = 1.456$. On peut ainsi placer le point P_2 (voir Figure 4.1-(b)). En reproduisant le même calcul, on obtient le point P_3 (Figure 4.2).

On organise ainsi les résultats aux temps $t_n = nh$ et on note x_n les approximations des valeurs exactes $x(t_n)$ et on a $x'_n = (1 - 2t_n)x_n$. En résumé, on a

$n = 0$	$t_0 = 0$ $x_0 = 1$ $x'_0 = 1$	$n = 1$	$t_1 = t_0 + h = 0.3$ $x_1 = x_0 + hx'_0 = 1.3$ $x'_1 = (1 - 2t_1)x_1 = 0.52$
$n = 2$	$t_2 = y_1 + h = 0.6$ $x_2 = x_1 + hx'_1 = 1.456$ $x'_2 = (1 - 2t_2)x_2 = -0.2912$	$n = 3$	$t_3 = t_2 + h = 0.9$ $x_3 = x_2 + hx'_2 = 1.3686$ $x'_3 = (1 - 2t_3)x_3 = 2.0949$



(a) (b)

FIGURE 4.1 – Évolution par la méthode d'Euler

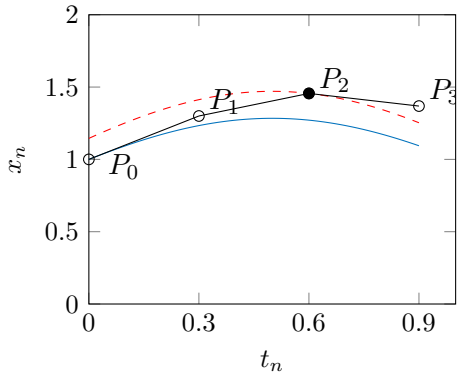


FIGURE 4.2 – Évolution par la méthode d'Euler

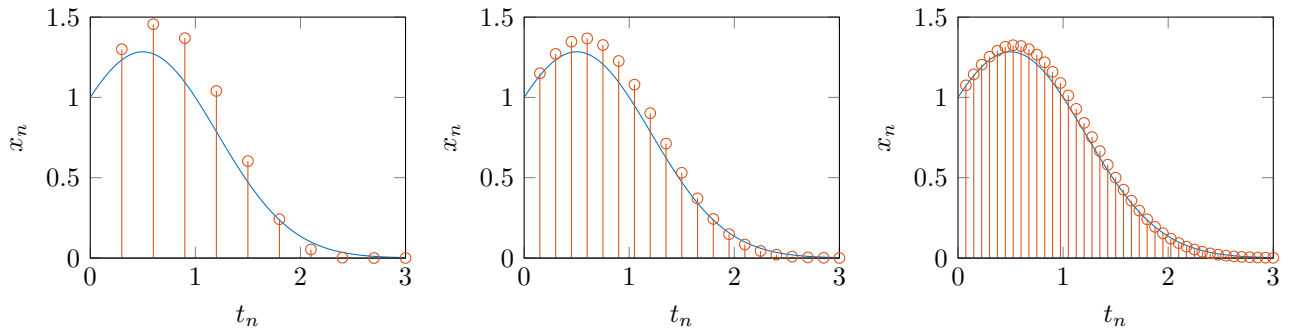
Les points $P + n(t_n, x_n)$ sont représentés sur les figures 4.1 et 4.2. Pour tout n, la ligne qui lie P_n à P_{n+1} est tangente à la solution de l'EDO qui passe par les points $x(t) = x_n$ et $t = t_n$. Comme on le voit, les points P_n sont à bonne distance de la courbe solution exacte de l'EDO. On voit cependant qu'en réduisant $h = 0.3$, 0.15 et 0.075, les points s'approchent de plus en plus près de la solution exacte (voir Figures 4.3). Ceci illustre la notion que la solution numérique *converge* vers la solution exacte quand $h \to 0$. Afin d'avoir une preuve plus concrète, on calcule la différence entre la solution numérique x_n et $x(t_n)$ au temps t_f (ici 0.9). On évalue ainsi l'erreur globale en $t = t_n$ par $e_n = x(t_n) - x_n$. On a l'évolution suivante

h	x_n	Erreur globale	e_n/h
0.3	$x_3 = 1.3686$	$x(0.9) - x_3 = -0.2745$	-0.91
0.15	$x_6 = 1.2267$	$x(0.9) - x_6 = -0.1325$	-0.89
0.075	$x_9 = 1.1591$	$x(0.9) - x_9 = -0.0649$	-0.86
Exacte	$x(0.9) = 1.0942$		

La table suggère donc que e_n soit proportionnelle à h et que la constante de proportionalité soit environ 0.9 : $e_n \approx -0.9h$ quand $nh = 0.9$. Ainsi, si on désire une précision de trois chiffres après la virgule, on va chercher h tel que $|e_n| < 0.0005$ ce qui se traduit par $h < 0.0005/0.9$ soit $h < 0.00055$. Par conséquent, pour intégrer jusqu'à $t = 0.9$ prendra environ $n = 0.9/h \approx 1620$ étapes. Ainsi, réduire $|e_n|$ par 10 nécessite donc de diviser h par 10 et donc multiplier n par 10.

Pour étudier l'erreur de façon générique, on utilisera les notations de Landau dont on rappelle ci-dessous le principe. On écrit $z = \mathcal{O}(h^p)$ si il existe deux constantes positives h_0 et C telles que

$$|z| \leq Ch^p, \quad \forall\ 0 < h < h_0.$$



$h = 0.3$ $h = 0.15$ $h = 0.075$

FIGURE 4.3 – Évolution par la méthode d'Euler

On dit que z est d'ordre h^p. Par exemple, on sait que

$$e^h = 1 + h + \frac{h^2}{2!} + \frac{h^3}{3!} + \cdots + \frac{1}{n!}h^n + \cdots$$

et donc on a

$$\begin{aligned} e^h &= 1 + \mathcal{O}(h) \\ &= 1 + h + \mathcal{O}(h^2) \\ &= 1 + h + h^2/2 + \mathcal{O}(h^3). \end{aligned}$$

$\mathcal{O}(h^3)$ est toujours plus petit que $\mathcal{O}(h^2)$: cela permet de comparer les tailles des différents termes.

4.2 Le cas général

On sait que si $x \in C^{p+1}([t_0, t_f])$,

$$x(t+h) = x(t) + hx'(t) + \frac{h^2}{2}x''(t) + \cdots + \frac{h^p}{p!}x^{(p)}(t) + \frac{h^{p+1}}{(p+1)!}x^{(p+1)}(\xi), \quad \xi \in]t, t+h[.$$

Ainsi, $x(t+h) = x(t) + hx'(t) + R_1(t)$ où R_1 est le terme de reste appelé erreur locale de troncature (ELT) et on a

$$R_1(t) = \frac{h^2}{2}x''(\xi), \quad \xi \in]t, t+h[.$$

Si il existe une constante positive M telle que $|x''(t)| \leq M$ pour tout $t \in]t_0, t_f[$, on a $|R_1(t)| \leq Mh^2/2$ et donc $R_1(t) = \mathcal{O}(h^2)$.

Pour obtenir la méthode d'Euler, on substitue $x'(t) = f(t, x)$ dans la série de Taylor

$$x(t+h) = x(t) + hf(t, x(t) + R_1(t)$$

et en introduisant les nœuds de la grille $t_n = t_0 + nh$, $n = 0, \cdots, N$, où N, définie comme la partie inférieure de $(t_f - t_0)/h$, désigne la nombre de pas de longueur h pour atteindre (sans dépasser) $t = t_f$. Ainsi, pour tout $n < N$, on a

$$\begin{cases} x(t_{n+1}) = x(t_n) + hf(t_n, x(t_n)) + R_1(t_n), \\ x(t_0) = \eta. \end{cases}$$

On a donc une méthode pour calculer les approximations x_n et $x(t_n)$ si on supprime le terme (ELT). En effet, le terme $R_1(t) = \mathcal{O}(h^2)$ peut être rendu aussi petit que l'on veut en diminuant h. La méthode d'Euler est ainsi

$$\begin{cases} x_{n+1} = x_n + hf(t_n, x_n) := x_n + hf_n, \quad n = 0, \cdots, N-1 \\ x(t_0) = \eta. \end{cases}$$

R On a vu que l'on peut écrire la solution de l'EDO

$$\begin{cases} x'(t) = f(t, x(t)), & t \geq t_0 \\ x(t_0) = \eta \end{cases}$$

comme

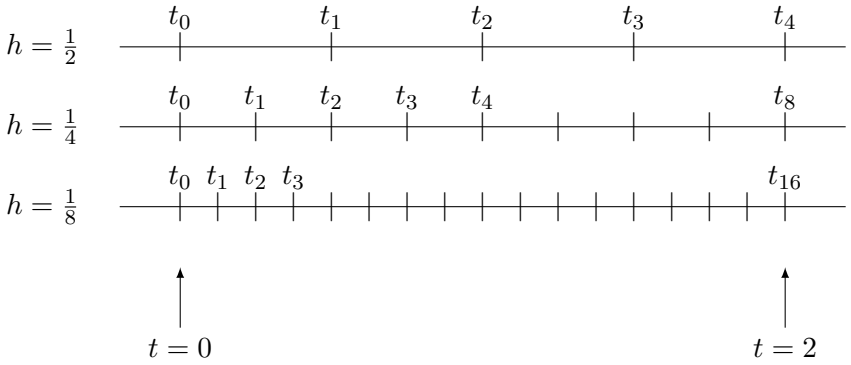
$$x(t) = \eta + \int_{t_0}^{t} f(s, x(s))ds.$$

Ainsi, $x(t+h) = x(t) + \int_t^{t+h} f(s, x(s))ds$ et en approchant l'intégrale par la méthode de quadrature des rectangles à gauche, on retrouve la méthode d'Euler.

4.3 Analyse de la méthode

On souhaite conprendre le comportement de $\lim_{h\to 0} x_n$. On peut espérer que $|e_n| = |x(t_n) - x_n|$ décroisse. Cependant, si $h \to 0$, à n^* fixé, on a $t_{n^*} = t_0 + n^*h \to t_0$. Par exemple, si $h = 1/2$, $1/4$ et $1/8$, on a respectivement $t_4 = t_0 + 2$, $t_0 + 1$, $t_0 + 0.5$. Ainsi, t_4 se rapproche de t_0 quand $h \to 0$. On doit donc comparer $x(t_n)$ et x_n au même temps $t = t^*$ fixe

$$t_{n^*} = t_0 + n^*h = t^* \iff n^* = \frac{t^* - t_0}{h}.$$



Pour le premier exemple que l'on a traité dans ce chapitre, on a vu que e_n était proportionnel à h quand $nh = 0.9$. Cela suggère que $t^* = 0.9$.

Définition 4.3.1 On appelle erreur locale de troncature ε_n définie par

$$\varepsilon_n = x(t_{n+1}) - x(t_n) - h_n f(t_n, x(t_n)).$$

Définition 4.3.2 On appelle erreur globale de convergence

$$e_n = x_n - x(t_n).$$

Définition 4.3.3 On dit que le schéma d'Euler est convergence si, en prenant $x_0 = x(t_0)$ et $x(t)$ la solution de (3.3) en $t = t^*$, on a $\lim_{h\to 0} |e_n| = 0$ en $t_n = t^*$ soit encore

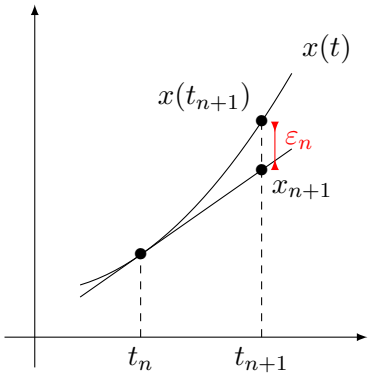
$$\lim_{h\to 0} \sup_{0\leq n\leq N} |e_n| = 0.$$

Pour Euler, on a $\varepsilon_n = x(t_{n+1}) - x(t_n) - (t_{n+1} - t_n)x'(t_n)$. Mais, on sait que

$$x(t_{n+1}) = x(t_n) + (t_{n+1} - t_n)x'(t_n) + \int_{t_n}^{t_{n+1}} (s - t_n)x''(s)ds.$$

Ainsi,

$$\varepsilon = \int_{t_n}^{t_{n+1}} (s - t_n)x''(s)ds,$$
$$|\varepsilon| \leq \|x''\|_\infty \frac{(t_{n+1} - t_n)^2}{2} \leq \frac{h_n^2}{2}\|x''\|_\infty.$$



Définition 4.3.4 La méthode à un pas est dite d'ordre $p \in \mathbb{N}$ si l'erreur locale satisfait $\exists h^* > 0$ tel que $\varepsilon = \mathcal{O}(h_n^{p+1})$, $h_n \leq h^*$. La méthode est consistante si et seulement si $p \geq 1$.

R L'ordre est défini comme étant l'entier p tel que $x(t_0 + h)$ et x_1 coïncide jusqu'à (en incluant) le terme h^p. La méthode d'Euler est d'ordre 1 car $R_1(t) = h^2 x''(\xi)/2$.

Définition 4.3.5 On dit qu'un schéma numérique est consistant si

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

Afin de démontrer la convergence, nous avons besoin de la consistance et de la notion de stabilité. La consistance nous assure qu'à chaque étape $n \to n+1$, l'erreur produite (erreur locale de troncature) est petite. Quand on applique Euler, $x_0 \to x_1$ puis $x_1 \to x_2$ mais pour construire x_2, on n'et pas repartie de la solution exacte $x(t_1)$. On n'est plus sur la courbe de la solution exacte. Une erreur est donc présente dès le départ de la deuxième étape. Par exemple, si on considère l'EDO

$$\begin{cases} x'(t) = t^2 \cos(2\pi t), & t \in [1, 3] \\ x(1) = 0. \end{cases}$$

alors, la suite de points générés par le schéma d'Euler et les valeurs des ε_n sont visibles sur la figure 4.4.

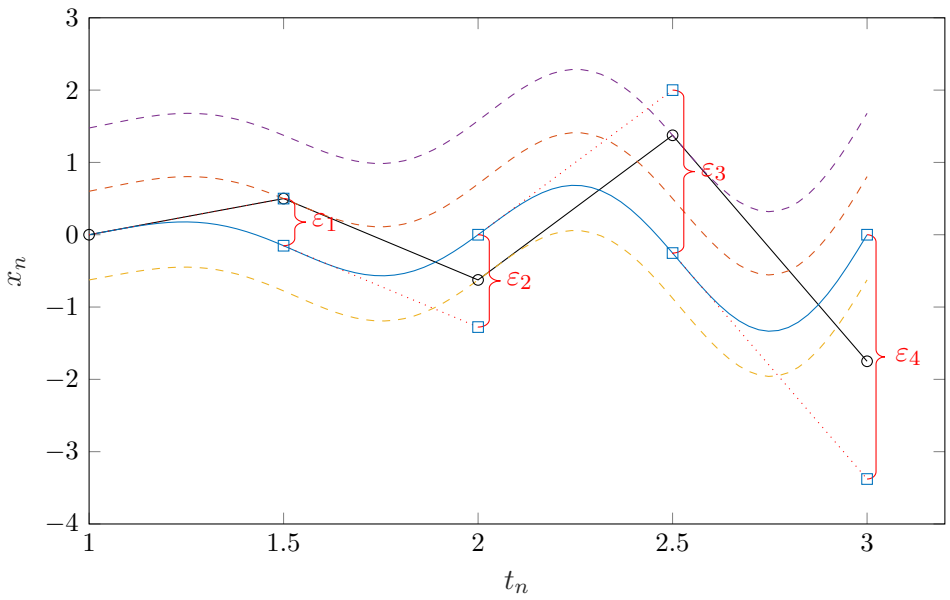


FIGURE 4.4 – Évolution par la méthode d'Euler et valeur des ε_n

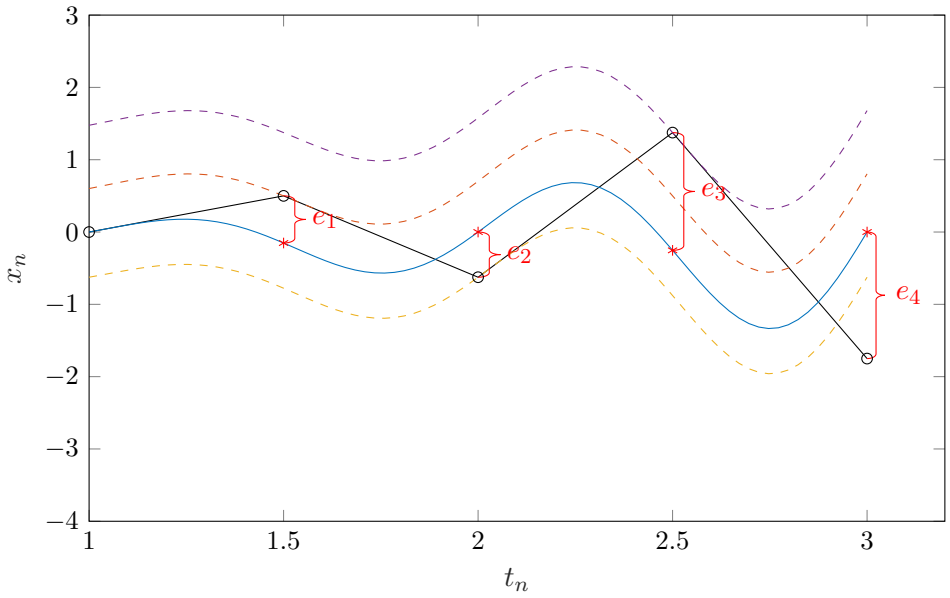


FIGURE 4.5 – Évolution par la méthode d'Euler et valeur des e_n

De même, les valeurs des e_n sont visibles sur la figure 4.5.
Considérons maintenant l'EDO

$$\begin{cases} x'(t) = x(t)(1 - x(t))(1 + x(t)), & t \in [0, 3] \\ x(0) = \eta, \end{cases}$$

dont la solution exacte est $x(t) = \eta/\sqrt{\eta^2 + \exp(-2t) - \exp(-2t)\eta^2}$. Cette EDO possède trois équilibvres 0, -1 et 1. Les solutions ± 1 sont stables, et la solution 0 est instable (voir figure 4.6). Si on change $f(t, x(t))$ en $-f(t, x(t))$, la situation s'inverse. Si la solution que l'on cherche à calculer par la méthode

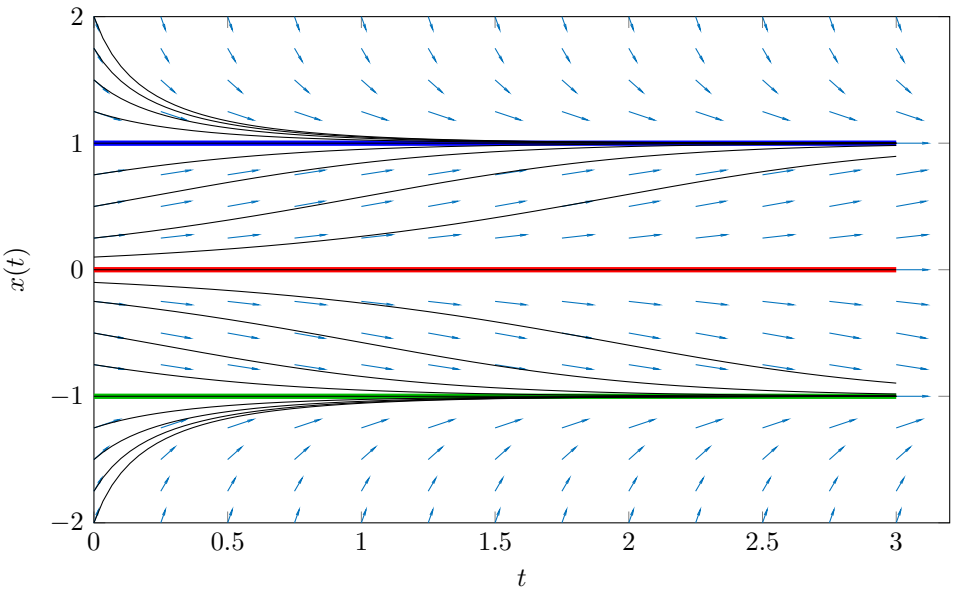


FIGURE 4.6 – Solutions stables et instables

d'Euler est la solution instable, et si on s'écarte de ε de la donnée initiale 0, alors on converge vers un des deux équilibres stables. Ainsi, même une petite ELT peut conduire à une solution radicalement différentes de celle recherchée.

La stabilité répond donc à la question : soit x solution de (3.3)

$$\begin{cases} x'(t) = f(t, x(t)), & t \geq t_0 \\ x(t_0) = x_0, \end{cases}$$

et z la solution de

$$\begin{cases} z'(t) = f(t, z(t)) + g(t), & t \geq t_0 \\ x(t_0) = z_0. \end{cases}$$

Si $\|g\| \ll 1$ et $\|x_0 - z_0\| \ll 1$, a-t-on $\|x(t) - z(t)\| < \varepsilon$? Pour pouvoir répondre à cette question, on rappelle ci-dessous quelques résultats de la théorie des EDOs.

On dit que x est solution de (3.3) si et seulement si $x \in C^1([t_0, t_0 + T])$, $x(t_0) = x_0$ et pour tout $t \in [t_0, t_0 + T]$, $x' = f(t, x)$.

Théorème 4.3.1 Soit $f : \begin{array}{ccc} \mathbb{R} \times \mathbb{R}^d & \longrightarrow & \mathbb{R}^d \\ (t, x) & \longmapsto & f(t, x) \end{array}$ une fonction globalement Lipschitz par rapport à x. Ainsi, il existe une constante $L \in \mathbb{R}$ de Lipschitz telle que pour tout $t \in \mathbb{R}$, $\forall (x_1, x_2) \in \mathbb{R}^{2d}$, $\forall t \in [t_0, t_0 + T]$,

$$\|f(t, x_1) - f(t, x_2)\| \leq L\|x_1 - x_2\|.$$

Alors, le problème (3.3) admet une unique solution $x \in C^1([t_0, t_0 + T])$.

Si on considère l'EDO

$$\begin{cases} u'(t) = a(t)u(t) + b(t), & a, b \in C^0([0, T]), \\ u(0) = u_0, \end{cases} \tag{4.1}$$

alors, so $b = 0$, $u'(t) = a(t)u(t)$ et donc $u(t) = u_0 e^{\int_0^T a(s)ds}$. Si b est non nulle, on applique la méthode de variation de la constante en posant $u(t) = C(t)e^{\int_0^T a(s)ds}$. On a $b = u' - au = C'(t)e^{\int_0^T a(s)ds}$ avec $c(0) = u_0$. On trouve ainsi

$$C(t) = u_0 + \int_0^t b(\sigma) \exp\left(-\int_0^\sigma a(s)ds\right) d\sigma$$

et donc

$$u(t) = u_0 \exp\left(\int_0^t a(s)ds\right) + \int_0^t b(\sigma) \exp\left(\int_\sigma^t a(s)ds\right) d\sigma.$$

Lemme 4.3.2 — Lemme de Gronwall. Soit $u \in C^1([0, T])$ telle que $|u'(t)| \leq a(t)|u(t)| + b(t)$ avec $a(t) \geq 0$, $b(t) \geq 0$ deux fonctions continues sur $[0, T]$, alors

$$|u(t) \leq |u(0)| \exp\left(\int_0^t a(s)ds\right) + \int_0^t b(\sigma) \exp\left(\int_\sigma^t a(s)ds\right) d\sigma.$$

Preuve Considérons $v \in C^1$ telle que $v'(t) \leq a(t)v(t) + b(t)$, avec $v' \geq 0$ et $v \geq 0$.
Cas $b = 0$ $v' \leq av$. On multiplie cette inégalité par $\exp(-\int_0^t a(s)ds)$ et on intègre ce qui conduit à

$$\int_0^t \left(v'(\sigma)e^{-\int_0^\sigma a(s)ds} - a(\sigma)v(\sigma)e^{-\int_0^\sigma a(s)ds}\right) d\sigma \geq 0.$$

Or

$$\begin{aligned} \int_0^t v'(\sigma)e^{-\int_0^\sigma a(s)ds} d\sigma &= -\int_0^t v(\sigma)\left(e^{-\int_0^\sigma a(s)ds}\right)' d\sigma + \left[ve^{-\int_0^\sigma a(s)ds}\right]_0^t \\ &= \int_0^t a(\sigma)v(\sigma)e^{-\int_0^\sigma a(s)ds} d\sigma + v(t)e^{-\int_0^t a(s)ds} - v(0). \end{aligned}$$

Ainsi

$$v(t) \leq v_0 e^{\int_0^t a(s)ds}$$

Cas $b \neq 0$ On reproduit les calculs précédents en appliquant la méthode de la variation de la constante. □

On peut alors montrer le résultat de stabilité

Proposition 4.3.3 Pour tout $t \in [0,T]$, on a

$$\|z(t) - x(t)\| \leq \|z_0 - x_0\| e^{Lt} + \int_0^t e^{L(t-s)} \|g(s)\| ds.$$

Preuve On rappelle que $z' = f(t,z) + g$ et $x' = f(t,x)$. On a ainsi l'estimation

$$\begin{aligned} \|z' - x'\| &= \|f(t,z) - f(t,x) + g\| \\ &\leq \|f(t,z) - f(t,x)\| + \|g(t)\| \\ &\leq L\|z - x\| + \|g(t)\|. \end{aligned}$$

On pose $u(t) = z(t) - x(t)$ de sorte que $\|u'\| \leq L\|u\| + \|g(t)\|$ pour tout $t \in [0,T]$. On applique le lemme de Gronwall et on obtient $\|u\| \leq \|u_0\| e^{Lt} + \int_0^t \|g(s)\| e^{L(t-s)} ds$. □

Revenons à la stabilité numérique

Théorème 4.3.4 Le schéma d'Euler $x_{n+1} = x_n + h_n f(t_n, x_n)$ est stable c'est à dire $\exists M \geq 0$ indépendante de n telle que $\forall (z_n)_{n\in\mathbb{N}}$ vérifiant $z_{n+1} = z_n + f(t_n, z_n) + g_n$, alors l'inégalité

$$\|x_n - z_n\| \leq M \left[\|x_0 - z_0\| + \sum_{k=0}^{n-1} |g_k| \right], \quad 0 \leq n \leq N$$

est satisfaite.

Preuve D'après leurs définitions, on a

$$\begin{aligned} z_{n+1} &= z_n + h_n f(t_n, z_n) + g_n, \\ x_{n+1} &= x_n + h_n f(t_n, x_n). \end{aligned}$$

On a immédiatement

$$\begin{aligned} \|z_{n+1} - x_{n+1}\| &= \|z_n - x_n + h_n(f(t_n, z_n) - f(t_n, x_n)) + g_n\| \\ &\leq \|z_n - x_n\| + h_n L \|z_n x_n\| + \|g_n\| \\ &\leq (1 + Lh_n)\|z_n - x_n\| + \|g_n\|. \end{aligned}$$

Soit d_n la solution de

$$\begin{cases} d_0 = \|x_0 - z_0\| \\ d_{n+1} = (1 + Lh_n) d_n + \|g_n\|. \end{cases}$$

Alors, pour tout $n \in [0,N]$, $\|x_n - z_n\| = d_n$ et

$$d_n \leq \prod_{\ell=0}^{n-1} (1 + h_\ell L) d_0 + \sum_{\ell=0}^{n-1} \|g_\ell\| \prod_{j=l+1}^{n-1} (1 + Lh_j)$$

et comme $1 + x \leq e^x$ pour tout $x \geq 0$,

$$\prod_{\ell=0}^{n-1} (1 + h_\ell L) \leq e^{(t_n - t_0)L}.$$

La preuve se fait par récurrence avec la convention $\prod_{\ell=0}^{-1} = 1$ et $\sum_{\ell=0}^{-1} = 0$. La cas $n = 0$ est immédiat. Supposons maintenant l'inégalité vraie au rang n. On a

$$\begin{aligned} d_{n+1} &= (1 + Lh_n) d_n + \|g_n\| \\ &\leq d_0 \prod_{\ell=0}^{n} (1 + Lh_\ell) + \sum_{\ell=0}^{n-1} \|g_\ell\| \prod_{j=l+1}^{n} (1 + Lh_j) + \|g_n\| \\ &\leq d_0 \prod_{\ell=0}^{n} (1 + Lh_\ell) + \sum_{\ell=0}^{n} \|g_\ell\| \prod_{j=l+1}^{n} (1 + Lh_j). \end{aligned}$$

Or

$$\prod_{\ell=0}^{n-1}(1+h_\ell L) \leq \prod_{\ell=0}^{n-1} e^{h_\ell L} = e^{\sum_{\ell=0}^{n-1} h_\ell L} = e^{(t_n-t_0)L}$$

et

$$\prod_{k=l+1}^{n-1}(1+h_k L) \leq e^{(t_n-t_{l+1})L}.$$

Ainsi, $d_n \leq d_0 e^{(t_n-t_0)L} + \sum_{\ell=0}^{n-1} \|g_\ell\| e^{(t_n-t_{\ell+1})L}$. Or $t_n - t_0 \leq T$ car $t_n \in [t_0, t_0+T]$ d'où

$$d_n \leq e^{LT}(d_0 + \sum_{\ell=0}^{n-1} \|g_\ell\|)$$

ce qui conclut la démonstration. □

En conclusion, on a

Consistance $\|\varepsilon_n\| \leq \frac{n_n^2}{2}\|x''\|_\infty$ où ε_n est définie par $x(t_{n+1}) = x(t_n) + h_n f(t_n, x(t_n)) + \varepsilon_n$.

Stabilité Si $x_0 = x(t_0)$, d'après le théorème, on a

$$\|x_n - x(t_n)\| \leq e^{LT}\left[\underbrace{\|x_0 - x(t_0)\|}_{=0} + \sum_{\ell=0}^{n-1}\|\varepsilon_l\|\right].$$

En utilisant la consistance, on en déduit

$$\|x_n - x(t_n)\| \leq \frac{e^{LT}}{2}\|x''\|_\infty \sum_{\ell=0}^{n-1} h_\ell^2.$$

En définissant $h \max_\ell h_\ell$, ceci conduit à

$$\|x_n - x(t_n)\| \leq \frac{he^{LT}}{2}\|x''\|_\infty \underbrace{\sum_{\ell=0}^{n-1} h_\ell}_{t_n - t_0 \leq T}$$

d'où

$$\|e_n\| \leq \frac{hTe^{Lt}}{2}\|x''\|_\infty, \quad \forall n$$

et donc, le schéma d'Euler est convergent à l'ordre 1

$$\sup_{0\leq n\leq N}\|e_n\| \leq \frac{hTe^{LT}}{2}\|x''\|_\infty.$$

Donc, la stabilité et la consistance implique la convergence.

Regardons maintenant quand le schéma d'Euler, malgré son caractère convergent, fournit de bonnes solutions. Considérons pour cela l'EDO linéaire

$$\begin{cases} x'(t) = -10x(t), & t \in [0,1] \\ x(0) = 1. \end{cases}$$

La solution est $x(t) = e^{10t}$ et on a $x'' = 100e^{-10t}$. Ainsi, $\|x''\|_\infty = 100$. L'estimation précédente devient $\|e_n\| \leq 100e^1 0(h/2) \approx 10^6 h$ (la constante de Lipschitz est ici 10). Donc, demander une erreur $|e_n| \leq 10^{-1}$ nécessite $h \leq 10^{-7}$ doit N de l'ordre de 10^7 application du schéma d'Euler. Il faut donc 10^7 points de discrétisation pour assurer un erreur de l'ordre de 10% pour un temps final 1. Le nombre de calculs nécessaires est donc trop important.

4.4 Le schéma d'Euler implicite

On considère toujours

$$\begin{cases} x'(t) = f(t, x(t)), & t \geq t_0, \\ x(t_0) = \eta, \end{cases}$$

et on a

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

La méthode d'Euler explicite repose sur une quadrature basée sur la méthode des rectangles à gauche. On peut donc tenter de remplacer cette quadrature par la méthode des rectanbles à droite

$$x(t_{n+1}) - x(t_n) = (t_{n+1} - t_n)f(t_{n+1}, x(t_{n+1})) + E(f).$$

On a donc défini un nouveau schéma numérique connu sous le nom de schéma d'Euler implicite

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

4.4.1 Consistance

Calculons l'erreur locale de troncature. Soit x la solution de l'EDO. On a

$$\begin{aligned} \varepsilon_n &= x(t_{n+1}) - x(t_n) - h_n f(t_{n+1}, x(t_{n+1})) \\ &= x(t_{n+1}) - x(t_n) - h_n x'(t_{n+1}). \end{aligned}$$

Mais, le développement de Taylor nous donne

$$f(x) = f(a) + (x-a)f'(a) + \frac{(x-a)^2}{2}f''(a) + \cdots + \frac{(x-a)^n}{n!}f^{(n)}(a) + R_n(x)$$

avec $R_n(x) = \dfrac{f^{(n+1)}(\xi)}{n!}(x-a)(x-\xi)^n$ pour $a < \xi < x$ ou encore $R_n(x) = \displaystyle\int_a^x \frac{f^{(n+1)}(t)}{n!}(x-t)^n dt$.
Ainsi

$$f(x-h) = f(x) - hf'(x) + \int_{x-h}^{x} \frac{f''(t)}{2}(x-h-t)dt$$

et

$$x(t_n) = x(t_{n+1}) - h_n x'(t_{n+1}) + \int_{t_n}^{t_{n+1}} \frac{x''(t)}{2}(t_n - t)dt$$

d'où

$$\varepsilon_n = \int_{t_n}^{t_{n+1}} \frac{x''(t)}{2}(t - t_n)dt \quad \text{et} \quad \|\varepsilon_n\| \leq \frac{h_n^2}{2}\|x''\|_\infty.$$

On a donc $\varepsilon_n = \mathcal{O}(h_n^{p+1})$ avec $p = 1$. Le schéma est ainsi consistant d'ordre 1.

4.4.2 Stabilité

Lemme 4.4.1 Soit h_* telle que $h_* L < 1$. On suppose $0 < h_n \leq h_*$. Alors, si z_n est solution de $z_{n+1} = z_n + h_n f(t_{n+1}, z_{n+1}) + g_n$, alors il existe une constante $M > 0$ telle que

$$\|z_n - x_n\| \leq e^{MT}\left(\|z_0 - x_0\| + \sum_{\ell=0}^{n-1}\|g_\ell\|\right).$$

Preuve Pour simplifier la preuve, on suppose $h_* L < 1/2$ de sorte que $h_n L < h_* L < 1/2$. D'après leurs définitions, on a

$$\begin{aligned} z_{n+1} &= z_n + h_n f(t_{n+1}, z_{n+1}) + g_n, \\ x_{n+1} &= x_n + h_n f(t_{n+1}, x_{n+1}). \end{aligned}$$

On pose $\xi_n = \|z_n - x_n\|$. On a

$$\xi_{n+1} \leq \xi_n + h_n L \xi_{n+1} + \|g_n\|$$

d'où

$$\xi_{n+1} \leq \frac{1}{1-Lh_n}\xi_n + \frac{\|g_n\|}{1-Lh_n}.$$

Par un développement en série entière, on a l'estimation

$$\begin{aligned} 1 < \frac{1}{1-h_nL} &= \sum_{k=0}^{\infty}(h_nL)^k \\ &= 1 + h_nL + (h_nL)^2\sum_{k=0}^{\infty}(h_nL)^k \\ &= 1 + h_nL + \frac{(h_nL)^2}{1-h_nL} \end{aligned}$$

Comme $h_nL < 1/2$, une étude de fonction permet de s'assurer que

$$\frac{(h_nL)^2}{1-h_nL} < h_nL$$

d'où en utilisant le fait que $1 + x < e^x$ pour tout x

$$1 < \frac{1}{1-h_nL} < 1 + 2h_nL < e^{2h_nL}.$$

En outre, $1/(1-h_nL) \leq 2$ ce qui permet d'affirmer

$$\xi_{n+1} \leq \frac{1}{1-Lh_n}\xi_n + \frac{\|g_n\|}{1-Lh_n} < e^{2h_nL}\xi_n + 2\|g_n\|.$$

En appliquant le même calcul que pour la démonstration de la stabilité pour la méthode d'Euler explicite, on a

$$\|z_n - x_n\| \leq e^{MT}\left(\frac{\|z_0 - x_0\|}{1-Lh_*} + \sum_{\ell=0}^{n-1}\|g'_n\|\right).$$

□

4.4.3 Convergence

Si x est solution de classe C^2 et $x'(t) = f(t, x(t))$ avec $x_0 = x(t_0)$, alors

$$\|e_n\| \leq \frac{e^{(M+L)T}}{2}Th\|x''\|_\infty.$$

Comme dans le cas explicite, par une preuve similaire, on a que consistance + stabilité implique la convergence.

4.5 Étude générale de l'erreur des méthodes à un pas

Le principe des méthodes à un pas est de calculer une approximation x_n et $x(t_n)$ par une formule du type

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

Cette formule met en jeu $h_{n-1}, t_{n-1}, x_{n-1}, \cdots$ On note $\tilde{x}_n = x(t_n)$.

Définition 4.5.1 On appelle erreur locale de troncature ε_n définie par

$$\varepsilon_n = \tilde{x}_{n+1} - \tilde{x}_n - h_n\Phi(t_n, \tilde{x}_n, h_n).$$

Définition 4.5.2 On appelle erreur globale de convergence

$$e_n = \tilde{x}_n - x_n.$$

Définition 4.5.3 Un schéma à un pas associé à Φ est stable si il vérifie la propriété : $\exists h^* > 0$, $\exists K \geq 0$ indépendante de n telle que pour toute suite $(z_n)_{n\in\mathbb{N}}$ vérifiant $z_{n+1} = z_n + h_n\Phi(t_n, z_n, h_n) + \eta_n$, on a l'estimation

$$\|x_n - z_n\| \leq K\left[\|x_0 - z_0\| + \sum_{k=0}^{n-1}\|\eta_k\|\right], \quad 0 \leq n \leq N,$$

pour tout $0 < h \leq h^*$ où $h = \sup_{0\leq n\leq N-1} h_n$.

Définition 4.5.4 Un schéma à un pas associé à Φ est convergent si

$$\lim_{h=0} \sup_{0\leq n\leq N} \|e_n\| = 0$$

pourvu que $e_0 = 0$.

Théorème 4.5.1 Tout schéma à un pas *consitant* et **stable** est **convergent** et on a l'estimation $\|e_n\| \leq M[\|e_0\| + C_pTh^p]$.

Preuve On a vu dans la définition de l'erreur locale de troncature que $\tilde{x}_{n+1} = \tilde{x}_n + h_n\Phi(t_n, \tilde{x}_n, h_n) + \varepsilon_n$. Dans la définition de la stabilité, on prend donc $z_n = \tilde{x}_n$ et $\eta_n = \varepsilon_n$. On a donc

$$\|\tilde{x}_n - x_n\| \leq K\left[\|x_0 - \tilde{x}_0\| + \sum_{k=0}^{n-1}\|\varepsilon_k\|\right].$$

Or le schéma est consitant donc $\|\varepsilon_k\| \leq Ch_k^{p+1} \leq Ch^p h_k$ avec $h = \sup_k h_k$. Ainsi,

$$\sum_{k=0}^{n-1}\|\varepsilon_k\| \leq Ch^p \sum_{k=0}^{n-1} h_k = Ch^p(t_n - t_0) = CTh^p.$$

□

Théorème 4.5.2 Si Φ satisfait une condition de Lipschitz

$$\|\Phi(t, x, h) - \Phi(t, z, h)\| \leq \Lambda\|x - z\|,$$

avec $h \in [0, h^*]$, $t \in [t_0, t_0 + T]$, $(x, z) \in (\mathbb{R}^d)^2$, alors la méthode est stable et on a l'estimation

$$\|z_n - x_n\| \leq e^{\Lambda T}\left[\|z_0 - x_0\| + \sum_{k=0}^{n-1}\|\eta_k\|\right].$$

Pour démontrer ce théorème, on a besoin de la version discrète du lemme de Gronwall.

Lemme 4.5.3 — Lemme de Gronwall discret. Si un suite de termes positifs u_n vérifie

$$u_{n+1} \leq e^{h_n\Lambda}u_n + \alpha_n,$$

alors

$$u_n \leq u_0 \prod_{\ell=0}^{n-1} e^{h_\ell\Lambda} + \sum_{\ell=0}^{n-1}\alpha_\ell \prod_{k=\ell}^{n-1} e^{h_k L}$$

avec les conventions $\prod_{\ell=0}^{-1} = 1$ et $\sum_{\ell=0}^{-1} = 0$.

Preuve du lemme La preuve de fait par récurrence. Le rang 0 est évidemment vrai. On suppose donc l'inégalité vraie au rang n. Par hypothèse, on a

$$u_{n+1} \leq e^{h_n\Lambda}u_n + \alpha_n,$$

ce qui devient

$$u_{n+1} \leq e^{h_n\Lambda}\left(u_0\prod_{\ell=0}^{n-1}e^{h_\ell\Lambda} + \sum_{\ell=0}^{n-1}\alpha_\ell\prod_{k=\ell}^{n-1}e^{h_kL}\right) + \alpha_n,$$

ou encore

$$u_{n+1} \leq u_0\prod_{\ell=0}^{n}e^{h_\ell\Lambda} + \sum_{\ell=0}^{n-1}\alpha_\ell\prod_{k=\ell}^{n}e^{h_kL} + \alpha_n.$$

Ainsi,

$$u_{n+1} \leq u_0\prod_{\ell=0}^{n}e^{h_\ell\Lambda} + \sum_{\ell=0}^{n}\alpha_\ell\prod_{k=\ell}^{n}e^{h_kL}.$$

□

Preuve du théorème Par définition, on a

$$\begin{aligned} z_{n+1} &= z_n + h_n\Phi(t_n, z_n, h_n) + \eta_n \\ x_{n+1} &= x_n + h_n\Phi(t_n, x_n, h_n). \end{aligned}$$

Donc,

$$z_{n+1} - x_{n+1} = z_n - x_n + h_n[\Phi(t_n, z_n, h_n) - \Phi(t_n, x_n, h_n)] + \eta_n.$$

On pose $\xi_n = \|z_n - x_n\|$. D'après les hypothèses du théorème, on a

$$\xi_{n+1} \leq (1 + h_n\Lambda)\xi_n + \|\eta_n\|.$$

Or, on a déjà vu que $1 + x \leq e^x$ pour tout $x \in \mathbb{R}$, donc

$$\xi_{n+1} \leq e^{h_n\Lambda}\xi_n + \|\eta_n\|.$$

On applique alors le lemme de Gronwall discret et en utilisant $\prod e^{h_\ell\Lambda} = e^{\sum h_\ell\Lambda}$, on a

$$\xi_n \leq \xi_0\underbrace{e^{(t_n-t_0)\Lambda}}_{\leq e^{\Lambda T}} + \sum_{\ell=0}^{n-1}\|\eta_\ell\|\underbrace{e^{(t_n-t_\ell)\Lambda}}_{\leq e^{\Lambda T}}$$

et donc

$$\xi_n \leq e^{\Lambda T}\left[\xi_0 + \sum_{k=0}^{n-1}\|\eta_k\|\right].$$

□

4.6 Les méthodes de prédicteur-correcteur

On a vu que l'on peut obtenir les deux méthodes d'Euler en considérant

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

et en appliquant un méthode de quadrature. L'idée naturelle est d'appliquer une méthode de quadrature d'ordre plus élevé. Appliquons par exemple la méthode des trapèzes. Ceci conduit à

$$x_1^* = x_0 + \frac{h}{2}\left(f(t_0, x_0) + f(t_0 + h, x(t_0 + h))\right). \tag{4.2}$$

Comme nous avons appliqué une méthode de quadrature d'ordre 2, on a l'estimation

$$\|x(t_0+h)-x_1^*\| \leq Ch^3.$$

En effet, comme $x(t_0+h)=x(t_0)+\int_{t_0}^{t_0+h} f(s,x(s))ds$, on a

$$\|x(t_0+h)-x_1^*\| \leq \| \int_{t_0}^{t_0+h} f(s,x(s))ds - \frac{h}{2}\left(f(t_0,x_0)+f(t_0+h,x(t_0+h))\right) \| \leq Ch^3.$$

Hélas, on a besoin de la valeur exacte de $x(t_0+h)$ qui n'est évidemment pas connue. On fait alors le choix de prédire cette valeur manquante, qu'on note u_2, par un pas de la méthode d'Euler et on utilise la formule (4.2) pour « corriger » la solution. Ceci conduit au schéma des trapèzes explicites

$$\begin{aligned} u_2 &= x_0 + hf(t_0,x_0), \\ x_1 &= x_0 + \tfrac{h}{2}\left(f(t_0,x_0)+f(t_0+h,u_2)\right). \end{aligned}$$

Si au lieu des trapèzes, on utilise la formule de quadrature d'ordre du point milieu $\int_{t_0}^{t_0+h} f(s)ds \approx hf(t_0+h/2)$, on a le schéma du point milieu

$$\begin{aligned} u_2 &= x_0 + hf(t_0,x_0), \\ x_1 &= x_0 + \tfrac{h}{2}\left(f(t_0+h/2,u_2)\right). \end{aligned}$$

Ces méthodes à un pas ont été inventées par Runge (1895).

R Si on remplace $x(t_0+h)$ dans (4.2), on obtient le schéma implicite de Crank-Nicolson ou encore des trapèzes implicites.

$$x_1 = x_0 + \frac{h}{2}((f(t_0,x_0)+f(t_1,x_1)).$$

Théorème 4.6.1 Tout en restant explicites, les deux méthodes de Runge ci-dessus sont d'ordre 2

$$\|x(t_0+h)-x_1\| \leq Ch^3$$

Preuve

$$\|x_1^*-x_1\| = \frac{h}{2}\|f(t_0+h,x)t_0+h)) - f(t_0+h,u_2)\| \leq \frac{Lh}{2}\|x(t_0+h)-u_2\| \leq \frac{Lh}{2}Ch^2$$

où L est la constante de Lipschitz de f. On a donc

$$\|x(t_0+h)-x_1\| \leq \|x(t_0+h)-x_1^*\| + \|x_1^*-x_1\| \leq Ch^3.$$

□

Méthode de Heun (1900) d'ordre 3 : on utilise ici la quadrature de Radau d'ordre 3

$$\int_0^1 g(t)dt \approx \frac{1}{4}g() + \frac{3}{4}g(\frac{2}{3}).$$

On doit ici prédire $x(t_0+\frac{2}{3}h)$ au moins à l'ordre 2 si on veut conserver l'ordre. On applique alors la méthode du point milieu où on remplce h par $2h/3$.

$$\begin{aligned} u_2 &= x_0 + \frac{h}{3}hf(t_0,x_0), \\ u_3 &= x_0 + \frac{2h}{3}f(t_0+\frac{h}{3},u_2), \\ x_1 &= x_0 + \frac{h}{2}\left(\frac{1}{4}f(t_0,x_0)+\frac{3}{4}f(t_0+\frac{2h}{3},u_3)\right). \end{aligned}$$

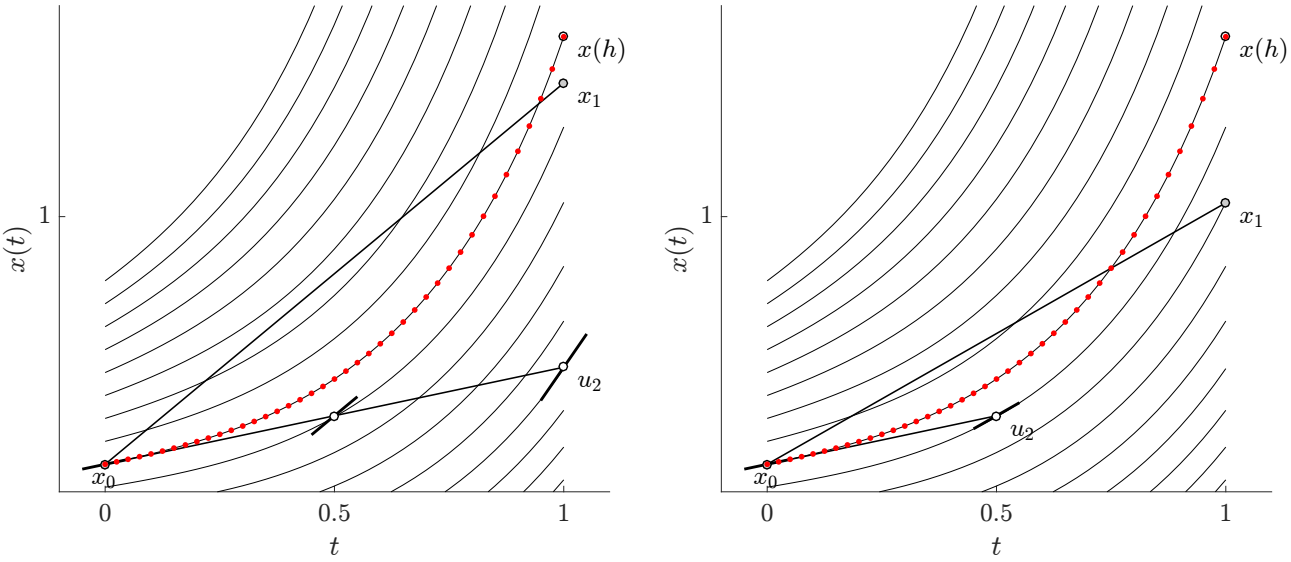
La preuve d'ordre est similaire à la précédente.

Un autre interprétation consiste à écrire

$$x_1^* = x_0 + h\left(\frac{1}{4}f(t_0, x_0) + \frac{3}{4}f(t_0 + \frac{2h}{3}, x(t_0 + \frac{2h}{3}))\right)$$

et on remplace l'inconnue $x(t_0 + 2h/3)$ par la méthode de Runge.

Il est intéressant de constater que ces méthodes ont une interprétation géométrique. Les figures suivantes sont construites pour l'EDO $x'(t) = t^2 + x(t)^2$, avec $x(0) = 0.46$ et $h = 1$. La solution exacte est en pointillés rouges.



trapèzes explicites point milieu

FIGURE 4.7 – Interprétation géométrique

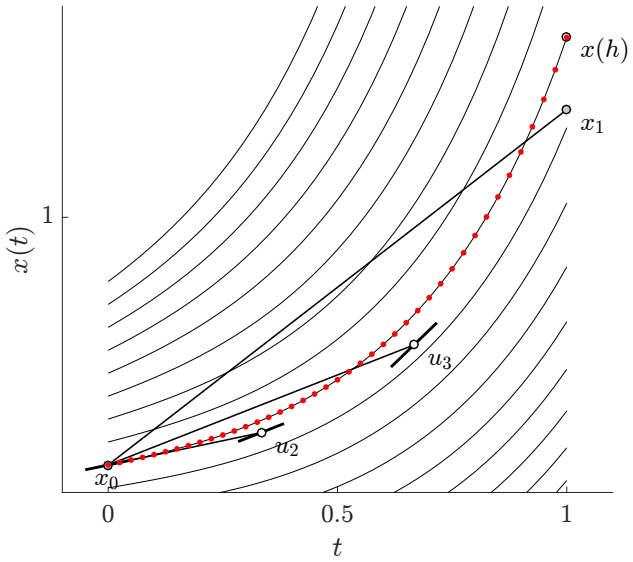


FIGURE 4.8 – Interprétation géométrique de Heun

4.7 Exercices

Exercice 4.1 Soit $\lambda \in \mathbb{C}$. Montrer que l'équation différentielle ordinaire $x'(t) = \lambda x(t)$ peut s'écrire comme un système du premier ordre composé de deux EDO réelles. On pourra pour cela écrire $x(t) = u(t) + iv(t)$ et $\lambda = a + i\,b$.

Quand $a = 0$, utiliser la règle de dérivation des fonctions composées pour vérifier que

$$\frac{d}{dt}(u^2(t) + v^2(t)) = 0$$

de telle sorte que les solutions vivent sur une famille de cercles d'équation $u^2(t) + v^2(t) =$ constante dans le plan $u-v$. ■

Exercice 4.2 Montrer que le problème de Cauchy

$$x''(t) - ax'(t) - bx(t) = f(t), \quad x(0) = \xi, \quad x'(0) = \eta$$

peut s'écrire comme un système du premier ordre $\mathbf{x}'(t) = A\mathbf{x}(t) + \mathbf{g}(t)$, où A est une matrice 2×2, $\mathbf{x} = [x, x']^t$ et les composantes du vecteur à deux composantes $\mathbf{g}$ sont reliés à la fonction de forçage $f(t)$. Quel est le polynôme caractéristique de A ? ■

Exercice 4.3 Soit $x(t)$ la solution de l'EDO

$$x'(t) = 1 + x^2(t). \tag{4.3}$$

1. Trouver la solution générale de cette équation dépendant d'une constante arbitraire.
2. Utiliser le changement de variables $x(t) = -y'(t)/y(t)$ pour montrer que $x(t)$ vérifie (4.3) si $y(t)$ est solution d'une certaine équation différentielle linéaire du second ordre. Déterminer la solution générale pour $y(t)$.
3. En déduire la solution générale de $x(t)$ et expliquer pourquoi elle contient deux constantes arbitraires.

■

Exercice 4.4 Utiliser la méthode d'Euler explicite avec $h = 0.2$ pour montrer que la solution de l' EDO $x'(t) = t^2 - x(t)^2$, $t > 0$, avec $x(0) = 1$ est approximativement $x(0.4) \approx 0.68$.

Montrer que cette estimation devient $x(0.4) \approx 0.708$ si le calcul est répété avec $h = 0.1$. ■

Exercice 4.5 Déterminer la relation de récurrence pour obtenir x_{n+1} en fonction de x_n si on applique la méthode d'Euler explicite à l'EDO $x'(t) = \lambda x(t)$, $x(0) = 1$ et $\lambda = -10$. Dans chacun des cas $h = 1/6$ et $h = 1/12$:

1. calculer x_1, x_2 et x_3.
2. tracer les points (t_0, x_0), (t_1, x_1), (t_2, x_2) et (t_3, x_3) et comparer avec un tracé de la solution exacte.

Commenter les résultats. Quelle est la plus grande valeur de h qu'il est possible d'utiliser quand $\lambda = -10$ pour assurer que $x_n > 0$ pour tout $n = 1, 2, 3, \cdots$? ■

Exercice 4.6 Appliquer la méthode d'Euler explicite à

$$\begin{cases} x'(t) = 1 + t - x(t), & t > 0, \\ x(0) = 0 \end{cases} \tag{4.4}$$

1. Calculer x_1, x_2, $\cdots$ et en déduire une expression de x_n en fonction de $t_n = nh$.
2. Calculer l'erreur locale de troncature.
3. Expliquer pourquoi $x_n = x(t_n)$ où $x(t)$ est la solution de (4.4).

■

Exercice 4.7 Déterminer la méthode d'Euler explicite pour le système du premier ordre

$$\begin{aligned} u'(t) &= -2u(t) + v(t) \\ v'(t) &= -u(t) - 2v(t) \end{aligned}$$

avec les conditions initiales $u(0) = 1$, $v(0) = 0$. Utiliser $h = 0.1$ pour calculer les valeurs approchées de $u(0.2)$ et $v(0.2)$. ■

Exercice 4.8 On s'intéresse aux approximations de l'EDO

$$\begin{cases} x''(t) + 3x'(t) + 2x(t) = t^2, \quad t > 0, \\ x(0) = 1, \ x'(0) = 0. \end{cases} \tag{4.5}$$

1. Écrire le système (4.5) comme un système du premier ordre et déterminer la méthode d'Euler explicite pour calculer les approximations de $x(t_{n+1})$ et $x'(t_{n+1})$ en fonction de $x(t_n)$ et $x'(t_n)$.
2. En éliminant y, montrer que le système

$$\begin{cases} x'(t) = y(t) - 2x(t) \\ y'(t) = t^2 - y(t) \end{cases} \tag{4.6}$$

possède la même solution $x(t)$ que (4.5) si $x(0) = 1$ et $y(0)$ est bien choisi. Quelle est cette valeur ?
3. Appliquer la méthode d'Euler explicite à (4.6) et donner les formules pour calculer les approximations de $x(t_{n+1})$ et $y(t_{n+1})$ en fonction des approximations de $x(t_n)$ et $y(t_n)$.
4. Montrer que les approximations de $x(t_2)$ produitent par les méthodes des questions 1) et 3) sont identiques si les méthodes utilisent le même paramètre de discrétisation h.

■

Exercice 4.9 **Un schéma d'Euler modifié**

On considère l'EDO $x'(t) = f(t, x(t))$ avec $x(0) = x^0$. La fonction f est Lipschitzienne par rapport à la deuxième variable, de constante de Lipschitz L. On suppose dans cet exercice que la solution de l'EDO $x \in C^3([0, T])$.

Montrer que le schéma suivant est consistant

$$\begin{cases} x_{n+1/2} = x_n + \dfrac{h}{2} f(t_n, x_n), \\ x_{n+1} = x_n + h f(t_{n+1/2}, x_{n+1/2}), \quad t_{n+1/2} = t_n + h/2, \\ x_0 = x^0. \end{cases}$$

et que l'erreur locale de troncature vérifie $\|\varepsilon_n\| = \mathcal{O}(h^3)$. ■

Exercice 4.10 On s'intéresse à la résolution numérique de l'équation différentielle ordinaire

$$\begin{cases} x'(t) = f(t, x(t)), \quad t \in I_0 = [t_0, t_0 + T], \\ x(t_0) = x_0. \end{cases}$$

On suppose que f est continue de $\mathbb{R} \times I_0$ dans $\mathbb{R}$ et est Lipschtiz. On introduit une subdivision uniforme $t_0 < t_1 < \cdots < t_N = t_0 + T$ et on pose $h = t_{n+1} - t_n$ et $t_n = nt + t_0$.

1. Trouver a, b et c pour que la formule de quadrature suivante soit d'ordre maximal

$$\int_{t_n}^{t_{n+2}} \psi(t)dt + 4 \int_{t_n}^{t_{n+1}} \psi(t)dt = h\big(a\psi(t_{n+2}) + b\psi(t_{n+1}) + c\psi(t_n)\big).$$

2. Montrer que l'on peut approcher l'équation différentielle par le schéma numérique

$$\begin{cases} X_0,\ X_1 \text{ donnés,} \\ X_{n+2} + 4X_{n+1} - 5X_n = 6h\Big(\frac{2}{3}f(t_{n+1}, X_{n+1}) + \frac{1}{3}f(t_n, X_n)\Big). \end{cases}$$

3. On suppose désormais $f \equiv 0$. Calculer X_n en fonction de X_0 et X_1.
4. On suppose, dans cette question uniquement, que $X_1 = X_0$. Montrer que $\forall 0 \leq n \leq N$, $e_n = X(t_n) - X_n = 0$.
5. Que se passe-t'il lorsque $X_1 = X_0 + Ch^p$?

■

Exercice 4.11 **Étude d'une chute dans l'atmosphère**

Un corps de faible densité et d'extérieur rêche (par exemple un flocon de neige), se déplaçant près de la surface de la terre subit une force de resistivité due à l'air qui est proportionnelle à la vitesse v mais qui agit en opposition de mouvement. Donc, si un tel corps de masse m est laché à une hauteur x_0 avec une vitesse initiale v_0 dans la direction verticale, il subit une force due à la gravité et à la resistance de l'air égale à $F = -mg - kv$, où g est l'accéleration de la pesanteur, et $k > 0$ est une constante de proportionalité. La seconde loi de la dynamique nous donne donc

$$mv' = mg - kv, \quad v(0) = v_0.$$

Dans d'autres situations telle que la chute d'un objet dense, la force de resistivité de l'air peut être proportionnelle au carré de la vitesse et agir dans le sens opposé au mouvement. Alors, l'équation du mouvement est de la forme $mv' = -mg + kv^2$ ou $mv' = -mg - kv^2$ avec $k > 0$ est le coefficient, avec le signe $+$ si l'objet est en train de tomber, et le signe $-$ si il monte. L'équation peut alors s'écrire $mv' = -mg - kv|v|$. Nous supposons dans la suite de l'exercice que l'équation différentielle considérée est

$$v' = -1 - v|v|.$$

1. Montrer que si $v_0 \leq 0$, alors $v(t) < 0$, $\forall t \in \mathbb{R}^+$.
2. Montrer que v est décroissante si $v(0) > -1$.
3. v admet elle une limite ?
4. Écrirer le schéma d'Euler explicite associé. Doit on avoir une limitation sur le pas de temps pour que le schéma préserve les propriétés des solutions exactes.
5. Application : au temps $t = 0$, un parachutiste de poids 85kg ouvre son parachute à une hauteur sufficante pour atteindre un régime constant avec une vitesse initiale supposée nulle. La résistance de l'air est donnée par $k = 20$kg/m. À quelle vitesse le parachutiste touche-t'il le sol ?

■

Exercice 4.12 On se propose d'étudier l'influence des erreurs d'arrondi sur la méthode d'Euler explicite. On considère donc l'équation différentielle

$$x' = f(t, x(t)),\ t \in [0, T], \quad x(0) = x_0,$$

où f est une fonction continue de $[0, T] \times \mathbb{R}$ à valeurs dans $\mathbb{R}$ et L-Lipschitz par rappors à la variable x. Soit X_n la solution du schéma d'Euler associé à une discrétisation uniforme $0 = t_0 < t_1 < \cdots < t_N = T$ de pas h.
Lorsqu'on effectue des opérations algébriques sur un ordinateur on commet des erreurs d'arrondi. L'essentiel de ces erreurs sont générées lorsqu'on effectue une addition ou un soustraction. On peut considérer que les opérations de multiplication ou de division se font sans aucune erreur d'arrondi. On suppose que la solution est de classe C^2.

1. Soit $(\theta_n)_n$ une suite de réels positifs qui vérifient $\theta_{n+1} \leq (1 + A)\theta_n + B$ où $A > 0$ et $B > 0$

sont des constantes. Montrer que

$$\theta_n \leq \theta_0 e^{nA} + \frac{e^{nA}-1}{A} B.$$

2. La solution du schéma d'Euler calculée par la machine, notée X_n^* vérifie le schéma pertubé

$$\left\{ \begin{array}{rcl} X_{n+1}^* & = & X_n^* + hf(t_n, X_n^*) + h\mu_n + \rho_n, \\ X_0^* & = & x_0 \end{array} \right.$$

où $0 \leq |\mu_n| \leq \mu$ et $0 \leq |\rho_n| \leq \rho$.
Pourquoi ? Que représente μ_n ? ρ_n ?

3. On pose $e_n^* = x(t_n) - X_n^*$. Montrer qu'il existe A, B et C indépendantes de h et ρ telles que

$$|e_n^*| \leq A + Bh + \frac{C}{h}\rho = \varphi(h).$$

En déduire l'existence d'un pas optimal pour lequel l'erreur soit minimale. Interpréter.

■

5. Les méthodes multi-pas

5.1 Introduction

On peut être tenté d'utiliser les développements de Taylor pour augmenter l'ordre des méthodes numériques pour les EDO. Par exemple, on a

$$x(t+h) = x(t) + hx'(t) + \frac{h^2}{2}x''(t) + R_2(t), \quad R_2(t) = \mathcal{O}(h^3).$$

On peut écrire ainsi

$$x_{n+1} = x_n + h \underbrace{x'_n}_{f(t_n,x_n)} + \frac{h^2}{2}x''_n.$$

Il nous faut une valeur pour x''_n. On différencie $f(t,x)$.

■ **Exemple 5.1** Considérons l'EDO

$$\begin{aligned} x'(t) &= (1-2t)x(t), \\ x(0) &= 1. \end{aligned}$$

Alors, la dérivée seconde de x est aisément accessible

$$\begin{aligned} x''(t) &= -2x(t) + (1-2t)x'(t) \\ &= -2x(t) + (1-2t)^2 x(t) \\ &= [(1-2t)^2 - 2]x(t) \end{aligned}$$

et donc on a un schéma

$$x_{n+1} = x_n + h(1-2t_n)x_n + \frac{1}{2}h^2[(1-2t_n)^2 - 2]x_n, \quad b \geq 0.$$

■

On peut généraliser et monter ainsi à un ordre p. Le problème de cette méthode est qu'il fait pouvoir différencier $f(t,x(t))$ et expliciter en fonction de $x(t)$. On va donc chercher des alternatives qui ne nécessitent pas l'utilisation de dérivées d'ordre supérieur.

L'idée des méthodes multipas est d'utiliser l'historique disponible, c'est à dire les valeurs de x et x' calculées aux pas de temps précédents. Si on revient à

$$x(t+h) = x(t) + hx'(t) + \frac{h^2}{2}x''(t) + \mathcal{O}(h^3),$$

on cherche donc une approximation de x''.

5.1.1 La règle des trapèzes

On sait que $x'(t+h) = x'(t) + hx''(t) + \mathcal{O}(h^2)$ et donc $hx''(t) = x'(t+h) - x'(t) + \mathcal{O}(h^2)$. On obtient ainsi

$$\begin{aligned} x(t+h) &= x(t) + hx'(t) + \frac{1}{2}h[x'(t+h) - x'(t) + \mathcal{O}(h^2)] + \mathcal{O}(h^3) \\ &= x(t) + \frac{h}{2}[x'(t+h) + x'(t)] + \mathcal{O}(h^3) \end{aligned}$$

d'où

$$x(t+h) = x(t) + \frac{h}{2}[f(t+h, x(t+h) + f(t, x(t))] + \mathcal{O}(h^3).$$

En négligeant le terme de reste, on a le schéma numérique

$$x_{n+1} = x_n + \frac{h}{2}[\underbrace{f(t_{n+1}, x_{n+1})}_{:=f_{n+1}} + \underbrace{f(t_n, x_n)}_{:=f_n}]$$

et on retrouve la méthodes des trapèzes implicites.

5.1.2 Méthode de Adams-Bashforth à 2 étapes AB(2)

On utilise cette fois plutôt $x'(t-h) = x'(t) - hx''(t) + \mathcal{O}(h^2)$ ce qui conduit à $hx''(t) = x'(t) - x'(t-h) + \mathcal{O}(h^2)$ et aini

$$\begin{aligned} x(t+h) &= x(t) + hx'(t) + \frac{1}{2}h[x'(t) - x'(t-h) + \mathcal{O}(h^2)] + \mathcal{O}(h^3) \\ &= x(t) + \frac{h}{2}[3x'(t) - x'(t-h)] + \mathcal{O}(h^3) \\ &= x(t) + \frac{h}{2}[3f(t, x(t)) - f(t-h, x(t-h))] + \mathcal{O}(h^3). \end{aligned}$$

En négligeant le terme de reste, on a le schéma

$$x_{n+1} = x_n + \frac{h}{2}(3f_n - f_{n-1}).$$

Ainsi, on a besoin des deux étapes précédentes pour calculer x_{n+1}. Il est donc plutôt préférable de noter cette méthode

$$x_{n+2} = x_{n+1} + \frac{h}{2}(3f_{n+1} - f_n).$$

Pour démarrer l'algorithme, on dispose de $x_0 = x(t_0)$. Il nous manque donc x_1. On peut par exemple faire le choix de la méthode d'Euler explicite ou d'une autre méthode à un pas.

5.2 Les méthodes à deux pas

On se concentre sur les méthodes qui mettent en jeu les trois temps discrets t_n, t_{n+1} et t_{n+2}. On doit trouver les coefficients α_0, α_1, β_0, β_1 et β_2 tels que

$$x(t+2h) + \alpha_1 x(t+h) + \alpha_0 x(t) = h[\beta_2 x'(t+2h) + \beta_1 x'(t+h) + \beta_0 x'(t)] + \mathcal{O}(h^{p+1})$$

ce qui conduit en négligeant le terme de reste à

$$x_{n+2} + \alpha_1 x_{n+1} + \alpha_0 x_n = h[\beta_2 f_{n+2} + \beta_1 f_{n+1} + \beta_0 f_n].$$

Une méthode multipas est dite explicite si $\beta_2 = 0$ et implicite si $\beta_2 \neq 0$.

R Les méthodes à un pas sont un sous ensemble des méthodes multipas.

■ **Exemple 5.2** — Méthode de Adams-Bashforth à 2 étapes, $p = 2$

$$x_{n+2} - x_{n+1} = \frac{h}{2}(3f_{n+1} - f_n).$$

— Méthode de Adams-Moulton à 2 étapes, $p = 2$

$$x_{n+2} - x_{n+1} = \frac{h}{12}(5f_{n+2} + 8f_{n+1} - f_n).$$

— Règle de Simpson, $p = 4$

$$x_{n+2} - x_n = \frac{h}{3}(f_{n+2} + 4f_{n+1} + f_n).$$

— Méthode de Dahlquist, $p = 3$

$$x_{n+2} + 4x_{n+1} - 5x_n = h(4f_{n+1} + 2f_n).$$

■

5.2.1 Consistance

Définition 5.2.1 On définit l'opérateur linéaire aux différences $\mathscr{L}_h$ associé à une méthode multipas à deux étapes pour un fonction différentiable $z(t)$ par

$$\mathscr{L}_h z(t) = z(t+2h) + \alpha_1 z(t+h) + \alpha_0 z(t) - h[\beta_2 z'(t+2h) + \beta_1 z'(t+h) + \beta_0 z'(t)].$$

$\mathscr{L}_h$ est dite être consistant à l'ordre p si $\mathscr{L}_h = \mathcal{O}(h^{p+1})$ pour tout entier $p > 0$.
Si $\mathscr{L}_h$ est consistant, alors la méthode multipas est dite consistante.

■ **Exemple 5.3 — Méthode de Dahlquist.** L'opératuer aux différences est

$$\mathscr{L}_h z(t) = z(t+2h) + 4z(t+h) - 5z(t) - h[4z'(t+h) + 2z'(t)].$$

Or, par Taylor

$$\begin{array}{lcccccccccc}
z(t+2h) & = & z(t) & + & 2hz'(t) & + & 2h^2 z''(t) & + & \frac{4}{3}h^3 z''' & + & \frac{2}{3}h^4 z^{(iv)} & + & \mathcal{O}(h^5), \\
z(t+h) & = & z(t) & + & hz'(t) & + & \frac{h^2}{2} z''(t) & + & \frac{1}{6}h^3 z''' & + & \frac{1}{24}h^4 z^{(iv)} & + & \mathcal{O}(h^5), \\
z'(t+h) & = & & & z'(t) & + & hz''(t) & + & \frac{1}{2}h^2 z''' & + & \frac{1}{6}h^3 z^{(iv)} & + & \mathcal{O}(h^4).
\end{array}$$

Alors,

$$\begin{array}{ll}
\mathscr{L}_h z(t) + (1+4-5)z(t) & +h(2+4-(4+2))z'(t) \\
& +h^2(2+2-4)z''(t) \\
& +h^3(4/3+4/6-4/2)z'''(t) \\
& +h^4(2/3+4/24-4/6)z^{(iv)}(t) + \mathcal{O}(h^5).
\end{array}$$

Ainsi,

$$\mathscr{L}_h z(t) = \frac{h^4}{6} z^{(iv)}(t) + \mathcal{O}(h^5)$$

et donc $\mathscr{L}_h z(t) = \mathcal{O}(h^4)$ et la méthode est donc consistante à l'ordre $p = 3$. ■

5.2.2 Construction

A partir de la définition de $\mathscr{L}_h$ et des développements de Taylor ci-dessus, on peut donc en fixant p déterminer les coefficients α_0, α_1, β_0, β_1 et β_2. Par exemple, si on souhaite une méthode d'ordre au moins $p = 1$,

$$\mathscr{L}_h z(t) = (1 + \alpha_1 + \alpha_0)z(t) + h(2 + \alpha_1 - (\beta_2 + \beta_1 + \beta_0))z'(t) + \mathcal{O}(h^2).$$

On désire que $\mathscr{L}_h z(t) = \mathcal{O}(h^2)$ ce qui implique $1 + \alpha_1 + \alpha_0 = 0$ et $2 + \alpha_1 = \beta_2 + \beta_1 + \beta_0$.

Définition 5.2.2 Les premier et deuxième polynômes caractéristiques d'une méthode multipas à 2 niveaux sont définis par

$$\rho(r) = r^2 + \alpha_1 r + \alpha_0 \quad \text{et} \quad \sigma(r) = \beta_2 r^2 + \beta_1 r + \beta_0$$

respectivement.

Théorème 5.2.1 Une méthode multipas à 2 étapes $x_{n+2}+\alpha_1 x_{n+1}+\alpha_0 x_n = h[\beta_2 f_{n+2}+\beta_1 f_{n+1}+\beta_0 f_n]$ est consistante avec l'EDO $x'(t) = f(t, x(t))$ si et seulement si

$$\rho(1) = 0 \quad \text{et} \quad \rho'(1) = \sigma(1).$$

Théorème 5.2.2 Une méthode multipas convergente est consistante.

Preuve On suppose que la méthode multipas est convergente. Ainsi, si $t^* = t_n$, $x(t_{n+2}) = x(t^* + 2h)$, $x(t_{n+1}) = x(t^* + h)$ et $x(t_n) = x(t^*)$. Or, t_{n+2} et t_{n+1} tendent vers t^*. Mais,

$$x_{n+2} + \alpha_1 x_{n+1} + \alpha_0 x_n = h[\beta_2 f_{n+2} + \beta_1 f_{n+1} + \beta_0 f_n].$$

En prenant la limite $h \to 0$, on obtient $\rho(1)x(t^*) = 0$. En général, $x(t^*) \neq 0$ et donc $\rho(1) = 0$. En outre

$$\frac{x_{n+2} + \alpha_1 x_{n+1} + \alpha_0 x_n}{h} = \beta_2 f_{n+2} + \beta_1 f_{n+1} + \beta_0 f_n.$$

On a $\lim_{h\to 0} \beta_2 f_{n+2} + \beta_1 f_{n+1} + \beta_0 f_n = \sigma(1) f(t^*, x(t^*))$.

Par la règle de l'Hospital, si $f, g \in C^1$,

$$\lim_{x\to a} \frac{f(x)}{g(x)} = \frac{f'(a)}{g'(a)}.$$

Ainsi,

$$\lim_{h\to 0} \frac{x'(t^* + 2h) + \alpha_1 x'(t^* + h) + \alpha_0 x(t)}{h} = \frac{2x'(t^*) + \alpha_1 x'(t^*)}{1} = (2 + \alpha_1)x'(t^*).$$

Donc, à la limite, on a

$$\underbrace{(2 + \alpha_1)}\rho'(1)x'(t^*) = \sigma(1) f(t^*, x(t^*))$$

ce qui donne $\rho'(1) = \sigma(1)$ d'où la consistance. □

Comme on l'a vu pour les méthodes à un pas, ce théorème ne dit pas qu'une méthode multipas consistante est convergente. Il faut un critère supplémentaire.

5.3 Méthodes à k étapes

La forme générale d'une méthode à k étapes est

$$x_{n+k} + \alpha_{k-1} x_{n+k-1} + \cdots + \alpha_0 x_n = h(\beta_k f_{n+k} + \beta_{k-1} f_{n+k-1} + \cdots + \beta_0 f_n).$$

La méthode est implicite si $\beta_k \neq 0$. Les polynômes caractéristiques de degré 1 et 2 sont respectivement

$$\rho(r) = \sum_{\ell=0}^{k} \alpha_\ell r^\ell, \quad \sigma(r) = \sum_{\ell=0}^{k} \beta_\ell r^\ell,$$

avec la convention $\alpha_k = 1$. L'opérateur linéaire aux différences est donné par

$$\mathscr{L}_h z(t) = \sum_{\ell=0}^{k} \alpha_\ell z(t + \ell h) - h\beta_\ell z'(t + \ell h).$$

Si l'on fait un développement de Taylor en $h \approx 0$, on trouve

$$\mathscr{L}_h z(t) = C_0 z(t) + C_1 h z'(t) + \cdots + C_p h^p z^{(p)}(t) + C_{p+1} h^{p+1} z^{(p+1)}(t) + \mathcal{O}(h^{p+2})$$

aec $C_0 = \rho(1)$, $C_1 = \rho'(1) - \sigma(1)$, et les coefficients C_j, $j \geq 2$ sont des combinaisons linéaires des coefficients $\alpha_0, \cdots, \alpha_{k-1}, \beta_0, \cdots, \beta_k$.

La méthode est d'ordre p si $C_0 = C_1 = \cdots = C_p = 0$. La constante C_{p+1} est dite constante d'erreur.

5.4 Convergence et (zéro)-stabilité

On s'intéresse comme dans les chapitres précédents à la convergence des schémas numériques. On considère le problème de Cauchy

$$\begin{cases} x'(t) = f(t, x(t)), \quad t \in]t_0, t_f], \\ x(t_0) = \eta, \end{cases} \tag{5.1}$$

pour lequel on applique une méthode multipas (MMP)

$$\sum_{\ell=0}^{k} \alpha_\ell x_{n+\ell} = h \sum_{\ell=0}^{k} \beta_\ell f_{n+\ell}$$

avec $\alpha_k = 1$ et les valeurs initiales $x_0 = \eta_0$, $x_1 = \eta_1$, $\cdots$, $x_{k-1} = \eta_{k-1}$, $0 \leq n \leq N - k$, $Nh = t_f - t_0$.

Comme les valeurs initiales η_j ne sont pas connues, on fait dès le début une erreur sur celles ci puisqu'on utilise des méthodes numériques pour les construires. On suppose cependant que $\lim_{h\to 0} \eta_j = \eta$, $0 \leq j \leq k-1$.

Définition 5.4.1 — Convergence. La méthode (MMP) est dite convergente si pour tout problème (5.1) qui admet une unique solution $x(t)$ pour $t \in [t_0, t_f]$, on a

$$\lim_{\substack{h \to 0 \\ nh = t^* - t_0}} x_n = x(t^*), \quad \forall t^* \in [t_0, t_f].$$

Comme on l'a vu, la consistance est nécessaire pour obtenir la convergence, c'est à dire

$$\rho(1) = 0 \quad \text{et} \quad \rho'(1) = \sigma(1)$$

ou encore

$$\sum_{j=0}^{k} \alpha_j = 0 \quad \text{et} \quad \sum_{j=0}^{k} j\alpha_j = \sum_{j=0}^{k} \beta_j.$$

■ **Exemple 5.4** Considérons la méthode d'ordre 3

$$x_{n+2} + 4x_{n-1} - 5x_n = h(4f_{n+1} + 2f_n)$$

La méthode est bien consistante ($1 + 4 - 5 = 0$ et $2 + 1 \times 4 = 4 + 2$).

Appliquons cette méthode à $x'(t) = 0$, $x(0) = 1$. On prend $x_0 = 1$ et $x_1 = 1 + h$. On a bien $\lim_{h\to 0} x_1 = 1$. La méthode devient pour cette EDO

$$x_{n+2} + 4x_{n-1} - 5x_n = 0. \tag{5.2}$$

Rappelons que si l'on considère une relation de récurrence $u_{n+2} + au_{n+1} + bu_n = 0$, alors trouver les solutions revient à résoudre l'équation du second ordre $r^2 + ar + b = 0$. Si l'équation possède deux racines distinctes r_1 et r_2, alors la solution est donnée par $u_n = Ar_1^n + Br_2^n$. Si une seule racine existe, alors $u_n = (A + Bn)r_1^n$. Pour notre schéma, l'équation du second ordre est $r^2 + 4r - 5 = (r-1)(r+5)$. La solution est donc

$$x_n = A + B(-5)^n.$$

On identifie A et B d'après les conditions initiales et qui conduit à $A = 1 + h/6$ et $B = -h/6$ d'où

$$x_n = 1 + \frac{h}{6}(1 - (-5)^n).$$

Si on suppose $t^* = 1$, de sorte que $nh = 1$, alors $h|(-5)^n| = 5^n/n$ qui tend vers $+\infty$ quand $h \to 0$ ($n \to \infty$) et donc la suite x_n diverge.

Le terme de l'équation auxiliaire (5.2) est le premier polynôme caractéristique $\rho(t)$. Or, $\rho(1) = 0$ implique que $r = 1$ est racine de $\rho(r)$ pour toute méthode consistante. On a donc nécessairement la factorisation $\rho(r) = (r-1)(r-a)$ pour les méthodes à 2 étapes.

Une telle méthode appliquée à $x'(t) = 0$ donnera la solution générale $x_n = A + Ba^n$. Donc, pour espérer une convergence, il faut que $|a| \leq 1$. ■

■ **Exemple 5.5** Considérons maintenant la méthode

$$x_{n+3} + x_{n+2} - x_{n+1} - x_n = 4hf_n$$

appliquée à $x'(t) = 0$, $x(0) = 1$. On choisit $x_0 = 1$, $x_1 = 1 - h$ et $x_2 = 1 - 2h$. L'équation aux différences homogène est

$$x_{n+3} + x_{n+2} - x_{n+1} - x_n = 0.$$

L'équation auxiliaire associée est $\rho(r) = (r-1)(r+1)^2$ et on a la solution générale $x_n = A+(B+Cn)(-1)^n$. Ainsi

$$x_n = 1 - h + (-1)^n(h - t_n),$$

et la solution exacte est $x(t) = 1$. Si $t^* = t_n = 1$, on a $x(1) = 1$ et $x_n = (1 - h) + (-1)^n(h - 1)$, de sorte que

$$|x_n - 1 + h| = 1 - h, \quad \forall h \text{ tel que } nh = 1.$$

Ainsi, l'erreur globale $|x(1) - x_n| = |1 - x_n|$ tend vers 1 quand $h \to 0$ et donc ne converge pas. Donc, demander $|a| \leq 1$ ne suffit pas. ■

Définition 5.4.2 Un polynôme est dit satisfaire la condition de racines si toutes ses racines sont incluses dans $\overline{B}(0, 1)$, les racines sur la frontière étant simples. Donc, pour tout racine r, $|r| \leq 1$ et si $|r| = 1$, la racine est simple.

■ **Exemple 5.6**
— $\rho(r) = r^2 - 1$ vérifie la condition de racines.
— $\rho(r) = (r - 1)^2$ ne vérifie pas la condition de racines.

■

Définition 5.4.3 Une méthode multipas est dite (zero)-stable si son premier polynôme caractéristique $\rho(r)$ satisfait la condition de racines.

R Toutes les méthodes multipas à une étape consistantes vérifie la condition de racines car $\rho(r) = r-1$. Elles sont donc (zero)-stable, Euler étant ainsi (zero)-stable.

Théorème 5.4.1 — Dahlquist 1956. Une méthode multipas est convergente si et seulement si elle est consistante et (zero)-stable.

La zero stabilité implique une restriction significative sur l'ordre atteignable pour une méthode multipas.

Théorème 5.4.2 — Première barrière de Dahlquist 1959. L'ordre p d'une méthode multipas à k étapes satisfait

1. $p \leq k + 2$ si k est pair
2. $p \leq k + 1$ si k est impair
3. $p \leq k$ si $\beta_k \leq 0$ (en particulier pour toutes les méthodes explicites).

5.5 Familles classiques

5.5.1 Adams-Bashforth 1883

On a $\rho = r^k - r^{k-1}$. On a deux racines : $r = 1$ qui est racine simple et $r = 0$ qui est racine multiple de multiplicité $k - 1$. Ceci conduit à des méthodes explicites

$$x_{n+k} - x_{n+k-1} = h(\beta_{k-1}f_{n+k-1} + \cdots + \beta_0 f_n)$$

où les coefficients $\beta_{k-1}, \cdots, \beta_0$ sont choisis tels que $C_0 = C_1 = \cdots = C_{k-1} = 0$. L'ordre est donc $p = k$. Ce sont des méthodes multipas explicites d'ordre le plus élevé.
$k = 1$ méthodes d'Euler
$k = 2$ méthode AB(2)
$k = 3$ $x_{n+3} - x_{n+2} = \frac{h}{12}(23f_{n+2} - 48f_{n+1} + 5f_n)$ méthode AB(3), ordre $p = 3$ de constante d'erreur $C_3 = 3/8$.

5.5.2 Famille Adams-Moulton 1926

C'est la version implicite de Adams-Bashforth.

$$x_{n+k} - x_{n+k-1} = h(\beta_k f_{n+k} + \cdots + \beta_0 f_n)$$

$k = 1$ trapèze implicite
$k = 2$ AM(2) $x_{n+2} - x_{n+1} = h(5f_{n+2} + 8f_{n+1} - f_n)/12$. L'ordre est $p = 3$ (on aurait pu espérer $p = 4$).
$k = 3$ AM(3) $x_{n+3} - x_{n+2} = h(9f_{n+3} + 19f_{n+2} - 5f_{n+1} + f_n)/24$, ordre $p = 4$.

5.5.3 Méthodes de Nyström 1925

Ce sont des méthodes explicites avec $k \geq 2$. On prend $\rho(r) = r^k - r^{k-2}$ de sorte que

$$x_{n+k} - x_{n+k-2} = h(\beta_{k-1} f_{n+k-1} + \cdots + \beta_0 f_n).$$

L'ordre est $p = k$

5.5.4 Milne-Simpson 1926

C'est la version implicite de Nyström.

$$x_{n+k} - x_{n+k-2} = h(\beta_k f_{n+k} + \cdots + \beta_0 f_n).$$

Par exemple, la règle de Simpson est donnée pour $k = 2$ et l'ordre est $p = 4$

5.5.5 Backward Differentiation Formulas (BDF) 1952

Elles constituent une généralisation de la méthode d'Euler implcite La formule la plus simple est $\sigma(r) = \beta_k r^k$. Ainsi, le second polynôme caractéristique est consistant.

$$x_{n+k} + \alpha_{k-1} x_{n+k-1} + \cdots + \alpha_0 x_n = h\beta_k f_{n+k}.$$

On choisit les $(k+1)$ coefficients $\alpha_{k-1}, \cdots, \alpha_0, \beta_k$ tels que l'ordre $p = k$. On peut montrer

$$\rho(r) = \frac{1}{C} \sum_{j=1}^{k} \frac{1}{j} r^{k-j} (r-1)^j, \quad C = \sum_{j=1}^{k} \frac{1}{j}.$$

5.6 Exercices

Exercice 5.1 La méthode multipas suivante est-elle consistante ?

$$x_{n+2} - x_n = \frac{1}{4} h(3f_{n+1} - f_n).$$

Pour quelles valeurs des paramètres a et b les méthode multipas suivantes sont consistantes ?

1. $x_{n+2} - a x_{n+1} - 2x_n = hbf_n$,
2. $x_{n+2} + x_{n+1} + a x_n = h(f_{n+2} + bf_n)$.

■

Exercice 5.2 On applique la méthode multipas

$$x_{n+1} = x_n + 2hf_n$$

au problème de Cauchy

$$\begin{cases} x'(t) = 1, & t > 0, \\ x(0) = 0. \end{cases}$$

1. Montrer que la méthode multipas n'est pas consistante.
2. Montrer que $x_n = 2nh$.

3. Calculer l'erreur de troncature globale $x(t_n) - x_n$ à $t_n = 1$ et vérifier que cette méthode multipas n'est pas convergente.

■

Exercice 5.3 On considère la méthode multipas

$$x_{n+2} + \alpha_1 x_{n+1} - a x_n = h\beta_2 f_{n+2}.$$

1. Quel est l'ordre maximal atteint avec a qui reste un paramètre ?
2. Quel est l'ordre maximal général ?
3. Commenter la méthode (connu comme schéma "Backward Differentiation Formula à deux pas BDF(2)")
$$3x_{n+2} - 4x_{n+1} + x_n = 2hf_{n+2}.$$

■

Exercice 5.4 En général on écrit

$$\mathcal{L}_h z(t) = C_0 z(t) + C_1 h z'(t) + \ldots + C_p h^p z^{(p)}(t) + \mathcal{O}(h^{p+1}),$$

on sait que la méthode multipas est consistante d'ordre p si

$$C_0 = C_1 = \ldots = C_p = 0$$

et donc

$$\mathcal{L}_h z(t) = C_{p+1} h^{p+1} z^{(p+1)}(t) + \mathcal{O}(h^{p+2}).$$

On a aussi

$$C_m = \sum_{j=0}^{2} \left[\frac{1}{m!} j^m \alpha_j - \frac{1}{(m-1)!} j^{m-1} \beta_j \right],$$

avec $\alpha_2 = 1$. Utiliser la dernière formule pour calculer l'ordre et trouver les constantes d'erreurs de

(1) $x_{n+2} - x_{n+1} = \frac{1}{12} h \Big(5f_{n+2} + 8f_{n+1} - f_n \Big)$ (méthode d'Adams-Moulton à deux pas "AM(2)").

(2) $x_{n+2} - x_n = \frac{1}{3} h \Big(f_{n+2} + 4f_{n+1} + f_n \Big)$ (règle de Simpson : $p = 4$, $C_5 = -1/90$).

■

Exercice 5.5 Étudier la zero-stabilité des méthode multipas

(a) $x_{n+2} - 4x_{n+1} + 3x_n = -2hf_n$;
(b) $3x_{n+2} - 4x_{n+1} + x_n = ahf_n$.

Y a-t-il des valeurs de a telles que (b) soit convergente ?

■

Exercice 5.6 On considère la méthode multipas

$$x_{n+2} + (b-1)x_{n+1} - b x_n = \frac{1}{4} h \Big[(b+3) f_{n+2} + (3b+1) f_n \Big]$$

1. Montrer que l'ordre est 2 si $b \neq -1$ et 3 si $b = -1$;
2. Montrer que si $b = -1$ la méthode n'est pas zero-stable ;

3. Que se passe-t-il pour le problème de Cauchy $x'(t) = 0$, $x(0) = 0$ et $x_0 = 0$, $x_1 = h$?

■

Exercice 5.7 On considère la méthode multipas

$$x_{n+1} = x_{n-2} + \frac{3}{4}h\Big(f_{n+1} + f_n + f_{n-1} + f_{n-2}\Big).$$

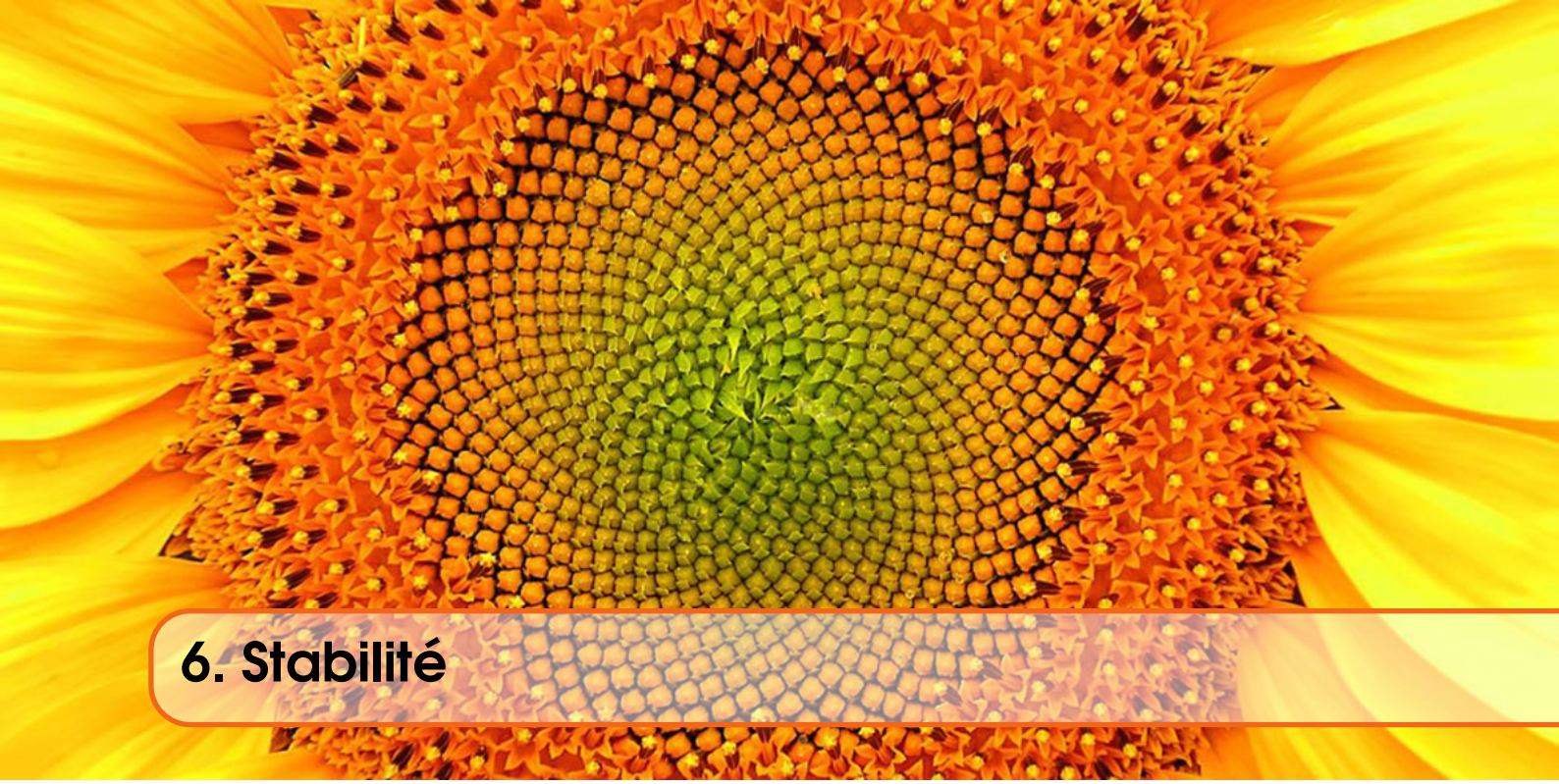
1. Combien de pas a-t-elle ?
2. Est-ce une méthode explicite ou implicite ?
3. Écrire les 1^{er} et 2^{nd} polynômes caractéristiques ;
4. Étudier la consistance et la zero-stabilité.

■

Exercice 5.8 La méthode multipas suivante est-elle convergente ?

$$x_{n+3} - x_{n+2} + x_{n+1} - x_n = \frac{1}{2}h\Big(f_{n+3} + f_{n+2} + f_{n+1} + f_n\Big)$$

■



6. Stabilité

6.1 Stabilité absolue - motivations

On suppose être en possession d'une méthode convergente pour résoudre

$$\begin{cases} x'(t) = f(t, x(t)), & t \in]t_0, t_f], \\ x(t_0) = \eta, \end{cases}$$

Soit $h = \sup_n h_n$ avec $h_n = t_{n+1} - t_n$. Comme les méthodes considérées sont convergentes, alors on a $\lim_{h\to 0} \sup_n |e_n| = 0$. Ainsi, si h est très petit, x_{n+1} est très proche de $x(t_{n+1})$. On s'intéresse dans ce chapitre au cas où h est « raisonnablement » petit (c'est à dire $h = 10^{-1}$ jusqu'à $h = 10^{-3}$ et non pas $h = 10^{-6}$ ou plus petit).

■ **Exemple 6.1** Considérons l'équation $x'(t) = -8x(t) + 40(3e^{-t/8} + 1)$ avec $x(0) = 100$. La solution exacte est donnée par

$$x(t) = \frac{1675}{21} e^{-8t} + \frac{320}{21} e^{-t/8} + 5.$$

Nous représentons sur la figure 6.1 la solution exacte.

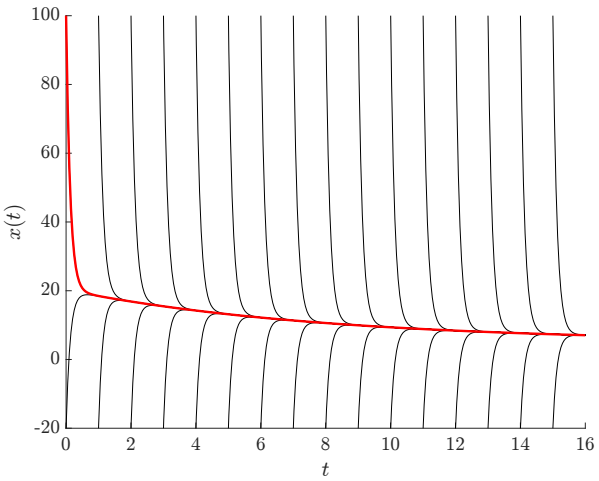


FIGURE 6.1 – Solution exacte

On va mener plusieurs expériences numériques avec les schémas d'Euler explicite et implicite pour

trois valeurs de h : 1/3, 1/5 et 1/9. La méthode d'Euler explicite donne le schéma

$$\begin{cases} x_{n+1} = (1-8h)x_n + h(120e^{-t_n/8} + 40), \quad n \geq 0, \\ x_0 = 100, \quad t_n = nh. \end{cases}$$

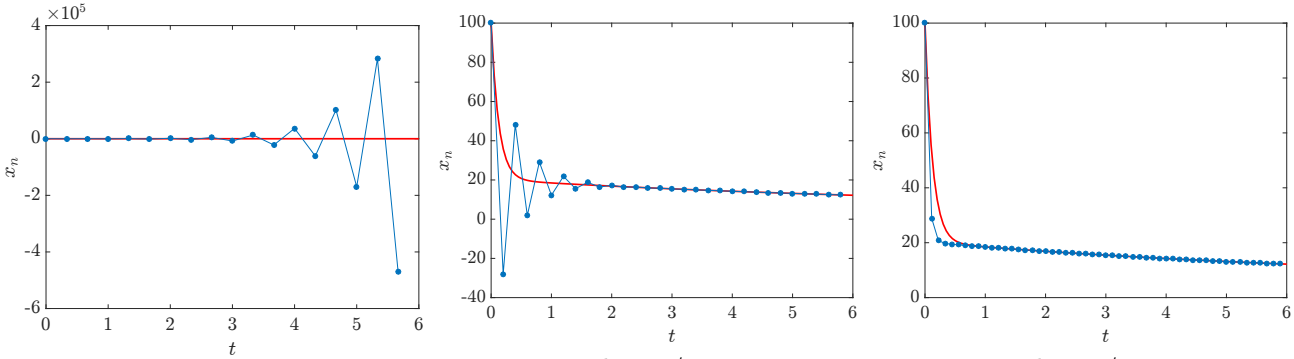
La méthode d'Euler implicite conduit à

$$x_{n+1} = x_n - 8hx_{n+1} + h(120e^{-t_{n+1}/8} + 40)$$

soit encore

$$x_{n+1} = \frac{1}{1+8h}(x_n + h(120e^{-t_{n+1}/8} + 40).$$

Observons tout d'abord les solutions générées par le schéma d'Euler explicite sur la figure 6.2 pour $0 \leq t \leq 6$.



$h = 1/3$ $h = 1/5$ $h = 1/9$

FIGURE 6.2 – Solutions par le schéma d'Euler explicite

On constate pour le cas $h = 1/3$ que la solution est oscillante avec des oscillations dont l'amplitude augmente jusqu'à 10^5. Ceci est le symptome d'une instabilité. Pour $h = 1/5$, un effet important sur la solution numérique subsiste. Elle continue à produire des oscillations mais qui s'attenuent. À partir de $t = 2$, la représentation graphique semble les faire disparaitre et la solution semble devenir une courbe régulière qui s'approche de la solution exacte. Enfin, pour $h = 1/9$, la solution ressemble à la solution exacte mais les courbes ne deviennent proches qu'à partir de $t = 0.5$.

Lorsque la solution numérique est proche de la solution exacte, les représenter toutes les deux sur la même figure ne suffit pas à les distinguer. On préfère pour visualiser la précision avec laquelle le solution numérique approche la solution exacte représenter l'évolution de l'erreur globale $e_n = |x(t_n) - x_n|$ en échelle log (voir figure 6.3).

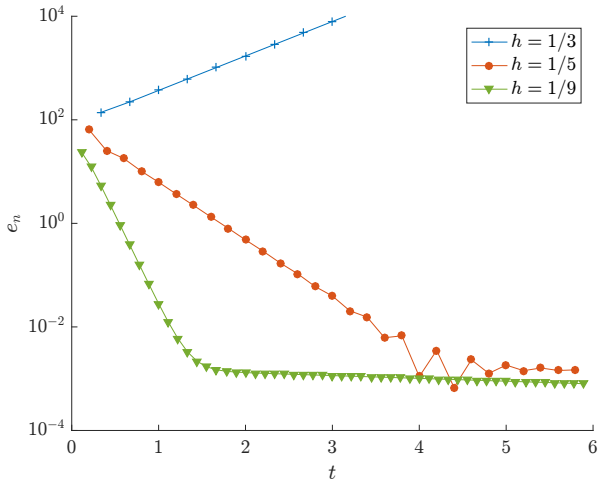
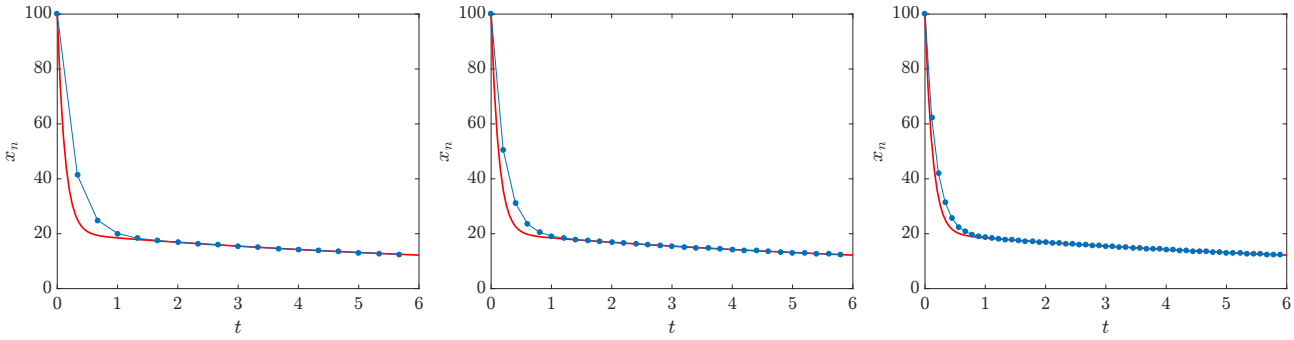


FIGURE 6.3 – Erreur globale produite par le schéma d'Euler explicite

On constate une croissane linéaire sur le graphe en échelle log pour $h = 1/3$. On est donc en présence d'une croissance exponentielle et donc d'un phénomène d'explosion. Pour $h = 1/5$, on voit une décroissance exponentielle de l'erreur jusqu'à $t = 4$ puis une stabilisation à environ 10^{-3}. De même, pour $h = 1/9$, il y a décroissance exponentielle plus rapide puis un plateau à environ 10^{-3}.

Nous allons montrer que la méthode d'Euler explicite souffre d'un phénomène d'instabilité pour $h \geq h_0$. C'est typique de l'équation différentielle considérée qui présente un phénomène d'évolution très rapide : la méthode d'Euler explicite n'est pas d'un intérêt pratique pour des solutions exponentiellement décroissantes à moins que h soit assez petit. Cette contrainte s'appelle **condition de stabilité**. En outre, même si h est assez petit pour éviter la croissance exponentielle de l'erreurm le niveau résiduel d'erreur est souvent plus élevé que celui que l'on peut espérer.

On reproduit ces cas tests avec la méthode d'Euler implicite. On constate sur la figure 6.4 que quelque soit h, on n'a plus d'oscillations et le comportement est raisonnable.



$h = 1/3$ $h = 1/5$ $h = 1/9$

FIGURE 6.4 – Solutions par le schéma d'Euler implicite

Les courbes d'erreur e_n sont maintenant en accord avec nos attentes (voir figure 6.5), mais toujours avec un phénomène de stabilisation.

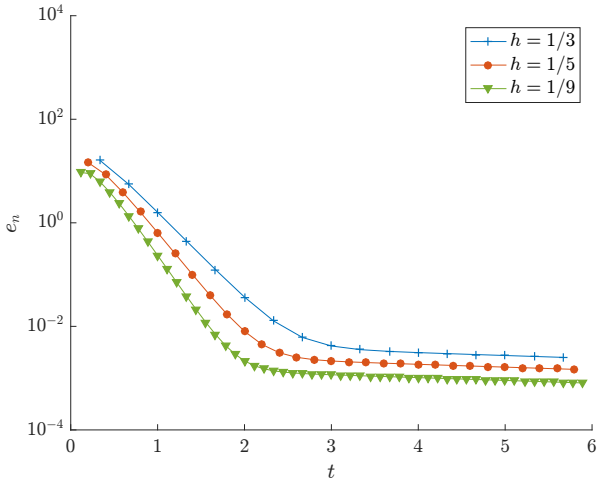


FIGURE 6.5 – Erreur globale produite par le schéma d'Euler implicite

Ce qui différencie les deux schémas considérés ici, c'est que les erreurs locales de troncature ε_n pour Euler explicite et implicite sont égales mais de signe opposé. ∎

6.2 Stabilité absolue

Les exemples vus ci-dessus nous conduisent à la théorie de la stabilité absolue, dans laquelle on examine l'effet de l'application des méthodes numériques convergentes au problème scalaire modèle

$$x'(t) = \lambda x(t), \quad \lambda \in \mathbb{C},\ \mathrm{Re}(\lambda) < 0. \tag{6.1}$$

La solution exacte est donnée par $x(t) = ce^{\lambda t}$, où c est une constante arbitraire. Comme $\mathrm{Re}(\lambda) < 0$, $\lim_{t\to+\infty} x(t) = 0$ pour tout c.

Notre but est de trouver quelles sont les méthodes qui appliquées à (6.1) donne une suite $(x_n)_n$ qui tend vers 0 quand $t_n \to +\infty$ avec h fixé. C'est différent de la convergence où on fixe le temps d'observation et $h \to 0$ avec $n \to \infty$.

Définition 6.2.1 Une méthode multipas est dite **absolument stable** si quand on l'applique à (6.1) avec un pas $\hat{h} = \lambda h$, sa solution tend vers 0 quand $n \to \infty$ pour toute donnée initiale.

Prenons une méthode multipas à 2 étapes. Comme $x' = \lambda x$, elle devient

$$x_{n+2} + \alpha_1 x_{n+1} + \alpha_0 x_n = h\lambda(\beta_2 x_{n+2} + \beta_1 x_{n+1} + \beta_0 x_n)$$

soit encore

$$(1 - \hat{h}\beta_2)x_{n+2} + (\alpha_1 - \hat{h}\beta_1)x_{n+1} + (\alpha_0 - \hat{h}\beta_0)x_n = 0.$$

Les solutions sont de la forme $x_n = ar^n$ où r est solution de

$$\underbrace{(1 - \hat{h}\beta_2)r^2 + (\alpha_1 - \hat{h}\beta_1)r + (\alpha_0 - \hat{h}\beta_0)}_{p(r)=\rho(r)-\hat{h}\sigma(r)} = 0.$$

On appelle $p(r) = \rho(r) - \hat{h}\sigma(r)$ le **polynôme de stabilité** de la méthode multipas. Ici, p à deux racines r_1 et r_2. Donc, $x_n = ar_1^n + br_2^n$ avec $a \neq b$ si $r_1 \neq r_2$.

Si on veut $|x_n| \to 0$ quand $n \to \infty$ pour tout a et b, il faut que $|r_1| < 1$ et $|r_2| < 1$. Donc, le polynôme p doit satisfaire la condition de racines stricte.

Lemme 6.2.1 Une méthode multipas est absolument stable pour $\hat{h} = \lambda h$ si et seulement si son polynôme de stabilité vérifie la condition de racines stricte.

Une méthode multipas n'est pas en général absolument stable pour tout choix de $\hat{h}$. On est donc amené à introduire la notion de région de stabilité.

Définition 6.2.2 — Région de stabilité absolue. L'ensemble $\mathscr{R}$ des valeurs dans le plan complexe pour lesquelles une méthode est absolument stable forme la région de stabilité absolue.

La question est donc : pour quelles valeurs de $\hat{h}$ les racines de p vérifient $|r| < 1$? Les cas $\lambda \in \mathbb{R}$, $\lambda < 0$ sont plus simples à analyser.

Définition 6.2.3 — Intervalle de stabilité absolue. L'intervalle de stabilité absolue pour une méthode multipas est l'intervalle le plus grand de la forme $\mathscr{R}_0 = (\hat{h}_0, 0)$, avec $\hat{h}_0 < 0$, pour lequel la méthode est absolument stable pour toutes les valeurs réelles $\hat{h} \in \mathscr{R}_0$. Ainsi, $\mathscr{R}_0 = \mathscr{R} \cap \mathbb{R}$.

■ **Exemple 6.2** Examinons le cas de la méthode d'Euler explicite. Le schéma appliqué à (6.1) est $x_{n+1} = (1+\hat{h})x_n$. Le polynôme de stabilité est $p(r) = r - (1+\hat{h})$ qui admet une racine simple $r_1 = 1+\hat{h}$. La région de stabilité absolue est donc

$$\mathscr{R} = \{\hat{h} \in \mathbb{C} |\ |1 + \hat{h}| < 1\}.$$

C'est l'intérieur du disque (sans le bord) de centre $\hat{h} = -1$ et de rayon 1. Si $\hat{h} \in \mathbb{R}$, l'intervalle de stabilité est défini par $-1 < 1 + \hat{h} < 1$ soit $\hat{h} \in]-2, 0[$. Par exemple, si $\lambda = -8$, alors on a la contrainte $h \in]0, 1/4[$ pour avoir stabilité absolue. La région de stabilité est représentée sur la figure 6.6.

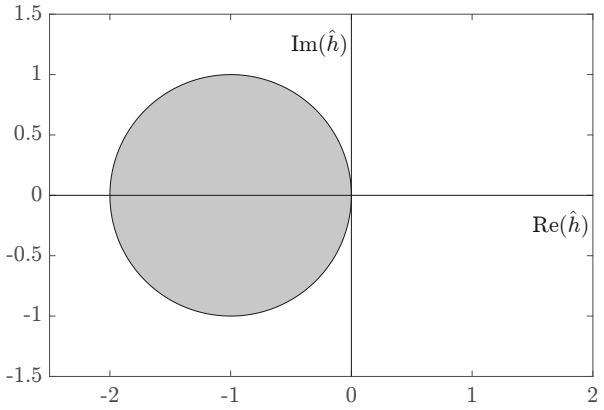


FIGURE 6.6 – Région de stabilité pour le schéma d'Euler explicite

■

■ **Exemple 6.3** Quelle est la valeur limite de h que l'on peut utiliser telle que Euler explicite soit absolument stable pour $x' = \lambda x$ avec $\lambda = -4+3i$? On a $\hat{h} = h(-4+3i)$. Donc, $|1+\hat{h}| = |1+h(-4+3i)|$. Demander $|1+\hat{h}| < 1$ revient à chercher $|1+\hat{h}|^2 < 1$ ce qui équivaut à $|1+\hat{h}|^2 - 1 = h(-8+25h) < 0$. Donc, il faut que $h < 8/25$. ■

■ **Exemple 6.4** Région de stabilité absolue pour la règle des trapèzes $x_{n+1} - x_n = (\hat{h}/2)(x_{n+1} + x_n)$. Le polynôme de stabilité est $p(r) = r - 1 - (\hat{h}/2)(r+1)$. Sa racine est simple $r_1 = (1+\hat{h}/2)/(1-\hat{h}/2)$. On a donc $|r_1| < 1$ pour tout $\hat{h}$, $\mathrm{Re}(\hat{h}) < 0$. Ainsi, $\mathscr{R} = \{\hat{h} | \mathrm{Re}(\hat{h}) < 0\}$. ■

Le lemme suivant est intéressant quand p est quadratique et $\hat{h} \in \mathbb{R}$.

Lemme 6.2.2 — Conditions de Jury. Le polynôme $q(r) = r^2 + ar + b$, $(a,b) \in \mathbb{R}^2$, vérifie la condition de racines stricte si et seulement si
1. $b < 1$
2. $1 + a + b > 0$
3. $1 - a + b > 0$

Les trois conditions précédentes définissent la région triangulaire ci-dessous

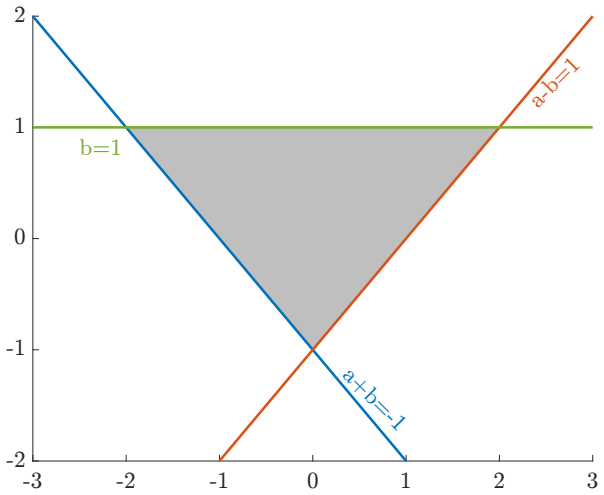


FIGURE 6.7 – Représentation des conditions de jury

Preuve On a $q(r) = r^2 + ar + b$. Les deux racines sont données par $r_{1,2} = (-a \pm \sqrt{a^2 - 4b})/2$. Ainsi

$$q(t) = (r - r_1)(r - r_2) = r^2 - (r_1 + r_2)r + r_1 r_2.$$

— Si $a^2 < 4b$, les racines sont complexes conjuguées. Mais, $b = r_1 r_2 = |r_1|^2 = |r_2|^2$. On a donc une condition de racines stricte si et seulement si $b < 1$. D'autre part, on a l'égalité (a et b sont réels)

$$a^2 - 4b = (|a| - 2)^2 - 4(1 + b - |a|), \forall (a,b) \in \mathbb{R}^2.$$

Comme $a^2 < 4b$, on a l'estimation

$$4(1+b-|a|) = (|a|-2)^2 - (a^2-4b) > 0$$

de sorte que $1+b-|a| > 0$ et donc $|a|-1 < b$.

— Si $a^2 \geq 4b$, la racine de plus grande amplitude R est

$$R = \max\{|r_1|, |r_2|\} = frac12(|a| + \sqrt{a^2-4b}).$$

C'est une fonction croissante de $|a|$ et on a $R = 1$ quand $|a| = 1+b$. Donc, $0 \leq R < 1$ est équivalent à $0 \leq |a| < 1+b$. On a aussi $|r_1 r_2| < 1$ et donc $|b| > 1$.

En combinant les résultats pour les racines réelles ou complexes, on trouve que la condition de racines stricte est vérifiée si et seulement si $|a|-1 < b < 1$. □

R Ces conditions sont équivalentes à $q(0) < 1$ et $q(\pm 1) > 0$ ce qui est plus simple à se souvenir

Si on veut appliquer ce lemme au polynôme de stabilité $p(r) = (1-\hat{h}\beta_2)r^2 + (\alpha_1 - \hat{h}\beta_1)r + (\alpha_0 - \hat{h}\beta_0)$, il faut diviser p par le coefficient du terme r^2 pour être de la forme $q(r)$, de sorte que

$$a = \frac{\alpha_1 - \hat{h}\beta_1}{1-\hat{h}\beta_2}, \quad b = \frac{\alpha_0 - \hat{h}\beta_0}{1-\hat{h}\beta_2}.$$

On peut montrer que si $\hat{h} \in \mathscr{R}_0$, alors a et b sont strictement positifs et donc la condition $q(\pm 1) > 0$ peut être remplacée par $p(\pm 1) > 0$, c'est à dire sans faire la division. Si la méthode multipas est explicite, alors $\beta_2 = 0$ et $q(r)$ coïncide avec $p(r)$.

■ **Exemple 6.5** Recherche de l'intervalle de stabilité absolue de la méthode $x_{n+2} - x_{n+1} = hf_n$. Le polynôme de stabilité est $p(r) = r^2 - r - \hat{h}$ avec $\hat{h} \in \mathbb{R}$. Comme les coefficients sont réels, on peut utiliser le lemme et $p(r) = q(r)$. A-t-on $p(\pm 1) > 0$ et $p(0) < 1$?

$$\left.\begin{array}{lll} p(0) < 1 \ : & -\hat{h} < 1 & \Rightarrow \ \hat{h} > -1 \\ p(1) > 0 \ : & -\hat{h} > 0 & \Rightarrow \ \hat{h} < 0 \\ p(-1) > 0 \ : & 2-\hat{h} > 0 & \Rightarrow \ \hat{h} < 2 \end{array}\right\} \Rightarrow -1 < \hat{h} < 0$$

Donc, $\mathscr{R}_0 =]-1, 0[$. ■

6.3 Méthode de localisation de la frontière

Il est en général difficile de déterminer la région de stabilité absolue car on doit décider pour quels $\hat{h} \in \mathbb{C}$ les racines du polynôme de stabilité vérifie la condition de racines stricte ($|r| < 1$). Il est plus aisé de déterminer la frontière de cette région car au moins une des racines de p à la frontière est de module 1. On a donc $|r| = 1$ sur $\partial\mathscr{R}$. La frontière de $\mathscr{R}$ est un sous ensemble des points $\hat{h} \in \mathbb{C}$ pour lesquels $r = e^{i\theta}$, $\theta \in \mathbb{R}$. On remplace $r = e^{i\theta}$ dans p. Comme $e^{i\theta}$ est racine, $p(e^{i\theta}) = 0$ et on résoud l'équation. On obtient $\hat{h} = \hat{h}(\theta)$ ce qui donne une courbe dans $\mathbb{C}$ en fonction de l'angle θ, ce qui décrit donc une courbe polaire.

■ **Exemple 6.6** On considère le schéma $x_{n+2} - x_{n+1} = hf_n$. Le polynôme de stabilité est $p(r) = r^2 - r - \hat{h}$. Si r est solution de $p(r) = 0$, on a $\hat{h} = r^2 - r$ et on a

$$\hat{h}(\theta) = e^{2i\theta} - e^{i\theta} = \underbrace{\cos(2\theta) - \cos\theta}_{x(\theta)} + i\underbrace{\sin(2\theta) - \sin\theta}_{y(\theta)}.$$

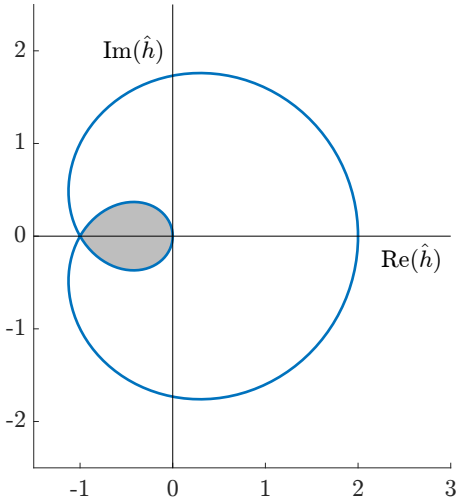


FIGURE 6.8 – Région de stabilité

La courbe représentée en figure 6.8 décompose le plan complexe $\mathbb{C}$ en trois régions. Il reste à décider quelles régions correspondent à cette zone de stabilité absolue. Il suffit de tester un point par région. On sait déjà que $\mathscr{R}_0 =]-1, 0[$. Vérifions le en prenant $\hat{h} = -1/2$. On a $p(r) = r^2 - r + 1/2$ et les racines sont $r = (1 \pm i)/2$ et donc $|r| = 1/\sqrt{2} < 1$ ce qui donne l'absolue stabilité. Si $\hat{h} = 1$, $p(r) = r^2 - r - 1$ de racines $r = (1 \pm \sqrt{5})/2$ et $|r| > 1$, et donc pas absolue stabilité. De même, $\hat{h} = -2$ donne $p(r) = r^2 - r + 2$ de racines $r = (1 \pm i\sqrt{7})/2$ et $|r| > 1$ d'où la non absolue stabilité. ∎

6.4 A-stabilité

Quelques méthode multipas (par exemple celle de la règle des trapèzes) appliquées à $x' = \lambda x$, $\mathrm{Re}(\lambda) < 0$ vérifient $\lim_{n\to\infty} x_n = 0$ quand $\lim_{t\to\infty} x(t) = 0$ et cela pour tout h. C'est un meilleur comportement que celui des méthodes où il faut imposer des restrictions à h pour avoir cette propriété.

Définition 6.4.1 — A-stabilité. Une méthode numérique est dite A-stable si sa région de stabilité absolue $\mathscr{R}$ inclut entièrement le demi plan $\mathrm{Re}(\hat{h}) < 0$.

Théorème 6.4.1 — Seconde barrière de Dahlquist.
1. Il n'y a aucune méthode multipas explicite A-stable.
2. une méthode multipas A-stable (implicite) ne peut avoir un ordre $p > 2$
3. La méthode multipas A-stable de constante d'erreur la plus petite est celle liée à la règle des trapèzes.

On peut relacher sensiblement les contraintes quand $\lambda \in \mathbb{R}$.

Définition 6.4.2 — A_0-stabilité. Une méthode numérique est dite A_0-stable si son intervalle de stabilité absolue inclut entièrement l'axe réel négatif $\mathbb{R}^-$, soit $\mathrm{Re}(\hat{h}) < 0$ et $\mathrm{Im}(\hat{h}) = 0$.

R Les méthodes A-stables sont étudiées sur une équation linéaire $x' = \lambda x$. Il est remarquable qu'en général, ce sont les meilleurs méthodes également pour les équations non linéaires.

6.5 Extension aux systèmes d'EDO

Comme pour les EDO, on considère un système linéaire

$$u'(t) = Au(t), \quad u(t) \in \mathbb{R}^m, \quad A \in \mathbb{R}^{m\times m},$$

avec A diagonalisable. La première chose à faire est de diagonaliser a. Soit les vecteurs propres $v_1, \cdots, v_m$ et les valeurs propres associées $\lambda_1, \cdots, \lambda_m$, c'est à dire

$$Av_j = \lambda_j v_j, \quad \lambda_j \in \mathbb{C}.$$

Soit V la matrice dont les colonnes sont les v_i. Alors, on a

$$V^{-1}AV = \Lambda, \quad \text{où } \Lambda = \text{diag}(\lambda_i)_{i=1}^m.$$

Soit $u(t) = Vx(t)$. Comme $u'(t) = Au(t) = V\Lambda V^{-1}u(t)$, on a

$$\left(V^{-1}u\right)'(t) = \Lambda V^{-1}u(t).$$

On a donc $x'(t) = \Lambda x(t)$ et donc pour chaque composante $x_i' = \lambda_i x_i$ qui sont des EDOs linéaires indépendantes. On a donc réduit le système d'EDOs $u' = Au$ en un système de m équations linéaires indépendantes.

■ **Exemple 6.7** On considère la matrice $A = \begin{pmatrix} 1 & 3 \\ -2 & -4 \end{pmatrix}$. Les valeurs propres sont $\lambda_1 = -1$ et $\lambda_2 = -2$ de vecteurs propres associés respectifs $v_1 = (3, -2)^t$ et $v_2 = (-1, 1)^t$. Les matrices intervenant dans le problème sont donc

$$V = \begin{pmatrix} 3 & -1 \\ -2 & 1 \end{pmatrix}, \quad \Lambda = \begin{pmatrix} -1 & 0 \\ 0 & -2 \end{pmatrix}.$$

Le système différentiel est

$$x'(t) = \begin{pmatrix} -1 & 0 \\ 0 & -2 \end{pmatrix} x(t)$$

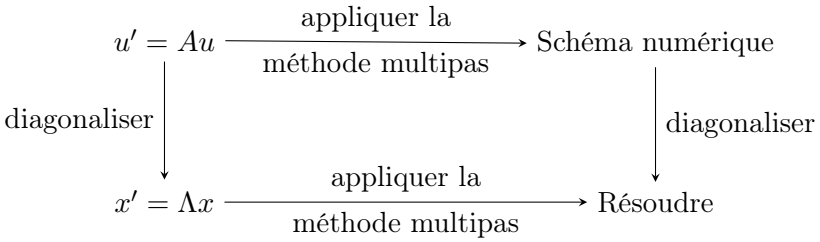
de solution $x_1(t) = \alpha e^{-t}$ et $x_2 = \beta e^{-2t}$ d'où

$$u(t) = Vx(t) = \alpha e^{-t} \begin{pmatrix} 3 \\ -2 \end{pmatrix} + \beta e^{-2t} \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

d'où $\lim_{t\to\infty} u(t) = 0$. ■

Théorème 6.5.1 Si A est diagonalisable de valeurs propres $\lambda_1, \cdots, \lambda_m$, alors les solutions de $u'(t) = Au(t)$ tendent vers 0 quand $t \to +\infty$ pour toute donnée initiale si et seulement si $\text{Re}(\lambda_i) < 0$ pour tout i.

Dans l'application des méthodes multipas aux systèmes d'EDOs diagonalisables, on peut suivre le diagramme suivant :



Définition 6.5.1 Soit le système différentiel $u' = Au$ dont toutes les solutions tendent vers 0 quand $t \to +\infty$ pour toute donnée initiale. Une méthodes multipas est dite absolument stable pour un pas h fixé si toutes ses solutions lorsque cette méthode est appliquée au système différentiel tendent vers 0 quand $n \to \infty$.

Proposition 6.5.2 Si A est diagonalisable, une méthode multipas est absolument stable si $\lambda_i h \in \mathscr{R}$, $\forall \lambda_i \in \text{Spectre}(A)$.

■ **Exemple 6.8** Considèrons le système différentiel

$$\begin{cases} u'(t) = -11u(t) + 100v(t) \\ v'(t) = u(t) - 11v(t) \end{cases}$$

La matrice associée est $A = \begin{pmatrix} -11 & 100 \\ 1 & -11 \end{pmatrix}$. Les valeurs propres sont -1 et -21. La méthode d'Euler explicite appliquée à ce système est

$$\begin{cases} u_{n+1} = u_n + h(-11u_n + 100v_n) \\ v_{n+1} = v_n + h(u_n - 11v_n) v'(t) = u(t) - 11v(t) \end{cases}$$

On a vu pour la méthode d'Euler que l'intervalle de stabilité est $]-2, 0[$. On demande donc $h\lambda_i \in]-2, 0[$, $i = 1, 2$ ce qui se traduit

$$-2 < -h < -0 \quad \text{et} \quad -2 < -21h < 0$$

soit encore

$$0 < h < \frac{2}{21} \approx 0,0952.$$

La solution exacte est

$$\begin{pmatrix} u \\ v \end{pmatrix} = \frac{1}{20} \begin{pmatrix} 10((u_0 + 10v_0)e^{-t} + (u_0 - 10v_0)e^{-21t}) \\ (u_0 + 10v_0)e^{-t} - (u_0 - 10v_0)e^{-21t} \end{pmatrix}.$$

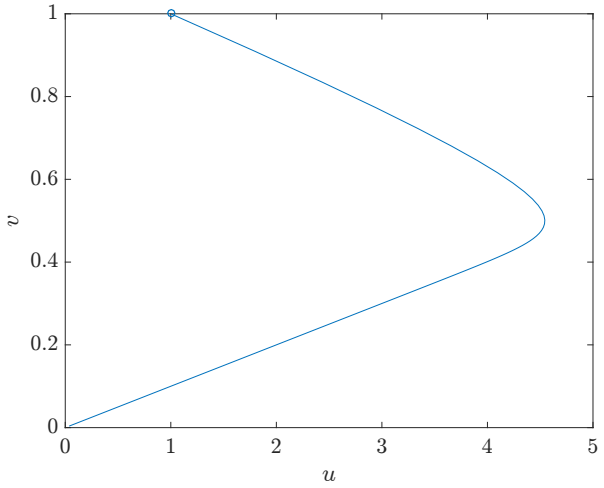


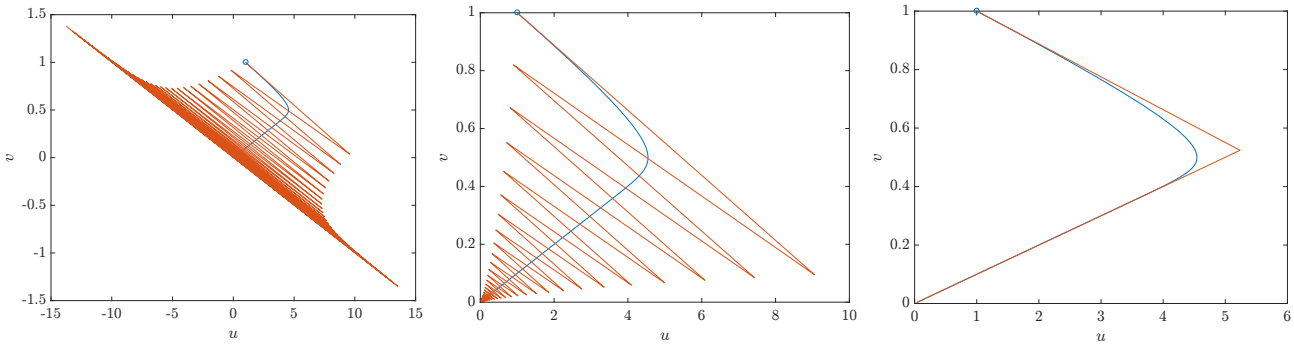
FIGURE 6.9 – Solution exacte dans le plan de phase

Si $u_0 = v_0 = 1$, on a donc

$$\begin{pmatrix} u \\ v \end{pmatrix} = \underbrace{\frac{11}{20} \begin{pmatrix} 10 \\ 1 \end{pmatrix} e^{-t}}_{\substack{\text{dcroissance} \\ \text{lente}}} - \underbrace{\frac{9}{20} \begin{pmatrix} 10 \\ -1 \end{pmatrix} e^{-21t}}_{\substack{\text{phase transitoire} \\ \text{rapide}}}.$$

■

La représentation graphique se fait dans le plan de phase. Comme c'était prévisible, pour $h = 0.096$ qui dépasse de 1% la limite de stabilité $h = 2/21$, l'amplitude des solutions croit et il existe des oscillations rapides qui sont le signe de l'instabilité. Pour $h = 0.0905$ qui est 5% sous la limite, la solution tend bien vers $(0, 0)$, mais continue à comporter de fortes oscillations. Pour $h = 0.0476$, soit 50% en dessous de la limite, on a un comportement correct.



$h = 0.0962$ $h = 0.0905$ $h = 0.0476$

FIGURE 6.10 – Solutions par le schéma d'Euler explicite

R Comme on le voit sur cet exemple, la plus grande valeur propre en module contraint très fortement la méthode numérique à cause d'une forte contrainte sur h alors que la plus petite valeur propre en module n'a que peu d'effet : on dit que le système est **raide**. Pour préciser, on dira qu'un système est raide si

$$\frac{\max -\mathrm{Re}(\lambda_j)}{\min -\mathrm{Re}(\lambda_j)} \gg 1.$$

6.6 Exercices

Exercice 6.1 On considère la méthode des trapèzes, le polynôme de stabilité associé et sa racine r_1

$$x_{n+1} - x_n = \frac{1}{2}h(f_{n+1} - f_n), \qquad p(r) = r - 1 - \frac{1}{2}\widehat{h}(r+1), \qquad r_1 = \frac{1 + \frac{1}{2}\widehat{h}}{1 - \frac{1}{2}\widehat{h}}.$$

En utilisant $\widehat{h} = 2X + 2iY$ prouver que

$$|r_1|^2 - 1 = \frac{4X}{(1 - X^2) + Y^2}$$

et déduire que pour tout $Re(\widehat{h}) < 0$ on a $|r_1| < 0$.

Qu'est-ce que on peut en déduire sur l'intervalle de stabilité absolue de la méthode de trapèzes ?

■

Exercice 6.2 Déterminer l'intervalle de stabilité absolue de la méthode multipas

$$x_{n+2} - x_n = \frac{1}{2}h(f_{n+1} + 3f_n).$$

■

Exercice 6.3 Soit $a \in \mathbb{R}$ un paramètre pour la méthode multipas

$$x_{n+2} - 2ax_{n+1} + (2a-1)x_n = h\Big(af_{n+2} + (2-3a)f_{n+1}\Big).$$

1. Calculer les 1^{er} et 2^{nd} polynômes caractéristiques ;
2. Étudier la consistance ;
3. Étudier la zero-stabilité ;
4. Étudier la convergence ;
5. Quels sont l'ordre et la constante d'erreur ?
6. Y a-t-il de schémas A_0-stable ?
7. Commenter sur le schéma "BDF(2)" $\quad 3x_{n+2} - 4x_{n+1} + x_n = 2hf_{n+2}$;

8. Le théorème de la “Seconde barrière de Dahlquist” est-t-il respecté ?

■

Exercice 6.4 Calculer l'intervalle de stabilité absolue pour la méthode d'Euler appliquée au problème de Cauchy

$$\begin{cases} u'(t) = -8(u(t) - v(t)), & u(0) = 100, \\ v'(t) = -(v(t) - 5)/8, & v(0) = 20. \end{cases}$$

■

Exercice 6.5 On considère le problème de Cauchy

$$\begin{cases} x'(t) = (1 - 2t)x(t), & t > 0, \\ x(0) = 1 \end{cases}$$

et sa solution exacte $x(t) = \exp[\frac{1}{4} - (\frac{1}{2} - t)^2]$.

1. Calculer x_1 à une précision de six chiffres décimales en utilisant $h = 0.1$ et les schémas suivants

2 étapes

0	0	
a	a	0
	$1-\theta$	θ

Heun 3

0	0		
$\frac{1}{3}$	$\frac{1}{3}$	0	
$\frac{2}{3}$	0	$\frac{2}{3}$	0
	$\frac{1}{4}$	0	$\frac{3}{4}$

Kutta 3

0	0		
$\frac{1}{2}$	$\frac{1}{2}$	0	
1	-1	2	0
	$\frac{1}{6}$	$\frac{4}{3}$	$\frac{1}{6}$

RK 4

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

TABLE 6.1 – Schémas à 2 étapes : $a = 1/(2\theta)$; Improved Euler $\theta = 1/2$, Modified Euler $\theta = 1$.

2. Comparer les résultats (table 6.2) avec la solution exacte.

Improved Euler	Modified Euler	Heun 3	Kutta 3	RK 4
1.094000	1.094500	1.094179	1.094187	1.094174

TABLE 6.2 – Résultats

■

Exercice 6.6 Appliquer la méthode de RK d'ordre 2 à deux étapes (cf. table 6.1) à l'EDO

$$x'(t) = \lambda x(t).$$

1. Comparer x_{n+1} au développement de Taylor de $x(t_{n+1})$ et montrer que la différence vaut $\mathcal{O}(h^3)$;
2. En déduire que l'ordre de la méthode est supérieur ou égal à 2 ;
3. Conclure (à l'aide les résultats théoriques) que l'ordre est exactement 2.

■

Exercice 6.7 Vérifier que les deux schémas de Runge-Kutta à 3 étapes “Heun 3” et “Kutta 3” (cf. table 6.1) ont la même fonction de stabilité $R(\widehat{h})$. ■

7. Les méthodes de Runge-Kutta

7.1 Description de la méthode

On considère le problème

$$\begin{cases} x'(t) = f(t, x(t)), & t \in]t_0, t_f], \\ x(t_0) = \eta, \end{cases}$$

et une subdivision $t_0 < t_1 < \cdots < t_N = t_0 + T$.

Sur l'intervalle $[t_n, t_{n+1}]$, on a $x(t_{n+1}) = x(t_n) + \int_{t_n}^{t_{n+1}} f(s, x(s))ds$ soit encore

$$x(t_{n+1}) = x(t_n) + \int_0^1 h_n f(t_n + \sigma h_n, x(t_n + \sigma h_n))d\sigma$$

ou bien

$$x(t_{n+1}) = x(t_n) + h_n \int_0^1 g(\sigma)d\sigma, \quad g(\sigma) = f(t_n + \sigma h_n, x(t_n + \sigma h_n)).$$

On a vu que pour construire une méthode numérique pour l'EDO ci-dessus, on peut approcher le calcul de $\int_0^1 g(\sigma)d\sigma$ par une formule de quadrature et on a déjà rencontré

Quadrature	Schéma numérique
Rectangle à gauche	Euler explicite
Rectangle à droite	Euler implicite
Trapèzes	Crank Nicolson

On veut maintenant utiliser une quadrature générale

$$\int_0^1 g(\sigma)d\sigma \approx \sum_{i=1}^{s} b_i g(c_i), \quad g(c_i) = f(t_n + c_i h_n, x(t_n + c_i h_n)).$$

Le problème est que l'on ne connait pas $x(t_n + c_i h_n) := x(t_{n,i})$. On a cependant déjà rencontré ce problème lors de l'étude des méthodes de prédicteur-correcteur. On suit donc la même idée : *on évalue la fonction x aux points $t_{n,i} = t_n + c_i h_n$ par une quadrature*

$$x(t_{n,i}) = x(t_n) + h_n \int_0^{c_i} g(\sigma)d\sigma.$$

Pour simplifier la présentation, on choisit ici des méthodes explicites où on approche $\int_0^{c_i} g(\sigma)d\sigma$ avec des valeurs de $g(c_j) = f(t_{n,j}, x(t_{n,j}))$, $j = 1, \cdots, i-1$, antérieurement calculées de sorte que

$$\int_0^{c_i} g(\sigma)d\sigma \approx \sum_{j=1}^{i-1} a_{i,j} g(c_j)$$

et on a donc

$$x(t_{n,i}) \approx x(t_n) + h_n \sum_{j=1}^{i-1} a_{ij} f(t_{n,j}, x(t_{n,j})),$$
$$x(t_{n+1}) \approx x(t_n) + h_n \sum_{i=1}^{s} b_i f(t_{n,i}, x(t_{n,i})),$$

avec $t_{n,i} = t_n + c_i h_n$.

La méthode de Runge-Kutta à s-étages explicite est donnée par

$$\begin{cases} k_i = f(t_{n,i}, x_n + h_n \sum_{j=1}^{i-1} a_{ij} k_j), \quad i = 1, \cdots, s \\ x_{n+1} = x_n + h_n \sum_{i=1}^{s} b_i k_i. \end{cases}$$

■ **Exemple 7.1 Méthode de Runge (trapèze)**

$$\begin{array}{llll} k_1 = f(t_n, x_n) & c_1 = 0 & a_{11} = 0 & a_{12} = 0 \\ k_2 = f(t_{n+1}, x_n + h_n k_1) & c_2 = 1 & a_{21} = 1 & a_{22} = 0 \\ x_{n+1} = x_n + h_n(\frac{1}{2}k_1 + \frac{1}{2}k_2) & & b_1 = \frac{1}{2} & b_2 = \frac{1}{2} \end{array}$$

Méthode de Runge (point milieu)

$$\begin{array}{llll} k_1 = f(t_n, x_n) & c_1 = 0 & a_{11} = 0 & a_{12} = 0 \\ k_2 = f(t_n + h_n/2, x_n + h_n k_1/2) & c_2 = 1/2 & a_{21} = 1/2 & a_{22} = 0 \\ x_{n+1} = x_n + h_n(0k_1 + 1k_2) & & b_1 = 0 & b_2 = 1 \end{array}$$

Méthode de Heun

$$\begin{array}{llllll} k_1 = f(t_n, x_n) & c_1 = 0 & a_{11} = 0 & a_{12} = 0 & a_{13} = 0 \\ k_2 = f(t_n + h_n/3, x_n + h_n k_1/3) & c_2 = 1/3 & a_{21} = 1/3 & a_{22} = 0 & a_{23} = 0 \\ k_3 = f(t_n + 2h_n/3, x_n + 2h_n k_2/3) & c_2 = 2/3 & a_{31} = 0 & a_{32} = 2/3 & a_{33} = 0 \\ x_{n+1} = x_n + h_n(\frac{1}{4}k_1 + 0k_2 + \frac{3}{4}k_3) & & b_1 = 1/4 & b_2 = 0 & b_3 = 3/4 \end{array}$$

■

L'usage veut que l'on représente les méthodes de Runge-Kutta sous la forme de tableaux de Butcher qui prennent la forme $\begin{array}{c|c} \mathbf{c} & \mathbf{A} \\ \hline & \mathbf{b}^t \end{array}$. On a donc pour les méthodes ci-dessus la représentation

Runge (trapèze)
$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & 1/2 & 1/2 \end{array}$$

Runge (point milieu)
$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1/2 & 1/2 & 0 \\ \hline & 0 & 1 \end{array}$$

Heun
$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 \\ 2/3 & 0 & 2/3 & 0 \\ \hline & 1/4 & 0 & 3/4 \end{array}$$

Les méthodes de Runge-Kutta peuvent évidemment être implicites. Si l'on considère une méthode de quadrature de Gauss à 2 points

$$\int_0^1 g(\sigma)d\sigma \approx \frac{1}{2} g(\frac{1}{2} - \gamma) + \frac{1}{2} g(\frac{1}{2} + \gamma), \quad \gamma = \frac{\sqrt{3}}{6}$$

alors on a la méthode de Runge-Kutta

$$\begin{aligned} k_1 &= f\left(t_n + (\frac{1}{2}-\gamma)h_n, x_n + h_n(\frac{1}{4}k_1 + (\frac{1}{4}-\gamma)k_2)\right), \\ k_2 &= f\left(t_n + (\frac{1}{2}+\gamma)h_n, x_n + h_n((\frac{1}{4}+\gamma)k_1 + \frac{1}{4}k_2)\right), \\ x_{n+1} &= x_n + h_n(\frac{1}{2}k_1 + \frac{1}{2}k_2). \end{aligned}$$

C'est une méthode d'ordre 4, A-stable et de tableau de Butcher

$$\begin{array}{c|cc} 1/2-\gamma & 1/4 & 1/4-\gamma \\ 1/2+\gamma & 1/4+\gamma & 1/4 \\ \hline & 1/2 & 1/2 \end{array}$$

Définition 7.1.1 — Méthodes de Runge-Kutta générales. Soient b_i, a_{ij} des nombres réels et $c_i = \sum_{j=1}^{s} a_{ij}$. Une méthode de Runge-Kutta à s étages est donnée par

$$\begin{cases} k_i = f\left(t_n + c_i h_n, x_n + h_n \sum_{j=1}^{s} a_{ij} k_j\right), \quad i = 1, \cdots, s \\ x_{n+1} = x_n + h_n \sum_{i=1}^{s} b_i k_i. \end{cases}$$

Si $a_{ij} = 0$, $j \geq i$, alors la méthode est explicite.

Définition 7.1.2 Une méthode de Runge-Kutta a un ordre p si pour toutes solutions suffisamment régulières de $x' = f(t,x)$, l'erreur locale de troncature $\varepsilon_1 = x_1 - x(t_1)$ vérifie $x_1 - x(t_1) = \mathcal{O}(h^{p+1})$ quand $h_n \to 0$.

7.2 Consistance

Considérons pour (beaucoup) simplifier que les méthodes de Runge-Kutta sont explicites.

7.2.1 Méthodes RK à une étape

On a $s = 1$ et $x_{n+1} = x_n + h_n b_1 k_1$ avec $k_1 = f(t_n + c_1 h_n, x_n + a_{11} k_1)$. Mais, la méthode est explicite donc $a_{11} = 0$ et $c_1 = 0$ (car $c_i = \sum a_{ij}$). Ainsi, $x_{n+1} = x_n + h_n b_1 f(t_n, x_n)$ et donc $x_1 = x_0 + b_1 f(t_0, x_0)$. Par Taylor, on a

$$x(t_0 + h) = x(t_0) + h \underbrace{x'(t_0)}_{f(t_0,x_0)} + \frac{h^2}{2} x''(t_0) + \mathcal{O}(h^3).$$

Comme $x'(t) = f(t, x(t))$,

$$\begin{aligned} x''(t) = \tfrac{d}{dt} x'(t) &= \tfrac{d}{dt} f((t, x(t))) \\ &= \partial_t f + x'(t) \partial_x f \\ &= f_t + f f_x \end{aligned}$$

d'où

$$x(t_0 + h) = x(t_0) + hf + \frac{h^2}{2}(f_t + f f_x) + \mathcal{O}(h^3).$$

L'erreur de troncature est donc

$$\begin{aligned} \varepsilon_1 &= x(t_0 + h) - x_1 \\ &= x(t_0) + hf + \frac{h^2}{2}(f_t + f f_x) + \mathcal{O}(h^3) - x_0 - h b_1 f \\ &= h(1 - b_1) f_{|t=t_0} + \frac{h^2}{2}(f_t + f f_x)_{|t=t_0} + \mathcal{O}(h^3). \end{aligned}$$

La méthode est donc d'ordre 1 à condition que $b_1 = 1$. La seule méthode de Runge-Kutta d'ordre 1 est la méthode d'Euler explicite.

7.2.2 Méthodes RK à deux étapes

Elles prennent la forme

$$\begin{cases} k_1 = f(t_n, x_n), \\ k_2 = f(t_n + ah_n, x_n + ah_n k_1), \quad c_2 = a_{21} := a, \\ x_{n+1} = x_n + h_n(b_1 k_1 + b_2 k_2). \end{cases}$$

Pour étudier l'erreur de troncature, il nous faut utiliser la formule de Taylor pour les fonctions de plusieurs variables $f : \mathbb{R}^p \to \mathbb{R}$

$$f(x) = f(a) + (\nabla f(a))^t (x-a) + \frac{1}{2}(x-a)^t Hf(a)(x-a) + \mathcal{O}(\|x-a\|^3), \quad x, a \in \mathbb{R}^p,$$

avec la gradient et la hessienne

$$\nabla f(a) = \begin{pmatrix} \partial_{x_1} f(a) \\ \vdots \\ \partial_{x_p} f(a) \end{pmatrix}, \qquad Hf(a) = \begin{pmatrix} \frac{\partial^2 f}{\partial_{x_1}^2} & \frac{\partial^2 f}{\partial_{x_1}\partial_{x_2}} & \cdots & \frac{\partial^2 f}{\partial_{x_1}\partial_{x_p}} \\ \vdots & & & \vdots \\ \frac{\partial^2 f}{\partial_{x_p}\partial_{x_1}} & \frac{\partial^2 f}{\partial_{x_p}\partial_{x_2}} & \cdots & \frac{\partial^2 f}{\partial_{x_p}^2} \end{pmatrix}.$$

Ainsi

$$\begin{aligned} f(t+\alpha h, x+\beta h) &= f(t,x) + \begin{pmatrix} \partial_t f(t,x) \\ \partial_x f(t,x) \end{pmatrix} \cdot \begin{pmatrix} \alpha h \\ \beta h \end{pmatrix} + \mathcal{O}(h^2) \\ &= f(x,t) + h(\alpha \partial_t f + \beta \partial_x f) + \mathcal{O}(h^2). \end{aligned}$$

On a $k_1 = f(t_0, x_0)$ et donc

$$\begin{aligned} f(t_0 + ah, x_0 + ahk_1) &= f(t_0,x_0) + h(a\partial_t f + ak_1 \partial_x f) + \mathcal{O}(h^2) \\ &= f(t_0,x_0) + ha(f_t + f f_x) + \mathcal{O}(h^2). \end{aligned}$$

De même

$$\begin{aligned} x_1 &= x_0 + h(b_1 k_1 + b_2 k_2) \\ &= x_0 + hb_1 f_0 + hb_2(f_0 + ah(f_t + f f_x)_{|t=t_0} + \mathcal{O}(h^2)) \\ &= x_0 + h(b_1+b_2) f_0 + ab_2 h^2 (f_t + f f_x)_{|t=t_0} + \mathcal{O}(h^3)). \end{aligned}$$

D'où

$$\begin{aligned} x(t_0+h) - x_1 &= x(t_0) + hf_0 + \tfrac{h^2}{2}(f_t + f f_x)_{|t=t_0} + \mathcal{O}(h^3)) \\ &\quad -x_0 - h(b_1+b_2) f_0 - ab_2 h^2 (f_t + f f_x)_{|t=t_0} + \mathcal{O}(h^3)) \\ &= h(1-(b_1+b_2)) f_0 + h^2(\tfrac{1}{2} - ab_2)(f_t + f f_x)_{|t=t_0} + \mathcal{O}(h^3)). \end{aligned}$$

Pour avoir l'ordre 2, il faut

$$\begin{cases} b_1 + b_2 = 1 \\ ab_2 = 1/2 \Leftrightarrow c_2 b_2 = 1/2 \end{cases}$$

Si on pousse le calcul avec la hessienne, on a le terme d'ordre 3 qui s'écrit

$$\mathcal{O}(h^3) = h^3 \left[(\frac{1}{6} - b_2 a^2)(f_{tt} + 2f f_x + f^2 f_{xx} + f_x(f_t + f f_x)) + \frac{1}{6} f_x (f_t + f f_x) \right] + \mathcal{O}(h^4).$$

Il n'est donc pas possible de trovuer des valeurs de a, b_1 et b_2 pour annuler ce terme.

Le tableau de Butcher pour une méthode d'ordre 2 est

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

7.2.3 Méthodes RK à trois étapes

Elles prennent la forme

$$\begin{cases} k_1 = f(t_n, x_n) \\ k_2 = f(t_n + c_2 h, x_n + h a_{21} k_1) \\ k_3 = f(t_n + c_3 h, x_n + h(a_{31} k_1 + a_{32} k_2)) \\ x_{n+1} = x_n + h(b_1 k_1 + b_2 k_2 + b_3 k_3) \end{cases}$$

avec $c_2 = a_{21}$ et $c_3 = a_{31} + a_{32}$. Il reste donc 6 paramètres libres. On répète les mêmes manipulations algébriques mais qui sont plus fastidieuses. On trouve des relations algébriques pour obtenir les ordres souhaités

$$b_1 + b_2 + b_3 = 1 \qquad \text{condition d'ordre 1} \qquad \sum_i b_i = 1$$

$$b_2 c_2 + b_3 c_3 = 1/2 \qquad \text{condition d'ordre 2} \qquad \sum_i b_i c_i = 1/2$$

$$\left.\begin{array}{l} b_2 c_2^2 + b_3 c_3^2 = 1/3 \\ c_2 a_{32} b_3 = 1/6 \end{array}\right\} \quad \text{condition d'ordre 3} \quad \left\{\begin{array}{l} \sum_i b_i c_i^2 = 1/3 \\ \sum_{ij} b_i a_{ij} c_j = 1/6 \end{array}\right.$$

■ **Exemple 7.2** — méthode de Heun (voir précédemment)
— méthode de Kutta

0	0	0	0
1/2	1/2	0	0
1	−1	2	0
	1/6	4/3	1/6

■

7.2.4 Méthodes RK à quatre étapes

Pour obtenir l'ordre 4, il faut rajouter des conditions. Pour les ordres 1 à 3, ce sont les mêmes que pour les méthodes à 3 étapes. On rajoute les conditions suivantes pour l'ordre 4

$$\sum_i b_i c_i^3 = 1/4$$
$$\sum_{ij} b_i c_i a_{ij} c_j = 1/8$$
$$\sum_{ij} b_i a_{ij} c_j^2 = 1/12$$
$$\sum_{ijk} b_i a_{ij} a_{jk} c_k = 1/24$$

On montre $c_4 = 1$. Les deux méthodes les plus standards sont

RK41

0				
1/2	1/2			
1/2	0	1/2		
1	0	0	1	
	1/6	1/3	1/3	1/6

RK42

0				
1/4	1/4			
1/2	0	1/2		
1	1	−2	2	
	1/6	0	2/3	1/6

On peut bien entendu monter en ordre, mais à un coût exhorbitant. Par exemple, pour obtenir l'ordre 9, 12 à 17 étapes sont nécessaires, 486 relations algébriques sont à vérifier.

7.2.5 Méthodes implicites

L'avantage des méthodes implicites par rapport aux méthodes explicites est leur capacité à obtenir un ordre plus élevé avec moins d'étapes.

■ **Exemple 7.3** DIRK : Diagonaly Implicit Runge-Kutta d'ordre 3

1/3	1/3	0
1	1	0
	3/4	1/4

■

Concernant la convergence de ces méthodes, il suffit d'utiliser les résultats vus pour les méthodes à un pas avec un Φ lié à la méthode de Runge-Kutta étudiée.

7.3 Stabilité absolue

On peut appliquer la même critique que pour les méthodes multipas. On considère pour cela l'EDO $x' = \lambda x$, $\mathrm{Re}(\lambda) < 0$ et on regarde si $\lim_{n\to\infty} x_n = 0$ avec h fixé.

Prenons l'exemple du cas

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

et posons $\hat{h} = \lambda h$. Alors

$$hk_1 = hf(t_n, x_n) = \hat{h}x_n$$

et

$$\begin{aligned} hk_2 &= hf(t_n + ah, x_n + ahk_1) \\ &= \hat{h}(x_n + ahk_1) \\ &= \hat{h}(1 + a\hat{h})x_n. \end{aligned}$$

Ainsi,

$$\begin{aligned} x_{n+1} &= x_n + (1-\theta)hk_1 + \theta hk_2 \\ &= (1 + \hat{h}(1 + \theta a\hat{h}))x_n \\ &= (1 + \hat{h} + \hat{h}/2)x_n := R(\hat{h})x_n. \end{aligned}$$

La fonction $R(\hat{h})$ s'appelle fonction de stabilité. C'est l'équivalent du polynôme de stabilité dans le cas explicite mais cela devient une expression rationnelle de $\hat{h}$ quand les méthodes sont implicites.

Résoudre cette équations aux différences revient à étudier les racines de $q(r) = r - (1 + \hat{h} + \hat{h}^2/2)$. La seule racine est $r = R(\hat{h})$. La fonction de stabilité est la même quelque soit θ. On sait que

$$e^{\hat{h}} = 1 + \hat{h} + \frac{\hat{h}^2}{2} + \mathcal{O}(\hat{h}^3)$$

et donc

$$R(\hat{h}) = e^{\hat{h}} + \mathcal{O}(\hat{h}^{p+1}).$$

Pour la famille de schéma considérée ici, $p = 2$.

Avoir $x_n \to 0$ est équivalent à $|R(\hat{h})| < 1$ qui est la condition de stabilité absolue. L'intervalle de stabilité absolue est donnée par $-1 < 1 + \hat{h}^2/2 < 1$ de sorte que

$$\mathscr{R}_0(\hat{h}) =]-2, 0[.$$

Pour connaitre la région de stabilité $\mathscr{R}$, on cheche sa frontière. On pose $\hat{h} = p + iq$. On cherche p et q tels que $|1 + \hat{h} + \hat{h}^2/2| = 1$. On a

$$4|1 + \hat{h} + \hat{h}^2/2|^2 = 4(1 + p + \frac{1}{2}(p^2 - q^2))^2 + 4q^2(1+p)^2.$$

Ainsi

$$\begin{aligned} 4|1 + \hat{h} + \hat{h}^2/2|^2 &= (2 + 2p + p^2 - q^2)^2 + 4q^2(1+p)^2 \\ &= (1 + (1+p)^2 - q^2)^2 + 4q^2(1+p)^2 \\ &= 1 + (1+p)^4 + q^4 + 2(1+p)^2 - 2q^2 - 2q^2(1+p)^2 + 4q^2(1+p)^2 \\ &= 1 + (1+p)^4 + q^4 + 2(1+p)^2 + 2q^2(1+p)^2 + 2q^2 - 4q^2 \\ = (1 + (1+p)^2 + q^2)^2 - 4q^2 \end{aligned}$$

et donc $4|1 + \hat{h} + \hat{h}^2/2|^2 = 4$ équivaut à

$$(1 + (1+p)^2 + q^2)^2 = 4(1 + q^2)$$

soit encore

$$1 + (1+p)^2 + q^2 = 2\sqrt{1+q^2}.$$

En développant, on a

$$(1+p)^2 + (1+q)^2 - 2\sqrt{1+q^2} + 1 = 1$$

et finalement

$$(1+p)^2 + (\sqrt{1+q^2} - 1)^2 = 1.$$

La paramétrisation est donnée par $1+p = cos\phi$ et $\sqrt{1+q^2} - 1 = \sin\phi$ avec $0 \leq \phi \leq pi$, soit encore

$$p = \cos\phi - 1 \quad \text{et} \quad q = \pm\sqrt{(2+\sin\phi)\sin\phi}.$$

Les résultats exprimés ci-dessus pour le schéma de Runge-Kutta d'ordre 2 se généralise aux méthodes à s-étapes d'ordre s

1. Si on applique une méthode de Runge-Kutta explicite à $x' = \lambda x$, alors $x_{n+1} = R(\hat{h})x_n$ et R est un polynôme.
2. En utilisant un développement asymptotique de $x(t_{n+1})$ au temps $t = t_n$, et en utilisant $x' = \lambda x$, $x'' = \lambda^2 x$, et ainsi de suite, on a

$$x(t_{n+1}) = (1 + \hat{h} + \frac{\hat{h}^2}{2!} + \cdots + \frac{\hat{h}^s}{s!})x(t_n) + \mathcal{O}(\hat{h}^{s+1})$$

3. $R(\hat{h}) = 1 + \hat{h} + \frac{\hat{h}^2}{2!} + \cdots + \frac{\hat{h}^s}{s!}$ et donc $R(\hat{h}) = e^{\hat{h}} + \mathcal{O}(\hat{h}^{s+1})$.

Ce dernier résultat indique donc que

$$\lim_{\hat{h}\to-\infty} |R(\hat{h})| = \infty,$$

et donc aucune méthode de Runge-Kutta explicite n'est A ou A_0 stable. On a

s	$R(\hat{h})$	Intervalle de stabilité absolue $\mathscr{R}_0$
1	$1 + \hat{h}$	$]-2, 0[$
2	$1 + \hat{h} + \frac{\hat{h}^2}{2}$	$]-2, 0[$
3	$1 + \hat{h} + \frac{\hat{h}^2}{2} + \frac{\hat{h}^3}{6}$	$]-2.513, 0[$
4	$1 + \hat{h} + \frac{\hat{h}^2}{2} + \frac{\hat{h}^3}{6} + \frac{\hat{h}^4}{24}$	$]-2.785, 0[$

Les régions de stabilité correspondantes sont représentées ci-dessous.

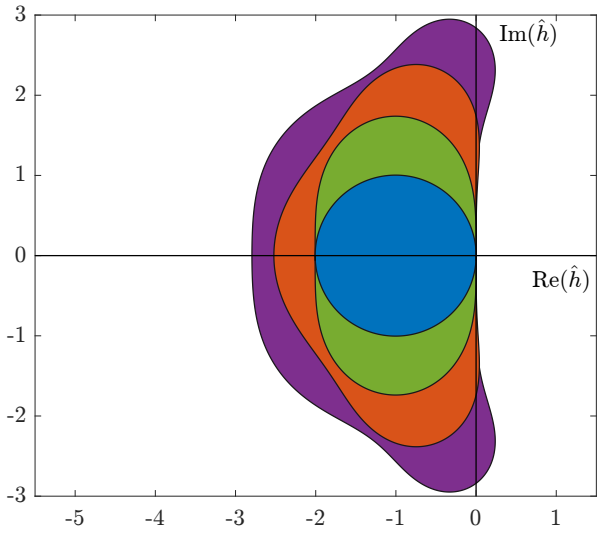


FIGURE 7.1 – Région de stabilité

7.4 Méthodes implicites

Nous avons vu dans l'exemple du schéma de Runge-Kutta basé sur la méthode de Gauss à 2 points est d'ordre 4. Ce résultat est général :

Théorème 7.4.1 L'ordre maximal d'une méthodes de Runge-Kutta implicite à s étages est $2s$.

Étudions la stabilité. Prenons l'exemple du schéma $\begin{array}{c|c} 1/2 & 1/2 \\ \hline & 1 \end{array}$. Alors, la méthode est

$$\begin{aligned} k_1 &= f(t_n + h/2, x_n + hk_1/2) \\ x_{n+1} &= x_n + hk_1. \end{aligned}$$

Si on considère $x' = \lambda x$, alors $k_1 = \lambda(x_n + hk_1/2)$ soit encore $k_1(1 - \lambda h/2) = \lambda x_n$ et donc

$$x_{n+1} = x_n + \frac{\lambda h}{1 - \lambda h/2} x_n = \left(\frac{1 + \hat{h}/2}{1 - \hat{h}/2} \right) x_n := R(\hat{h}) x_n.$$

Comme on l'avait annoncé, $R(\hat{h})$ est cette fois une fraction rationnelle. Demander $|R(\hat{h})| = 1$ revient donc à chercher $R(\hat{h}) = e^{i\theta}$ ce qui équivaut à

$$\hat{h} = 2 \frac{e^{i\theta/2} - e^{-i\theta/2}}{e^{i\theta/2} + e^{-i\theta/2}} = 2i \tan(\theta/2).$$

Ainsi, la frontière de la zone de stabilité est l'axe imaginaire pur. Si $\hat{h} = -1$, $R(-1) = 1/3 < 1$ et donc $|R(\hat{h})| < 1$ équivaut à $\mathrm{Re}(\hat{h}) < 0$. La méthode est donc A-stable.

Si on étudie le cas général

$$\left\{ \begin{array}{rl} k_i &= f\left(t_n + c_i h_n, x_n + h_n \displaystyle\sum_{j=1}^{s} a_{ij} k_j \right), \\ &= \lambda x_n + \lambda h \displaystyle\sum_{j=1}^{s} a_{ij} k_j, \end{array} \right.$$

ce qui conduit à

$$(k_i - \lambda h \sum_j a_{ij} k_j) = \lambda x_n$$

et

$$x_{n+1} = x_n + h_n \sum_{i=1}^{s} b_i k_i.$$

On peut réécrire ces égalités avec

$$(\mathbf{I} - (\lambda h)\mathbf{A})\,\mathbf{K} = \lambda x_n \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} := \lambda \mathbf{e}, \quad \mathbf{K} = \begin{pmatrix} k_1 \\ \vdots \\ k_s \end{pmatrix}$$

et si on définit $\mathbf{b} = \big(b_1, \cdots, b_s\big)\,\hat{}\,t$, on a

$$\begin{aligned} x_{n+1} &= x_n + h\mathbf{b}\hat{}\,t\mathbf{K} \\ &= (\underbrace{\mathbf{I} + \hat{h}\mathbf{b}\hat{}\,t(\mathbf{I} - \hat{h}\mathbf{A})^{-1}\mathbf{e}}_{R(\hat{h})}) x_n. \end{aligned}$$

On a vu précedemment que $R(\hat{h}) = e^{\hat{h}} + \mathcal{O}(\hat{h}^{p+1})$.

Définition 7.4.1 On appelle approximation de Padé (j, k), notée R_{jk}, de e^z l'approximation d'ordre maximal de la forme (fraction rationnelle)

$$R_{jk}(z) = \frac{P_k(z)}{Q_j(z)} = \frac{p_0 + p_1 z + \cdots + p_k z^k}{q_0 + q_1 z + \cdots + q_j z^j}$$

où P_k et Q_j n'ont aucun facteur commun et

$$Q_j(0) = q_0 = 1 \quad \text{et} \quad R_{jk}(z) = e^z + \mathcal{O}(z^{k+j+1}).$$

■ **Exemple 7.4** Cherchons l'approximation de Padé $(2,0)$ de e^z, soit $j = 2$ et $k = 0$. On souhaite

$$e^z = R_{jk}(z) + \mathcal{O}(z^{k+j+1})$$

ce qui donne

$$\sum_{i=0}^{k+j} \frac{z^i}{i!} = \frac{\sum_{i=0}^{k} p_i z^i}{\sum_{i=0}^{j} q_i z^i} + \mathcal{O}(z^{k+j+1}).$$

Pour l'ordre considéré, ceci équivaut à

$$1 + z + \frac{z^2}{2} = \frac{p_0}{1 + q_1 z + q_2 z^2} + \mathcal{O}(z^3)$$

ou encore

$$(1 + z + \frac{z^2}{2})(1 + q_1 z + q_2 z^2) = p_0 + \mathcal{O}(z^3).$$

On a donc

$$1 + (1 + q_1)z + (\frac{1}{2} + q_2 + q_1)z^2 + \mathcal{O}(z^3) = p_0 + \mathcal{O}(z^3).$$

On obtient ainsi les coefficients $p_0 = 1$, $q_1 = -1$ et $q_2 = 1/2$ et

$$R_{20}(z) = \frac{1}{1 - z + z^2/2} \quad \text{et} \quad e^z = R_{20}(z) + \mathcal{O}(z^3).$$

■

Les premières approximations de Padé sont données dans le tableau suivant

	$k = 0$	1	2
$j = 0$	1	$1 + z$	$1 + z + z^2/2$
$j = 1$	$\dfrac{1}{1 - z}$	$\dfrac{1 + z/2}{1 - z/2}$	$\dfrac{1 + 2z/3 + z^2/6}{1 - z/3}$
$j = 2$	$\dfrac{1}{1 - z + z^2/2}$	$\dfrac{1 + z/3}{1 - 2z/3 + z^2/6}$	$\dfrac{1 + z/2 + z^2/12}{1 - z/3 + z^2/12}$

TABLE 7.1 – Approximants de Padé

Donc, la méthode d'Euler explicite est encodé par l'approximation de Padé R_{01}, Euler implicite par R_{10} et la méthode du point milieu par R_{11}.

Théorème 7.4.2 Il y a une et une seule méthode de Runge-Kutta implicite à s étages d'ordre $2s$. Elle correspond à l'approximation de Padé (s, s).

Théorème 7.4.3 Les méthodes qui correspondent à la diagonale ou aux deux premières sous-diagonales du tableau de Padé 7.1 pour l'approximation de e^z sont A-stable.

7.5 Exercices

Exercice 7.1 On considère la méthode RK définie par le tableau

$$\begin{array}{c|c} \mathbf{c} & \mathbf{A} \\ \hline & \mathbf{b}^t \end{array}$$

où $\mathbf{b} = (b_1, \ldots, b_s)^t$, $\mathbf{c} = (c_1, \ldots, c_s)^t$ et $\mathbf{A} = (a_{ij})$ est une matrice $s \times s$. On a aussi le vecteur des étapes $\mathbf{k} = (k_1, \ldots, k_s)^t$ et on note $\mathbf{e} = (1, \ldots, 1)^t$.

En appliquant ce schéma RK à l'équation test $x'(t) = \lambda x(t)$, montrer que

$$\mathbf{k} = \lambda(I - \widehat{h}\mathbf{A})^{-1}\mathbf{e}\, x_n$$

et que $x_{n+1} = R(\widehat{h})x_n$, avec $R(\widehat{h})$ la fonction de stabilité

$$R(\widehat{h}) = 1 + \widehat{h}\mathbf{b}^t(I - \widehat{h}\mathbf{A})^{-1}\mathbf{e}.$$

■

Exercice 7.2 Montrer que toute méthode RK d'ordre 2 à deux étapes appliquées au système $\mathbf{u}' = A\mathbf{u}$, avec

$$A = \begin{pmatrix} 5 & 2 \\ -2 & 5 \end{pmatrix},$$

est absolument stable si $h < 0.392$. Quel est le résultat si $A = \begin{pmatrix} 50 & 20 \\ -20 & 50 \end{pmatrix}$?

■

Exercice 7.3 On suppose que la matrice $A \in \mathbb{R}^{s\times s}$ soit diagonalisable par la matrice V, i.e.

$$V^{-1}AV = \Lambda$$

avec Λ une matrice diagonale. Montrer que toute puissance positive de A est diagonalisable par V et on a

$$V^{-1}A^kV = \Lambda^k.$$

En déduire que $V^{-1}R(hA)V = R(h\Lambda)$, où R est un polynôme de degré s.

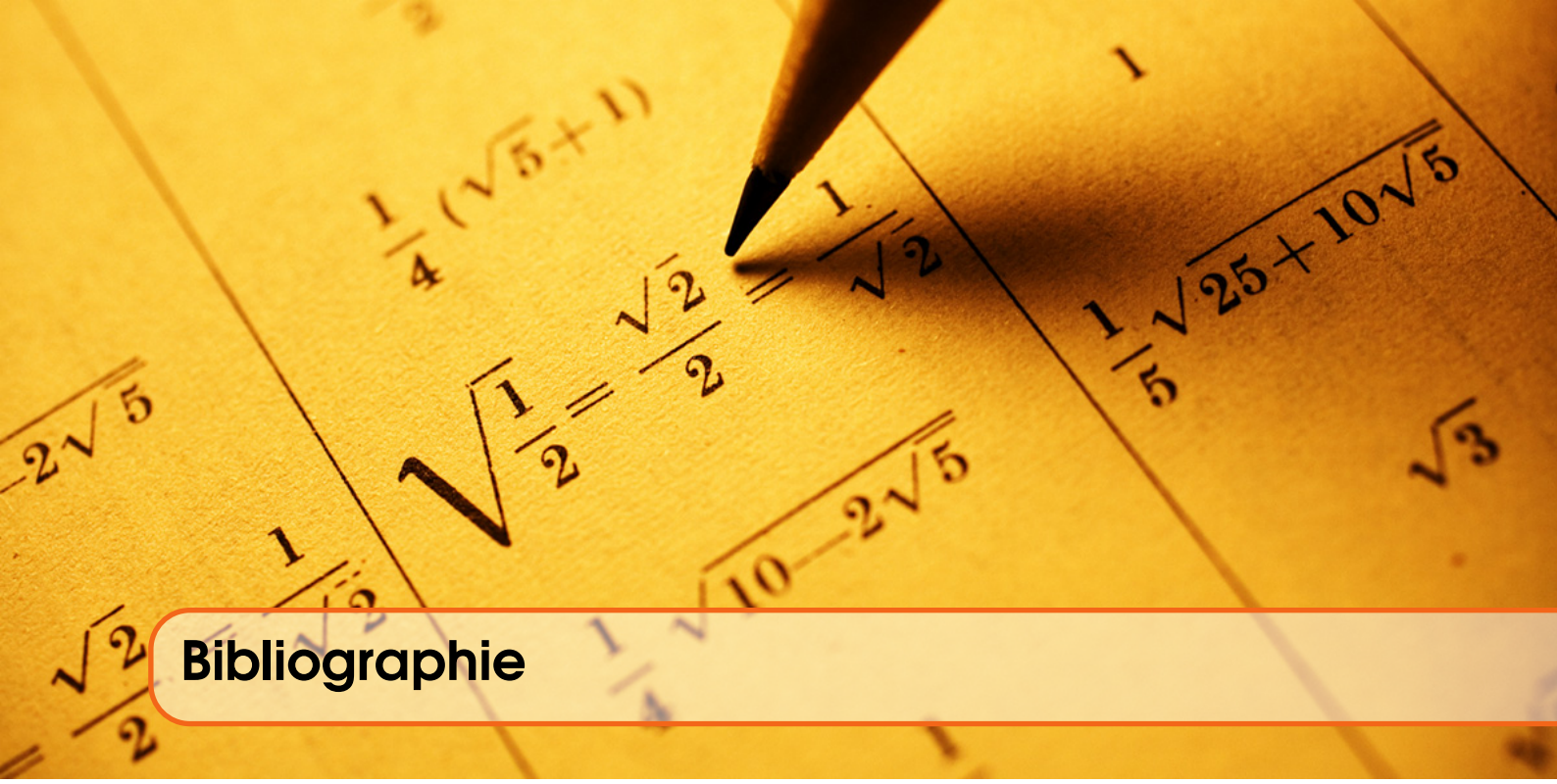
■

Exercice 7.4 On considère la méthode RK semi-implicite donnée par le tableau

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{5}{24} & \frac{1}{3} & -\frac{1}{24} \\ 1 & 0 & 1 & 0 \\ \hline & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \end{array}$$

Appliquer cette méthode à l'équation test $x'(t) = \lambda x(t)$ et déterminer le rapport x_{n+1}/x_n.

En déduire que la méthode ne peut pas être A-stable. ■



Bibliographie

Livres

[But08] J.C. Butcher. *Numerical Methods for Ordinary Differential Equations*. 2e édition. John Wiley & Sons, 2008.

[GH10] D.F. Griffiths et D.J. Higham. *Numerical Methods for Ordinary Differential Equations, Initial Value Problems*. Springer Undergraduate Mathematics Series. Springer, 2010.

[HLW05] E. Hairer, C. Lubich et G. Wanner. *Geometric Numerical Integration*. Springer Series in Computational Mathematics. Springer, 2005.

[HNW08] E. Hairer, S.P. Norsett et G. Wanner. *Solving Ordinary Differential Equation I, Nonstiff Problems*. 3e édition. Springer Series in Computational Mathematics. Springer, 2008.

[QSS07] A. Quarteroni, R. Sacco et F. Saleir. *Méthodes Numériques, Algorithms, analyse et applications*. Springer, 2007.

Index

O

P

R

S

Z