Analyse numérique

INSA de Rouen
GM3
2019-2020

N. Forcadel, A. Tonnoir

5 janvier 2021

Chapitre 1

Interpolation polynomiale

I Introduction

Dans ce chapitre, nous nous intéressons au problème d'interpolation suivant : étant donné $n+1$ points de $\mathbb{R}^2$ de coordonnées $(x_i, y_i)_{i=0,\cdots,n}$, trouver $p \in V$ vérifiant :

$$p(x_i) = y_i, \quad \forall i = \{0, \cdots, n\}$$

où V est un espace vectoriel de dimension $n+1$. Plus précisément, si les $(\varphi_i)_{i=0,\cdots,n}$ forment une base de V et qu'on décompose p ainsi :

$$p(x) = \sum_{j=0}^{n} \alpha_j \varphi_j(x)$$

l'objectif est de déterminer les coefficients α_i solutions du système linéaire :

$$\sum_{j=0} \alpha_i \varphi_j(x_i) = y_i, \quad \forall i = \{0, \cdots, n\}$$

Notons qu'il s'agit d'un système de $n+1$ équations à $n+1$ inconnues, et que ce dernier peut se réécrire sous forme matricielle comme suit :

$$\underbrace{\begin{bmatrix} \varphi_0(x_0) & \varphi_1(x_0) & \cdots & \varphi_n(x_0) \\ \varphi_0(x_1) & \varphi_1(x_1) & \cdots & \varphi_n(x_1) \\ \vdots & \vdots & & \vdots \\ \varphi_0(x_n) & \varphi_1(x_n) & \cdots & \varphi_n(x_n) \end{bmatrix}}_{:=G} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix} \tag{1.1}$$

où la matrice G est appelée la matrice de Gram. Sous cette forme, il est facile d'obtenir la condition nécessaire et suffisante d'existence et unicité d'une solution au problème d'interpolation :

Théorème 1.1. *Le problème d'interpolation admet une unique solution si et seulement si la matrice de Gram est inversible.*

Remarque 1.2. *Soulignons que ce n'est pas parce qu'on a V de dimension $n+1$ et $n+1$ points que le problème d'interpolation admettra une unique solution. En effet, si on considère $V = vec\{1, x^2\}$, de dimension 2, et les points $(-1, 1)$ et $(1, 1)$, alors il n'y a pas unicité de la solution (1 et x^2 sont solutions). Dans cet exemple, on notera que la matrice de Gram n'est pas inversible et on a :*

$$G = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \Rightarrow \det(G) = 0$$

De manière plus générale, on dira que l'ensemble $\Sigma = \{x_0, \cdots, x_n\}$ est V-unisolvant si la matrice de Gram associée est inversible.

Les problèmes d'interpolation interviennent dans plusieurs contextes. Typiquement, si on souhaite reconstruire des données en des points non connus (par exemple pour retrouver la topographie d'un fond marin à l'aide de données ponctuelles, voir figure 1.1), on pourra utiliser des techniques d'interpolation. Un autre exemple d'application se retrouve dans les logiciels de Dessin Assisté par Ordinateur (DAO) où on souhaitera définir des courbes à l'aide uniquement de certain point de contrôle.

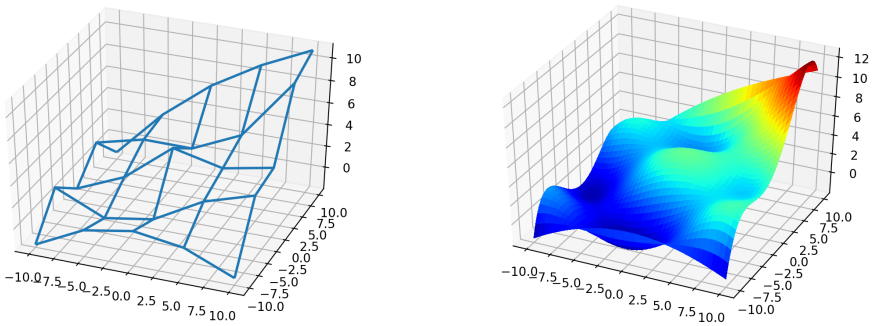

FIGURE 1.1 – Illustration de la reconstruction d'une topographie (à droite) à l'aide de données de topographie ponctuelles sur une grille cartésienne (à gauche)

Dans ce chapitre, nous nous intéresserons à l'interpolation polynomiale, c'est à dire que l'espace vectoriel que nous considérerons est $V = \mathcal{P}_n$.

II Interpolation polynomiale et base de Lagrange

Base canonique

Dans la base canonique, on a $\varphi_i(x) = x^i$ et la matrice de Gram est alors donné par :

$$G = \begin{bmatrix} 1 & x_0 & \cdots & x_0^n \\ 1 & x_1 & \cdots & x_1^n \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \cdots & x_n^n \end{bmatrix} \tag{1.2}$$

Cette matrice est une matrice dite de Vandermonde et on peut montrer qu'elle est inversible si et seulement si les points sont distincts, c'est à dire $x_i \neq x_j$ pour tout $i \neq j$. On déduit aisément ce résultat du théorème suivant :

Théorème 1.3. *Le déterminant de la matrice de Vandermonde est donné par :*

$$\det(G) = \Pi_{0 \leq i < j \leq n}(x_j - x_i)$$

La base canonique présente néanmoins un défaut important. En effet, pour déduire le polynôme d'interpolation dans cette base, il faut résoudre le système linéaire (1.1) qui est plein et mal conditionné. Nous allons donc chercher à construire des bases plus appropriées pour résoudre le problème d'interpolation en rendant la matrice G :

- diagonale avec la base de Lagrange,
- triangulaire inférieure avec la base de Newton (étudié dans la section suivante).

Base Lagrange

La base de Lagrange est construite de sorte d'assurer $\varphi_i(x_j) = \delta_{ij}$. On déduit par conséquent que, pour i fixé, les points x_j où $j \neq i$ sont racines de $\varphi_i(x)$ d'où :

$$\varphi_i(x) = Q_i(x) \underbrace{\prod_{j \neq i}(x - x_j)}_{\text{degré } n}$$

Le polynôme φ_i étant un élément de $\mathcal{P}_n$ est de degré n au plus, et par conséquent $Q_i(x) = C_i^{ste}$ doit être constant. Pour déterminer cette constante, on utilise le fait que $\varphi_i(x_i) = 1$, donc :

$$\varphi_i(x_i) = 1 \;\Leftrightarrow\; C_i^{ste} \prod_{j \neq i}(x_i - x_j) = 1 \;\Leftrightarrow\; C_i^{ste} = \frac{1}{\prod_{j \neq i}(x_i - x_j)}$$

Ainsi, on obtient les éléments de la base de Lagrange, qu'on notera L_i^n :

$$L_i^n(x) = \frac{\prod_{j \neq i}(x - x_j)}{\prod_{j \neq i}(x_i - x_j)}$$

Remarque 1.4. *Les $\{L_i^n\}_{i=0,\cdots,n}$ forment une base de $\mathcal{P}_n$. En effet, ayant $n+1$ éléments, il suffit de montrer qu'ils forment une famille libre :*

$$\sum_{j=0}^{n} \alpha_j L_j^n(x) = 0 \quad \Leftrightarrow \quad \alpha_i = 0 \quad \forall i \in \{0, \cdots, n\}$$

On obtient ce résultat facilement en évaluant le terme $\sum_{i=0}^{n} \alpha_i L_i^n(x)$ en $x = x_i$ et en utilisant le fait que, par construction $L_j^n(x_i) = \delta_{ij}$.

Dans la base de Lagrange, dont on remarquera qu'elle est bien définie ssi les points d'inteprolation x_i sont tous distincts, l'expression du polynôme d'interpolation est très simple car pour tout $i \in \{0, \cdots, n\}$:

$$p(x_i) = y_i \quad \Leftrightarrow \quad \sum_{j=0}^{n} \alpha_j L_j^n(x_i) = y_i \quad \Leftrightarrow \quad \alpha_i = y_i$$

d'où

$$p(x) = \sum_{j=0}^{n} \alpha_j L_j^n(x)$$

III Schéma de Neville-Aitken

L'expression du polynôme d'interpolation dans la base de Lagrange est très simple à obtenir comme nous venons de le voir. Cependant, elle ne permet pas d'évaluer facilement le polynôme p en un point $a \in [min(x_0, \ldots, x_n), \max(x_0, \ldots, x_n)]$ (si $a \neq x_i$). Pour cela, nous allons utiliser un schéma particulier appelé le schéma de Neville-Aitken.

Notons par T_k^i le polynôme de $\mathcal{P}_k$ vérifiant $T_k^i(x_j) = y_j$ pour $j = \{i, \cdots, i+k\}$. On a alors la propriété suivante :

$$T_{k+1}^i(x) = \frac{(x_{i+k+1} - x)T_k^i(x) - (x_i - x)T_k^{i+1}(x)}{x_{i+k+1} - x_i} \tag{1.3}$$

Pour montrer ce résultat, il suffit de remarquer que pour tout $j \in \{i+1, \cdots, i+k\}$ on a :

$$T_{k+1}^i(x_j) = \frac{(x_{i+k+1} - x_j)y_j - (x_i - x_j)y_j}{x_{i+k+1} - x_i} = y_j$$

car $T_k^i(x_j) = T_k^{i+1}(x_j) = y_j$ par définition de T_k^i. Reste à vérifier que la propriété est vrai pour en x_i et x_{i+k+1}. En $x = x_i$, on a :

$$T_{k+1}^i(x_i) = \frac{(x_{i+k+1} - x_i)y_i - (x_i - x_i)T_k^{i+1}(x_i)}{x_{i+k+1} - x_i} = y_i$$

et de même en $x = x_{i+k+1}$:

$$T_{k+1}^i(x_i) = \frac{(x_{i+k+1} - x_{i+k+1})T_k^i(x_{i+k+1}) - (x_i - x_{i+k+1})T_k^{i+1}(x_{i+k+1})}{x_{i+k+1} - x_i} = y_{i+k+1}$$

ce qui prouve notre résultat. On en déduit alors le schéma de calcul suivant :

Schéma de Neville-Aitken

$$\begin{array}{c|cccccccc}
x_0 & y_0 = & T_0^0(a) & \to & T_1^0(a) & \cdots & T_{n-1}^0(a) & \to & T_n^0(a) \\
 & & & \nearrow & & \nearrow & & \nearrow & \\
x_1 & y_1 = & T_0^1(a) & \to & T_1^1(a) & \dots & T_{n-1}^1(a) & & \\
\vdots & \vdots & \vdots & \iddots & \vdots & \iddots & & & \\
x_{n-1} & y_{n-1} = & T_0^{n_1}(a) & \to & T_1^{n-1}(a) & & & & \\
 & & \vdots & \nearrow & & & & & \\
x_n & y_n = & T_0^n(a) & & & & & &
\end{array}$$

On lira alors dans ce tableau la valeur du polynôme d'interpolation p en a au bout de la première ligne puisque $T_n^0(a) = p(a)$. Soulignons que dans ce schéma, l'ordre des points x_i n'a pas d'importance. De plus, ajouter un point d'interpolation ne nécessite pas de recalculer tout le tableau pour obtenir $p(a)$. En revanche, si on change le point d'évaluation a, tout le tableau doit être recalculer.

Complexité
Le calcul de $T_n^0(a)$ nécessite le calcul de $\frac{n(n+1)}{2}$ terme $T_k^i(a)$. Pour chacun de ces termes, nous avons d'après la formule (1.3) 4 additions, 2 multiplications et 1 divisions, soit 7 opérations. On a alors un cout total de $\frac{7(n+1)n}{2}$ opérations.

Notons qu'on peut optimiser ce calcul en pré-calculant les $x_i - a$ pour chaque i, et économiser ainsi 2 additions à chaque étape. Ainsi le nombre total d'opération devient $\frac{5(n+1)n}{2}$.

IV Base de Newton : méthodes des différences divisées

La base de Newton est définie de la manière suivante :

$$\begin{aligned}
&N_0(x) = 1 \\
&N_{i+1}(x) = (x - x_i)N_i(x), \quad i = 0, \dots, n-1
\end{aligned}$$

En remarquant que N_i est de degré i, on en déduit que $\{N_i\}_{i=0,\dots,n}$ forme une base de $\mathcal{P}_n$. Dans cette base, le polynôme d'interpolation s'écrit alors

$$p(x) = \sum_{i=0}^{n} \beta_i N_i(x) \quad \text{avec } p(x_j) = y_j = \sum_{i=0}^{n} \beta_i N_i(x_j),\ j = 0, \dots n.$$

En utilisant que $N_j(x_i) = 0$ si $i < j$, on est alors ramené à résoudre le système suivant :

$$\begin{bmatrix} N_0(x_0) & \cdots & N_n(x_0) \\ \vdots & & \vdots \\ N_0(x_n) & \cdots & N_n(x_n) \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_n \end{bmatrix} = \begin{bmatrix} N_0(x_0) & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ N_0(x_{n-1}) & \cdots & N_n(x_{n-1}) & 0 \\ N_0(x_n) & \cdots & \cdots & N_n(x_n) \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_n \end{bmatrix} = \begin{bmatrix} y_0 \\ \vdots \\ y_n \end{bmatrix} \tag{1.4}$$

où

$$G := \begin{bmatrix} N_0(x_0) & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ N_0(x_{n-1}) & \cdots & N_n(x_{n-1}) & 0 \\ N_0(x_n) & \cdots & \cdots & N_n(x_n) \end{bmatrix}$$

est la matrice de Gram. Le coût de résolution de ce système par une méthode de descente serait alors de $8n^2$. On va voir dans la suite que l'on peut faire mieux. Pour cela, on introduit les différences divisées :

- d'ordre 0 : $f[x_i] = y_i,\ i = 0, \dots, n$
- d'ordre 1 : $f[x_i, x_j] = \frac{f[x_j] - f[x_i]}{x_j - x_i}$
- d'ordre k : soient $\sigma_i \in \{1, \dots, n\}$ k indices tels que les x_{σ_i} soient tous distincts. Alors

$$f[x_{\sigma_1}, \dots, x_{\sigma_k}] = \frac{f[x_{\sigma_1}, \dots, x_{\sigma_{k-2}}, x_{\sigma_k}] - f[x_{\sigma_1}, \dots, x_{\sigma_{k-1}}]}{x_{\sigma_k} - x_{\sigma_{k-1}}}$$

On a alors la propriété suivante :

Théorème 1.5. *On a, pour tout $j = 0, \dots, n$*

$$\beta_j = f[x_0, \dots, x_j].$$

Démonstration. Nous alors démontrer par récurrence que

$$\beta_k = f[x_0, \dots, x_k] \quad \text{pour } k = 1, \dots n$$

et que

$$\frac{y_j - \sum_{i=0}^{k-1} \beta_i N_i(x_j)}{N_k(x_j)} = f[x_0, \dots x_{k-1}, x_j] \quad \text{pour } j \geq k, k = 1, \dots n$$

Tout d'abord, la première ligne de (1.4) nous donne que $\beta_0 = y_0 = f[x_0]$. Montrons que les 2 propriétés sont vraies pour $k = 1$. Pour $j \geq 1$, on a

$$\frac{y_j - \beta_0 N_0(x_j)}{N_1(x_j)} = \frac{f[x_j] - f[x_0]}{x_j - x_0} = f[x_0, x_j].$$

En prenant $j = 1$ et en utilisant la 2ème ligne de (1.4), on obtient

$$\beta_1 = \frac{y_1 - \beta_0 N_0(x_1)}{N_1(x_1)} = f[x_0, x_1].$$

La propriété est donc vraie au rang 1. On suppose maintenant qu'elle est vraie au rang k et montrons qu'elle reste vraie au rang $k+1$. Tout d'abord, pour $j \geq k+1$, on a

$$\begin{aligned}
\frac{y_j - \sum_{i=0}^{k} \beta_i N_i(x_j)}{N_{k+1}(x_j)} &= \frac{y_j - \sum_{i=0}^{k-1} \beta_i N_i(x_j) - \beta_k N_k(x_j)}{N_{k+1}(x_j)} \\
&= \frac{f[x_0, \ldots, x_{k-1}, x_j] N_k(x_j) - f[x_0, \ldots, x_k] N_k(x_j)}{N_{k+1}(x_j)} \\
&= \frac{f[x_0, \ldots, x_{k-1}, x_j] - f[x_0, \ldots, x_k]}{x_j - x_k} \\
&= f[x_0, \ldots, x_k, x_j]
\end{aligned}$$

où pour la 2ème ligne nous avons utilisé les hypothèses de récurrence. En utilisant la $(k+1)$-ème ligne de (1.4) et en prenant $j = k+1$ dans l'équation précédente, on obtient

$$\beta_{k+1} = \frac{y_{k+1} - \sum_{i=0}^{k} \beta_i N_i(x_{k+1})}{N_{k+1}(x_{k+1})} = f[x_0, \ldots, x_k, x_{k+1}]$$

ce qui montre que la propriété est héréditaire et termine la preuve. □

La définition proposée pour les différences divisées fait que ce sont des fonctions symétriques des deux derniers arguments. En fait, elles sont symétriques par rapport à tous leurs arguments. Ceci est une conséquence du théorème suivant :

Théorème 1.6. *Soient $\sigma_i \in \{1, \ldots, n\}$ k indices tels que les x_{σ_i} soient tous distincts. On pose*

$$M(x) = \prod_{i=0}^{k} (x - x_{\sigma_i}) \quad \text{et} \quad M_{\sigma_j}(x) = \frac{M(x)}{x - x_{\sigma_j}}.$$

Alors

$$f[x_{\sigma_0}, \ldots, x_{\sigma_k}] = \sum_{j=0}^{k} \frac{y_{\sigma_j}}{M_{\sigma_j}(x_{\sigma_j})}. \tag{1.5}$$

La relation (1.5) montre donc que la différence divisée d'ordre k, est indépendante de l'ordre des points x_i. C'est une fonction symétrique de ses arguments, c'est à dire que si $(\rho_0, \ldots, \rho_k)$ est une permutation quelconque du $(k+1)$-uplet $(\sigma_0, \ldots, \sigma_k)$, alors

$$f[x_{\rho_0}, \ldots, x_{\rho_k}] = \sum_{j=0}^{k} \frac{y_{\rho_j}}{M_{\rho_j}(x_{\rho_j})} = \sum_{j=0}^{k} \frac{y_{\sigma_j}}{M_{\sigma_j}(x_{\sigma_j})} = f[x_{\sigma_0}, \ldots, x_{\sigma_k}].$$

Une conséquence directe de cette propriété est la formule suivante pour les différences divisées :

Corollaire 1.7. *On a*

$$f[x_q, \ldots, x_{k+q}] = \frac{f[x_{q+1}, \ldots, x_{k+q}] - f[x_q, \ldots x_{k+q-1}]}{x_{k+q} - x_q}.$$

Démonstration. On a

$$\begin{aligned} f[x_q, \ldots, x_{k+q}] &= f[x_{q+1}, \ldots, x_{k+q-1}, x_q, x_{k+q}] \\ &= \frac{f[x_{q+1}, \ldots, x_{k+q-1}, x_{k+q}] - f[x_{q+1}, \ldots, x_{k+q-1}, x_q]}{x_{k+q} - x_q} \\ &= \frac{f[x_{q+1}, \ldots, x_{k+q}] - f[x_q, \ldots x_{k+q-1}]}{x_{k+q} - x_q}. \end{aligned}$$

□

Calcul des différences divisées

x_i		β_0		β_1		β_{n-1}		β_n
x_0	$y_0 =$	$f[x_0]$	$\to$	$f[x_0, x_1]$	$\cdots$	$f[x_0, \cdots, x_{n-1}]$	$\to$	$f[x_0, \cdots, x_n]$
			$\nearrow$		$\nearrow$		$\nearrow$	
x_1	$y_1 =$	$f[x_1]$	$\to$	$f[x_1, x_2]$	$\ldots$	$f[x_1, \ldots, x_n]$		
		$\vdots$	$\iddots$	$\vdots$	$\iddots$			
x_{n-1}	$y_{n-1} =$	$f[x_{n-1}]$	$\to$	$f[x_{n-1}, x_n]$				
		$\vdots$	$\nearrow$					
x_n	$y_n =$	$f[x_n]$						

Sur ce schéma, on notera qu'à la ligne l, on peut directement lire les coefficients de la décomposition dans la base de Newton du polynôme d'interpolation associé aux points $\{x_l, \cdots, x_n\}$.

Schéma de Hörner
Pour évaluer le polynôme d'interpolation p en un point a, on utilise l'algorithme de Hörner. On a

$$p(a) = \sum_{i=0}^{n} \beta_i N_i(a) \quad \text{et} \quad N_{i+1}(a) = (a - x_i)N_i(x).$$

On peut donc écrire

$$p(a) = \beta_0 + (a - x_0)\left(\beta_1 + (a - x_1)\left(\cdots\left(\beta_{n-2} + (a - x_{n-2})\left(\beta_{n-1} + (a - x_{n-1})\,\beta_n\right)\right)\right)\right).$$

On pose donc

$$\begin{aligned} b_n &= \beta_n \\ b_{n-1} &= \beta_{n-1} + (a - x_{n-1})b_n \\ &\vdots \\ b_j &= \beta_j + (a - x_j)b_{j+1} \\ &\vdots \\ b_0 &= \beta_0 + (a - x_0)b_1. \end{aligned}$$

La valeur de $p(a)$ est alors donnée par b_0.

Complexité
Calcul des β_n :
Il y a $\frac{n(n+1)}{2}$ différences divisées à calculer. Pour chaque différence divisée, il y a 2 soustractions et 1 division. Le nombre total d'opérations est donc $\frac{3}{2}n(n+1)$.

Calcul de $p(a)$:
Il y a n b_j à calculer. Pour chaque b_j, il y a 2 additions et 1 multiplication, soit $3n$ opérations au total.

Remarque 1.8.

1. *L'ordre des points est toujours indifférents. On peut donc ajouter de nouveaux points avec une complexité de calcul faible.*
2. *Si on change la valeur de a pour l'évaluation de $p(a)$:*
   - *Schéma de Neville-Aitken : on recommence tous les calculs ($\frac{5}{2}n^2$ opérations).*
   - *Différences divisées : les β_j restent inchangés, on réutilise l'algorithme de Hörner ($3n$ opérations)*

V Interpolation polynomiale d'une fonction

Il est plus pratique de manipuler (dériver, additionner, multiplier, intégrer,...) et de stocker un polynôme plutôt qu'une fonction . Il est donc naturel de remplacer une fonction par le polynôme d'interpolation associé aux valeurs prises par la fonction en des noeuds choisis.

Définition 1.9. *Soient n un entier positif, $(x_i)_{i=0,\dots,n}$, $n+1$ points distincts et f une fonction réelle donnée et définie aux points x_i.*

On appelle polynôme d'interpolation (ou interpolant) de Lagrange de degré n de la fonction f, notée $P_n f$, le polynôme d'interpolation de degré n associé aux points $(x_i, f(x_i))_{i=0,\dots,n}$.

Le polynôme d'interpolation $P_n f$ peut être vu comme le polynôme de degré n minimisant l'erreur d'approximation $\|f - p_n\|$, $p_n \in \mathcal{P}_n$, mesurée avec la semi-norme

$$\|f\| = \sum_{i=0}^{n} |f(x_i)|.$$

On peut alors étudier l'erreur d'interpolation $f - P_n f$.

Théorème 1.10. *Soient f une fonction de classe C^{n+1} et $(x_i)_{i=0,\dots,n}$, $n+1$ noeuds contenus dans $[a,b]$. Alors, pour tout réel $x \in [a,b]$, $\exists \xi \in I_x := [min(x_0,\dots,x_n,x), \max(x_0,\dots,x_n,x)]$ tel que l'erreur d'interpolation au point x soit donnée par*

$$f(x) - P_n f(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{j=0}^{n} (x - x_j).$$

Démonstration. Soit $x \in [a,b]$. Si $x = x_i$, $i \in \{0,\dots,n\}$, le résultat est trivial. On suppose que $x \neq x_i$ pour tout i et on définit

$$\varphi(t) = f(t) - P_n f(t) - \frac{f(x) - P_n f(x)}{\prod_{j=0}^{n} (x - x_j)} \prod_{j=0}^{n} (t - x_j).$$

Par hypothèse, φ est de classe C^{n+1} sur I_x et s'annule en $n+2$ points (car $\varphi(x) = \varphi(x_0) = \cdots = \varphi(x_n) = 0$). D'après le Théorème de Rolle, φ' possède au moins $n+1$ zéros distincts dans l'intervalle I_x.

Par récurrence, on peut montrer que $\varphi^{(j)}$, $0 \leq j \leq n+1$ possède au moins $n+2-j$ zéros distincts sur I_x. Ainsi, il existe $\xi \in I_x$ tel que $\varphi^{(n+1)}(\xi) = 0$, i.e.

$$f^{(n+1)}(\xi) - \frac{f(x) - P_n f(x)}{\prod_{j=0}^{n} (x - x_j)} (n+1)! = 0$$

car $(P_n f)^{(n+1)} = 0$ et $(\prod_{j=0}^{n}(t - x_j))^{(n+1)} = (n+1)!$. Ceci peut se réécrire

$$f(x) - P_n f(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod (x - x_j).$$

□

Corollaire 1.11. *On a la majoration suivante :*

$$\begin{aligned} \|f - P_n f\|_{L^\infty} &\leq \frac{1}{(n+1)!} \left\|f^{(n+1)}\right\|_{L^\infty} \left\|\prod_{j=0}^{n}(\cdot - x_j)\right\|_{L^\infty} \\ &\leq \frac{1}{(n+1)!} \left\|f^{(n+1)}\right\|_{L^\infty} |b-a|^{n+1} \end{aligned}$$

Convergence des polynômes d'interpolation
On s'intéresse à la convergence uniforme du polynôme d'interpolation lorsque le nombre de points d'interpolation tend vers $+\infty$. On suppose pour l'instant que les points d'interpolation sont équi-répartis, i.e.

$$x_i = a + i\frac{b-a}{n}, \quad i = 0, \ldots, n.$$

D'après le Théorème 1.10, la convergence de la suite $(P_n f)_{n\in\mathbb{N}^*}$ est liée au comportement de $\|f^{(n+1)}\|_{L^\infty}$ lorsque n augmente. En effet, si

$$\lim_{n\to+\infty} \frac{1}{(n+1)!} \left\|f^{(n+1)}\right\|_{L^\infty} \left\|\prod_{j=0}^{n}(\cdot - x_j)\right\|_{L^\infty} = 0,$$

alors, on a immédiatement

$$\lim_{n\to+\infty} \|f - P_n f\|_{L^\infty} = 0,$$

i.e. $P_n f$ converge uniformément sur $[a, b]$ vers f.

Malheureusement, pour certaines fonctions, le produit $\left\|f^{(n+1)}\right\|_{L^\infty} \left\|\prod_{j=0}^{n}(\cdot - x_j)\right\|_{L^\infty}$ tend vers l'infini plus rapidement que $(n+1)!$ quand $n \to +\infty$. Un célèbre exemple est dû à Runge pour la fonction

$$f(x) = \frac{1}{1+x^2} \quad \text{sur } [-5, 5]$$

et montre que $P_n f$ ne converge pas vers f uniformément. Ce phénomène est explicité dans la figure 1.2.

Comme nous allons le voir dans la section suivante, cela ne vient pas de la régularité de la fonction f mais du choix des noeuds.

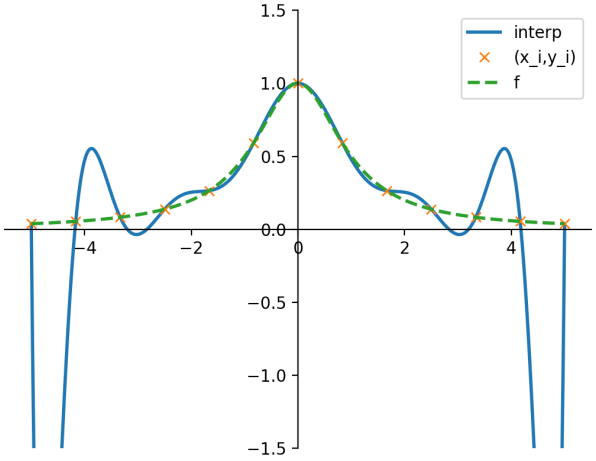

FIGURE 1.2 – Phénomène de Runge avec les points équidistants

VI Meilleure interpolation

But : Trouver les meilleurs x_i pour minimiser l'erreur d'interpolation.

On note $I = [a, b]$ et on suppose que $x_i \in I$ pour tout i. On pose

$$v_n(x) = \prod_{j=0}^{n}(x - x_j), \quad \|v_n\|_{L^\infty} = \max_{x\in[a,b]} |v_n(x)|.$$

Le problème à résoudre est donc le suivant

$$\min_{x_j\in[a,b], j=0,\dots,n} \|v_n\|_{L^\infty} \tag{1.6}$$

En utilisant l'application $x \mapsto x\frac{b-a}{2} + \frac{b+a}{2}$, on peut se ramener à travailler sur l'intervalle $[-1, 1]$. On va voir que la solution du problème va correspondre aux racines du polynôme de Tchebychev.

Polynômes de Tchebychev

Définition 1.12. *On appelle polynôme de Tchebychev de première espèce de degré k, l'application T_k de $[-1, 1]$ dans lui-même définie par*

$$T_k(x) := \cos(k \arccos(x)).$$

Théorème 1.13. *Les T_k satisfont une relation de récurrence à trois termes :*

$$T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x), \quad k \geq 1, \tag{1.7}$$

avec $T_0(x) = 1$ et $T_1(x) = x$.

Démonstration. On a $T_0(x) = \cos(0) = 1$ et $T_1(x) = \cos(\arccos(x)) = x$. Pour $k \geq 1$, on a

$$\begin{aligned}
2xT_k(x) - T_{k-1}(x) =& 2x\cos(k\arccos(x)) - \cos((k-1)\arccos(x)) \\
=& 2\cos(\arccos(x))\cos(k\arccos(x)) \\
& - \cos(\arccos(x))\cos(k\arccos(x)) \\
& - \sin(\arccos(x))\sin(k\arccos(x)) \\
=& \cos(\arccos(x))\cos(k\arccos(x)) \\
& - \sin(\arccos(x))\sin(k\arccos(x)) \\
=& \cos((k+1)\arccos(x)) \\
=& T_{k+1}(x)
\end{aligned}$$

□

Remarque 1.14.

- *D'après la relation de la récurrence, T_k est un polynôme de degré k*
- *Pour $k \geq 1$, le monôme de plus haut degré de T_k est $2^{k-1}x^k$, i.e., $T_k(x) = 2^{k-1}x^k + \dots$.*

Racines de T_k

On regarde les points x_j qui annule T_k. On a

$$\begin{aligned}
T_k(x_j) = 0 &\iff \cos(k\arccos(x_j)) = 0 \\
&\iff k\arccos(x_j) = \frac{\pi}{2} + j\pi, \quad j \in \mathbb{Z} \\
&\iff \arccos(x_j) = \frac{\pi}{2k} + \frac{j}{k}\pi, \quad j \in \mathbb{Z} \\
&\iff \arccos(x_j) = \frac{2j+1}{2k}\pi, \quad j = 0, \dots k-1 \quad (\text{rappel : } \arccos \in [0, \pi]) \\
&\iff x_j = \cos\left(\frac{2j+1}{2k}\pi\right), \quad j = 0, \dots k-1
\end{aligned}$$

Les racines de T_k sont donc toutes réelles et distinctes.

Points d'annulation de la dérivée
On cherche les points $\hat{x}_j$ qui annulent la dérivée de T_k. On a, pour $x \in]-1,1[$,

$$T_k'(x) = \frac{d}{dx}(\cos(k\arccos(x))) = \frac{k}{\sqrt{1-x^2}}\sin(k\arccos(x)),$$

et donc

$$\begin{aligned} T_k'(\hat{x}_j) = 0 &\iff \sin(k\arccos(\hat{x}_j)) = 0 \\ &\iff k\arccos(\hat{x}_j) = j\pi \\ &\iff \hat{x}_j = \cos\left(\frac{j\pi}{k}\right), \quad j = 1,\dots,k-1 \quad \text{(pour être dans }]-1,1[\text{)}. \end{aligned}$$

On a donc $T_k(\hat{x}_j) = \cos(j\pi) = (-1)^j$. On a donc une succession de minima et de maxima locaux, voir Figure 1.3.

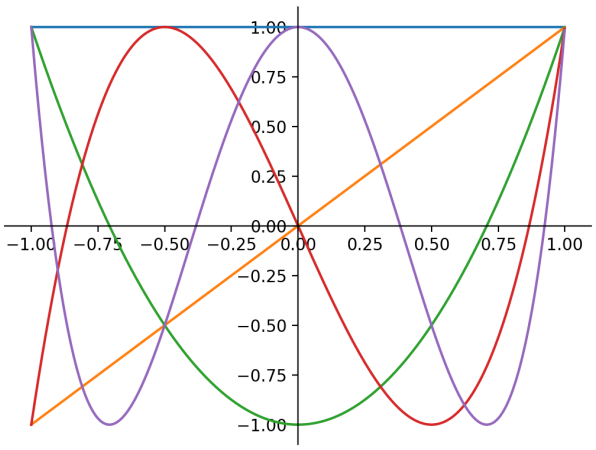

FIGURE 1.3 – Polynômes de Tchebychev pour $k = 0,\dots,4$

On remarque en particulier que les racines de T_k séparent les racines de T_{k+1}.

Meilleurs points d'interpolation
Soit Γ_{n+1} l'ensemble des polynômes de degré au plus $n+1$ tels que les polynômes soient unitaires ($p(x) = x^{n+1} + \dots$) et aient leurs racines réelles et distinctes dans $[-1,1]$. On veut donc résoudre le problème (équivalent à (1.6))

$$\min_{v\in\Gamma_{n+1}} \|v\|_{L^\infty}. \tag{1.8}$$

Remarque 1.15. *Le polynôme* $\frac{1}{2^n}T_{n+1}$ *appartient à* Γ_{n+1}.

Théorème 1.16. *Le polynôme* $\frac{1}{2^n}T_{n+1}$ *est la solution du problème* (1.8)*, i.e., les meilleurs* x_j *sont les racines de* T_{n+1} *:*

$$x_j = \cos\left(\frac{2j+1}{2n+2}\pi\right), \quad j = 0, \dots n.$$

Démonstration. Par l'absurde, on suppose qu'il existe $p \in \Gamma_{n+1}$ tel que

$$\|p\|_{L^\infty} < \left\|\frac{T_{n+1}}{2^n}\right\|_{L^\infty}.$$

On note $r(x) = \frac{T_{n+1}(x)}{2^n} - p(x)$. On a en particulier $\deg(r) \leq n$ (car c'est la différence de 2 polynômes unitaires de degré $n+1$). De plus, pour

$$\hat{x}_j = \cos\left(\frac{j\pi}{n+1}\right), \quad j = 0, \dots, n+1,$$

on a

$$\frac{T_{n+1}(\hat{x}_j)}{2^n} = \frac{(-1)^j}{2^n}$$

et

$$\max_{x\in[-1,1]} |p(x)| \leq \max_{x\in[-1,1]} \left|\frac{T_{n+1}}{2^n}\right| = \frac{1}{2^n}.$$

On en déduit donc que

$$r(\hat{x}_j) = \frac{(-1)^j}{2^n} - p(\hat{x}_j) \begin{cases} > 0 & \text{si } j \text{ est pair} \\ < 0 & \text{si } j \text{ est impair} \end{cases}$$

Ceci implique que r s'annule au moins $n+1$ fois (sur chaque intervalle $]\hat{x}_j, \hat{x}_{j+1}[$. Or $\deg(r) \leq n$, donc $r(x) = 0$ pour tout x. On en déduit que $p = \frac{T_{n+1}}{2^n}$, ce qui est absurde. □

Dans la Figure 1.4, on compare l'interpolation de la fonction

$$f(x) = \frac{1}{1+x^2} \quad \text{sur } [-5, 5]$$

pour 13 points d'interpolation, respectivement avec les points équi-distants et les points de Tchebychev.

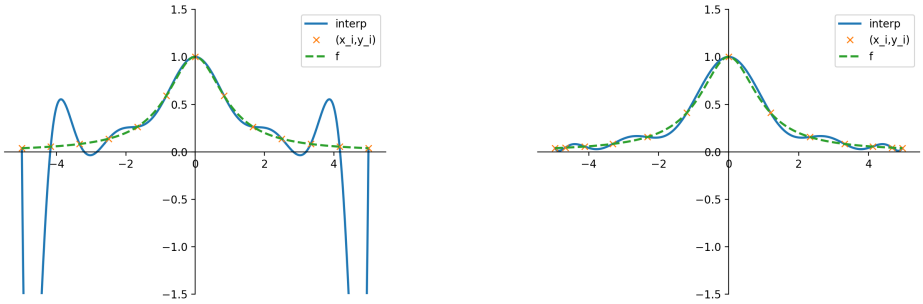

FIGURE 1.4 – Interpolation de f avec les points équi-distants (à gauche) et les points de Tchebychev (à droite).

Convergence

La situation pour la convergence n'est pas si simple. On a vu que la convergence ne pouvait pas avoir lieu pour n'importe quel choix de points x_i (phénomène de Runge). Les 2 théorèmes suivants illustre la difficulté du choix des points x_i :

Théorème 1.17. *Pour chaque choix de séquence de points $x_i^{(n)}$, $i = 0, \dots, n$, il existe au moins une fonction $g \in C^0([a,b])$ pour laquelle il n'y a pas convergence, i.e.*

$$\lim_{n \to +\infty} \|g - P_n g\|_{L^\infty} > 0.$$

Théorème 1.18. *Soit $f \in C^0([a,b])$. Alors, il existe un choix de séquence de points $x_i^{(n)}$, $i = 0, \dots, n$ pour lequel on a convergence, i.e.*

$$\lim_{n \to +\infty} \|f - P_n f\|_{L^\infty} = 0.$$

Remarque 1.19. *Cela illustre le Théorème de Weierstrass qui dit que toute fonction continue peut être approchée par un polynôme. Tout le problème est que l'on ne peut pas savoir a priori quel choix de points d'interpolation (et donc de polynôme) il faut faire.*

VII Interpolation d'Hermite

Le problème d'interpolation d'Hermite se formule ainsi : étant donné une fonction $f \in C^M([a,b])$, avec $a < b$ et $M \in \mathbb{N}^*$, et $n+1$ points d'interpolation $\{x_0, x_1, \cdots, x_n\}$ distincts, on cherche le polynôme $p \in \mathcal{P}_N$ vérifiant :

$$f^{(j)}(x_i) = p^{(j)}(x_i) \quad \forall i \in \{0, \cdots, n\}, \forall j \in \{0, \cdots, M\} \tag{1.9}$$

où l'indice supérieur indique l'ordre de dérivation. À la différence de l'interpolation de Lagrange, on demande en plus aux dérivées du polynôme p d'être égales aux dérivées de f jusque l'ordre M en chaque point. En un sens, l'interpolation d'Hermite est un mixte entre l'interpolation de Lagrange et les développements limités où le polynôme p égale f et ses dérivées d'ordre successif en un même point.

Pour résoudre le problème (1.9), la première question qui apparaît est de savoir quel degré N du polynôme nous devons prendre. A priori, ayant $(M+1)\times(n+1)$ conditions à satisfaire, le polynôme d'interpolation devra être de degré $N = (M+1)\times(n+1) - 1$:

Théorème 1.20. *Le problème d'interpolation d'Hermite* (1.9) *admet une unique solution* $p \in \mathcal{P}_N$ *avec* $N = (M+1)\times(n+1) - 1$.

Démonstration. Pour montrer ce résultat, commençons par montrer que si une solution existe, alors elle est unique. Pour ce faire, on considère p_1 et p_2 deux polynômes de degré $N = (M+1)\times(n+1) - 1$ solutions de (1.9). En posant $v = p_1 - p2$, on déduit que v est de degré N (au plus) et que :

$$v^{(j)}(x_i) = 0 \quad \forall i \in \{0,\cdots,n\}, \forall j \in \{0,\cdots,M\}$$

Par conséquent, les x_i sont racines de v et elles sont racines de multiplicité $M+1$. On a alors nécessairement v de la forme :

$$v(x) = \underbrace{\prod_{i=0}^{n}(x-x_i)^M\, Q(x)}_{\text{degré } N+1}$$

Or, v étant de degré N, on déduit que nécessairement $Q(x) \equiv 0$ d'où $p_1 = p_2$. On prouve ainsi l'unicité.

Pour montrer l'existence d'une solution, on remarque que chercher $p \in \mathcal{P}_N$, c'est à dire de la forme :

$$p(x) = \alpha_0 + \alpha_1 x + \cdots + \alpha_N x^n$$

solution des $N+1$ équations (1.9) revient à résoudre un système linéaire de la forme $A\underline{\alpha} = \underline{b}$ où $\underline{\alpha}$ est le vecteur des α_i. Or, on sait par le théorème du rang que si A est injectif, alors elle est surjective. L'unicité montrer juste avant nous donne l'injectivité de A et permet de conclure. □

Remarque 1.21. *Une formulation plus général du problème d'interpolation d'Hermite peut être donné ainsi : Trouver le polynôme p vérifiant :*

$$f^{(j)}(x_i) = p^{(j)}(x_i) \quad \forall i \in \{0, \cdots, n\}, \forall j \in \{0, \cdots, M_i\}$$

où les ordres de dérivations $M_i \in \mathbb{N}^*$ sont différents en chaque point. En reprenant la démonstration ci-dessus, on montre que ce problème admet une unique solution dans $\mathcal{P}_N$ avec $N = \left[\sum_{i=0}^{n}(M_i + 1)\right] - 1$.

Le problème (1.9) étant bien posé, l'objectif maintenant est de voir comment les algorithmes construits pour l'interpolation de Lagrange (Neville-Aitken, différences divisées, Horner) peuvent s'adapter pour évaluer ou décomposer dans une base notre polynôme d'interpolation. L'idée, relativement simple (mais géniale !), est de "dédoubler" les points x_i en posant :

$$x_0 = y_0 = \cdots = y_M < x_1 = y_{M+1} = \cdots = y_{2M+1} < \cdots < x_n = y_{n\times M+n} = \cdots = y_{(n+1)\times M+n}$$

On aura ainsi $(n+1) \times (M+1)$ points y_l.

Généralisation de l'algorithme de Neville-Aitken
Pour généraliser l'algorithme, on utilise la propriété suivante :

$$T_{k+1}^i(x) = \begin{cases} \dfrac{(y_{i+k+1} - x)T_k^i(x) - (y_i - x)T_k^{i+1}(x)}{y_{i+k+1} - y_i} & \text{si} \quad y_i \neq y_{i+k+1} \\ T_k^i(x) + f^{(k+1)}(y_i)\frac{(x-y_i)^{k+1}}{(k+1)!} & \text{si} \quad y_i = y_{i+k+1} \end{cases} \tag{1.10}$$

où $T_0^i = f(y_i)$ et T_k^i est le polynôme d'interpolation d'Hermite vérifiant :

$$\left(T_k^i\right)^{(j)}(y_l) = f^{(j)}(y_l) \quad \forall l \in \{i, \cdots, i+k\}, \forall j \in \{0, \cdots, k_l = (l-i) \mod M\} \tag{1.11}$$

Pour retrouver ce résultat, la démonstration repose sur une récurrence sur k. On pourra noter par ailleurs que dans l'expression de T_{k+1}^i dans le cas où $y_i = y_{i+k+1}$, on retrouve exactement le D.L. de f à l'ordre $k+1$ au voisinage de y_i.

Ainsi, à l'aide de la relation (1.10), on retrouve directement un schéma de calcul similaire à l'algorithme de Neville-Aitken pour évaluer le polynôme d'interpolation en un point x arbitraire, avec comme seule adaptation l'évaluation de $T_{k+1}^i(x)$ lorsque $y_i = y_{i+k+1}$.

Généralisation de l'algorithme des différences divisées
Avant généraliser l'algorithme des différences divisées, il est utile de préciser à quoi correspond la base de Newton dans ce cas. Ici, la base de Newton prendra la forme suivante :

$$N_i(x) = \prod_{j=0}^{i-1}(x - y_j), \quad \text{et} \quad N_0(x) = 1.$$

Dans cette base, comme pour l'interpolation de Lagrange, on pourra appliquer directement l'algorithme de Horner pour évaluer le polynôme en un point x donné. Reste à voir comment obtenir la décomposition, c'est à dire les coefficients α_j t.q.

$$p(x) = \sum_{j=0}^{(n+1)\times(M+1)-1} \alpha_j N_j(x).$$

Théorème 1.22. *Le polynôme d'interpolation d'Hermite dans la base de Newton est donnée par :*

$$p(x) = \sum_{j=0}^{K-1} f[y_0, \cdots, y_j] N_j(x)$$

où $K = (n+1) \times (M+1)$ *et*

$$f[y_i, \cdots, y_{i+j}] = \begin{cases} \frac{f[y_{i+1}, \cdots, y_{i+j}] - f[y_[i], \cdots, y_{i+j-1}]}{y_{i+j} - y_i} & si \quad y_i \neq y_{i+j} \\ \frac{f^{(j)}(y_i)}{j!} & si \quad y_i = y_{i+j} \end{cases}$$

Démonstration. La preuve de ce théorème repose sur le résultat (1.10) et sur une preuve par récurrence. Prouvons par récurrence que :

$$T_k^i(x) = f[y_i] + f[y_i, y_{i+1}](x - y_i) + \cdots + f[y_i, \cdots, y_{i+k}] \prod_{j=i}^{i+k-1} (x - y_j)$$

où on rappelle que T_k^i est défini par (1.10). Pour $k = 0$, la proposition est vraie car $f[y_i] = f(y_i) = T_0^i(x)$. Supposons le résultat vrai pour k donné et montrons l'hérédité. Notons :

$$T_{k+1}^i(x) = \alpha_0 + \alpha_1(x - y_i) + \cdots + \alpha_{k+1} \prod_{j=i}^{i+k} (x - y_j)$$

la décomposition de T_{k+1}^i dans la base $\{1, (x - y_i), \cdots, (x - y_i) \cdots (x - y_{i+k})\}$. On sait que T_{k+1}^i et T_k^i vérifient (1.11) aux points y_j pour $j \in \{i, \cdots, i + k\}$,

donc on peut déduire que $\alpha_j = f[y_i, \cdots, y_{i+j}]$. Il reste à prouver que $\alpha_{k+1} = f[y_i, \cdots, y_{i+k+1}]$. Pour cela, on utilise la relation (1.10) en regardant les termes de plus au degré (en x^{k+1}) pour déduire :

$$\alpha_{k+1} = \frac{f[y_{i+1}, \cdots, y_{i+k+1}] - f[y_i, \cdots, y_{i+k}]}{y_{i+k+1} - y_i} = f[y_i, \cdots, y_{i+k+1}]$$

si $y_{i+k+1} \neq y_i$, et

$$\alpha_{k+1} = \frac{f^{(k+1)}(y_i)}{(k+1)!}$$

sinon.

□

Estimation d'erreur

Enfin, étudions l'erreur d'interpolation d'Hermite. Utilisant plus d'informations sur la fonction que pour l'interpolation de Lagrange, ses valeurs et dérivées aux points x_i, on s'attend à avoir une meilleure approximation que l'interpolant de Lagrange pour le même nombre de points, comme illustré sur la Figure 1.5.

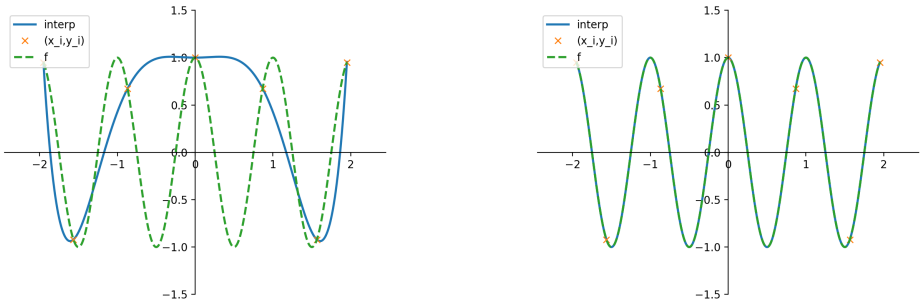

FIGURE 1.5 – Interpolation de $f(x) = \cos(2\pi x)$ avec 7 points de Chebychev, interpolation de Lagrange (à gauche) et d'Hermite avec $m = 2$ (à droite).

Théorème 1.23. *Soient f une fonction de classe C^K avec $K = (M+1)\times(n+1)$ et $(x_i)_{i=0,\cdots,n}$, $n+1$ points de $[a,b]$, alors pour tout $x \in [a,b]$, il existe $\xi \in I_x := [min(x_0, \ldots, x_n, x), \max(x_0, \ldots, x_n, x)]$ tel que l'erreur d'interpolation au point x soit donnée par*

$$f(x) - p(x) = \frac{f^{(K)}(\xi)}{K!} \prod_{j=0}^{(M+1)(n+1)} (x - y_j)$$

où les y_j sont définis par :

$$x_0 = y_0 = \cdots = y_M < x_1 = y_{M+1} = \cdots = y_{2M+1} < \cdots < x_n = y_{n\times M+n} = \cdots = y_{(n+1)\times M+n}$$

Démonstration. La preuve de ce théorème est similaire à celle du théorème 1.10. On supposera $x \neq x_i$ sinon le résultat est direct. Posons $v_K(x) = \prod_{j=0}^{M(n+1)}(x - y_j)$, et introduisons la fonction $\varphi(t)$ définie par :

$$\varphi(t) = f(t) - p(t) - v_K(t)\frac{f(x) - p(x)}{v_K(x)}$$

On remarque que $\varphi(t)$ s'annule en chaque x_i avec une multiplicité M et en x, par conséquent, sa dérivée (d'après le Thm. de Rolle) s'annulera (au moins) $n + 1$ fois (dans chaque intervalle $]x_i, x_{i+1}[$ et dans les intervalles $]x_{i_0}, x[$ et $]x, x_{i_0+1}[$ où i_0 est l'indice t.q. $x \in]x_{i_0}, x_{i_0+1}[$). De plus, chaque racine étant de multiplicité M, $\varphi'(t)$ s'annulera aussi en x_i. Au total, $\varphi'(t)$ s'annulera $2n + 2$ fois. On répète alors l'opération en dérivant de nouveau pour déduire le résultat énoncé. □

Chapitre 2

Quadrature numérique de type interpolation

I Introduction

Soient $[a,b]$ un intervalle (borné de $\mathbb{R}$) et $f \in C([a,b])$. Le but de ce chapitre est de proposer et d'étudier des méthodes numériques pour calculer

$$I := \int_a^b f(x)dx.$$

L'idée est de remplacer la fonction f par son polynôme d'interpolation $P_n f$ sur des points x_i tous distincts. On note

$$E_n(f,x) = f(x) - P_n f(x).$$

Dans la base de Lagrange, on rappelle que le polynôme $P_n f$ s'écrit

$$P_n f(x) = \sum_{j=0}^{n} f(x_j) L_j^{(n)}(x).$$

On a donc

$$\begin{aligned} I = \int_a^b f(x)dx &= \int_a^b P_n f(x)dx + \int_a^b E_n(f,x)dx \\ &= \int_a^b \sum_{j=0}^{n} f(x_j) L_j^{(n)}(x)dx + \int_a^b E_n(f,x)dx \\ &= \sum_{j=0}^{n} f(x_j) \int_a^b L_j^{(n)}(x)dx + \int_a^b E_n(f,x)dx \end{aligned}$$

On note

$$A_j^{(n)} = \int_a^b L_j^{(n)}(x)dx \quad \text{(indépendant de } f \text{ et ne dépendant que des } x_j\text{)}$$

et

$$R_n(f) = \int_a^b E_n(f,x)dx.$$

On a donc la formule suivante pour l'intégrale I :

$$I = \sum_{j=0}^{n} A_j^{(n)} f(x_j) + R_n(f).$$

Les $A_j^{(n)}$ son appelées les coefficients de quadrature et les x_j sont appelés les noeuds de quadrature. L'approximation de I sera alors donnée par

$$I \simeq \sum_{j=0}^{n} A_j^{(n)} f(x_j)$$

et $R_n(f)$ est appelée l'erreur de quadrature. On caractérise souvent ces formules de quadrature par leur exactitude sur des sous ensembles de $\mathcal{P}$

Définition 2.1. *Une formule de quadrature est dite exacte sur $\mathcal{P}_k$ (on dit aussi qu'elle a un degré d'exactitude k) si*

$$R_n(f) = 0 \quad \forall f \in \mathcal{P}_k.$$

Remarque 2.2. *Pour les formules de quadratures de type interpolation, on va voir que $R_n(f) = 0$ si $f \in \mathcal{P}_n$ donc les formules de quadrature de type interpolation ont un degré d'exactitude au moins égal à n.*

Pour clore cette introduction, notons que le calcul numérique d'intégrale apparait, par exemple, lorsqu'on ne peut pas connaître de primitive de la fonction f (typiquement $f(x) = e^{-x^2}$). Également, les formules de quadrature peuvent servir à construire des schémas numériques pour les EDO du type $y'(t) = f(t, y(t))$ en remarquant que :

$$y(t) = y(0) + \int_0^t f(s, y(s))\, ds.$$

II Formules de Newton-Côtes

On considère le cas des points equi-distants :

$$x_0 = a, \quad x_i = a + ih, \quad x_n = b \quad \text{et} \quad h = \frac{b-a}{n}.$$

On a donc

$$\begin{aligned} A_i^{(n)} &= \int_a^b L_i^{(n)}(x)dx \\ &= \int_a^b \prod_{j=0, j\neq i}^{n} \frac{x - x_j}{x_i - x_j} dx \\ &= \int_a^b \prod_{j=0, j\neq i}^{n} \frac{x - (a + jh)}{(i - j)h} dx \end{aligned}$$

On pose $x = a + th$, $t \in [0, n]$. Ainsi

$$\begin{aligned} A_i^{(n)} &= h \int_0^n \prod_{j=0, j\neq i}^{n} \frac{t - j}{i - j} dt \\ &= h \int_0^n \prod_{j=0}^{i-1} \frac{t - j}{i - j} \prod_{i+1}^{n} \frac{t - j}{i - j} dt \\ &= \frac{h(-1)^{n-i}}{i!(n - i)!} \int_0^n \prod_{j=0, j\neq i}^{n} (t - j)dt. \end{aligned}$$

A partir de cette expression on peut en particulier montrer que

$$A_{n-i}^{(n)} = A_i^{(n)}.$$

Détermination de $R_n(f)$
On a le théorème suivant :

Théorème 2.3. *1. Si n est pair et si $f \in C^{n+2}([a, b])$, alors il existe $\eta \in [a, b]$ tel que*

$$R_n(f) = h^{n+3} \frac{f^{(n+2)}(\eta)}{(n + 2)!} \int_0^n t^2(t - 1) \dots (t - n)dt.$$

2. Si n est impair et si $f \in C^{n+1}([a, b])$, alors il existe $\eta \in [a, b]$ tel que

$$R_n(f) = h^{n+2} \frac{f^{(n+1)}(\eta)}{(n + 1)!} \int_0^n t(t - 1) \dots (t - n)dt.$$

Remarque 2.4. *Quand n est pair, la formule de quadrature est exacte sur $\mathcal{P}_{n+1}$ alors que quand n est impair, elle est exacte sur $\mathcal{P}_n$.*

Applications On va maintenant expliciter les formules pour différentes valeurs de n.

***n* = 1 :** On a

$$R_1(f) = h^3 \frac{f''(\eta)}{2} \int_0^1 t(t-1)dt = -\frac{h^3}{12} f''(\eta)$$

et

$$A_0^{(1)} = A_1^{(1)} = h(-1) \int_0^1 (t-1)dt = \frac{h}{2}.$$

Ainsi

$$\boxed{\int_a^b f(x)dx = h\frac{f(a)+f(b)}{2} - \frac{h^3}{12} f''(\eta).}$$

Cette formule est appelée la formule des Trapèzes.

***n* = 2 :** Dans ce cas, on a

$$\begin{aligned} R_2(f) &= h^5 \frac{f^{(4)}(\eta)}{4!} \int_0^2 t^2(t-1)(t-2)dt \\ &= h^5 \frac{f^{(4)}(\eta)}{4!} \left[\frac{t^5}{5} - \frac{3t^4}{4} + \frac{2t^3}{3}\right]_0^2 \\ &= h^5 \frac{f^{(4)}(\eta)}{4!} \left(\frac{32}{5} - 12 + \frac{16}{3}\right) \\ &= -\frac{h^5}{90} f^{(4)}(\eta). \end{aligned}$$

De plus

$$A_0^{(2)} = A_2^{(2)} = \frac{h}{2} \int_0^2 (t-1)(t-2)dt = \frac{h}{2}\left[\frac{t^3}{3} - \frac{3t^2}{2} + 2t\right]_0^2 = \frac{h}{3}$$

et

$$A_1^{(2)} = h(-1) \int_0^2 t(t-2)dt = -h\left[\frac{t^3}{3} - t^2\right]_0^2 = \frac{4h}{3}.$$

La formule de quadrature est donc la suivant

$$\boxed{\int_a^b f(x)dx = \frac{h}{3}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right) - \frac{h^5}{90} f^{(4)}(\eta).}$$

Cette formule est appelé formule de Simpson.

***n* = 3 :** Inutile car ca n'apporte pas de précision par rapport à n=2

***n* = 4 :** Trop compliqué.

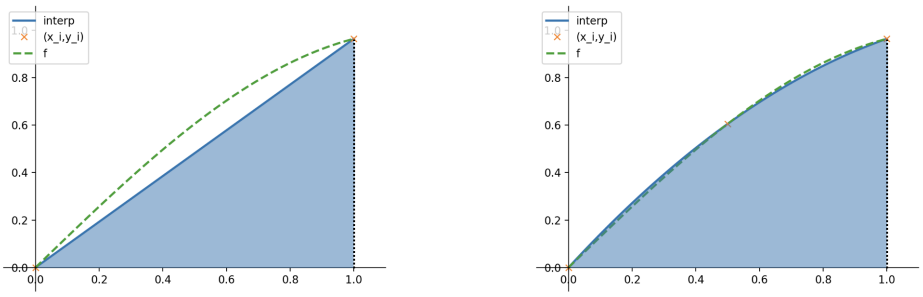

FIGURE 2.1 – Illustration des formules des Trapèzes (à gauche) et de Simpson (à droite).

Convergence et stabilité

La notion de stabilité est liée à la faible variation des résultats lorsque l'on perturbe les données $f(x_i)$. Ainsi, si on calcule une intégrale avec des valeurs $f(x_i) + \varepsilon_i$ au lieu de $f(x_i)$, on commet une erreur de $\sum_{i=0}^{n} A_i^{(n)} \varepsilon_i$ qui doit rester faible si la formule de quadrature est stable.

Définition 2.5. *On dira que la formule de quadrature est stable s'il existe une constante M_1 telle que*

$$\left| \sum_{i=0}^{n} A_i^{(n)} \varepsilon_i \right| \leq M_1 \max |\varepsilon_i| \quad \forall n.$$

Théorème 2.6. *Une condition nécessaire et suffisante pour qu'il existe une constante M_1 telle que $\left|\sum_{i=0}^{n} A_i^{(n)} \varepsilon_i\right| \leq M_1 \max|\varepsilon_i|$ $\forall n$ est qu'il existe une constante M telle que*

$$\sum_{i=0}^{n} \left| A_i^{(n)} \right| \leq M \quad \forall n.$$

Définition 2.7. *On dit que la formule de quadrature est convergente si* $\lim_{n \to +\infty} |R_n(f)| = 0$.

Théorème 2.8. *Une condition nécessaire et suffisante pour qu'une formule de quadrature soit convergente pour toute fonction $f \in V$ (V espace de Banach) est que :*

- *La formule de quadrature soit convergente sur un sous-espace W dense dans V.*
- *Il existe M telle que* $\sum_{i=0}^{n} \left| A_i^{(n)} \right| \leq M \ \forall n.$

Cas des formules de Newton-Côtes
Pour les formules de Newton-Côtes, la première condition du théorème précédent est toujours vérifiée. En effet, si $f \in \mathcal{P}$ avec $\deg(f) = k$, on a $E_n(f) = 0$ si $n \geq k$ et donc $R_n(f) = 0$ si $n \geq k$. Ainsi, on a convergence sur $\mathcal{P}$ qui est un sous-espace dense de $C^0([a,b])$.

Par contre, la deuxième condition n'est pas satisfaite ! Les formules de Newton-Côtes ne sont donc ni stables ni convergentes.

III Formules composites

Principe : L'idée des méthodes composites est de subdiviser $[a,b]$ en p sous-intervalles. Ensuite, sur chaque sous-intervalle, on va utiliser une formule de Newton-Côtes avec $n+1$ points, n petit ($n = 1, 2$), comme illustré sur la Figure 2.2.

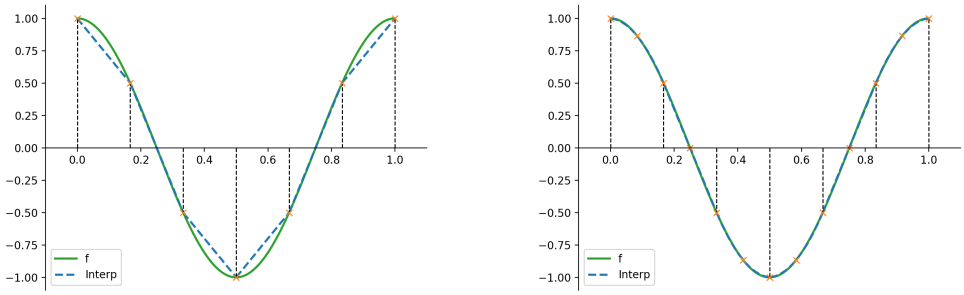

FIGURE 2.2 – Illustration des formules composites Trapèzes (à gauche) et de Simpson (à droite).

Si $n = 1$: On a la formule des trapèzes.

$$\int_a^b f(x)dx = \sum_{i=0}^{p-1} \int_{x_i}^{x_{i+1}} f(x)dx$$
$$= \sum_{i=0}^{p-1} (x_{i+1} - x_i) \left[\frac{f(x_i) + f(x_{i+1})}{2} \right] - \sum_{i=0}^{p-1} \frac{(x_{i+1} - x_i)^3}{12} f''(\eta_i)$$

avec $\eta_i \in [x_i, x_{i+1}]$. Si $x_i = a + ih$ (où $b - a = ph$), alors

$$\int_a^b f(x)dx = \sum_{i=0}^{p-1} h \left[\frac{f(x_i) + f(x_{i+1})}{2} \right] - \sum_{i=0}^{p-1} \frac{h^3}{12} f''(\eta_i)$$
$$= h \frac{f(a) + f(b)}{2} + h \sum_{i=1}^{p-1} f(x_i) - \frac{h^3}{12} \sum_{i=0}^{p-1} f''(\eta_i).$$

On peut alors utiliser le Théorème de la moyenne : si $f \in C^2([a,b])$, $\exists \xi \in [a,b]$ tel que

$$\sum_{i=0}^{p-1} \frac{f''(\eta_i)}{p} = f''(\xi).$$

On a donc

$$\int_a^b f(x)dx = h\frac{f(a)+f(b)}{2} + h\sum_{i=1}^{p-1} f(x_i) - \frac{h^3}{12} p f''(\xi).$$

et donc (comme $b-a = ph$)

$$\boxed{\int_a^b f(x)dx = h\frac{f(a)+f(b)}{2} + h\sum_{i=1}^{p-1} f(x_i) - \frac{(b-a)h^2}{12} f''(\xi).}$$

On obtient donc une méthode d'ordre 2 (erreur $= O(h^2)$).
Stabilité : On a

$$A_i = \begin{cases} \frac{h}{2} & \text{si} \quad i = 0, p \\ h & \text{si} \quad i = 1, \ldots, p-1 \end{cases}$$

Ainsi

$$\sum_{i=0}^{p} |A_i| = ph = (b-a) = M \text{ (indépendant de } n).$$

La méthode est donc stable et convergente.

Si $n = 2$: On a la formule de Simpson. On utilise p sous-intervalles de la forme $[x_{2i}, x_{2i+2}]$. On note également x_{2i+1} le point milieu de cet intervalle. On obtient donc

$$\begin{aligned} \int_a^b f(x)dx &= \sum_{i=0}^{p-1} \int_{x_{2_i}}^{x_{2i+2}} f(x)dx \\ &= \sum_{i=0}^{p-1} \frac{x_{2i+1} - x_{2i}}{3} \left(f(x_{2_i}) + 4f(x_{2i+1}) + f(x_{2i+2})\right) \\ &\quad - \sum_{i=0}^{p-1} \frac{(x_{2i+1} - x_{2i})^5}{90} f^{(4)}(\eta_i) \end{aligned}$$

avec $\eta_i \in [x_{2i}, x_{2i+2}]$. Si $x_i = a + ih$ (avec $h = \frac{b-a}{2p}$), on a alors

$$\begin{aligned} \int_a^b f(x)dx &= \sum_{i=0}^{p-1} \frac{h}{3} \left(f(x_{2_i}) + 4f(x_{2i+1}) + f(x_{2i+2})\right) - \sum_{i=0}^{p-1} \frac{h^5}{90} f^{(4)}(\eta_i) \\ &= \frac{h}{3}(f(a)+f(b)) + \frac{2h}{3}\sum_{i=1}^{p-1} f(x_{2_i}) + \frac{4h}{3}\sum_{i=0}^{p-1} f(x_{2i+1}) - \frac{h^5}{90}\sum_{i=0}^{p-1} f^{(4)}(\eta_i). \end{aligned}$$

Par le théorème de la moyenne, il existe $\xi \in [a,b]$ tel que

$$\sum_{i=0}^{p-1} \frac{f^{(4)}(\eta_i)}{p} = f^{(4)}(\xi).$$

On a donc

$$\frac{h^5}{90}\sum_{i=0}^{p-1} f^{(4)}(\eta_i) = \frac{h^5}{90} p f^{(4)}(\xi) = \frac{(b-a)h^4}{180} f^{(4)}(\xi)$$

et

$$\boxed{\int_a^b f(x)dx = \frac{h}{3}(f(a)+f(b)) + \frac{2h}{3}\sum_{i=1}^{p-1} f(x_{2_i}) + \frac{4h}{3}\sum_{i=0}^{p-1} f(x_{2i+1}) - \frac{(b-a)h^4}{180} f^{(4)}(\xi).}$$

La méthode est donc d'ordre 4.
Stabilité : On a

$$A_i = \begin{cases} \frac{h}{3} & \text{si} \quad i = 0, 2p \\ \frac{2h}{3} & \text{si} \quad i = 2k, k = 1, \dots p-1 \\ \frac{4h}{3} & \text{si} \quad i = 2k+1, k = 0, \dots p-1 \end{cases}$$

Ainsi

$$\sum_{i=0}^{p} |A_i| = \frac{h}{3}(2 + 2(p-1) + 4p) = 2hp = (b-a) = M \text{ (indépendant de } n).$$

La méthode est donc stable et convergente.

Remarque 2.9. *Une autre façon pour montrer la stabilité est de remarquer que $A_i > 0$ pour tout i donc $\sum |A_i| = \sum A_i$. En utilisant le fait que la formule de quadrature est exacte sur $\mathcal{P}_3$, on a donc, avec $f = 1$*

$$\int_a^b 1dx = b - a = \sum_{i=0}^{2p} A_i f(x_i) = \sum_{i=0}^{2p} A_i$$

donc $M = b - a$ et la méthode est stable. Cette démarche est d'ailleurs généralisable pour d'autres formules de quadrature composite.

Méthode générale de construction et d'analyse

Comme nous l'avons mentionné, le principe d'une méthode composite est de diviser l'intervalle $[a,b]$ d'intégration en p sous intervalles $[x_i, x_{i+h}]$

(avec pour rappel $x_i = ih + a$ et $h = \frac{b-a}{p}$), et d'appliquer sur chaque intervalle $[x_i, x_{i+1}]$ une formule de quadrature "simple". Considérons une méthode de quadrature (qu'on qualifiera de) simple sur l'intervalle $[-1, 1]$ à $n+1$ points, c'est à dire de la forme :

$$\int_{-1}^{1} f(x)dx \simeq \sum_{j=0}^{n} f(\xi_j)A_j \tag{2.1}$$

où les ξ_j sont $n+1$ points distincts de l'intervalle $[-1, 1]$ et les A_j sont les poids de quadrature.

Proposition 2.10. *Si la méthode de quadrature* (2.1) *est exacte sur* $\mathcal{P}_n$*, alors nécessairement on a*

$$A_j = \int_{-1}^{1} L_j^n(x)dx$$

.

La preuve de ce résultat est directe en choisissant $f(x) = L_j^n(x)$. Ce premier résultat nous donne donc une méthode a priori pour calculer les poids de quadrature A_j étant donné les points ξ_j. Une autre façon de caractériser les A_j est de remarquer que la formule est exacte sur $\mathcal{P}_n$ si et seulement si :

$$\int_{-1}^{1} x^k dx = \sum_{j=0}^{n} \xi_j^k A_j, \quad \forall k \in \{0, \cdots, n\}.$$

Les deux manières conduisent bien sûr au même résultat mais celle ci-dessus présente l'avantage de mener à la résolution d'un système linéaire (simple pour n petit).

Voyons maintenant la formule de quadrature associée à notre formule de quadrature simple. Pour ce faire, on note qu'à l'aide du changement de variable $x = \frac{(X+1)h}{2} + x_i$ où $X \in [-1, 1]$, on a :

$$\int_{x_i}^{x_i+h} f(x)dx = \frac{h}{2} \int_{-1}^{1} f\left(\frac{(X+1)h}{2} + x_i\right) dx$$

et on peut donc appliquer notre formule de quadrature simple. On a ainsi :

$$\begin{aligned} \int_a^b f(x)dx &\simeq \sum_{i=0}^{p-1} \left[\frac{h}{2} \sum_{j=0}^{n} A_j f\left(\frac{(\xi_j+1)h}{2} + x_i\right)\right] \\ &\simeq \boxed{\sum_{i=0}^{p-1} \sum_{j=0}^{n} \frac{A_j h}{2} f(x_{i,j}) \quad \text{où } x_{i,j} = \frac{(\xi_j+1)h}{2} + x_i} \end{aligned} \tag{2.2}$$

Voyons maintenant comment analyser la convergence de cette méthode :

Théorème 2.11. *Si la formule de quadrature simple* (2.1) *est exacte sur* $\mathcal{P}_k$ *avec* $k \geq n$ *et que* $f \in C^{k+1}$*, alors la formule composite associée* (2.2) *est convergence et on a l'estimation suivante :*

$$|E| = \left| \int_a^b f(x)dx - \sum_{i=0}^{p-1} \sum_{j=0}^{n} \frac{A_j h}{2} f(x_{i,j}) \right| \leq C^{ste} h^{k+1}$$

Démonstration. On commence par décomposer l'erreur E ainsi :

$$E = \sum_{i=0}^{p-1} \underbrace{\int_{x_i}^{x_i+h} f(x)dx - \sum_{j=0}^{n} \frac{A_j h}{2} f(x_{i,j})}_{:=E_i}$$

Étudions chaque E_i et posons $P_i(f) \in \mathcal{P}_k$ le polynôme d'interpolation de f sur l'intervalle $[x_i, x_i + h]$ associé à $k+1 \geq n+1$ points d'interpolation $(\tilde{x}_{i,l})_{l=0,\cdots,k}$ distincts. Parmi ces $k+1$ points, on choisira notamment les $(x_{i,j})_{j=0,\cdots,n}$. Alors, on remarque que

$$E_i = \underbrace{\int_{x_i}^{x_i+h} f(x) - P_i(f)(x)dx}_{:=E_{i,1}} + \underbrace{\int_{x_i}^{x_i+h} P_i(f)(x)dx - \sum_{j=0}^{n} \frac{A_j h}{2} f(x_{i,j})}_{:=E_{i,2}}$$

D'une part, la formule de quadrature étant exacte sur $\mathcal{P}_k$ avec $k \geq n$, on a :

$$\int_{x_i}^{x_i+h} P_i(f)(x)dx = \sum_{j=0}^{n} \frac{A_j h}{2} P_i(f)(x_{i,j})$$

et sachant que $P_i(f)$ interpole f aux points $x_{i,j}$, on déduit que $E_{i,2} = 0$. D'autre part, on utilise l'estimation vue au chapitre précédent :

$$f(x) - P_i(f)(x) = f^{(k+1)}(\eta) \frac{\prod_{l=0}^{k} (x - \tilde{x}_{i,l})}{(k+1)!}$$

où on rappelle que η dépend de x. On en déduit :

$$|E_{i,1}| \leq \int_{x_i}^{x_i+h} \left| f^{(k+1)}(\eta) \right| \frac{\prod_{l=0}^{k} |x - \tilde{x}_{i,l}|}{(k+1)!}$$

En notant que pour $x \in [x_i, x_i + h]$ on a $|x - \tilde{x}_{i,l}| \leq h$ pour tout l, on obtient la majoration :

$$|E_{i,1}| \leq \frac{h^{k+1}}{(k+1)!} \int_{x_i}^{x_i+h} \left| f^{(k+1)}(\eta) \right| dx \leq C_1^{ste} h^{k+2}$$

La deuxième inégalité est obtenue en utilisant le fait que $f^{(k+1)}$ est une fonction continue et on peut donc la majorer sur le compact $[x_i, x_i + h]$. Pour finir, il ne reste qu'à sommer les erreurs E_i :

$$|E| \leq \sum_{i=0}^{p-1} |E_i| \leq C_1^{ste} p h^{k+2}$$

ce qui conduit au résultat annoncé en se rappelant que $p = \frac{(b-a)}{h}$. □

Remarque 2.12. *Au travers de la preuve ci-dessus, on montre directement la convergence de la méthode composite, sans avoir à étudier la stabilité. En fait, toute méthode composite basée sur une méthode de quadrature simple à $n+1$ points et exacte sur $\mathcal{P}_k$ avec $k \geq n$ est stable.*

IV Formules de quadrature optimales

On veut désormais fixer les x_i de telle sorte que le degré d'exactitude soit le plus grand possible. Pour cela, nous ajoutons une fonction de poids w dans l'intégrale et nous cherchons à calculer

$$\mathcal{I} = \int_a^b f(x) w(x) dx.$$

La fonction poids w vérifie les propriétés suivantes :

$$w(x) \geq 0 \ \forall x \in [a, b] \quad \text{et} \quad w \in L^1([a, b]).$$

En écrivant f sous la forme $f(x) = P_n f(x) + E_n(f, x)$, on a donc

$$\int_a^b f(x) w(x) dx = \sum_{i=0}^{n} f(x_i) \int_a^b L_i^{(n)}(x) w(x) dx + \int_a^b E_n(f) w(x) dx.$$

On pose alors

$$A_i^{(n)} = \int_a^b L_i^{(n)}(x) w(x) dx \quad \text{et} \quad R_n(f) = \int_a^b E_n(f) w(x) dx,$$

et on obtient la formule de quadrature (appelée quadrature de Gauss) :

$$\boxed{\int_a^b f(x) w(x) dx = \sum_{i=0}^{n} A_i^{(n)} f(x_i) + R_n(f).}$$

On cherche alors $\mathcal{I}$ sous la forme

$$\mathcal{I} \simeq \sum_{i=0}^{n} A_i^{(n)} f(x_i)$$

et l'erreur de quadrature est donnée par $R_n(f)$.

Théorème 2.13. *Une condition nécessaire et suffisante pour qu'une formule de quadrature de Gauss soit exacte sur $\mathcal{P}_{2n+1}$ est que*

$$\int_a^b x^j \prod_{i=0}^{n}(x-x_i)w(x)dx = 0 \quad \text{pour tout } j = 0,\dots,n. \tag{2.3}$$

Démonstration. On note $v_n(x) = \prod_{i=0}^{n}(x-x_i)$. En particulier v_n est un polynôme de degré $n+1$.

Si la formule est exacte sur $\mathcal{P}_{2n+1}$ alors pour tout $j = 0,\dots,n$, on a

$$R_n(x^j v_n) = 0$$

et

$$\int_a^b x^j v_n(x)w(x)dx = \sum_{i=0}^{n} A_i^{(n)} x_i^j v_n(x_i) = 0$$

car $v_n(x_i) = 0$ pour tout i.

Réciproquement, on suppose que (2.3) est vérifiée. Soit $Q \in \mathcal{P}_n$. On a $Q(x) = \sum_{i=0}^{n} \alpha_i x^i$ et donc

$$\int_a^b Q(x)v_n(x)w(x)dx = \sum_{i=0}^{n} \alpha_i \int x^i v_n(x)w(x) = 0. \tag{2.4}$$

Soit maintenant $P \in \mathcal{P}_{2n+1}$. On fait la division euclidienne de P par v_n. On appelle $Q \in \mathcal{P}_n$ le quotient et $R \in \mathcal{P}_n$ le reste. On a donc

$$P(x) = v_n(x)Q(x) + R(x)$$

et

$$\int_a^b P(x)w(x)dx = \int_a^b Q(x)v_n(x)w(x)dx + \int_a^b R(x)w(x)dx.$$

Comme $Q \in \mathcal{P}_n$, par (2.4), on a

$$\int_a^b Q(x)v_n(x)w(x)dx = 0.$$

De plus, comme $R \in \mathcal{P}_n$ et en utilisant le fait qu'une formule de quadrature de type interpolation est exacte sur $\mathcal{P}_n$, on a

$$\int_a^b R(x)w(x)dx = \sum_{i=0}^{n} A_i^{(n)} R(x_i).$$

En utilisant que $P(x_i) = v_n(x_i)Q(x_i) + R(x_i) = R(x_i)$, on en déduit que

$$\int_a^b P(x)w(x)dx = \sum_{i=0}^{n} A_i^{(n)} P(x_i)$$

et donc la formule de quadrature est exacte sur $\mathcal{P}_{2n+1}$. □

Remarque 2.14. *La formule n'est par contre pas exacte sur $\mathcal{P}_{2n+2}$. En effet, on a*

$$0 < \int_a^b v_n^2(x)w(x)dx = \sum_{i=0}^{n} v_n^2(x_i)A_i^{(n)} + R_n(v_n^2) = R_n(v_n^2)$$

car $v_n(x_i) = 0$ pour tout i. On en déduit que $R_n(v_n^2) > 0$ et comme $v_n^2 \in \mathcal{P}_{2n+2}$, on en déduit que la formule n'est pas exacte sur $\mathcal{P}_{2n+2}$.

IV.1 Recherche des noeuds de quadrature

On rappelle que les noeuds de quadrature satisfont

$$\int_a^b x^j \prod_{i=0}^{n}(x - x_i)w(x)dx = 0 \quad \text{pour } j = 0, \dots, n. \tag{2.5}$$

On pose $v_n = \prod_{i=0}^{n}(x - x_i)$. Il faut donc montrer que v_n, solution de (2.5), possède des racines réelles et distinctes dans $[a, b]$.

Théorème 2.15. *Les points x_i d'une méthode de Gauss sont réels, distincts, situés dans $[a, b]$ et uniques.*

Démonstration. On a déjà démontré que (2.5) était équivalent à

$$\int_a^b Q(x)v_n(x)w(x)dx = 0 \quad \forall Q \in \mathcal{P}_n. \tag{2.6}$$

— Supposons que v_n possède une racine α au moins double. Ainsi

$$v_n(x) = (x - \alpha)^2 q(x)$$

avec $q \in \mathcal{P}_{n-1}$. On a donc

$$0 = \int_a^b q(x)v_n(x)w(x)dx = \int_a^b (x - \alpha)^2 (q(x))^2 w(x)dx > 0,$$

ce qui est absurde. Donc v_n n'a que des racines simples.

— Supposons que v_n possède une racine complexe. On note $k \leq n$ le nombre de racines réelles dans $[a, b]$ et $x_0, \dots, x_{k-1}$ ces racines. Ainsi

$$v_n(x) = (x - x_0) \dots (x - x_{k-1})p(x)$$

où p est un polynôme de degré $n + 1 - k$ et de signe constant dans $[a, b]$. On pose $Q(x) = \prod_{i=0}^{k-1}(x - x_i)$. On a alors

$$0 = \int_a^b Q(x)v_n(x)w(x)dx = \int_a^b (Q(x))^2 p(x)w(x) \begin{array}{ll} > 0 & \text{si } p > 0 \\ < 0 & \text{si } p < 0 \end{array}$$

ce qui est absurde et montre que v_n ne possède que des racines réelles.

— Il ne reste plus qu'à montrer l'unicité de v_n. Par l'absurde, on suppose qu'il existe $\bar{v}_n \in \mathcal{P}_{n+1}$ unitaire et vérifiant les conditions (2.5). On a alors

$$\int_a^b Q(x)(v_n(x) - \bar{v}_n(x))w(x)dx = 0$$

pour tout $Q \in \mathcal{P}_n$. En prenant $Q = v_n - \bar{v}_n$, on obtient

$$\int_a^b (v_n(x) - \bar{v}_n(x))^2 w(x)dx = 0$$

ce qui est absurde. Ainsi v_n est unique.

□

IV.2 Stabilité et convergence des méthodes de Gauss

Théorème 2.16. *Les méthodes de Gauss sont stables et convergentes sur* $C^\infty([a,b])$.

Démonstration. Les méthodes de Gauss sont convergentes sur $\mathcal{P}$ car ce sont des méthodes de type interpolation. Il suffit donc de montrer que

$$\sum_{i=0}^{n} |A_i^{(n)}| \leq M \quad \forall n$$

avec M indépendant de n. On montre tout d'abord que $A_i^{(n)} > 0$ pour tout i et pour tout n. Comme $\left(L_i^{(n)}\right)^2 \in \mathcal{P}_{2n}$ et que la formule est exacte sur $\mathcal{P}_{2n+1}$, on a

$$0 < \int_a^b \left(L_i^{(n)}\right)^2 w(x)dx = \sum_{j=0}^{n} A_j^{(n)} \underbrace{\left(L_i^{(n)}(x_j)\right)^2}_{=\delta_{ij}} = A_i^{(n)}.$$

Ainsi $A_i^{(n)} > 0$ et $\sum_{i=0}^n |A_i^{(n)}| = \sum_{i=0}^n A_i^{(n)}$. De plus,

$$\int_a^b w(x)dx = \sum_{i=0}^{n} A_i^{(n)}$$

car la formule est exacte sur $\mathcal{P}_0$. Comme $w \in L^1([a,b])$, on en déduit que $\sum_{i=0}^n A_i^{(n)}$ est fini et constant pour tout n.

□

IV.3 Etude de l'erreur

Théorème 2.17. *Si $f \in C^{2n+2}([a,b])$, alors l'erreur de quadrature de la méthode de Gauss est donnée par*

$$R_n(f) = \frac{f^{(2n+2)}(\xi)}{(2n+2)!}\int_a^b v_n^2(x)w(x)dx \quad \text{où } \xi \in [a,b].$$

Démonstration. Soit P_{2n+1} le polynôme d'interpolation de Hermite de f aux points $x_0,\dots,x_n$ (interpolant f et sa dérivée). Par l'erreur d'interpolation, on a

$$f(x) = P_{2n+1}(x) + v_n^2(x)\frac{f^{(2n+2)}(\eta)}{(2n+2)!}$$

où $\eta \in [a,b]$. Ceci implique que

$$\int_a^b f(x)w(x)dx = \int_a^b P_{2n+1}(x)w(x)dx + \int_a^b \frac{f^{(2n+2)}(\eta)}{(2n+2)!}v_n^2(x)w(x)dx.$$

Or, comme $P_{2n+1} \in \mathcal{P}_{2n+1}$ et que la formule est exacte sur $\mathcal{P}_{2n+1}$, on en déduit que

$$\int_a^b P_{2n+1}(x)w(x)dx = \sum_{i=0}^n P_{2n+1}(x_i)A_i^{(n)} = \sum_{i=0}^n f(x_i)A_i^{(n)}.$$

Ainsi, en utilisant le théorème de la moyenne, il existe $\xi \in [a,b]$ tel que

$$\begin{aligned} R_n(f) &= \int_a^b f(x)w(x)dx - \sum_{i=0}^n f(x_i)A_i^{(n)} \\ &= \int_a^b \frac{f^{(2n+2)}(\eta)}{(2n+2)!}v_n^2(x)w(x)dx \\ &= \frac{f^{(2n+2)}(\xi)}{(2n+2)!}\int_a^b v_n^2(x)w(x)dx \end{aligned}$$

□

IV.4 Calcul des polynômes orthogonaux v_n

Le but de cette sous-section est de calculer les polynômes v_n solutions de (2.5). On cherche v_n sous la forme

$$v_n(x) = x^{n+1} + \sum_{i=0}^n \alpha_i x^i.$$

On note $C_k = \int_a^b x^k w(x)dx$ pour $k \in \mathbb{N}$. On a donc

$$\int_a^b x^{j+n+1}w(x) + \sum_{i=0}^n \alpha_i \int_a^b x^{i+j}w(x) = 0.$$

Ainsi

$$C_{j+n+1} + \sum_{i=0}^{n} \alpha_i C_{i+j} = 0 \quad j \in \{0, \dots, n\}$$

et

$$\underbrace{\begin{bmatrix} C_0 & C_1 & \dots & C_n \\ C_1 & C_2 & \dots & C_{n+1} \\ \vdots & \vdots & & \vdots \\ C_n & C_{n+1} & \dots & C_{2n} \end{bmatrix}}_{\text{Matrice de Hankel}} \times \begin{bmatrix} \alpha_0 \\ \vdots \\ \alpha_n \end{bmatrix} = - \begin{bmatrix} C_{n+1} \\ \vdots \\ C_{2n+1} \end{bmatrix}$$

Cela donne donc un moyen pour calculer le polynôme v_n. Néanmoins, cela s'avère assez lourd numériquement. On va maintenant voir que les polynômes v_n forme une famille de polynôme orthogonaux, ce qui permet de les calculer plus facilement.

Tout d'abord, comme

$$\int_a^b Q(x) v_n(x) w(x) dx = 0 \quad \forall Q \in \mathcal{P}_n$$

et v_k est un polynôme de degré $k+1$, on obtient que

$$\int_a^b v_k(x) v_n(x) w(x) dx = 0 \quad \forall k \neq n.$$

De plus

$$\int_a^b v_k^2(x) w(x) dx > 0 \quad \forall k \in \mathbb{N}.$$

On dit alors que la famille $\{v_k\}_{k \in \mathbb{N}}$ forme un système de polynômes orthogonaux.

Théorème 2.18. *Les $\{v_k\}$ satisfont une relation de récurrence à 3 termes :*

$$v_k = (x + B_k) v_{k-1} - C_k v_{k-2}$$

avec v_{-1} et v_0 donnés.

Démonstration. La famille $\{v_i\}_{i=-1,\dots,k}$ forme une base de $\mathcal{P}_{k+1}$. On peut donc exprimer $x v_{k-1}$ dans cette base :

$$x v_{k-1}(x) = \sum_{i=-1}^{k} \beta_i v_i(x).$$

En multipliant par $v_j(x) w(x)$ puis en intégrant, on obtient :

$$\int_a^b x v_{k-1}(x) v_j(x) w(x) dx = \sum_{i=-1}^{k} \beta_i \int_a^b v_i(x) v_j(x) w(x) dx. \tag{2.7}$$

En prenant $j = -1, \ldots, k-3$, on voit que le membre de gauche vaut 0 (car $xv_j \in \mathcal{P}_{k-1}$) et, en utilisant que $\{v_k\}$ forme un système de polynômes orthogonaux, que le membre de droite vaut β_j. Ainsi $\beta_j = 0$ pour $j = -1 \ldots k-3$.

Il nous reste donc une relation entre $xv_{k-1}, v_{k-2}, v_{k-1}$ et v_k :

$$xv_{k-1} = \beta_{k-2}v_{k-2} + \beta_{k-1}v_{k-1} + \beta_k v_k.$$

En prenant $j = k-2$ dans (2.7), on a alors

$$\int_a^b xv_{k-1}(x)v_{k-2}(x)w(x)dx = \beta_{k-2}\int_a^b v_{k-2}^2(x)w(x)dx. \tag{2.8}$$

On pose $h_i = \int_a^b v_i^2(x)w(x)dx$. Par la propriété d'orthogonalité, on a

$$h_i = \int_a^b x^i v_i(x)w(x)dx = \int_a^b xv_{i-1}v_i(x)w(x)dx.$$

Ainsi, (2.8) implique que

$$\beta_{k-2} = \frac{h_{k-1}}{h_{k-2}}.$$

En prenant maintenant $j = k-1$ dans (2.7), on a

$$\int_a^b v_{k-1}^2(x)w(x)dx = \beta_{k-1}\int_a^b v_{k-1}^2(x)w(x)dx$$

et donc

$$\beta_{k-1} = \frac{1}{h_{k-1}}\int_a^b v_{k-1}^2(x)w(x)dx.$$

Enfin, pour $j = k$ dans (2.7), on a

$$\int_a^b xv_{k-1}(x)v_k(x)w(x)dx = \beta_k\int_a^b v_k^2(x)w(x)dx,$$

i.e.

$$\beta_k = 1.$$

On obtien finalement

$$v_k = (x - \beta_{k-1})v_{k-1} - \beta_{k-2}v_{k-2}$$

avec

$$\beta_{k-2} = \frac{h_{k-1}}{h_{k-2}} \quad \text{et} \quad \beta_{k-1} = \frac{1}{h_{k-1}}\int_a^b v_{k-1}^2(x)w(x)dx.$$

□

IV.5 Quelques exemples classiques de polynômes orthogonaux

Polynômes de Hermite : $\{H_n\}_{n\in\mathbb{N}}$ obtenus pour

$$w(x) = e^{-x^2}.$$

Polynômes de Laguerre : $\{L_n^\alpha\}_{n\in\mathbb{N}}$ obtenus pour

$$w(x) = e^{-x}x^\alpha \quad \alpha > -1.$$

Polynômes de Jacobi : $\{P_n^{\alpha,\beta}\}_{n\in\mathbb{N}}$ obtenus pour

$$w(x) = (1-x)^\alpha(1+x)^\beta \quad \alpha > -1,\ \beta > -1.$$

- Pour $\alpha = \beta = 0$, on obtient les polynômes de Legendre.
- Pour $\alpha = \beta = -\frac{1}{2}$, on a $w(x) = \frac{1}{\sqrt{1-x^2}}$ et on obtient les polynômes de Chebyschev de première espèce ($P_k = \cos(k \arccos(x))$).
- Pour $\alpha = \beta = \frac{1}{2}$, on a $w(x) = \sqrt{1-x^2}$ et on obtient les polynômes de Chebyschev de deuxième espèce.

Chapitre 3

Résolution numérique d'EDO

I Introduction

Dans ce chapitre, nous allons nous intéresser à la résolution (de systèmes) d'équations différentielles ordinaires (EDO). Plus précisément, on étudiera le problème suivant :

Définition 3.1 (Problème de Cauchy)**.** *Le problème dit de Cauchy consiste à trouver* $\underline{y} : t \in \mathbb{R} \to \underline{y}(t) \in \mathbb{R}^n$ *une fonction* C^1 *solution de :*

$$\left| \begin{array}{l} \underline{y}'(t) = \underline{f}(t, \underline{y}(t)) \quad \textit{pour tout} \quad t \in [a,b] \\ \underline{y}(t_0) = \underline{y}_0 \end{array} \right. \tag{3.1}$$

où $t_0 \in [a,b]$ *et* $\underline{f} : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ *est une fonction continue.*

Nous verrons ci-après sous quelles conditions sur la fonction f ce problème admet une unique solution. Notons que ce type de problème apparait dans de nombreuses situations de la physique, de la biologie ou encore de l'économie. Typiquement en physique, la deuxième loi de Newton indique que l'accélération $\underline{x}''$ d'un object à la position $\underline{x}$ vérifie :

$$m\, \underline{x}''(t) = \underline{F}$$

où $\underline{F}$ est la somme des forces appliquées à cette objet. On peut alors se ramener à une formulation dite d'ordre 1 en posant $\underline{y} = [\underline{x}, \underline{x}']$ qui vérifiera alors :

$$\underline{y}'(t) = \frac{1}{m} [\underline{x}', \underline{F}] =: \widetilde{\underline{F}}$$

Pour retrouver exactement un problème de Cauchy (3.1), il faut ajouter une condition "initiale" $\underline{y}(t_0) = \underline{y}_0$. En admettant que $\widetilde{\underline{F}}$ vérifie les bonnes hypothèses assurant existence et unicité de la solution du problème de Cauchy, on

retrouve ici que pour déterminer de manière unique la trajectoire d'un objet, il faudra connaître sa position $\underline{x}$ et sa vitesse $\underline{x}'$ à un instant t_0.

Remarque 3.2. *On peut toujours ramener une équation différentielle d'ordre n de la forme :*

$$x^{(n)}(t) = F(t, x(t), \cdots, x^{(n-1)}(t))$$

à un système d'ordre 1 en posant $\underline{y}(t) = [x, \cdots, x^{(n-1)}]$.

Sauf pour des cas particulier, connaître analytiquement la solution d'une EDO n'est pas possible. Il est alors nécessaire de recourir à une méthode numérique. Mais avant d'utiliser de telles méthodes, il faut s'assurer que notre problème est bien posé, c'est à dire qu'il admet une solution et que cette dernière est unique.

II Le théorème de Cauchy-Lipschitz

L'object de cette section sera de prouver le théorème suivant :

Théorème 3.3 (Cauchy-Lipschitz)**.** *Si la fonction $\underline{f}$ du problème de Cauchy (3.1) est continue et Lipschitzienne par rapport à sa deuxième variable, c'est à dire qu'il existe $L > 0$ t.q. pour tout $t \in [a, b]$ et pour tout $\underline{y}_1 \in \mathbb{R}^n$ et $y_2 \in \mathbb{R}^n$ on ait :*

$$\|\underline{f}(t, \underline{y}_1) - \underline{f}(t, \underline{y}_2)\|_{\mathbb{R}^n} \leq L\|\underline{y}_1 - \underline{y}_2\|_{\mathbb{R}^n}$$

alors le problème de Cauchy admet une unique solution.

Pour simplifier les explications, nous allons nous attacher à montrer le résultat dans la situation où $y(t) \in \mathbb{R}$. L'extension au cas $\mathbb{R}^n$ ne pose pas de réelles difficultés. Tout d'abord, notons que y est solution du problème (3.1) ssi :

$$y(t) = y(t_0) + \int_{t_0}^{t} f(s, y(s))\, ds \tag{3.2}$$

Unicité

Commeçons par prouver que si une solution existe, alors elle est unique. Considérons y_1 et y_2 deux solutions du problème de Cauchy, alors la différence vérifie :

$$y_1(t) - y_2(t) = \int_{t_0}^{t} f(s, y_1(s) - y_2(s))\, ds \;\Rightarrow\; |y_1(t) - y_2(t)| \leq \int_{t_0}^{t} |f(s, y_1(s) - y_2(s))|\, ds$$

$$\Rightarrow\; |y_1(t) - y_2(t)| \leq L \int_{t_0}^{t} |y_1(s) - y_2(s)|\, ds$$

où $L > 0$ est la constante de Lipschitz de f. Pour conclure sur l'unicité, nous aurons besoin du Lemme de Grönwall suivant :

Lemme 3.4 (Grönwall)**.** *Soient* $\Phi : [t_0, t_1] \to \mathbb{R}$ *et* $\Psi : [t_0, t_1] \to \mathbb{R}$ *deux fonctions continues et positives vérifiant :*

$$\Phi(t) \leq K + L \int_{t_0}^{t} \Psi(s)\Phi(s)\, ds,$$

alors on a l'inégalité suivante :

$$\Phi(t) \leq K e^{L \int_{t_0}^{t} \Psi(s)\, ds}, \quad \forall t \in [t_0, t_1].$$

En appliquant ce lemme avec $K = 0$, $\Phi(t) = |y_1(t) - y_2(t)|$, $\Psi(t) = 1$ et $L = L$, on déduit aisément

$$|y_1(t) - y_2(t)| \leq 0$$

ce qui prouve l'unicité.

Existence

Commençons par rappeler que l'ensemble $C^0([a, b])$ muni de la norme infinie est complet. Posons alors $\mathcal{L} : (C^0([t_0, t_1]), \|\cdot\|_\infty) \to (C^0([t_0, t_1]), \|\cdot\|_\infty)$ l'application définie par :

$$\mathcal{L}y(t) = y(t_0) + \int_{t_0}^{t} f(s, y(s))\, ds \quad t \in [t_0, t_1]$$

Ainsi, trouver y solution de (3.2) revient à chercher un point fixe de $\mathcal{L}$. Montrons que $\mathcal{L}$ est une contraction :

$$\begin{aligned}
\|\mathcal{L}y_1 - \mathcal{L}y_2\|_\infty &= \| \int_{t_0}^{t} f(s, y_1(s)) - f(s, y_2(s))\, ds \|_\infty \\
&\leq \sup_{t \in [t_0, t_1]} \left| \int_{t_0}^{t} f(s, y_1(s)) - f(s, y_2(s))\, ds \right| \\
&\leq \sup_{t \in [t_0, t_1]} \int_{t_0}^{t} |f(s, y_1(s)) - f(s, y_2(s))|\, ds \\
&\leq L \sup_{t \in [t_0, t_1]} \int_{t_0}^{t} |y_1(s) - y_2(s)|\, ds \\
&\leq L(t_1 - t_0)\|y_1 - y_2\|_\infty
\end{aligned}$$

On déduit que si $t_1 (> t_0)$ est t.q. $L(t_1 - t_0) < 1$, alors on a une contraction. Il existe donc un (unique) point fixe $y^1 \in C^0([t_0, t_1])$ solution de (3.2). On a ainsi construit une solution du problème de Cauchy sur l'intervalle $[t_0, t_1]$.

Pour construire une solution globale, c'est à dire sur $\mathbb{R}$, l'idée est de remarquer qu'on peut construire de même une solution y^2 solution sur l'intervalle $\left[\frac{t_0+t_1}{2}, t_1 + \frac{t_1-t_0}{2}\right]$ avec comme condition "initiale" $y^2(\frac{t_0+t_1}{2}) = y^1(\frac{t_0+t_1}{2})$. Par unicité de la solution, on déduit que $y^1 = y^2$ sur l'intervalle $\left[\frac{t_0+t_1}{2}, t_1\right]$, et par conséquent on peut prolonger y_1 (par y_2) sur l'intervalle $\left[t_0, t_1 + \frac{t_1-t_0}{2}\right]$, et ainsi de suite.

Remarque 3.5. *Une situation importante pour les applications pratiques est le cas où f est localement Lipschitz. Dans ce cas, on ne peut plus prouver l'existence d'une solution globale, mais seulement d'une solution locale, c'est à dire qui existe sur un intervalle $[a, b]$ autour de t_0.*

III Méthodes à 1−pas

Nous venons de voir que, sous certaines conditions, le problème de Cauchy (3.1) admet une unique solution. Sauf pour des cas particuliers, déterminer la solution analytiquement est impossible. On a alors recourt à des méthodes numériques. L'idée directrice pour construire une méthode numérique est, partant de la relation :

$$\underline{y}(t) = \underline{y}(t_0) + \int_{t_0}^{t} \underline{f}(s, \underline{y}(s))\, ds, \tag{3.3}$$

d'appliquer une formule de quadrature pour approcher l'intégrale ci-dessus. Plus précisément, en s'inspirant des méthodes de quadrature composite, on pose $t_{i+1} = t_i + h_i$ avec $h_i > 0$ une suite de pas de discrétisation telle que $t_N = T$ où $[t_0, T]$ est l'intervalle d'intérêt. L'objectif sera de calculer $\underline{y}_i$ une approximation de la solution $\underline{y}(t_i)$ à l'instant t_i. Pour ce faire, nous verrons deux approches :

1. les méthodes à 1−pas : elles consistent à calculer $\underline{y}_{i+1}$ uniquement à l'aide de $\underline{y}_i$,
2. les méthodes multi-pas (ou à pas liés) : elles consistent à calculer $\underline{y}_{i+1}$ à l'aide de r−pas antérieures $\underline{y}_i, \cdots, \underline{y}_{i-r+1}$.

III.1 Éléments d'analyse des méthodes

Les méthodes à 1−pas s'écrivent sous la forme suivante :

$$\underline{y}_{i+1} = \underline{y}_i + h_i \underline{\Phi}(t_i, \underline{y}_i, h_i) \tag{3.4}$$

où $\underline{\Phi} : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ est ce qu'on appelle la *fonction d'incrément* et est une fonction continue. Bien évidemment, on devra choisir cette fonction de

sorte que $\underline{y}_i$ soit une bonne approximation de $\underline{y}(t_i)$. Pour étudier la qualité du schéma (3.4), nous devrons introduire trois notions : la *consistance*, la *stabilité* et la *convergence*.

Consistance
Intuitivement, la consistance d'un schéma consiste à vérifier que ce dernier est "cohérent" avec l'équation, c'est à dire que si on injecte la solution exacte, l'erreur est petite.

Définition 3.6. *On appelle erreur de consistance et on note ε_i la quantité :*

$$\underline{\varepsilon}_i = \underline{y}(t_{i+1}) - h_i \underline{\Phi}(t_i, \underline{y}(t_i), h_i)$$

On dira que la méthode à 1–pas est consistante ssi

$$\lim_{h\to 0} \sum_{i=0}^{N-1} \|\varepsilon_i\|_{\mathbb{R}^n} = 0 \quad \text{où} \quad h = \max_{0\le i\le N-1} h_i$$

Théorème 3.7. *Une méthode à 1–pas est consistante ssi $\underline{\Phi}(t, \underline{y}(t), 0) = \underline{f}(t, \underline{y}(t))$ où $\underline{y}$ est la solution du problème de Cauchy.*

Nous ferons la preuve de ce résultat avec la démonstration d'un autre théorème plus général donné ci-après.

Stabilité
Commençons par donner la définition :

Définition 3.8. *Soit $\underline{z}_i$ la suite définie par :*

$$\underline{z}_{i+1} = \underline{z}_i + h_i \underline{\Phi}(t_i, \underline{z}_i, h_i) + \underline{\widetilde{\varepsilon}}_i$$

où $(\underline{\widetilde{\varepsilon}}_i)_i$ est une suite de perturbation. On dit que la méthode est stable ssi il existe une constante M t.q.

$$\max_{0\le i\le N} \|z_i - y_i\|_{\mathbb{R}^n} \le M \left(\|z_0 - y_0\|_{\mathbb{R}^n} + \sum_{i=0}^{N-1} \|\widetilde{\varepsilon}_i\|_{\mathbb{R}^n} \right)$$

L'idée de la définition est de se dire que si on perturbe le schéma par $\underline{\widetilde{\varepsilon}}_i$, alors l'écart entre $\underline{y}_i$ et $\underline{z}_i$ reste borné par la somme des perturbations (et l'écart initiale $\|\underline{z}_0 - \underline{y}_0\|$).

Théorème 3.9. *Une condition suffisante pour garantir la stabilité d'une méthode à 1–pas est que la fonction $\underline{\Phi}$ soit Lipschitz par rapport à sa deuxième variable, i.e. il existe L^* t.q. $\forall t \in [t_0, T]$, $\forall (\underline{z}_1, \underline{z}_2) \in \mathbb{R}^n$, $\forall h \in [0, h^*]$:*

$$\|\Phi(t, z_2, h) - \Phi(t, z_1, h)\|_{\mathbb{R}^m} \le \|z_2 - z_1\|_{\mathbb{R}^n}$$

Démonstration. Pour la preuve, nous prendrons pour simplifier $n = 1$. On commence par noter que, la fonction Φ étant supposée Lipschitz, on a :

$$\begin{aligned}
|z_{i+1} - y_i| &\leq |z_i - y_i| + h_i|\Phi(t_i, z_i, h_i) - \Phi(t_i, y_i, h_i)| + |\widetilde{\varepsilon}_i| \\
&\leq |z_i - y_i| + h_i L^*|z_i - y_i| + |\widetilde{\varepsilon}_i| \\
&\leq (1 + L^*h)|z_i - y_i| + |\widetilde{\varepsilon}_i| \\
&\leq (1 + L^*h)^2|z_{i-1} - y_{i-1}| + |\widetilde{\varepsilon}_i| + (1 + L^*h)|\widetilde{\varepsilon}_{i-1}| \\
&\leq \cdots \\
&\leq (1 + L^*h)^{i+1}|z_0 - y_0| + \sum_{j=0}^{i} |\widetilde{\varepsilon}_j|(1 + L^*h)^{i-j}
\end{aligned}$$

où on rappelle que $h = \max_{0\leq i\leq N} h_i$. En remarquant que $(1 + u) \leq e^u$ et en notant $T^* = Nh$, on déduit $\forall i \leq N - 1$:

$$\begin{aligned}
|z_{i+1} - y_i| &\leq e^{L^*(i+1)h}|z_0 - y_0| + \sum_{j=0}^{i} |\widetilde{\varepsilon}_j| e^{L^*(i-j)h} \\
&\leq e^{L^*T^*}\left(|z_0 - y_0| + \sum_{j=0}^{N-1} |\widetilde{\varepsilon}_j|\right)
\end{aligned}$$

ce qui prouve la stabilité de la méthode (la majoration est indépendante de i). □

Convergence

Cette dernière notion est la plus importante car elle revient à étudier l'erreur entre notre solution y_i qu'on peut calculer avec notre schéma et $y(t_i)$ la solution exacte.

Définition 3.10. *On dit qu'une méthode à 1−pas est convergente ssi*

$$\lim_{h\to 0} \max_{0\leq i\leq n} \|y_i - y(t_i)\|_{\mathbb{R}^n} = 0 \quad où \quad h = \max_{0\leq i\leq N} h_i$$

Le théorème suivant fait le lien entre les notions de *consistance*, *stabilité* et *convergence* :

Théorème 3.11. *Si la méthode à 1−pas est stable et consistante, alors elle est convergente.*

Démonstration. Là encore, pour simplifier les explications nous considérerons $n = 1$. Considérons une méthode consistante et stable. Alors, par définition de l'erreur de consistance on a :

$$y(t_{i+1}) = y(t_i) + h_i\Phi(t_i, y(t_i), h_i) + \varepsilon_i$$

La méthode étant stable, on déduit par définition de la stabilité :

$$\max_{0\le i\le N}|y(t_i)-y_i| \le M\left(|y(t_0)-y_0|+\sum_{i=0}^{N-1}|\varepsilon_i|\right) = M\sum_{i=0}^{N-1}|\varepsilon_i| \underset{h\to 0}{\longrightarrow} 0$$

car la méthode est initialisé en prenant $y_0 = y(t_0)$ et étant consistante, on obtient déduit bien le résultat ci-dessus. □

Ce résultat est très important car il donne la démarche pour étudier une méthode à 1–pas. On cherchera d'abord à prouver qu'elle est *consistante*, puis *stable* pour déduire qu'elle est *convergente*. Insistons également sur le fait que la *convergence* est primordiale car elle nous assure que la solution approchée calculée à un sens puisqu'elle tend vers la solution exacte pourvu que l'effort de calculs est suffisant (et cela sans connaitre bien sûr la solution exacte !).

Ordre

Une question naturelle est alors celle de la vitesse de convergence. Ce sera un critère (pas le seul) déterminant la qualité d'un schéma numérique.

Définition 3.12. *On dira que la méthode à 1–pas est d'ordre p ssi il existe un réel K (dépendant uniquement de y solution du problème de Cauchy et Φ) t.q. :*

$$\sum_{i=0}^{N-1}|\varepsilon_i| \le Kh^p \quad\Leftrightarrow\quad \sum_{i=0}^{N-1}|\varepsilon_i| = O(h^p)$$

Remarque 3.13. *On notera que si la méthode est d'ordre 1 au moins, elle est alors consistante.*

Théorème 3.14. *Si la méthode est stable et d'ordre p, alors on a la majoration d'erreur :*

$$\max_{0\le i\le n}\|\underline{y}_i - \underline{y}(t_i)\|_{\mathbb{R}^n} \;\le\; C^{ste}h^p \quad\Leftrightarrow\quad \max_{0\le i\le n}\|\underline{y}_i - \underline{y}(t_i)\|_{\mathbb{R}^n} = O(h^p)$$

La preuve de ce résultat est directe en reprenant la démarche du Théorème 3.11. Voyons maintenant un théorème pratique pour déterminer l'ordre d'une méthode à 1–pas.

Théorème 3.15. *Supposons la fonction $\underline{f}$ de classe C^p par rapport à ses deux variables et la fonction $\underline{\Phi}$ également de classe C^p par rapport à sa dernière variable h. Alors, une condition nécessaire et suffisante pour qu'une méthode à 1–pas soit d'ordre p est que :*

$$\frac{1}{j+1}\frac{d^j}{dt^j}\underline{f}(t,\underline{y}(t)) \;=\; \frac{\partial^j}{\partial h^j}\underline{\Phi}(t,\underline{y}(t),h)\Big|_{h=0} \qquad \forall j\in\{0,\cdots,p-1\} \qquad (3.5)$$

Démonstration. On se place dans le cas $n = 1$. Rappelons que l'erreur de consistance est donnée par :

$$\varepsilon_i = y(t_i + h_i) - y(t_i) - h_i \Phi(t_i, y(t_i), h_i)$$

L'idée clé est d'utiliser un développement de Taylor à l'ordre p. D'une part on a :

$$\begin{aligned} y(t_i + h_i) &= \sum_{j=0}^{p} y^{(j)}(t_i) \frac{h_i^j}{j!} + O(h_i^{p+1}) \\ &= y(t_i) + \sum_{j=1}^{p} \frac{d^{j-1}}{dt^{j-1}} f(t, y(t))\Big|_{t=t_i} \frac{h_i^j}{j!} + O(h_i^{p+1}) \\ &= y(t_i) + \sum_{j=0}^{p-1} \frac{d^j}{dt^j} f(t, y(t))\Big|_{t=t_i} \frac{h_i^{j+1}}{(j+1)!} + O(h_i^{p+1}) \end{aligned}$$

où on a utilisé le fait que y est solution de $y'(t) = f(t, y(t))$, donc en particulier $y^{(j)}(t) = d_t^j f(t, y(t))$. D'autre part, on a :

$$\Phi(t_i, y(t_i), h_i) = \sum_{j=0}^{p} \frac{\partial^j}{\partial h^j} \Phi(t_i, y(t_i), h)\Big|_{h=0} \frac{h_i^j}{j!} + O(h_i^{p+1})$$

En revenant à l'expression de l'erreur de consistance et si la relation (3.5) est satisfaite, on déduit :

$$\begin{aligned} \varepsilon_i &= \sum_{j=0}^{p-1} \frac{d^j}{dt^j} f(t, y(t))\Big|_{t=t_i} \frac{h_i^{j+1}}{(j+1)!} - \frac{\partial^j}{\partial h^j} \Phi(t_i, y(t_i), h)\Big|_{h=0} \frac{h_i^{j+1}}{j!} + O(h_i^{p+1}) \\ &= \sum_{j=1}^{p-1} \frac{h_i^{j+1}}{j!} \underbrace{\left(\frac{1}{j+1} \frac{d^j}{dt^j} f(t, y(t))\Big|_{t=t_i} - \frac{\partial^j}{\partial h^j} \Phi(t_i, y(t_i), h)\Big|_{h=0} \right)}_{=0} + O(h_i^{p+1}) \end{aligned} \tag{3.6}$$

Ainsi, en notant que

$$\varepsilon_i = O(h_i^{p+1}) = h_i O(h^p) \quad \Leftrightarrow \quad |\varepsilon_i| \leq K h_i h^p$$

et en sommant les erreurs de consistance, on obtient :

$$\sum_{i=0}^{N-1} |\varepsilon_i| \leq K h^p \sum_{i=0}^{N-1} h_i = K(T - t_0) h^p.$$

On montre ainsi que les relations (3.5) sont suffisantes pour avoir une méthode d'ordre p. Montrons maintenant qu'elles sont également nécessaire.

Pour cela, on supposera le pas $h_i = h$ constant. Considérons que la méthode est d'ordre p et qu'il existe $k < p$ t.q. :

$$\frac{1}{k+1}\frac{d^k}{dt^k}\underline{f}(t,\underline{y}(t)) \quad \neq \quad \frac{\partial^k}{\partial h^k}\underline{\Phi}(t,\underline{y}(t),h)\Big|_{h=0}$$

On supposera également que k est le plus petit indice pour lequel la relation (3.5) n'est pas satisfaite. Posons Ψ_k la fonction définie par :

$$\Psi_k(t) = \frac{1}{k+1}\frac{d^k}{dt^k}\underline{f}(t,\underline{y}(t)) - \frac{\partial^k}{\partial h^k}\underline{\Phi}(t,\underline{y}(t),h)\Big|_{h=0}$$

qui par hypothèse est non identiquement nulle. En reprenant la formule (3.6), on obtient cette fois :

$$\varepsilon_i \quad = \quad h^{k+1}\frac{\Psi_k(t_i)}{k!} + O(h^{k+2})$$

On déduit alors en sommant les erreurs de consistance :

$$\begin{aligned} \sum_{i=0}^{N-1}|\varepsilon_i| \quad &= \quad h^k\sum_{i=0}^{N-1} h\frac{|\Psi_k(t_i)|}{k!} + O(h^{k+1}) \\ \Leftrightarrow \quad \frac{1}{h^k}\sum_{i=0}^{N-1}|\varepsilon_i| \quad &= \quad \sum_{i=0}^{N-1} h\frac{|\Psi_k(t_i)|}{k!} + O(h) \end{aligned}$$

Or, la méthode étant supposé d'ordre $p > k$, le terme :

$$\frac{1}{h^k}\sum_{i=0}^{N-1}|\varepsilon_i| \quad \underset{h\to 0}{\longrightarrow} \quad 0,$$

et le terme :

$$\sum_{i=0}^{N-1} h\frac{|\Psi_k(t_i)|}{k!} + O(h) \quad \underset{h\to 0}{\longrightarrow} \quad \int_{t_0}^{T}\frac{|\Psi_k(t)|}{k!}\,dt > 0$$

ce qui conduit à une contradiction. □

Remarque 3.16. *Suivant la démarche de la preuve, on peut montrer que si la méthode est consistante et que f est au moins C^1, alors elle est au moins d'ordre 1.*

III.2 Les méthodes de Runge-Kutta

Dans la suite de cette section, on se placera dans le cas $n = 1$ pour simplifier la présentation des méthodes.

Le schéma d'Euler

Voyons maintenant quelques méthodes à 1−pas, et commençons avec la plus simple, la *méthode d'Euler*. Pour faire le lien avec la formule (3.3) et les méthodes de quadrature, on doit choisir la fonction d'incrément Φ de sorte que :

$$h_i \Phi(t_i, y_i, h_i) \simeq \int_{t_i}^{t_i+h_i} f(s, y(s))\, ds. \tag{3.7}$$

Par exemple, si on choisit la formule des rectangles "à gauche" :

$$\int_{t_i}^{t_i+h_i} f(s, y(s))\, ds \simeq h_i f(t_i, y(t_i)),$$

on déduit le schéma dit d'Euler explicite :

$$y_{i+1} = y_i + h_i f(t_i, y_i).$$

Ce schéma revient à prendre $\Phi(t, z, h) = f(t, z)$. Notons que ne connaissant pas la solution exacte y, on remplace $y(t_i)$ par y_i. Cette méthode est illustrée sur la figure 3.1.

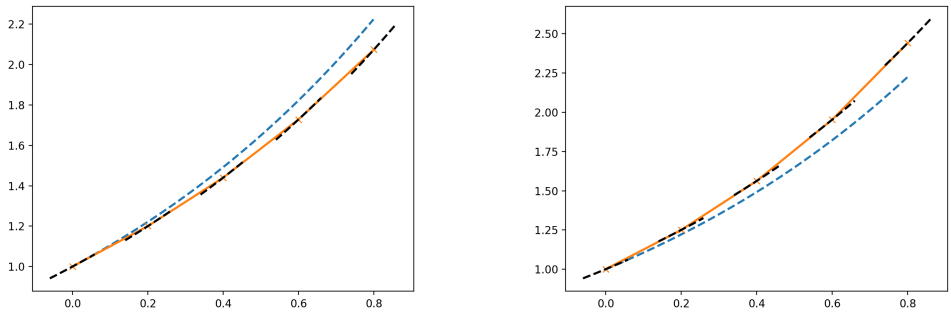

FIGURE 3.1 – Illustration des méthodes d'Euler explicite (à gauche) et Euler implicite (à droite) dans le cas $y'(t) = y(t)$. En pointillé bleue, on représente la solution exacte e^t. Les pointillés noires représentes les tangentes $f(t_i, y(t_i))$ aux points t_i.

Remarque 3.17. *De même, en prenant la formule des rectangles "à droite", on obtient la formule d'Euler dite implicite :*

$$y_{i+1} = y_i + h_i f(t_{i+1}, y_{i+1})$$

On parle de schéma implicite dans le sens qu'il faut résoudre une équation non linéaire pour déterminer y_{i+1}.

Analysons le schéma d'Euler à l'aide des outils présentés ci-dessus. Le schéma est :

— *consistant* : en effet, on a bien $\Phi(t, y(t), 0) = f(t, y(t))$. On déduit de plus que la méthode est d'ordre 1 au moins.

— *stable* : si la fonction f est Lipschitz (pour assurer existence et unicité de la solution), alors la fonction Φ le sera aussi.

Par conséquent, la méthode est *convergente*. Voyons si elle est d'ordre 2. On a d'une part :

$$\begin{aligned}\frac{d}{dt}f(t,y(t)) &= \left(\frac{\partial}{\partial t}f\right)(t,y(t)) + \left(\frac{\partial}{\partial y}f\right)(t,y(t))y'(t)\\ &= \left(\frac{\partial}{\partial t}f\right)(t,y(t)) + \left(\frac{\partial}{\partial y}f\right)(t,y(t))f(t,y(t))\end{aligned}$$

car $y'(t) = f(t, y(t))$, et d'autre part :

$$\frac{\partial}{\partial h}\Phi(t,y,h) = \frac{\partial}{\partial h}f(t,y(t)) = 0$$

Donc elle ne peut pas être d'ordre 2, et est donc exactement d'ordre 1.

Remarque 3.18. *Une façon de retrouver numériquement l'ordre d'une méthode consiste à calculer l'erreur*

$$\varepsilon(h) = \max_{0\le i\le N} |y(t_i) - y_i|,$$

et à tracer $\log(\varepsilon(h))$ *en fonction du* $\log(h)$. *En effet, si la méthode est d'ordre* p, *alors on doit retrouver une droite de coefficient directeur* p *car*

$$\varepsilon(h) \sim h^p \quad \Leftrightarrow \quad \log(\varepsilon(h)) \sim p\,\log(h).$$

Bien évidemment, pour réaliser le calcul de l'erreur, il faut avoir à disposition la solution exacte.

Construction de méthode d'ordre (plus) élevé

Pour construire des méthodes d'ordre plus élevé (et donc plus performantes), nous devons avoir une meilleure approximation de l'intégrale dans l'équation (3.7). Une idée naturelle est alors d'introduire r "pas intermédiaires" $t_{ij} = t_i + h_i\theta_j$, $j \in \{1, \cdots, r\}$, où on supposera $0 \le \theta_1 \le \cdots \le \theta_r \le 1$, pour construire la formule :

$$\int_{t_i}^{t_i+h_i} f(s, y(s))\,ds \simeq h_i \sum_{j=1}^{r} c_j f(t_{ij}, y(t_{ij})).$$

On retrouve ici une formule de quadrature où les t_{ij} sont les points de quadrature et les c_j les poids de quadrature. On en déduit le schéma :

$$y_{i+1} = y_i + h_i \sum_{j=1}^{r} c_j f(t_{ij}, y_{ij})), \tag{3.8}$$

où il faut déterminer les y_{ij} (qui sont des approximations de $y(t_{ij})$). Pour ce faire, on applique la même idée que pour calculer y_{i+1} :

$$y_{ij} = y_i + h_i \sum_{k=1}^{r} a_{jk} f(t_{ik}, y_{ik}), \tag{3.9}$$

où les coefficients a_{kj} sont des paramètres.

Remarque 3.19. *Cette formule est inspirée d'une formule de quadrature pour évaluer :*

$$y(t_{ij}) \;=\; y(t_i) + h_i\theta_j \int_{t_i}^{t_{ij}} f(s, y(s))\, ds.$$

Elle ne correspond pas exactement à une formule de quadrature dans le cas où $a_{ij} \neq 0$ pour $j > i$ puisqu'on évalue l'intégrande hors de l'intervalle $[t_i, t_{ij}]$. On notera par ailleurs, qu'a priori, pour chaque j on utilise une formule différente.

Définition 3.20. *Une méthode de Runge-Kutta est définie à l'aide des relations* (3.8) *et* (3.9)*, et des paramètres c_j, a_{jk} et θ_j pour $j \in \{1, \cdots, r\}$. Ces paramètres sont regroupés dans ce qu'on appelle le tableau de Butcher :*

$$\begin{array}{c|ccc} \theta_1 & a_{11} & \cdots & a_{1r} \\ \vdots & \vdots & \vdots & \vdots \\ \theta_r & a_{r1} & \cdots & a_{rr} \\ \hline & c_1 & \cdots & c_r \end{array}$$

Une façon plus compacte de formuler une méthode de Runge-Kutta est la suivante :

$$\begin{bmatrix} y_{i1} \\ \vdots \\ y_{ir} \\ y_{i+1} \end{bmatrix} = \begin{bmatrix} y_i \\ \vdots \\ y_i \\ y_i \end{bmatrix} + h_i \begin{bmatrix} a_{11} & \cdots & a_{1r} \\ \vdots & & \vdots \\ a_{r1} & \cdots & a_{rr} \\ c_1 & \cdots & c_r \end{bmatrix} \begin{bmatrix} f_{i1} \\ \vdots \\ f_{ir} \end{bmatrix} \quad \text{où} \quad f_{ij} = f(t_{ij}, y_{ij}) \tag{3.10}$$

Soulignons qu'il s'agit bien d'une méthode à 1−pas, même si la fonction Φ n'est pas évidente à expliciter. On se sert uniquement de y_i pour calculer y_{i+1} (via le calcul des y_{ij}).

Remarque 3.21. *Lorsque le choix des paramètres a_{jk} conduit à devoir résoudre un système non linéaire pour calculer les y_{ij}, on parlera de méthode de Runge-Kutta implicite.*

Remarque 3.22. *On notera que la méthode d'Euler explicite est bien une méthode de Runge Kutta définie par le tableau suivant :*

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

Consistance
Commeçons par détailler la fonction Φ :

$$\Phi(t,y,h) = \sum_{j=1}^{r} c_j f(t+\theta_j h, y_{ij}(t,y,h))$$

où

$$y_{ij}(t,y,h) = y + h\sum_{k=1}^{r} a_{jk} f(t+\theta_k h, y_{jk}(t,y,h)) \tag{3.11}$$

Soulignons ici aussi que l'équation ci-dessus définie $y_{ij}(t,y,h)$ de manière implicite.

Remarque 3.23. *On pourrait se demander si le système d'équations* (3.11) *admet bien une unique solution. Dans le cas où la méthode est explicite, c'est évident. Sinon, comme f est supposée Lipschitz, on peut prouver le résultat à l'aide du théorème de point fixe, pourvu que h soit suffisamment petit. Nous verrons dans un instant que cette condition revient à la stabilité de la méthode.*

D'après le théorème 3.7, nous avons vu qu'une méthode à 1−pas est consistante ssi $\Phi(t,y(t),0) = f(t,y(t))$. Bien que nous n'ayons pas dans le cas générale l'expression de Φ, il est facile de vérifier que $y_{ij}(t,y,0) = y$, et donc :

$$\Phi(t,y,h) = \sum_{j=1}^{r} c_j f(t,y(t)),$$

dont on déduit le résultat suivant :

Théorème 3.24. *Une méthode de Runge-Kutta est consistante ssi $\sum_{j=1}^{r} c_j = 1$.*

Stabilité
Nous avons vu avec le Théorème 3.9 qu'une condition suffisante pour que la méthode à 1−pas soit stable est que la fonction Φ soit Lipschitz. On en déduit le théorème suivant :

Théorème 3.25. *En notant A la matrice dont les coefficients sont $A_{ij} = a_{ij}$ et en considérant h^* t.q. $h^*\|A\|L < 1$, nous avons :*

1. *le système d'équations non linéaires* (3.10) *admet une unique solution (on peut donc calculer y_i pour tout i) pour tout $h \leq h^*$,*
2. *et la méthode est stable pour tout $h < h^*$.*

Démonstration. Pour commencer, rappelons que les y_{ij} doivent être solution du système non linéaire :

$$\begin{bmatrix} y_{i1} \\ \vdots \\ y_{ir} \end{bmatrix} = \begin{bmatrix} y_i \\ \vdots \\ y_i \end{bmatrix} + h_i \begin{bmatrix} a_{11} & \cdots & a_{1r} \\ \vdots & & \vdots \\ a_{r1} & \cdots & a_{rr} \end{bmatrix} \begin{bmatrix} f_{i1} \\ \vdots \\ f_{ir} \end{bmatrix} \quad \text{où} \quad f_{ij} = f(t_{ij}, y_{ij})$$

Si on sait les déterminer de manière unique, alors il est immédiat qu'on peut déduire y_{i+1}. Notons par $\underline{G} : \mathbb{R}^r \to \mathbb{R}^r$ l'application définie par :

$$\underline{G}(\underline{u}) = \begin{bmatrix} y_i \\ \vdots \\ y_i \end{bmatrix} + h_i \begin{bmatrix} a_{11} & \cdots & a_{1r} \\ \vdots & & \vdots \\ a_{r1} & \cdots & a_{rr} \end{bmatrix} \begin{bmatrix} f(t_{i1}, u_1) \\ \vdots \\ f(t_{i1}, u_r) \end{bmatrix}$$

Ainsi, les y_{ij} sont solutions de l'équation précédente ssi le vecteur $\underline{y}_i = [y_{i1}, \cdots, y_{ir}]$ est un point fixe de $\underline{G}$. Pour prouver l'existence et l'unicité du point fixe, nous allons prouver que l'application $\underline{G}$ est une contraction :

$$\begin{aligned} \|\underline{G}(\underline{u}) - \underline{G}(\underline{v})\| &\leq h_i \left\| \begin{bmatrix} a_{11} & \cdots & a_{1r} \\ \vdots & & \vdots \\ a_{r1} & \cdots & a_{rr} \end{bmatrix} \left(\begin{bmatrix} f(t_{i1}, u_1) \\ \vdots \\ f(t_{i1}, u_r) \end{bmatrix} - \begin{bmatrix} f(t_{i1}, v_1) \\ \vdots \\ f(t_{i1}, v_r) \end{bmatrix} \right) \right\| \\ &\leq h_i L \|A\| \|\underline{u} - \underline{v}\| \end{aligned} \tag{3.12}$$

Ainsi, si $h^*\|A\|L < 1$, on prouve que $\underline{G}$ est bien une contraction.

Pour prouver la deuxième partie du théorème, il nous reste à montrer que Φ est bien une application Lipschitz. On a :

$$\begin{aligned} |\Phi(t_i, y_i, h_i) - \Phi(t_i, z_i, h_i)| &\leq |y_i - z_i| + h_i \sum_{j=1}^{r} |c_j| |f(t_{ij}, y_{ij}) - f(t_{ij}, z_{ij})| \\ &\leq |y_i - z_i| + h_i L \sum_{j=1}^{r} |c_j| |y_{ij} - z_{ij}| \end{aligned}$$

où y_{ij} et z_{ij} sont des points fixe de $\underline{G}$ (associés à y_i et z_i respectivement). Étant point fixe de $\underline{G}$, on a en particulier, en notant $\underline{y}_i$ (resp. $\underline{z}_i$) le vecteur de composante y_{ij} (resp. z_{ij}) :

$$\begin{aligned}
& \underline{G}(\underline{y}_i) - \underline{G}(\underline{z}_i) = \underline{y}_i - \underline{z}_i \\
\Leftrightarrow \quad & (y_i - z_i)\underline{1} + h_i \begin{bmatrix} a_{11} & \cdots & a_{1r} \\ \vdots & & \vdots \\ a_{r1} & \cdots & a_{rr} \end{bmatrix} \left(\begin{bmatrix} f(t_{i1}, y_1) \\ \vdots \\ f(t_{i1}, y_r) \end{bmatrix} - \begin{bmatrix} f(t_{i1}, z_1) \\ \vdots \\ f(t_{i1}, z_r) \end{bmatrix} \right) = \underline{y}_i - \underline{z}_i \\
\Rightarrow \quad & \|\underline{y}_i - \underline{z}_i\| \leq |y_i - z_i| \|\underline{1}\| + h_i L \|A\| \|\underline{y}_i - \underline{z}_i\|
\end{aligned}$$

où $\underline{1}$ est le vecteur unitaire de $\mathbb{R}^r$. On peut alors déduire qu'il existe une constante $C > 0$ t.q. :

$$\|\underline{y}_i - \underline{z}_i\| \leq C|y_i - z_i|$$

Pour conclure, on notera que comme toutes les normes en dimension finie sont équivalentes, on a :

$$\begin{aligned}
|\Phi(t_i, y_i, h_i) - \Phi(t_i, z_i, h_i)| &\leq |y_i - z_i| + h_i L \max_{1 \leq j \leq r} |c_j| \|\underline{y}_i - \underline{z}_i\|_1 \\
&\leq |y_i - z_i| (1 + h_i L \max_{1 \leq j \leq r} |c_j| C^{ste})
\end{aligned}$$

ce qui prouve que Φ est bien Lipschitz. □

Remarque 3.26. *Soulignons que la condition $h^* L\|A\| < 1$ peut toujours être respectée. Elle contraint simplement à prendre un pas petit si $L\|A\|$ est grand. Par ailleurs, on peut montrer que pour tout $\varepsilon > 0$, il existe une norme matricielle t.q. $\|A\| \leq \rho(A) + \varepsilon$. On peut donc affiner la condition précédente en imposant $h^* L\rho(A) < 1$. La difficulté est alors de montrer l'inégalité* (3.12) *pour le choix de la norme assurant $h^* L\|A\| \leq h^* L\rho(A) + h^* L\varepsilon < 1$. La condition $h^* L\rho(A) < 1$ permet notamment de voir que pour une méthode où $a_{jk} = 0$ pour tout $k \geq j$ (méthode explicite), la méthode est systématiquement stable.*

Ordre

Pour clore notre analyse des méthodes de Runge-Kutta, étudions maintenant l'ordre de ces méthodes. Pour cela, on exploite la condition nécessaire et suffisante vue au Théorème 3.15. Nous savons déjà que pour que la méthode soit consistante (et d'ordre 1), on doit avoir :

$$\sum_{j=1}^{r} c_j = 1$$

Pour assurer l'ordre 2, on doit en plus de cette condition vérifier :

$$\frac{1}{2}\frac{d}{dt}f(t,y(t)) = \frac{\partial}{\partial h}\Phi(t,y(t),h)\Big|_{h=0}.$$

On a déjà vu que :

$$\frac{d}{dt}f(t,y(t)) = (\partial_t f)(t,y(t)) + (\partial_y f)(t,y(t))f(t,y(t)).$$

Voyons maintenant pour le deuxième terme :

$$\begin{aligned}\partial_h\Phi(t,y(t),h)\Big|_{h=0} &= \sum_{j=1}^{r} c_j f(t+\theta_j h, y_{ij}(t,y(t),h))\Big|_{h=0} \\ &= \sum_{j=1}^{r} c_j\left[\theta_j(\partial_t f)(t,y(t)) + (\partial_y f)(t,y(t))\partial_h y_{ij}(t,y(t),h)\Big|_{h=0}\right]\end{aligned}$$

où on rappelle que $y_{ij}(t,y,h)$ est définie par (3.11) et on a $y_{ij}(t,y,0) = y$. On a :

$$\partial_h y_{ij}(t,y,h) = \sum_{k=1}^{r} a_{jk} f(t+\theta_k h, y_{jk}(t,y,h)) + h\partial_h\left(\sum_{k=1}^{r} a_{jk} f(t+\theta_k h, y_{jk}(t,y,h))\right)$$

Pour calculer $\partial_h\Phi|_{h=0}$, il est inutile de poursuivre davantage le calcul ci-dessus puisqu'on cherche à l'évaluer en $h = 0$. En reprenant les deux dernières équations et en utilisant de nouveau que $y_{ij}(t,y,0) = y$, on déduit :

$$\begin{aligned}\partial_h\Phi(t,y(t),h)\Big|_{h=0} &= \sum_{j=1}^{r} c_j\theta_j(\partial_t f)(t,y(t)) + c_j(\partial_y f)(t,y(t))\sum_{j=1}^{r} a_{jk} f(t,y(t)) \\ &= (\partial_t f)(t,y(t))\underline{c}\cdot\underline{\theta} + (\partial_y f)(t,y(t))f(t,y(t))\underline{c}\cdot A\underline{1}\end{aligned}$$

où $\underline{c}$ est le vecteur des c_j, $\underline{\theta}$ est le vecteur des θ_j et $\underline{1}$ est le vecteur unitaire. On déduit de ce résultat les conditions suivantes pour avoir une méthode d'ordre 2 :

$$\left|\begin{array}{ll}\underline{c}\cdot\underline{1} = 1 & \text{(ordre 1)} \\ \underline{c}\cdot\underline{\theta} = \frac{1}{2} & \\ \underline{c}\cdot A\underline{1} = \frac{1}{2} & \end{array}\right.$$

En poursuivant cette démarche, on peut déterminer les conditions nécessaires et suffisantes pour avoir une méthode d'ordre p. Seulement, le nombre de conditions devient rapidement très grand (par exemple, pour $p = 5$, on a 58 conditions non linéaires). Afin de simplifier ces relations, on ajoute des conditions simplificatrices, ce qui même à des conditions suffisantes (et plus nécessaires). Par exemple, en imposant $A\underline{1} = \underline{\theta}$, on réduit le nombre de conditions à 17 pour $p = 5$. Ci-dessous, nous donnons un tableau de conditions

suffisantes pour assurer l'ordre p (avec la condition simplificatrice $A\underline{1} = \underline{\theta}$). Notons que pour que la méthode soit d'ordre p, il faut satisfaire les conditions $j = 1, \cdots, p$.

j=1	$\underline{c} \cdot \underline{1} = 1$
j=2	$\underline{c} \cdot \Theta\underline{1} = \frac{1}{2}$
j=3	$\underline{c} \cdot \Theta^2\underline{1} = \frac{1}{3}, \quad \underline{c} \cdot A\Theta\underline{1} = \frac{1}{6}$
j=4	$\underline{c} \cdot \Theta^3\underline{1} = \frac{1}{4}, \quad \underline{c} \cdot \Theta A\Theta\underline{1} = \frac{1}{8}$ $\underline{c} \cdot A\Theta^2\underline{1} = \frac{1}{12}, \quad \underline{c} \cdot A^2\Theta\underline{1} = \frac{1}{24}$

Dans ce tableau, on note par Θ la matrice diagonale t.q. $\Theta_{ii} = \theta_i$.

Quelques exemples de méthodes de Runge-Kutta

Méthode d'ordre 1

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

Cette méthode s'appelle le θ schéma, et elle est d'ordre 2 pour $\theta = \frac{1}{2}$. On notera qu'il s'agit d'une méthode implicite (sauf pour $\theta = 0$ où il s'agit de la méthode d'Euler).

Méthode d'ordre 2

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

Pour $\alpha = \frac{1}{2}$, la méthode s'appelle méthode d'Euler améliorée, et pour $\alpha = 1$ méthode de Heun.

Méthode d'ordre 4

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

III.3 Contrôle du pas de discrétisation

Estimation de l'erreur de consistance

Lorsqu'on résout une EDO (ou un système d'EDO), on peut souhaiter adapter le pas de discrétisation h_i en fonction de l'itération i. Typiquement, si on pouvait montrer que l'erreur est grande à l'instant t_i, on souhaiterait alors réduire le pas h_i. Dans l'autre sens, si on pouvait montrer que l'erreur est petite à l'instant t_i, pour réduire le coût de calculs, on pourrait vouloir agrandir le pas.

Le point est qu'on ne connait pas l'erreur à l'instant t_i. On va donc chercher à l'estimer à l'aide de deux méthodes à 1−pas d'ordre différents :

$$y_{i+1} = y_i + h_i\Phi(t_i, y_i, h_i) \quad \text{et} \quad y'_{i+1} = y'_i + h_i\Phi(t_i, y'_i, h_i).$$

Supposons qu'elles sont d'ordre p et p' respectivement avec $p < p'$. On a alors :

$$\varepsilon_i = y(t_{i+1}) - y(t_i) - h_i\Phi(t_i, y(t_i), h_i) = O(h_i^{p+1})$$

et

$$\varepsilon'_i = y(t_{i+1}) - y(t_i) - h_i\Phi'(t_i, y(t_i), h_i) = O(h_i^{p'+1})$$

Notre objectif va être d'avoir une estimation de ε_i à l'aide de ce qu'on connait, c'est à dire y_i et y'_i. On a :

$$\begin{aligned}\Phi(t_i, y(t_i), h_i) &= \Phi(t_i, y(t_i) - y_i + y_i, h_i)\\ &= \Phi(t_i, y_i, h_i) + (\partial_y\Phi)(t_i, y_i, h_i)(y(t_i) - y_i) + O((y(t_i) - y_i)^2).\end{aligned}$$

De même, on montre que :

$$\Phi'(t_i, y(t_i), h_i) = \Phi'(t_i, y_i, h_i) + (\partial_y\Phi')(t_i, y_i, h_i)(y(t_i) - y_i) + O((y(t_i) - y_i)^2).$$

On en déduit :

$$\begin{aligned}\varepsilon_i - \varepsilon'_i &= -h_i\left(\Phi(t_i, y(t_i), h_i) - \Phi'(t_i, y(t_i), h_i)\right)\\ &= -h_i\left(\Phi(t_i, y_i, h_i) - \Phi'(t_i, y_i, h_i)\right)\\ &\quad - h_i\left((\partial_y\Phi)(t_i, y_i, h_i) - (\partial_y\Phi')(t_i, y_i, h_i)\right)(y(t_i) - y_i) + O((y(t_i) - y_i)^2)\\ &= -h_i\left(\Phi(t_i, y_i, h_i) - \Phi'(t_i, y_i, h_i)\right)\\ &\quad - h_i\left((\partial_y\Phi)(t_i, y_i, h_i) - (\partial_y\Phi')(t_i, y_i, h_i)\right)(y(t_i) - y_i) + O(h_i^{2p})\end{aligned}$$

On rappelle que la première méthode étant convergente et d'ordre p, on a $|y(t_i) - y_i| = O(h_i^p)$, d'où la dernière égalité. Par ailleurs, en utilisant un deuxième D.L., cette fois par rapport à h, on a :

$$(\partial_y\Phi)(t_i, y_i, h_i) - (\partial_y\Phi')(t_i, y_i, h_i) = \underbrace{\partial_y(\Phi - \Phi')(t_i, y_i, 0)}_{=0} + \partial^2_{hy}\left(\Phi - \Phi'\right)h_i + O(h_i^2)$$

car, la méthode étant consistante, on a $\Phi(t,y,0) = \Phi'(t,y,0)$. En combinant ces résultats, déduit :

$$\begin{aligned} \varepsilon_i - \varepsilon'_i &= -h_i\left(\Phi(t_i,y_i,h_i) - \Phi'(t_i,y_i,h_i)\right) + O(h_i^{p+2}) \\ \Leftrightarrow\ \varepsilon_i &= \underbrace{y_i + h_i\Phi'(t_i,y_i,h_i)}_{=\widetilde{y}'_{i+1}} - y_{i+1} + \varepsilon'_i + O(h_i^{p+2}) \end{aligned}$$

Pour conclure, il faut rappeler que $p' > p$ et par conséquent $\varepsilon'_i = O(h_i^{p+2})$. Finalement, on a l'estimation de l'erreur de consistance :

$$\varepsilon_i = \widetilde{y}'_{i+1} - y_{i+1} + O(h_i^{p+2})$$

où on peut calculer $\widetilde{y}_{i+1}$ en appliquant le schéma d'ordre p' partant de y_i. L'idée est alors, à chaque itération i, de choisir le plus grand pas h_i t.q. l'approximation $|\widetilde{y}'_{i+1} - y_{i+1}|$ soit inférieure à une tolérance fixée. La méthode est illustrée sur la figure 3.2.

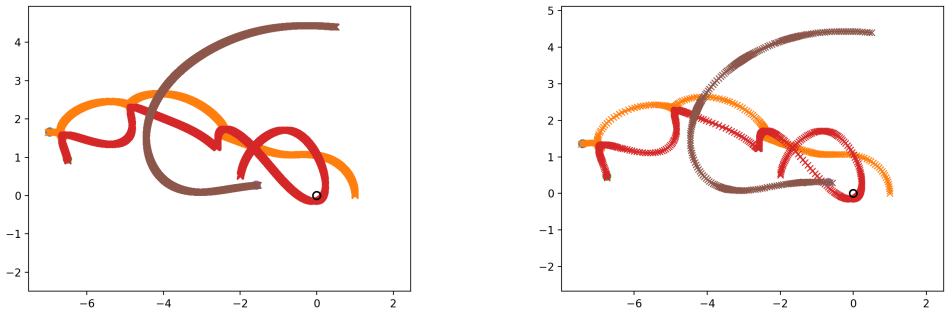

FIGURE 3.2 – Simulation du mouvement de 4 corps (un servant de référentiel) soumis aux lois de Newton avec une méthode de Runge-Kutta d'ordre 4 (à gauche) et une méthode à pas adaptatif $R24$ (à droite). On notera que dans le cas de la méthode à pas variable, le pas est réduit seulement lorsque deux astres sont proches.

Les méthodes de Runge-Kutta emboitée

Pour l'adaptation de pas, afin de ne pas rendre le coût de calculs à chaque itération prohibitif, on utilise généralement des méthodes dites de Runge-Kutta emboitée.

Définition 3.27. *Un couple de méthodes de Runge-Kutta respectivement d'ordre p et p' avec $p' > p$, et impliquant r et $r' = r + 1$ points intermédiaires,*

est dite emboitée ssi le tableau de la méthode d'ordre p' est donnée par :

$$\begin{array}{c|c|ccc|c} \theta_1 & a_{11} & \cdots & a_{1r} & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \theta_r & a_{r1} & \cdots & a_{rr} & 0 \\ \hline 1 & c_1 & \cdots & c_r & 0 \\ \hline\hline & c'_1 & \cdots & c'_r & c'_{r+1} \end{array}$$

où les coefficients a_{ij}, c_j et θ_j définissent la méthode d'ordre p.

L'intérêt d'une telle méthode est qu'elle réduit considérablement les calculs à effectuer pour évaluer $\widetilde{y}_{i+1}$ puisque :

$$\widetilde{y}_{i+1} = y_i + h_i \left(\sum_{j=1}^{r} c'_j f(t_{ij}, y_{ij}) + c_{r+1} f(t_{i+1}, y_{i+1}) \right)$$

où y_{ij} (et $f(t_{ij}, y_{ij})$) sont déjà calculées pour la méthode d'ordre p. Pour optimiser l'implémentation, on peut également noter que $f(t_{i+1}, y_{i+1})$ peut être conservé pour l'itération suivante, ce qui évite une évaluation de f dans le formule ci-dessus.

Quelques exemples de méthodes de Runge-Kutta emboitée :

Méthode de RK12 :

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

Méthode de RK24 :

$$\begin{array}{c|c|ccc|c} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ \hline 1 & 0 & 0 & 1 & 0 \\ \hline\hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array}$$

IV Méthodes multi-pas

Dans cette partie, on considèrera uniquement le cas d'un pas constant $h_i = h$. L'idée générale pour construire une méthode multi-pas est la suivante. On a :

$$y(t_{i+k}) = y(t_i) + \int_{t_i}^{t_{i+k}} f(s, y(s))\, ds$$

En utilisant une formule de quadrature sur l'intervalle $[t_i, t_{i+k}]$ exploitant les points $\{t_i, t_{i+1}, \cdots, t_{i+k}\}$ pour approcher l'intégrale, on déduit :

$$y(t_{i+k}) \simeq y(t_i) + h \sum_{j=0}^{k} \beta_j f(t_{i+j}, y(t_{i+j}))$$

où les β_j sont les poids de quadrature. On peut alors construire un schéma de la forme suivante :

$$y_{i+k} = y_i + h \sum_{j=0}^{k} \beta_j f(t_{i+j}, y_{i+k}) \tag{3.13}$$

En fait, on va généraliser en définissant les méthodes multi-pas ainsi :

Définition 3.28. *Une méthode multi-pas est de la forme par :*

$$\sum_{j=0}^{k} \alpha_j y_{i+j} = h \sum_{j=0}^{k} \beta_j f(t_{i+j}, y_{i+j})$$

où $\alpha_j \neq 0$.

Soulignons que si $k > 1$, il faut connaitre $y_0, \cdots, y_{k-1}$ pour initialiser la méthode multi-pas. On a alors recourt à une méthode à 1−pas pour les $k-1$ premiers itérés. Par ailleurs, on notera que lors du calcul de y_{i+k}, une seule évaluation de f est nécessaire en se servant des évaluations de $f(t_{i-1+j}, y_{i-1+j})$ précédentes. Cela permet de rendre les méthodes multi-pas très compétitives en terme de coût de calculs par rapport aux méthodes à 1−pas, pour le même ordre de précision. Néanmoins, on perd l'avantage de pouvoir adapter le pas de discrétisation.

Remarque 3.29. *La formule donnée dans la définition ci-dessus peut s'interpréter comme l'application de plusieurs formules de quadrature à :*

$$y(t_{i+j}) = y(t_i) + \int_{t_i}^{t_{i+j}} f(s, y(s))\, ds$$

dont on a fait la somme pondérée par α_j *pour* $j \in \{0, \cdots, r\}$.

Remarque 3.30. *Dans le cas où* $\beta_k \neq 0$*, on a une méthode implicite. On peut montrer, si* f *est Lipschitz, que l'on peut déterminer de manière unique* y_{i+k} *pour* h *suffisamment petit (comme nous l'avons vu avec les méthodes de Runge-Kutta implicite).*

IV.1 Analyse des méthodes multi-pas

Comme pour les méthodes à 1−pas, la question naturelle sera de savoir si le schéma numérique proposé est convergent. L'analyse des schémas suivra la même démarche avec quelques adaptations. Pour la suite de l'analyse, il sera utile d'introduire deux polynômes :

$$\alpha(t) = \sum_{j=0}^{k} \alpha_j t^j \quad \text{et} \quad \beta(t) = \sum_{j=0}^{k} \beta_j t^j.$$

Consistance

Définition 3.31. *On appelle erreur de consistance pour une méthode multi-pas :*

$$\varepsilon_i = \frac{1}{h} \sum_{j=0}^{k} \alpha_j y(t_{i+j}) - \sum_{j=0}^{k} \beta_j f(t_{i+j}, y(t_{i+j}))$$

et on dira que la méthode est consistante ssi :

$$\lim_{h \to 0} \max_{0 \le 0 \le N-k} |\varepsilon_i| = 0.$$

Théorème 3.32. *Une méthode multi-pas est consistante ssi* $\alpha(1) = 0$ *et* $\alpha'(1) = \beta(1)$.

La démonstration de ce résultat sera donnée avec l'étude de l'ordre ci-après. Notons tout de même que dans le cas du schéma (3.13), la première condition $\alpha(1) = 0$ est automatiquement vérifiée et la deuxième $\alpha'(1) = \beta(1) \Leftrightarrow k = \beta(1)$ revient à dire que la formule de quadrature est exacte sur $\mathbb{P}_0$ (et donc la formule composite convergente).

Stabilité

Définition 3.33. *Considérons la suite* z_i*, initialisée avec* $z_0, \cdots, z_{k-1}$ *et définie par :*

$$\sum_{j=0} \alpha_k z_{i+j} = h \left[\sum_{j=0}^{k} \beta_i f(t_{i+j}, z_{i+j}) + \widetilde{\varepsilon}_i \right]$$

On dira que la méthode est stable ssi il existe $M > 0$ *t.q. :*

$$\max_{0 \le i \le N} |y_i - z_i| \ \le \ M \left(\max_{0 \le i \le k-1} |y_i - z_i| + \max_{0 \le i \le N-k} |\widetilde{\varepsilon}_i| \right)$$

Théorème 3.34 (admis)**.** *Une condition nécessaire et suffisante pour qu'une méthode multi-pas soit stable est que les racines de* $\alpha(t)$ *soient dans le cercle unité, et celles de module* 1 *sont de multiplicité simple.*

Convergence et ordre

Définition 3.35. *Une méthode multi-pas est convergente ssi :*

$$\lim_{h\to 0} \max_{0\le i\le N} |y(t_i) - y_i| = 0$$

Théorème 3.36. *Une méthode multi-pas est convergente ssi elle est stable et consistante et :*

$$\lim_{h\to 0} \max_{0\le i\le k-1} |y(t_i) - y_i| = 0$$

Démonstration. Nous allons prouvons uniquement que si la méthode est stable et consistante, alors elle est convergente. La démarche est la même que pour les méthodes à 1−pas. On commence par noter que, par définition de l'erreur de consistance, on a :

$$\sum_{j=0}^{k} \alpha_j y(t_{i+j}) = h\left(\sum_{j=0}^{k} \beta_j f(t_{i+j}, y(t_{i+j})) + \varepsilon_i\right)$$

La méthode étant stable, on déduit :

$$\max_{0\le i\le N} |y_i - y(t_i)| \;\le\; M\left(\max_{0\le i\le k-1} |y_i - y(t_i)| + \max_{0\le i\le N-k} |\varepsilon_i|\right) \longrightarrow_{h\to 0} 0$$

□

Comme pour les méthodes à 1−pas, on s'intéressera à la vitesse de convergence à l'aide de la notion d'ordre.

Définition 3.37. *On dit que la méthode multi-pas est d'ordre p ssi :*

$$\max_{0\le i\le N-k} |\varepsilon_i| = O(h^p)$$

Dans le cas particulier où $\alpha(t) = t^k - 1$, ce qui correspond au schéma de la forme (3.13), une façon simple d'étudier l'ordre de la méthode revient à étudier l'ordre de la formule de quadrature :

$$\varepsilon_i = \int_{t_i}^{t_{i+k}} f(s, y(s))\, ds - h\sum_{j=0}^{k} \beta_j f(t_{i+j}, y(t_{i+j}))$$

Les points étant équirépartis, cela revient à une formule de Newton-Côtes qui, selon le nombre de points, est exacte sur $\mathcal{P}_k$ ou $\mathcal{P}_{k+1}$ (en choisissant bien sur convenablement les poids β_j). On en déduit alors que la méthode est

d'ordre k ou $k+1$.

Dans le cas général, on utilise une approche similaire aux méthodes à 1−pas, les développement limités. D'une part on a :

$$\begin{aligned} y(t_{i+j}) &= y(t_i + jh) \\ &= \sum_{l=0}^{p} y^{(l)}(t_i)\frac{(jh)^l}{l!} + O(h^{p+1}) \end{aligned}$$

et d'autre part :

$$\begin{aligned} f(t_{i+j}, y(t_{i+j})) &= y'(t_i + jh) \\ &= \sum_{l=0}^{p-1} y^{(l+1)}(t_i)\frac{(jh)^l}{l!} + O(h^p) \end{aligned}$$

car pour rappel y est solution de $y'(t) = f(t, y(t))$. En injectant ces expressions dans la définition de ε_i, on déduit :

$$\begin{aligned} \varepsilon_i &= \frac{1}{h}\sum_{j=0}^{r}\alpha_j \sum_{l=0}^{p} y^{(l)}(t_i)\frac{(jh)^l}{l!} - \sum_{j=0}^{r}\beta_j \sum_{l=0}^{p-1} y^{(l+1)}(t_i)\frac{(jh)^l}{l!} + O(h^p) \\ &= \sum_{l=0}^{p} C_l \frac{h^{l-1}}{l!} + O(h^p) \end{aligned}$$

où

$$\left| \begin{array}{l} C_0 = \sum_{j=0}^{r}\alpha_j \\ C_l = \sum_{j=0}^{r}\alpha_j \frac{j^l}{l!} - \beta_j \frac{j^{l-1}}{(l-1)!}, \quad \forall l \in \{1, \cdots, p\}. \end{array} \right.$$

On en déduit directement le théorème suivant :

Théorème 3.38. *Une méthode multi-pas est d'ordre p ssi $C_l = 0$ pour $l \in \{0, \cdots, p\}$.*

On remarquera qu'avoir la consistance revient à assurer $C_0 = 0$ et $C_1 = 0$, ce qui est bien équivalent à $\alpha(1) = 0$ et $\alpha'(1) = \beta(1)$.

Théorème 3.39. *Si une méthode multi-pas est stable et d'ordre p, et si*

$$\max_{0 \le i \le k-1} |y_i - y(t_i)| = O(h^p)$$

alors on a l'estimation d'erreur suivante :

$$\max_{0 \le i \le N} |y(t_i) - y_i| = O(h^p)$$

Soulignons une nouvelle fois que pour préserver l'ordre de convergence, on doit initialiser avec une méthode à 1−pas aussi précise que la méthode multi-pas.

Quelques exemples de méthodes multi-pas

Méthode de Adams-Bashforth 3 (ordre 4) :

$$\left|\begin{array}{l} \alpha_3 = 1,\ \ \alpha_2 = -1,\ \ \alpha_1 = 0,\ \ \alpha_0 = 0, \\ \beta_3 = \frac{9}{24},\ \ \beta_2 = \frac{19}{24},\ \ \beta_1 = \frac{-5}{24},\ \ \beta_0 = \frac{1}{24}. \end{array}\right.$$

Méthode de Nyström 3 (ordre 2) :

$$\left|\begin{array}{l} \alpha_2 = 1,\ \ \alpha_1 = 0,\ \ \alpha_0 = -1, \\ \beta_2 = 0,\ \ \beta_1 = 2,\ \ \beta_0 = 0. \end{array}\right.$$

IV.2 Méthodes de prédiction-correction

Dans le cas où $\beta_k \neq 0$, on a une méthode implicite. Pour déterminer y_{i+k}, il faut résoudre une équation non linéaire, par méthode de Newton par exemple. Se pose alors deux difficultés : comment initialiser l'algorithme (itératif) de Newton et à quelle itération arrêter l'algorithme. Afin de résoudre ces deux points, l'idée est de calculer une valeur approchée de y_{i+k} à l'aide d'un *prédicteur*, c'est à dire une méthode multi-pas explicite, et de d'utiliser cette valeur approchée dans le schéma implicite, ce qui conduit au schéma :

$$\left|\begin{array}{ll} \alpha'_k y'_{i+k} + \sum_{j=0}^{k-1} \alpha'_j y'_{i+j} = h \sum_{j=0}^{k-1} \beta'_j f(t_{i+j}, y'_{i+j}) & \text{(prédicteur)} \\ \sum_{j=0}^{k} \alpha_j y_{i+j} = h \left(\beta_k f(t_{i+k}, y'_{i+k}) + \sum_{j=0}^{k-1} \beta_j f(t_{i+j}, y_{i+j})\right) & \text{(correcteur)} \end{array}\right. \tag{3.14}$$

Supposons que le *prédicteur* est d'ordre q et la méthode implicite est d'ordre p. La question est alors de savoir si le schéma ci-dessus conduit bien à une méthode d'ordre p.

Théorème 3.40. *La méthode de prédiction / correction* (3.14) *est d'ordre p, c'est à dire :*

$$\max_{0\leq i\leq N} |y(t_i) - y_i| = O(h^p)$$

ssi le prédicteur est d'ordre $q \geq p-1$.

Démonstration. Considérons y_{i+k} donné par (3.14), et supposons que pour tous les itérés précédent y_{i+j} on a :

$$|y_{i+j} - y(t_{i+j})| = O(h^p), \quad \forall j \in \{0, \cdots, k-1\}$$

On prouvera le résultat par récurrence. Le prédicteur étant une méthode d'ordre q, on sait que :

$$|y'_{i+k} - y(t_{i+k})| = O(h^q)$$

Par ailleurs, la méthode implicite étant d'ordre p, on a :

$$\sum_{j=0}^{k} \alpha_j y(t_{i+j}) = h \sum_{j=0}^{k} \beta_j f(t_{i+j}, y(t_{i+j})) + O(h^{p+1})$$

dont on déduit :

$$\begin{aligned}\sum_{j=0}^{k} \alpha_j (y(t_{i+j}) - y_{i+j}) = h(&\sum_{j=0}^{k-1} \beta_j (f(t_{i+j}, y(t_{i+j})) - f(t_{i+j}, y_{i+j})) \\ &+ \beta_k (f(t_{i+k}, y(t_{i+k})) - f(t_{i+k}, y'_{i+k})) + O(h^{p+1})\end{aligned}$$

On déduit :

$$\begin{aligned}|\alpha_k||y_{i+k} - y(t_{i+k})| &\leq \sum_{j=0}^{k-1} |\alpha_j| \underbrace{|y(t_{i+j}) - y_{i+j}|}_{=O(h^p)} + h \sum_{j=0}^{k-1} |\beta_j||f(t_{i+j}, y(t_{i+j})) - f(t_{i+j}, y_{i+j})| \\ &\quad + h|\beta_k||f(t_{i+k}, y(t_{i+k})) - f(t_{i+k}, y_{i+k})| + O(h^{p+1}) \\ &\leq O(h^p) + hL \sum_{j=0}^{k-1} |\beta_j| \underbrace{|y(t_{i+j}) - y_{i+j}|}_{=O(h^p)} + h|\beta_k| \underbrace{|y(t_{i+k}) - y'_{i+k}|}_{=O(h^q)} \\ &\leq O(h^p) + LO(h^{p+1}) + O(h^{q+1}) = O(h^{q+1}) + O(h^p)\end{aligned}$$

On retrouve que pour avoir une majoration en $O(h^p)$, il faut avoir $q = p - 1$ au moins. □