CHAPITRE 1

COURS UM6P, ANALYSE APPLIQUEE

L'objectif de cette série de cours réalisés en janvier 2020 à l'UM6P est de présenter différentes méthodes de quadrature et de résolution approchée d'équations différentielles ordinaires.

1.1 Étude générale d'une méthode de quadrature

1.1.1 Définitions et notations

Une méthode de quadrature déterministe sur l'intervalle $]\alpha,\beta[$ ($(\alpha,\beta) \in \bar{\mathbf{R}}^2$) est caractérisée par la donnée d'un entier n et de deux familles, $(\lambda_i)_{0\leq i\leq n} \in \mathbf{R}^{n+1}$ et $(x_i)_{0\leq i\leq n} \in \mathbf{R}^{n+1}$. Elle consiste à approcher pour toute fonction f correctement définie, l'intégrale $\displaystyle\int_\alpha^\beta f(x)\omega(x)dx$ par la somme $\displaystyle\sum_{i=0}^{n}\lambda_i f(x_i)$. Dans l'intégrale précédente, $\omega \in C^0(]\alpha,\beta[,\mathbf{R}_+^*)$ représente un poids fixé dont tous les moments $x \mapsto x^k\omega(x)$ ($k \in \mathbf{N}$) sont intégrables. On utilise alors la notation

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

pour désigner la méthode de quadrature et on note $E(f)$ l'erreur commise :

$$E(f) = \int_\alpha^\beta f(x)\omega(x)dx - \sum_{i=0}^{n}\lambda_i f(x_i).$$

À toute méthode, on associe enfin un ordre :

Définition 1. *On dit qu'une méthode de quadrature est d'ordre $N \in \mathbf{N}$ si elle est exacte pour tous les polynômes de $\mathbf{P}_N$ (espace des polynômes de degré inférieur ou égal à N).*

Remarque : Il est préférable dans une méthode de quadrature d'avoir des coefficients λ_i tous positifs afin d'avoir une erreur dans le calcul approché de l'intégrale (ΔI) issue d'une erreur sur la fonction à intégrer (Δf) majorée par :

$$\Delta I \leq \sum_{i=0}^{n} |\lambda_i| \Delta f = \Delta f \sum_{i=0}^{n} \lambda_i = \Delta f \int_{\alpha}^{\beta} \omega(x) dx$$

en supposant la méthode au moins d'ordre 0 (à savoir exacte pour les constantes). Cette dernière valeur est en effet la plus petite pouvant être obtenue, étant égale à l'erreur maximale dans le calcul exact de l'intégrale sous les mêmes conditions.

1.1.2 Condition nécessaire et suffisante de convergence

Théorème 1. *On considère pour tout $n \in \mathbf{N}$ une méthode de quadrature définie par les familles $(\lambda_{i,n})_{0 \leq i \leq n} \in \mathbf{R}^{n+1}$ et $(x_{i,n})_{0 \leq i \leq n} \in \mathbf{R}^{n+1}$ (erreur associée : E_n) et telle que l'ensemble*

$$\bigcup_{n \in \mathbf{N}} \{x_{i,n},\ 0 \leq i \leq n\} \cup]\alpha, \beta[$$

soit contenu dans un intervalle borné I.

Une condition nécessaire et suffisante de convergence de la méthode pour toute fonction continue, à savoir

$$\forall f \in C^0(I, \mathbf{R}), \quad \lim_{n \to +\infty} E_n(f) = 0$$

est la suivante :

$$\begin{cases} (\alpha)\ \exists M \in \mathbf{R}_+^*, \quad \forall n \in \mathbf{N}, \quad \sum_{i=0}^{n} |\lambda_{i,n}| \leq M, \\ (\beta)\ \forall N \in \mathbf{N}, \quad \lim_{n \to +\infty} E_n(x \mapsto x^N) = 0. \end{cases}$$

Démonstration. : Condition suffisante : soit $f \in C^0(I, \mathbf{R})$ et $\varepsilon > 0$. Par le théorème de Stone Weierstrass (voir leçon A.III.), on sait que

$$\exists P \in \mathbf{R}[X], \quad ||f - P||_{L^\infty(I)} \leq \frac{\varepsilon}{2(\int_{\alpha}^{\beta} \omega(x) dx + M)}$$

avec M défini dans (α). L'hypothèse (β) appliquée par combinaison linéaire au polynôme P implique :

$$\exists n_0 \in \mathbf{N}, \quad \forall n \in \mathbf{N}, \quad n \geq n_0 \Rightarrow |E_n(P)| \leq \frac{\varepsilon}{2}.$$

On peut alors écrire pour tout $n \in \mathbf{N}$,

$$\begin{aligned} |E_n(f-P)| &\leq \int_\alpha^\beta |f(x)-P(x)|\omega(x)dx + \sum_{i=0}^{n} |\lambda_{i,n}||f(x_{i,n})-P(x_{i,n})| \\ &\leq \frac{\varepsilon}{2(\int_\alpha^\beta \omega(x)dx + M)}(\int_\alpha^\beta \omega(x)dx + M) = \frac{\varepsilon}{2} \end{aligned}$$

et pour $n \geq n_0$,

$$|E_n(f)| \leq |E_n(f-P)| + |E_n(P)| \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Condition nécessaire : la propriété (β) est une conséquence directe de la convergence de la méthode. Pour démontrer (α), on considère la famille des formes linéaires $(E_n)_{n\in\mathbf{N}}$

$$E_n : \begin{pmatrix} C^0(I,\mathbf{R}) \to \mathbf{R} \\ f \mapsto E_n(f) \end{pmatrix}.$$

Pour tout $n \in \mathbf{N}$, E_n est continue pour la norme infinie car

$$|||E_n||| = \sup_{f\in C^0(I,\mathbf{R})} \frac{|E_n(f)|}{||f||_\infty} \leq \int_\alpha^\beta \omega(x)dx + \sum_{i=0}^{n} |\lambda_{i,n}|.$$

En outre, comme la méthode est convergente

$$\forall f \in C^0(I,\mathbf{R}), \quad \exists C > 0, \quad \sup_{n\in\mathbf{N}} |E_n(f)| \leq C.$$

Les conditions sont donc réunies pour appliquer le théorème de Banach-Steinhaus (voir [Br]), à savoir :

$$\exists C > 0, \quad \sup_{n\in\mathbf{N}} |||E_n||| \leq C.$$

On considère alors une famille de fonctions $(f_n)_{n\in\mathbf{N}} \in C^0(I,\mathbf{R})^{\mathbf{N}}$ de norme infinie égale à 1 et telle que

$$\forall n \in \mathbf{N}, \quad \forall i \in \{0,...,n\}, \quad f_n(x_{i,n}) = \operatorname{sgn}(\lambda_{i,n}).$$

Pour tout entier n, on a

$$\begin{aligned}\sum_{i=0}^{n}|\lambda_{i,n}| = \sum_{i=0}^{n}\lambda_{i,n}f_n(x_{i,n}) = E_n(f_n) - \int_\alpha^\beta f_n(x)\omega(x)dx \\ \leq \sup_{n\in\mathbf{N}}|||E_n||| + \int_\alpha^\beta \omega(x)dx \\ \leq C + \int_\alpha^\beta \omega(x)dx = M \\ cr\end{aligned}$$

□

Le corollaire suivant donne une condition suffisante de convergence d'une méthode de quadrature pour toute fonction continue par morceaux sur I :

Corollaire 1. *Les hypothèses du Théorème 1 sont conservées. Une condition suffisante pour que*

$$\lim_{n\to+\infty} E_n(f) = 0$$

pour toute fonction réelle f continue par morceaux sur I est la suivante :

$$\begin{cases}(\alpha)'\ \exists n_0 \in \mathbf{N}, \quad \forall n \geq n_0, \quad \forall i \in \{0,...,n\}, \quad \lambda_{i,n} \geq 0, \\ (\beta)\ \forall N \in \mathbf{N}, \quad \lim_{n\to+\infty} E_n(x^N) = 0.\end{cases}$$

Démonstration. : On remarque que si la propriété (α)' est satisfaite, alors pour $n \geq n_0$

$$\sum_{i=0}^{n}|\lambda_{i,n}| = \int_\alpha^\beta \omega(x)dx - E_n(1)$$

et cette dernière valeur est bien bornée indépendamment de n d'après la propriété (β). La condition (α) du Théorème 1 est donc vérifiée. On en déduit ainsi dans un premier temps la convergence de la méthode de quadrature pour toute fonction continue sur I. Si f est seulement continue par morceaux, on peut démontrer facilement que

$$\forall \varepsilon > 0,\ \exists (s,S) \in C^0(I,\mathbf{R})^2, \quad s \leq f \leq S \quad \text{et} \quad \int_\alpha^\beta \big(S(x) - s(x)\big)\omega(x)dx \leq \varepsilon.$$

En particulier, grâce à (α)', on a

$$\forall n \in \mathbf{N}, \quad n \geq n_0 \Rightarrow \sum_{i=0}^{n}\lambda_{i,n}s(x_{i,n}) \leq \sum_{i=0}^{n}\lambda_{i,n}f(x_{i,n}) \leq \sum_{i=0}^{n}\lambda_{i,n}S(x_{i,n})$$

ce qui donne l'encadrement

$$E_n(S)+\int_\alpha^\beta \big(f(x)-S(x)\big)\omega(x)dx \le E_n(f) \le E_n(s)+\int_\alpha^\beta \big(f(x)-s(x)\big)\omega(x)dx$$

puis

$$E_n(S)-\varepsilon \le E_n(f) \le E_n(s)+\varepsilon.$$

En utilisant la convergence de la méthode pour les fonctions continues, on obtient finalement

$$\lim_{n\to+\infty} E_n(f)=0$$

□

1.1.3 Estimation de l'erreur

Théorème 2. *: on considère une méthode de quadrature sur un intervalle borné* $]\alpha,\beta[$*, à* $(n+1)$ *points tous situés dans un segment :*

$$]\alpha,\beta[\cup\{x_0,...,x_n\}\subset I=[a,b].$$

On suppose que la méthode est d'ordre $N\in\mathbf{N}$*. Alors, on a*

$$\forall f\in C^{(N+1)}([a,b],\mathbf{R}),\quad E(f)=\frac{1}{N!}\int_a^b K_N(t)f^{(N+1)}(t)dt \tag{1}$$

où on a noté pour tout $n\in\mathbf{N}$ *et* $t\in\mathbf{R}$ *:*

$$K_n(t)=E(x\mapsto(x-t)_+^n)$$

avec la convention usuelle $u_+=\max(u,0)$*.*

Démonstration. : On écrit la formule de Taylor intégrale à l'ordre $N+1$ pour la fonction f entre a et $x\in[a,b]$:

$$\begin{aligned} f(x) &= f(a)+(x-a)f'(a)+...+\frac{(x-a)^N}{N!}f^{(N)}(a)+\frac{1}{N!}\int_a^x \frac{(x-t)^N}{N!}f^{(N+1)}(t)dt \\ &= \underbrace{Q_N(x)}_{\in\mathbf{P}_N}+\int_a^b \frac{(x-t)_+^N}{N!}f^{(N+1)}(t)dt \end{aligned}$$

ce qui donne ensuite, la méthode étant d'ordre N, l'expression de l'erreur suivante :

$$\begin{aligned}
E(f) &= \frac{1}{N!}E\left(x \mapsto \int_a^b \frac{(x-t)_+^N}{N!} f^{(N+1)}(t)dt\right) \\
&= \frac{1}{N!}\Big[\int_\alpha^\beta \Big(\int_a^b (x-t)_+^N f^{(N+1)}(t)dt\Big)\omega(x)dx - \sum_{i=0}^n \lambda_i \int_a^b (x_i-t)_+^N f^{(N+1)}(t)dt\Big] \\
&= \frac{1}{N!}\int_a^b \left[\int_\alpha^\beta (x-t)_+^N \omega(x)dx - \sum_{i=0}^n \lambda_i (x_i-t)_+^N\right] f^{(N+1)}(t)dt \\
&= \frac{1}{N!}\int_a^b K_N(t) f^{(N+1)}(t)dt
\end{aligned}$$

□

Définition 2. *La famille $(K_n)_{n\in\mathbf{N}}$ s'appelle la famille des noyaux de Peano associée à la méthode considérée.*

Les deux corollaires suivant peuvent être immédiatement déduits :

Corollaire 2. *sous les mêmes hypothèses qu'au Théorème 3, on a*

$$\forall f \in C^{(N+1)}([a,b],\mathbf{R}), \quad |E(f)| \leq \frac{1}{N!}||f^{(N+1)}||_\infty \int_a^b |K_N(t)|dt.$$

Corollaire 3. *sous les mêmes hypothèses qu'au Théorème 3 et si de plus K_N garde un signe constant sur $[a,b]$, alors*

$$\exists \xi \in]a,b[, \quad E(f) = \frac{1}{(N+1)!} f^{(N+1)}(\xi) E(x \mapsto x^{N+1}).$$

Démonstration. : Le Corollaire 4 est immédiat. Pour démontrer le Corollaire 5, on utilise la deuxième formule de la moyenne à partir de la relation (1) :

$$\exists \xi \in]a,b[, \quad E(f) = \frac{1}{N!} f^{(N+1)}(\xi) \int_a^b K_N(t)dt.$$

On observe ensuite qu'en appliquant la formule (1) à la fonction $f(x) = x^{N+1}$, il vient

$$E(x \mapsto x^{N+1}) = \frac{1}{N!}\int_a^b (N+1)! K_N(t)dt = (N+1)\int_a^b K_N(t)dt$$

ce qui achève la démonstration □

1.2 Méthodes composées

1.2.1 Construction

La construction d'une méthode composée consiste d'abord à se donner une subdivision de l'intervalle d'intégration (supposé borné) :

$$\alpha = \alpha_0 < \alpha_1 < ... < \alpha_{k+1} = \beta$$

puis à remplacer sur chaque sous-intervalle $[\alpha_i, \alpha_{i+1}]$ la fonction à intégrer par son polynôme d'interpolation de Lagrange (voir leçon A.II.) associé à $(l+1)$ points ($l \in \mathbf{N}$) toujours répartis de manière identique.

Plus précisément, soient $(\tau_j)_{0\leq j\leq l}$, $(l+1)$ points distincts de $[-1,1]$ auxquels on associe après translation et homothétie les $(l+1)$ points de $[\alpha_i, \alpha_{i+1}]$:

$$\alpha_{i,j} = \frac{1}{2}(\alpha_i + \alpha_{i+1} + k_i\tau_j) \quad (0 \leq i \leq k, \quad 0 \leq j \leq l)$$

avec $k_i = \alpha_{i+1} - \alpha_i$. On remplace donc l'intégrale de f sur $[\alpha_i, \alpha_{i+1}]$,

$$\int_{\alpha_i}^{\alpha_{i+1}} f(x)dx = \frac{k_i}{2}\int_{-1}^{1} f\big(\frac{1}{2}(\alpha_i + \alpha_{i+1} + k_i y)\big)dy$$

par $\dfrac{k_i}{2}\displaystyle\int_{-1}^{1} P_{i,l}(y)dy$ où $P_{i,l}$ représente le polynôme d'interpolation de Lagrange de la fonction $f_i : \begin{pmatrix} [-1,1] \to \mathbf{R} \\ y \mapsto f(\frac{1}{2}(\alpha_i + \alpha_{i+1} + k_i y)) \end{pmatrix}$ associé aux points $(\tau_j)_{0\leq j\leq l}$. On a

$$P_{i,l}(y) = \sum_{j=0}^{l} f_i(\tau_j)l_j(y) = \sum_{j=0}^{l} f(\frac{1}{2}(\alpha_i + \alpha_{i+1} + k_i\tau_{i,j}))l_j(y) = \sum_{j=0}^{l} f(\alpha_{i,j})l_j(y)$$

où l_j désigne le j-ième polynôme de base de Lagrange associé aux mêmes points.

En notant

$$\omega_j = \frac{1}{2}\int_{-1}^{1} l_j(y)dy, \tag{2}$$

on a donc construit la méthode de quadrature suivante : Définition : On appelle méthode composée associée aux deux indices $(k,l) \in \mathbf{N}^* \times \mathbf{N}$ et à la famille $(\tau_i)_{0\leq i\leq l} \in [0,1]^{l+1}$, la méthode de quadrature définie par :

$$\int_{\alpha}^{\beta} f(x)dx \simeq \sum_{i=0}^{k} k_i \sum_{j=0}^{l} \omega_j f(\alpha_{i,j}) \tag{3}$$

où les coefficients k_i, ω_j et $\alpha_{i,j}$ ont été définis au cours de la construction.

On peut également définir la méthode élémentaire associée (correspondant à $k=0$ et $[\alpha,\beta]=[-1,1]$) :

$$\int_{-1}^{1} f(x)dx \simeq 2\sum_{j=0}^{l} \omega_j f(\tau_j).$$

1.2.2 Exemples de méthodes composées

(i) $l=0, \tau_0=-1$: on retrouve la méthode bien connue des rectangles à gauche :

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

(ii) $l=0, \tau_0=0$: on retrouve la méthode du point milieu :

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

(iii) $l=0, \tau_0=1$: on retrouve la méthode des rectangles à droite :

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

(iv) $l=1, \tau_0=-1, \tau_1=1$: dans ce cas,

$$\begin{cases} l_0(y)=\dfrac{1-y}{2} \Rightarrow \omega_0=\dfrac{1}{2}, \\ l_1(y)=\dfrac{y+1}{2} \Rightarrow \omega_1=\dfrac{1}{2} \end{cases}$$

et on retrouve la méthode des trapèzes :

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

1.2.3 Méthodes de Newton Cotes fermées

On peut construire pour tout $l \in \mathbf{N}^*$ une méthode composée en prenant $(l+1)$ points d'interpolation équirépartis sur $[-1,1]$:

$$\tau_j = -1 + \frac{2j}{l} \quad (0 \leq j \leq l).$$

On parle alors de la famille des méthodes de Newton Cotes fermées. Les cas particuliers importants sont les suivants : (i) $l = 1$: on retrouve la méthode des trapèzes. (ii) $l = 2$: dans ce cas on montre que $\omega_0 = \omega_2 = \frac{1}{6}$ et $\omega_1 = \frac{2}{3}$ et on construit la méthode de Simpson :

$$\int_\alpha^\beta f(x)dx \simeq \sum_{i=0}^{k} \frac{k_i}{6}\Big[f(\alpha_i) + 4f(\frac{\alpha_i + \alpha_{i+1}}{2}) + f(\alpha_{i+1})\Big].$$

(iii) $l = 4$: dans ce cas, $\omega_0 = \omega_4 = \frac{7}{90}$, $\omega_1 = \omega_3 = \frac{16}{45}$ et $\omega_2 = \frac{2}{15}$ et on construit la méthode de Boolle-Villarceau.

Remarque : On utilise également les méthodes de Newton Cotes pour $l = 6$ (Weddle-Hardy). Pour des valeurs supérieures de l, ces méthodes ne sont pas utilisées car elles font apparaitre des valeurs négatives de ω_i pouvant engendrer des problèmes numériques d'amplification d'erreurs d'arrondis (voir remarque au début du chapitre).

1.2.4 Ordre des méthodes composées

Théorème 3. *Les méthodes composées définies au paragraphe précédent sont d'ordre l. Ce résultat général peut être amélioré dans le cas de la méthode du point milieu (ordre 1) et des méthodes de Newton Cotes fermées lorsque l est pair (ordre $l+1$).*

Démonstration. : La première affirmation est une conséquence directe de la construction : si $f \in \mathbf{P}_l$ alors, en conservant les notations du paragraphe précédent,

$$\forall i \in \{0,...,l\}, \quad \forall y \in [-1,1], \quad f(\frac{1}{2}(\alpha_i + \alpha_{i+1} + k_i y)) = P_{i,l}(y)$$

et la formule (3) est exacte. Comme l'ordre de la méthode du point milieu est clairement égal à 1, il reste seulement à améliorer la valeur de l'ordre des méthodes

de Newton Cotes lorsque l est pair : on remarque que dans ce cas la méthode élémentaire

$$\int_{-1}^{1} f(y)dy \simeq 2\sum_{j=0}^{l} \omega_j f(\tau_j) \tag{4}$$

est non seulement exacte pour toute fonction f dans $\mathbf{P}_l$ mais aussi pour toute fonction impaire. En effet, on a alors

$$\forall j \in \{0,...,l\}, \quad \tau_{l-j} = -\tau_j$$

ce qui implique

$$\forall y \in [-1,1], \quad l_j(-y) = l_{l-j}(y)$$

puis $\omega_j = \omega_{l-j}$ et enfin

$$\sum_{j=0}^{l} f(\tau_j)\omega_j = \sum_{j=0}^{\frac{l}{2}} f(\tau_j)\omega_j + \sum_{j=0}^{\frac{l}{2}} f(-\tau_j)\omega_j = 0 = \int_{-1}^{1} f(y)dy.$$

La fonction $f(y) = y^{l+1}$ étant impaire lorsque l est pair, la méthode élémentaire est exacte pour cette fonction. Il en va de même par combinaison linéaire pour l'ensemble des fonctions dans $\mathbf{P}_{l+1}$. La méthode est donc bien d'ordre $l+1$ □

Remarque : La méthode de Simpson est en particulier d'ordre 3 et la méthode de Boole-Villarceau d'ordre 5.

1.2.5 Convergence des méthodes composées

Avant d'établir un théorème de convergence pour les méthodes composées, on démontre un résultat intermédiaire :

Proposition 1. *Les coefficients $(\omega_j)_{0\leq j\leq l}$ définis par la formule (2) vérifient :*

$$\sum_{j=0}^{l} \omega_j = 1. \tag{5}$$

Démonstration. : Il suffit de remaquer que

$$\sum_{j=0}^{l} \omega_j = \frac{1}{2}\sum_{j=0}^{l} \int_{-1}^{1} l_j(y)dy = \frac{1}{2}\int_{-1}^{1} 1.dy = 1$$

□

On peut alors démontrer le résultat de convergence suivant :

Théorème 4. *Soit $l > 0$ fixé. En notant $\delta_k = \max_{0\leq i\leq k} k_i$, on a pour toute fonction f intégrable au sens de Rieman sur $[\alpha, \beta]$:*

$$\lim_{\substack{k\to+\infty \\ \delta_k\to 0}} \sum_{i=0}^{k} k_i \sum_{j=0}^{l} \omega_j f(\alpha_{i,j}) = \int_{\alpha}^{\beta} f(x)dx.$$

Démonstration. : On écrit

$$\sum_{i=0}^{k} k_i \sum_{j=0}^{l} \omega_j f(\alpha_{i,j}) = \sum_{j=0}^{l} \omega_j I_{j,k}(f)$$

où $I_{j,k}(f) = \sum_{i=0}^{k} k_i f(\alpha_{i,j})$. Par définition de l'intégrale de Rieman, on a pour tout $j \in \{0, ..., l\}$

$$\lim_{\substack{k\to+\infty \\ \delta_k\to 0}} I_{j,k}(f) = \int_{\alpha}^{\beta} f(x)dx$$

et il suffit alors d'utiliser la Proposition 7 pour conclure la démonstration □

Remarque : La nécessité d'utiliser une formule composée vient du fait qu'il est impossible d'obtenir un résultat de convergence similaire pour la méthode élémentaire lorsque l tend vers l'infini (puisqu'il n'existe aucun résultat de convergence du polynôme d'interpolation vers la fonction interpolée lorsque le nombre de points tend vers l'infini).

1.2.6 Estimation de l'erreur

1.2.6.0.1 Cas général On définit tout d'abord le noyau de Peano (voir précédemment) correspondant à la méthode de quadrature élémentaire :

$$\mathcal{K}_n(t) = \int_{-1}^{1} (s-t)_+^n ds - 2\sum_{j=0}^{l} \omega_j (\tau_j - t)_+^n.$$

On peut alors exprimer le noyau de Peano K_n de la méthode composée à l'aide de $\mathcal{K}_n$:

Lemme 1. *Le noyau de Peano K_n d'une méthode composée peut s'écrire de la manière suivante :*

$$\forall n \in \mathbf{N}, \quad \forall t \in \mathbf{R}, \quad K_n(t) = \sum_{i=0}^{k} \left(\frac{k_i}{2}\right)^{n+1} \mathscr{K}_n\Big(\frac{2t-\alpha_i-\alpha_{i+1}}{k_i}\Big).$$

Démonstration. : On remarque d'une part que pour tout $i \in \{0, ..., k\}$

$$\begin{aligned}\int_{\alpha_i}^{\alpha_{i+1}} (x-t)_+^n dx &= \frac{k_i}{2}\int_{-1}^{1} \Big(\frac{1}{2}(\alpha_i+\alpha_{i+1}+k_i s)-t\Big)_+^n ds \\ &= \left(\frac{k_i}{2}\right)^{n+1} \int_{-1}^{1} \Big(s - \frac{2t-\alpha_i-\alpha_{i+1}}{k_i}\Big)_+^n ds\end{aligned}$$

et d'autre part que

$$k_i \sum_{j=0}^{l} \omega_j (\alpha_{i,j}-t)_+^n = 2\left(\frac{k_i}{2}\right)^{n+1} \sum_{j=0}^{l} \omega_j \Big(\tau_j - \frac{2t-\alpha_i-\alpha_{i+1}}{k_i}\Big)_+^n$$

ce qui donne bien par soustraction le résultat annoncé □

On peut à présent énoncer le théorème suivant d'estimation de l'erreur $E(f)$ d'une méthode composée :

Théorème 5. *: on suppose que la méthode élémentaire issue d'une méthode composée est d'ordre $N \in \mathbf{N}$. Alors,*

$$\forall f \in C^{(N+1)}([\alpha,\beta],\mathbf{R}),\ |E(f)| \le C_N(\beta-\alpha)\delta_k^{N+1}||f^{(N+1)}||_\infty$$

où on a noté

$$\begin{cases} C_N = \dfrac{1}{2^{N+2}N!}\displaystyle\int_{-1}^{1} |\mathscr{K}_{\mathscr{N}}(t)|dt, \\ \delta_k = \max\limits_{0\le i\le k} k_i. \end{cases}$$

Démonstration. : On commence par utiliser les résultats du Corollaire 4 et du Lemme 9 :

$$\begin{aligned}|E(f)| &\le \frac{1}{N!}||f^{(N+1)}||_\infty \int_\alpha^\beta |K_N(t)|dt \\ &\le \frac{1}{N!}||f^{(N+1)}||_\infty \sum_{i=0}^{k} \left(\frac{k_i}{2}\right)^{N+1} \int_\alpha^\beta \left|\mathscr{K}_N\Big(\frac{2t-\alpha_i-\alpha_{i+1}}{k_i}\Big)\right| dt.\end{aligned}$$

On montre ensuite que $\mathscr{K}_N(t) = 0$ si $|t| > 1$: en effet, si $t > 1$

$$\forall s \in [-1,1], \quad (s-t)_+^N = 0$$

tandis que si $t < -1$

$$\mathscr{K}_N(t) = E\big(x \mapsto (x-t)_+^N\big) = E\big(\underbrace{x \mapsto (x-t)^N}_{\in \mathbf{P}_N}\big) = 0$$

car la méthode élémentaire est d'ordre N. Cette remarque permet d'améliorer la majoration de $|E(f)|$ obtenue :

$$|E(f)| \leq \frac{1}{N!} ||f^{(N+1)}||_\infty \sum_{i=0}^{k} \left(\frac{k_i}{2}\right)^{N+2} \int_{-1}^{1} |\mathscr{K}_N(s)| ds$$

et il suffit alors pour conclure de remarquer que

$$\sum_{i=0}^{k} \left(\frac{k_i}{2}\right)^{N+2} \leq \left(\frac{\delta_k}{2}\right)^{N+1} \frac{\beta - \alpha}{2} \quad \square$$

Remarque : Ce résultat justifie la définition adoptée pour l'ordre d'une méthode en reliant sa valeur avec la précision de celle-ci pour des fonctions régulières.

Corollaire 4. *Sous les mêmes conditions qu'au Théorème 10 et si $\mathscr{K}_N$ garde un signe constant sur $[-1,1]$, alors*

$$\exists \xi \in]\alpha, \beta[, \quad E(f) = 2^{N+2} C_N f^{(N+1)}(\xi) \sum_{i=0}^{k} \left(\frac{k_i}{2}\right)^{N+2}.$$

Démonstration. : On utilise comme dans la preuve du Corollaire 5 la deuxième formule de la moyenne à partir de la relation

$$E(f) = \frac{1}{N!} \sum_{i=0}^{k} \left(\frac{k_i}{2}\right)^{N+2} \int_{-1}^{1} \mathscr{K}_{\mathscr{N}}(t) f^{(N+1)}\left(\frac{\alpha_i + \alpha_{i+1} + k_i t}{2}\right) dt \quad \square$$

1.2.6.0.2 Exemples (i) Dans le cas de la méthode du point milieu (d'ordre 1), on montre que

$$\mathscr{K}_1(t) = \begin{cases} \dfrac{(1+t)^2}{2} & \text{si} \quad -1 \le t \le 0, \\ \dfrac{(1-t)^2}{2} & \text{si} \quad 0 \le t \le 1. \end{cases}$$

$\mathscr{K}_1$ est de signe constant et par conséquent, pour toute fonction $f \in C^2([\alpha,\beta],\mathbf{R})$,

$$\exists \xi \in]\alpha,\beta[, \quad E(f) = \frac{1}{3} f''(\xi) \sum_{i=0}^{k} \left(\frac{k_i}{2}\right)^3.$$

(ii) Dans le cas de la méthode des trapèzes (d'ordre 1), on montre que

$$\forall t \in [-1,1], \quad \mathscr{K}_1(t) = -\frac{1}{2}(1-t)^2 \le 0.$$

De même, pour toute fonction $f \in C^2([\alpha,\beta],\mathbf{R})$ on a

$$\exists \xi \in]\alpha,\beta[, \quad E(f) = -\frac{2}{3} f''(\xi) \sum_{i=0}^{k} \left(\frac{k_i}{2}\right)^3.$$

(iii) Dans le cas de la méthode de Simpson (d'ordre 3), on montre que pour toute fonction $f \in C^4([\alpha,\beta],\mathbf{R})$,

$$\exists \xi \in]\alpha,\beta[, \quad E(f) = -\frac{1}{90} f^{(4)}(\xi) \sum_{i=0}^{k} \left(\frac{k_i}{2}\right)^5$$

(voir Exercice 1 pour une démonstration directe).

Remarque : Plus généralement, on peut montrer que tous les noyaux de Peano élémentaires $\mathscr{K}_n$ associés aux méthodes de Newton-Cotes fermées gardent un signe constant.

1.3 Méthodes de Gauss

1.3.1 Construction

La construction des méthodes de Gauss consiste à fixer un certain nombre de points à l'extérieur de l'intervalle d'intégration $]\alpha,\beta[$ puis à compléter cette famille par des points intérieurs choisis de manière à optimiser l'ordre final de la méthode.

Théorème 6. *: soient $(l,n) \in \mathbf{N}^{*2}$, $l \leq n+1$ et $(x_i)_{l\leq i\leq n}$ une famille (éventuellement vide lorsque $l = n+1$) de points distincts extérieurs à $]\alpha,\beta[$. Il existe un unique choix (à une permutation près) de points $(x_i)_{0\leq i\leq l-1}$, distincts et intérieurs à $]\alpha,\beta[$ et une unique famille de scalaires $(\lambda_i)_{0\leq i\leq n}$ telle que la méthode de quadrature*

$$\int_\alpha^\beta f(x)\omega(x)dx \simeq \sum_{i=0}^{n} \lambda_i f(x_i) \tag{6}$$

soit d'ordre $n+l$.

La famille de points intérieurs correspond aux racines du $(l+1)$-ième polynôme orthogonal sur $]\alpha,\beta[$ (voir leçon A.III.) associé au poids

$$\Theta(x) = \omega(x)\prod_{i=l}^{n}(x-x_i). \tag{7}$$

Démonstration. : Unicité : soit une méthode d'ordre $n+l$ à $(n+1)$ points (l points intérieurs et $(n-l+1)$ points extérieurs). On note

$$P_l(x) = \prod_{i=0}^{l-1}(x-x_i)$$

et on montre que P_l est le $(l+1)$-ième polynôme orthogonal sur $]\alpha,\beta[$ associé au poids Θ défini par (7) (Θ est bien une fonction continue, de signe constant sur $]\alpha,\beta[$ et dont tous les moments sont intégrables). Pour cela, on calcule pour tout $Q \in \mathbf{P}_{l-1}$

$$\begin{aligned}\int_\alpha^\beta P_l(x)Q(x)\Theta(x)dx &= \int_\alpha^\beta \overbrace{\prod_{i=0}^{n}(x-x_i)Q(x)}^{\in \mathbf{P}_{n+l}}\omega(x)dx \\ &= \sum_{j=0}^{n}\lambda_j\prod_{i=0}^{n}(x_j-x_i)Q(x_j) = 0\end{aligned}$$

car la méthode est d'ordre $n+l$. Ainsi est démontrée l'unicité (à une permutation près) des points $(x_i)_{0\leq i\leq l-1}$. Il reste à prouver l'unicité de la famille des scalaires $(\lambda_i)_{0\leq i\leq n}$: pour cela, on écrit la formule de quadrature pour chaque polynôme de base de Lagrange $(l_i)_{0\leq i\leq n}$ associé aux points $(x_i)_{0\leq i\leq n}$:

$$l_i(x) = \prod_{\substack{j=0\\ j\neq i}}^{n}\frac{x-x_j}{x_i-x_j}. \tag{8}$$

Comme $l_i \in \mathbf{P}_n \subset \mathbf{P}_{n+l}$, la méthode de quadrature est exacte pour l_i :

$$\forall i \in \{1, ..., n\}, \quad \int_\alpha^\beta l_i(x)\omega(x)dx = \sum_{j=0}^{n} \lambda_j l_i(x_j) = \lambda_i \tag{9}$$

et la famille des scalaires est bien unique. Existence : on choisit pour famille de points intérieurs $(x_i)_{0\leq i\leq l-1}$ la famille des l racines distinctes (voir leçon A.III.) du $(l+1)$-ième polynôme orthogonal (noté P_l) associé au poids Θ sur $]\alpha, \beta[$ et pour famille de scalaires $(\lambda_i)_{0\leq i\leq n}$, la famille donnée par la relation (9).

On montre tout d'abord que la méthode ainsi définie est d'ordre n : pour cela, on note $P_n(f)$ le polynôme de Lagrange de f associé aux points $(x_i)_{0\leq i\leq n}$. Alors

$$\sum_{i=0}^{n} \lambda_i f(x_i) = \int_\alpha^\beta \sum_{i=0}^{n} l_i(x)\omega(x)f(x_i)dx = \int_\alpha^\beta P_n(f)(x)\omega(x)dx$$

et la formule est bien exacte si $f \in \mathbf{P}_n$ (car alors $f \equiv P_n(f)$).

Soit ensuite $P \in \mathbf{P}_{l+n}$: grâce à la division euclidienne dans $\mathbf{R}[X]$, on sait qu'il existe $Q \in \mathbf{P}_{l-1}$ et $R \in \mathbf{P}_n$ tels que

$$P(X) = Q(X)\prod_{i=0}^{n}(X - x_i) + R(X).$$

En particulier,

$$\begin{aligned}
\int_\alpha^\beta P(x)\omega(x)dx &= \int_\alpha^\beta Q(x)\prod_{i=0}^{n}(x - x_i)\omega(x)dx + \int_\alpha^\beta R(x)\omega(x)dx \\
&= \int_\alpha^\beta Q(x)\prod_{i=0}^{l-1}(x - x_i)\Theta(x)dx + \int_\alpha^\beta R(x)\omega(x)dx \\
&= \int_\alpha^\beta \overbrace{Q(x)}^{\in \mathbf{P}_{l-1}} P_l(x)\Theta(x)dx + \sum_{i=0}^{n} \lambda_i R(x_i) \\
&= 0 + \sum_{i=0}^{n} \lambda_i R(x_i) \\
&= \sum_{i=0}^{n} \lambda_i P(x_i)
\end{aligned}$$

et la formule de quadrature est bien d'ordre $n + l$ (on a utilisé dans la troisième égalité le fait qu'elle était déjà d'ordre n) □

Définition 3. *La méthode construite dans le théorème précédent s'appelle la méthode de Gauss sur* $]\alpha,\beta[$ *associée aux points extérieurs* $(x_i)_{l\leq i\leq n}$ *et au poids* ω.

Remarque : Cette méthode est exactement d'ordre $n+l$. En effet, en considérant le polynôme $\Pi \in \mathbf{P}_{l+n+1}$ tel que

$$\Pi(x) = \prod_{i=0}^{l-1}(x-x_i)^2 \prod_{i=l}^{n}(x-x_i),$$

on remarque d'une part que

$$\sum_{i=0}^{n} \lambda_i \Pi(x_i) = 0$$

et d'autre part que

$$\int_{\alpha}^{\beta} \Pi(x)\omega(x)dx \neq 0.$$

car Π est non nul et garde un signe constant sur $]\alpha,\beta[$. Les deux corollaires suivants

précisent certaines caractéristiques des méthodes de Gauss pour deux familles particulières de points extérieurs :

Corollaire 5. *soit* $n \in \mathbf{N}$. *Il existe une et une seule méthode de quadrature (avec un poids* ω *fixé) à* $(n+1)$ *points dans* $]\alpha,\beta[$ *et qui soit d'ordre* $2n+1$ *: les points* $(x_i)_{0\leq i\leq n}$ *correspondent aux racines du* $(n+2)$*-ième polynôme orthogonal sur* $]\alpha,\beta[$ *pour le poids* ω. *De plus, les scalaires* $(\lambda_i)_{0\leq i\leq n}$ *définis par (9) sont strictement positifs.*

Démonstration. : Il suffit d'utiliser le Théorème 12 avec $l = n+1$ (ce qui correspond à une famille vide de points extérieurs). En outre, en appliquant la formule de quadrature à $l_i^2 \in \mathbf{P}_{2n} \subset \mathbf{P}_{2n+1}$, (où la famille $(l_i)_{0\leq i\leq n}$ est définie par (8)) on a la relation

$$0 < \int_{\alpha}^{\beta} l_i^2(x)\omega(x)dx = \sum_{i=0}^{n} \lambda_i l_i^2(x_i) = \lambda_i$$

ce qui achève la démonstration du corollaire □

Corollaire 6. *Soit* $n \in \mathbf{N}^*$. *Il existe une et une seule méthode de quadrature (avec un poids* ω *fixé) à* $(n+1)$ *points* $(x_i)_{0\leq i\leq n}$ *dans* $[\alpha,\beta]$ *tels que* $x_{n-1} = \alpha$ *et* $x_n = \beta$ *et qui soit d'ordre* $2n-1$ *: les points* $(x_i)_{0\leq i\leq n-2}$ *correspondent aux racines du n-ième polynôme orthogonal sur* $]\alpha,\beta[$ *pour un poids égal à* $\omega(x)(x-\alpha)(x-\beta)$. *De plus, les scalaires* $(\lambda_i)_{0\leq i\leq n}$ *définis par (9) sont strictement positifs.*

Démonstration. : Il suffit d'appliquer le Théorème 12 avec $l = n-1$ et en prenant la famille de points extérieurs $(x_{n-1}, x_n) = (\alpha, \beta)$. Pour démontrer la positivité des coefficients, on applique la formule de quadrature à la famille de polynômes dans $\mathbf{P}_{2n-1}$

$$P_i(x) = \begin{cases} \dfrac{(x-\alpha)}{(x_i-\alpha)}\dfrac{(\beta-x)}{(\beta-x_i)}\displaystyle\prod_{\substack{j=0\\ j\neq i}}^{n-2}\left(\frac{x-x_j}{x_i-x_j}\right)^2 & \text{si} \quad 0 \leq i \leq n-2, \\ \dfrac{(\beta-x)}{(\beta-\alpha)}\displaystyle\prod_{j=0}^{n-2}\left(\frac{x-x_j}{\alpha-x_j}\right)^2 & \text{si} \quad i = n-1, \\ \dfrac{(\alpha-x)}{(\alpha-\beta)}\displaystyle\prod_{j=0}^{n-2}\left(\frac{x-x_j}{\beta-x_j}\right)^2 & \text{si} \quad i = n \end{cases}$$

□

1.3.2 Exemples de méthodes de Gauss

À partir du Corollaire 13 et des familles usuelles de polynômes orthogonaux (voir leçon A.III.), on construit les méthodes suivantes : (i) lorsque $\omega(x) = 1$ et $]\alpha, \beta[=]-1, 1[$, on parle de la méthode de Gauss-Legendre. Les points $(x_i)_{0\leq i\leq n}$ correspondent aux racines du $(n+2)$-ième polynôme de Legendre. Par exemple, si $n = 1$, la méthode (d'ordre 3) s'écrit :

$$\int_{-1}^{1} f(x)dx \simeq f(-\frac{1}{\sqrt{3}}) + f(\frac{1}{\sqrt{3}}).$$

(ii) lorsque $\omega(x) = \dfrac{1}{\sqrt{1-x^2}}$ et $]\alpha, \beta[=]-1, 1[$, on parle de la méthode de Gauss-Tchebychev. Les points correspondent aux racines des polynômes de Tchebychev. On peut montrer dans ce cas (voir [CLFe]) que la méthode s'écrit :

$$\int_{\alpha}^{\beta} f(x)\frac{1}{\sqrt{1-x^2}}dx \simeq \frac{\pi}{n+1}\sum_{i=0}^{n} f\big(\cos(\frac{2i+1}{2n+2}\pi)\big).$$

(iii) lorsque $\omega(x) = e^{-x^2}$ et $]\alpha, \beta[= \mathbf{R}$, on parle de la méthode de Gauss-Hermite. Les points correspondent aux racines des polynômes d'Hermite. À partir du Corollaire 14, on construit la famille de méthodes suivante : (iv) lorsque $\omega(x) = 1$ et $\{\alpha, \beta\}$ sont quelconques dans $\mathbf{R}$, on construit pour tout $n \in \mathbf{N}^*$ la méthode de Gauss Lobatto à

$(n+1)$ points (tels que $x_{n-1} = \alpha$ et $x_n = \beta$) qui est d'ordre $2n-1$. Les points intérieurs sont pris égaux aux racines du n-ième polynôme orthogonal sur $]\alpha, \beta[$ pour le poids $(x-\alpha)(x-\beta)$.

En particulier, lorsque $n = 1$, on retrouve la méthode élémentaire des trapèzes et celle de Simpson lorsque $n = 2$.

Remarque : Les méthodes de Gauss possèdent rarement d'expression simple. En général, il est nécessaire d'avoir recours à des tables pour obtenir les valeurs approchées des coefficients $(x_i)_{0\le i\le n}$ et $(\lambda_i)_{0\le i\le n}$. Il est également possible de déterminer ceux-ci en utilisant Maple (voir [GH] pour un exemple de programme).

1.3.3 Convergence des méthodes de Gauss

Il est aisé après l'étude générale du paragraphe précédent de démontrer la convergence des méthodes de Gauss construites.

Théorème 7. *Les méthodes de Gauss à $(n+1)$ points tous situés dans l'intervalle réel $[\alpha, \beta]$ convergent lorsque n tend vers $+\infty$ pour toute fonction continue par morceaux.*

Démonstration. : On a vu que la propriété (α)' énoncée dans le Corollaire 2 était vérifiée pour les méthodes de Gauss construites dans les Corollaires 13 et 14. En outre, celles-ci étant d'ordre au moins égal à $2n-1$, on remarque que la condition (β) du Corollaire 2 est également satisfaite. Par conséquent, la convergence des méthodes de Gauss pour les fonctions continues par morceaux est bien assurée □

1.3.4 Estimation de l'erreur

Deux estimations d'erreur pour les méthodes de Gauss sont établies :

Théorème 8. *On considère une méthode de Gauss dont tous les points sont situés dans un intervalle borné $I = [a, b]$ contenant $]\alpha, \beta[$. On a*

$$\forall f \in C^{n+l+1}([a,b], \mathbf{R}),\ \exists \xi \in]a,b[,\ E(f) = \frac{f^{(n+l+1)}(\xi)}{(n+l+1)!} \int_\alpha^\beta \Pi(x)\omega(x)dx \qquad (10)$$

avec

$$\Pi(x) = \prod_{i=0}^{l-1} (x - x_i)^2 \prod_{i=l}^{n} (x - x_i).$$

Par ailleurs, le noyau de Peano K_{n+l} garde un signe constant sur $[a, b]$ et l'estimation du Corollaire 5 est donc également valable.

Démonstration. : Soit $P(f) \in \mathbf{P}_{n+l}$ le polynôme d'interpolation d'Hermite de f (voir leçon A.II.) tel que

$$\begin{cases} P(f)(x_i) = f(x_i) & \forall i \in \{0, ..., n\}, \\ P(f)'(x_i) = f'(x_i) & \forall i \in \{0, ..., l-1\}. \end{cases}$$

La méthode de Gauss étant d'ordre $n+l$, elle est exacte pour $P(f)$, soit

$$\int_\alpha^\beta P(f)(x)\omega(x)dx = \sum_{i=0}^{n} \lambda_i P(f)(x_i) = \sum_{i=0}^{n} \lambda_i f(x_i)$$

et ainsi

$$E(f) = \int_\alpha^\beta \big(f(x) - P(f)(x)\big)\omega(x)dx.$$

On peut montrer (voir leçon A.II. ou [CrMi]) que

$$\forall x \in [a,b], \quad \exists \xi \in [a,b], \quad f(x) - P(f)(x) = \frac{f^{(n+l+1)}(\xi)}{(n+l+1)!}\Pi(x). \tag{11}$$

Π étant de signe constant sur $]\alpha, \beta[$ (par exemple positif), il suffit de s'inspirer de la démonstration de la deuxième formule de la moyenne pour déduire la première estimation de $E(f)$.

Concernant le signe du noyau de Peano, supposons par l'absurde que celui-ci change sur $[a,b]$: en considérant alors la fonction ϕ continue sur $[a,b]$ telle que

$$\phi(t) = \begin{cases} (K_{n+l}(t))_+ & \text{si} \quad \displaystyle\int_a^b K_{n+l}(t)dt \le 0, \\ -(K_{n+l}(t))_- & \text{si} \quad \displaystyle\int_a^b K_{n+l}(t)dt > 0, \end{cases}$$

on remarque pour des raisons de signe que

$$\forall x \in [a,b], \quad \int_a^b K_{n+l}(t)\phi(t)dt \neq \phi(x)\int_a^b K_{n+l}(t)dt \tag{12}$$

(autrement dit, la deuxième formule de la moyenne n'est pas vérifiée avec ces deux fonctions). Soit alors $f \in C^{n+l+1}([a,b], \mathbf{R})$ telle que $f^{(n+l+1)} \equiv \phi$. Grâce au Théorème 3, la méthode étant d'ordre $n+l$, on a

$$E(f) = \frac{1}{(n+l)!}\int_a^b K_{n+l}(t)\phi(t)dt, \tag{13}$$

tandis que pour tout $x \in [a,b]$

$$\begin{aligned}\frac{\phi(x)}{(n+l)!}\int_a^b K_{n+l}(t)dt &= \frac{f^{(n+l+1)}(x)}{(n+l+1)!}E(x\mapsto x^{n+l+1}) \\ &= \frac{f^{(n+l+1)}(x)}{(n+l+1)!}\int_\alpha^\beta \Pi(t)\omega(t)dt\end{aligned} \tag{14}$$

après application consécutive des formules (13) et (10) à la fonction $g : x \mapsto x^{n+l+1}$.

En écrivant la formule (14) pour $x = \xi$ défini dans (10), et en utilisant successivement les relations (10) et (13) on obtient

$$\frac{\phi(\xi)}{(n+l)!}\int_a^b K_{n+l}(t)dt = E(f) = \frac{1}{(n+l)!}\int_a^b K_{n+l}(t)\phi(t)dt$$

ce qui contredit la propriété (12) □

Exemple : L'estimation (10) appliquée à la méthode de Gauss-Legendre à deux points donne

$$\forall f \in C^4([-1,1],\mathbf{R}), \quad |E(f)| \leq \frac{||f^{(4)}||_\infty}{135}.$$

Ce résultat est légèrement meilleur que celui obtenu pour la méthode de Simpson dans lequel 135 est remplacé par 90.

1.4 Autres méthodes de quadrature

1.4.1 Méthode de Romberg

La méthode de Romberg est construite dans les Exercices 2, 3 et 4 auxquels le lecteur est invité à se référer. Elle peut être interprétée comme un procédé d'accélération de convergence de la méthode des trapèzes : on construit une suite à double indice $(A_{m,k})_{0\leq k\leq m}$ en définissant d'abord les termes $(A_{m,0})_{m\in\mathbf{N}}$ comme étant égaux à l'approximation des trapèzes de $\int_\alpha^\beta f(x)dx$ sur 2^m+1 points équirépartis. Une relation de récurrence permet ensuite de construire les suites successives $(A_{m,1})_{1\leq m}$,$(A_{m,2})_{2\leq m}$,etc... pouvant être rangées en colonnes pour former un tableau triangulaire :

$$\forall m \in \mathbf{N}^*, \forall k \in \mathbf{N}, \quad k \leq m-1 \Rightarrow A_{m,k+1} = \frac{4^{k+1}A_{m,k} - A_{m-1,k}}{4^{k+1}-1}.$$

Le résultat essentiel est que, à k fixé et lorsque n tend vers $+\infty$,

$$A_{m,k} = \int_\alpha^\beta f(x)dx + O(\frac{1}{2^{m(2k+2)}}).$$

Pour démontrer cette estimation, on utilise le procédé d'extrapolation (ou d'accélération de convergence) de Richardson (voir Exercice 3) appliqué à la méthode des trapèzes dont une estimation d'erreur pour un pas de subdivision tendant vers 0 est fournie par la formule d'Euler Mac Laurin (Exercice 2).

Pour une démonstration complète de ces résultats, le lecteur pourra se référer à [Dem]. Une autre construction de la méthode de Romberg (avec estimation précise de l'erreur) est également possible (voir [CLFe]). Enfin, la programmation de celle-ci avec Maple est proposée dans [Fer].

Remarques : 1. La famille $(A_{m,1})_{1\leq m}$ (respectivement $(A_{m,2})_{2\leq m}$) correspond en fait à la méthode de Simpson (respectivement Boole Villarceau) sur $2^{m-1}+1$ (respectivement $2^{m-2}+1$) points équirépartis. Au delà (pour $k \geq 3$), on peut montrer que la méthode de Romberg ne correspond à aucune méthode de Newton Cotes fermée. 2. La formule d'Euler Mac Laurin (voir Exerice 2) fait intervenir la famille importante des polynômes de Bernoulli $(B_n)_{n\in\mathbf{N}}$ (par $B_n(x) = (-1)^n(\alpha_n - P_n(x))$). Elle permet également de démontrer que la méthode des trapèzes est d'ordre de précision infini appliquée au calcul de l'intégrale de période d'une fonction $f \in C^\infty(\mathbf{R})$ de période $\beta - \alpha$:

$$\forall m \in \mathbf{N}^*, \quad \exists C > 0, \quad \forall n \in \mathbf{N}, \quad |\int_\alpha^\beta f(x)dx - \sum_{i=0}^{n-1} f(\alpha + \frac{i}{n}(\beta-\alpha)| \leq \frac{C}{n^m}.$$

Enfin, la formule d'Euler Mac Laurin est également utilisée pour calculer des développements asymptotiques de sommes du type $f(1)+f(2)+...+f(n)$, par exemple pour estimer la constante d'Euler γ (algorithme de Sweeney).

1.4.2 Méthode de Monte Carlo

1.4.2.1 Définition et convergence

La méthode de Monte Carlo est une méthode de quadrature de type probabiliste consistant à interpréter une intégrale comme l'espérance d'une certaine variable aléatoire Y sur un espace de probabilités $(\Omega, \mathscr{A}, P)$ et à utiliser le théorème central limite pour approcher sa valeur : à partir d'une famille $(Y_i)_{1\leq i\leq N} \in \mathbf{R}^N$ de réalisations indépendantes de Y, on écrit :

$$E(Y) = \int_\Omega Y(\omega)P(d\omega) \simeq \frac{1}{N}\sum_{i=1}^N Y_i.$$

Ainsi, on pourra par exemple approcher l'intégrale $\int_{[0,1]^d} f(x)dx$ en considérant la variable $Y = f(X)$ où X est une variable aléatoire de loi uniforme sur $[0,1]^d$ (ce qui est très facilement programmable par exemple avec l'instruction `rand` sous Scilab). Comme pour une méthode déterministe, on définit l'erreur associée

$$\varepsilon_N = \int_\Omega Y(\omega)P(d\omega) - \frac{1}{N}\sum_{i=1}^{N} Y_i.$$

Le théorème central limite permet d'estimer la vitesse de convergence de ε_N vers 0 (assurée presque sûrement par la loi des grands nombres) : si $E(Y^2) < +\infty$, on note

$$\sigma^2 = \mathrm{Var}(Y) = E(Y^2) - E(Y)^2$$

la variance de Y, et on a

$$\lim_{N\to+\infty} P(\frac{\sigma}{\sqrt{N}}c_1 \leq \varepsilon_n \leq \frac{\sigma}{\sqrt{N}}c_2) = \int_{c_1}^{c_2} e^{-\frac{x^2}{2}}\frac{dx}{\sqrt{\pi}}.$$

L'erreur commise est donc en général de l'ordre de $O(\frac{1}{\sqrt{N}})$. Plus précisément, on doit donner le résultat de l'approximation sous la forme d'un intervalle de confiance (par exemple avec une probabilité de 95% pour que la valeur exacte soit bien dans celui-ci). Pour déterminer cet intervalle au cours du calcul, on peut montrer que l'espérance et la variance de Y, a priori inconnues, peuvent être remplacées par

$$I_N = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

et

$$\sigma_N^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - I_N)^2$$

sans changer la probabilité de confiance.

Remarque : La méthode de Monte Carlo présente l'avantage important que sa vitesse de convergence est indépendante de la dimension de l'espace d'intégration et de la régularité de la fonction à intégrer.

1.4.2.2 Accélération de convergence

Il existe des méthodes d'accélération de convergence qui consistent par exemple à écrire pour une fonction g bien choisie, positive et de mesure 1 sur $[0,1]^d$,

$$\int_{[0,1]^d} f(x)dx = \int_{[0,1]^d} \overbrace{\frac{f(x)}{g(x)}}^{h(x)} g(x)dx = E(h(Z))$$

où Z suit une loi de probabilité $g(x)dx$ sur $[0,1]^d$.

En prenant g tel que $\mathrm{Var}(h(Z)) < \mathrm{Var}(f(X))$ (X variable aléatoire uniforme sur $[0,1]^d$), la méthode de Monte Carlo appliquée avec $Y = h(Z)$ convergera en effet plus rapidement que celle avec $Y = f(X)$.

Une autre méthode consiste à effectuer une partition du domaine d'intégration :

$$[0,1]^d = \sqcup_{i=1}^m D_i,$$

puis à écrire

$$E(f(X)) = \sum_{i=1}^m P(X \in D_i) E(f(X)|X \in D_i) = \sum_{i=1}^m p_i I_i.$$

Si les quantités $p_i = P(X \in D_i)$ sont connues, la variance de l'erreur commise à partir de N points répartis judicieusement dans chaque sous-ensemble (par exemple $N_i = N p_i$) et des m intégrales $(I_i)_{1 \le i \le m}$ sera inférieure à la variance de l'erreur sans décomposition pour le même nombre de points.

Remarque : Une autre méthode d'accélération de convergence consiste à utiliser des suites déterministes de points $(x_i)_{1 \le i \le N}$ possédant de meilleures propriétés d'équirépartition que les suites aléatoires (on parle de suites à discrépance faible et de méthode de quasi Monte Carlo : voir des exemples dans [LPS]).

1.4.2.3 Exemples

Les méthodes de Monte Carlo sont utilisées dans de nombreuses applications en raison de leur simplicité de mise en œuvre et de leur robustesse. Elles peuvent aussi servir à calculer des solutions approchées d'équations différentielles en exprimant celles-ci en termes d'espérance (par exemple la solution de l'équation du transport linéaire ou celle de l'équation de Boltzmann). On les retrouve donc à la base de nombreux codes de simulation dans divers domaines industriels (aéronautique, neutronique) ou financiers.

1.5 Transformation de Fourier discrète et rapide

L'algorithme exposé brièvement ci-dessous consiste à calculer de manière efficace une approximation des N premières harmoniques d'une fonction f continue et périodique (de période 1) :

$$\hat{f}(k) = \int_0^1 f(x) e^{2ik\pi x} dx \quad (0 \leq k \leq N-1).$$

La formule d'Euler Mac Laurin (voir paragraphe précédent) indique qu'une excellente approximation (d'ordre infini) de cette quantité est fournie par la méthode des trapèzes composée (confondue dans ce cas avec la méthode des rectangles à gauche) :

$$\forall k \in \{0, ..., N-1\},\ \hat{f}(k) \simeq U_k = \frac{1}{N} \sum_{j=0}^{N-1} f(\frac{j}{N}) e^{\frac{2ikj\pi}{N}}.$$

La transformation associée

$$(u_k = f(\frac{k}{n}))_{0 \leq k \leq N-1} \mapsto (U_k)_{0 \leq k \leq N-1}$$

s'appelle la transformation de Fourier discrète. Sous la forme précédente, celle-ci demande un nombre d'opérations de l'ordre de N^2.

Pour les entiers N du type 2^p, il existe une méthode d'estimation plus efficace des coefficients $(U_k)_{0 \leq k \leq N-1}$ demandant $O(N \log_2 N)$ opérations et appelée transformation de Fourier rapide. Elle repose sur une remarque permettant de calculer la transformation de N coefficients en fonction de la transformation de $\frac{N}{2}$ coefficients. Les résultats sont alors obtenus rangés en indices pairs/indices impairs (on parle d'entrelacement fréquentiel). En itérant cette remarque, on obtient une transformation rapide de $(u_k)_{0 \leq k \leq N-1}$ en $(U_{\tau_N(k)})_{0 \leq k \leq N-1}$ pour $N = 2^p$ sous la forme d'un schéma en papillon. Le désentrelacement des harmoniques s'effectue en appliquant à nouveau la transposition τ_N qui consiste à inverser l'ordre des chiffres du développement binaire d'un entier. Le lecteur est invité à se référer à [Sch] et [DaLi] pour une justification complète et le comptage des opérations de cette méthode (ainsi qu'une extension à d'autres valeurs de N). En outre, plusieurs applications de la transformation de Fourier (rapide ou discrète) sont développées dans [DaLi] (résolution approchée de problèmes aux limites, régularisation de fonctions, recherche des valeurs propres de l'opérateur Laplacien discrétisé).

1.6 Résolution approchée des EDO : méthode à un pas

On part du problème de Cauchy suivant : trouver $y \in C^1([t_0, t_0+T], \mathbf{R}^m)$ tel que

$$\begin{cases} y'(t) = f(t, y(t)), \quad t \in [t_0, t_0+T], \\ y(t_0) \in \mathbf{R}^m \text{ fixé.} \end{cases} \tag{1}$$

Pour que ce problème possède une unique solution grâce au théorème de Cauchy-Lipschitz, nous supposerons par la suite (sauf mention contraire) que la fonction f est continue de $[t_0, t_0+T] \times \mathbf{R}^m$ dans $\mathbf{R}^m$ et est globalemement Lipschitzienne par rapport à sa seconde variable avec un coefficient de Lipschitz noté L pour la norme choisie sur $\mathbf{R}^m$:

$$\forall t \in [t_0, t_0+T], \quad \forall (y_1, y_2) \in (\mathbf{R}^m)^2, \quad ||f(t, y_2) - f(t, y_1)|| \leq L||y_2 - y_1||.$$

Une méthode de résolution approchée du problème (1) consiste d'abord à effectuer une subdvision de l'intervalle de définition de y

$$[t_0, t_0+T] = \bigcup_{n=0}^{N-1} [t_n^N, t_{n+1}^N], \quad N \in \mathbf{N}^*$$

puis à construire une suite $(y_n^N)_{0 \leq n \leq N}$ telle que y_n^N approche $y(t_n^N)$ pour tout $n \in \{0, .., N\}$.

Dans la suite de la construction, pour des raisons de clarté, les indices N seront omis. De plus, on se restreindra au cas scalaire $m = 1$ et aux méthodes à pas de temps constant, à savoir :

$$\forall n \in \{0, ..., N\}, \quad t_n = t_0 + n\Delta T \quad \text{avec} \quad \Delta T = \frac{T}{N}$$

(voir ultérieurement pour l'extension dans le cas particulier de la méthode d'Euler à $m > 1$ et au pas de temps variable). Définition : On appelle méthode de résolution

approchée explicite à un pas (et à pas de temps constant) du problème (1), la construction pour tout $N \in \mathbf{N}^*$ d'une suite finie $(y_n)_{0 \leq n \leq N}$ telle que

$$\begin{cases} y_0 \in \mathbf{R} \text{ fixé,} \\ y_{n+1} = y_n + \Delta T \Phi(t_n, y_n, \Delta T) \quad 0 \leq n \leq N-1, \end{cases} \tag{2}$$

où Φ désigne une fonction continue de $[t_0, t_0+T] \times \mathbf{R} \times [0, T]$ dans $\mathbf{R}$.

1.6.1 Consistance, stabilité, convergence et ordre d'une méthode

On définit ci-dessous les différentes propriétés importantes éventuellement satisfaites par une méthode du type précédent. Définition : On dit qu'une méthode à un pas

de type (2) est consistante avec problème (1) si pour toute solution y de celui-ci, on a

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

où

$$\varepsilon_n = y(t_{n+1}) - y(t_n) - \Delta T\Phi(t_n, y(t_n), \Delta T)$$

Définition 4. *On dit qu'une méthode à un pas de type (2) est stable si pour tout entier $N \in \mathbf{N}^*$ et pour toute famille $(z_n)_{0\le n\le N}$ solution de la relation de récurrence perturbée par le réel ε_n :*

$$\begin{cases} z_0 \in \mathbf{R} \text{ fixé}, \\ z_{n+1} = z_n + \Delta T\Phi(t_n, z_n, \Delta T) + \varepsilon_n \quad 0 \le n \le N-1, \end{cases}$$

on a la relation

$$\max_{0\le n\le N-1}|z_n - y_n| \le M|z_0 - y_0| + M'\sum_{n=0}^{N-1}|\varepsilon_n|$$

où M et M' sont des constantes indépendantes de y_0, z_0 et N.

Définition 5. *On dit qu'une méthode à un pas de type (2) est convergente si pour toute solution y du problème (1)*

$$\lim_{\substack{N\to+\infty \\ y_0\to y(t_0)}} \max_{0\le n\le N}|y(t_n) - y_n| = 0.$$

Définition 6. *On dit qu'une méthode à un pas de type (2) est d'ordre $p \in \mathbf{N}^*$ si Φ et f sont p-fois continûment différentiables sur leur ensemble de définition et si pour toute solution y du problème (1), il existe une constante $C > 0$ indépendante de N telle que*

$$\forall N \in \mathbf{N}^*, \quad \sum_{n=0}^{N-1}|\varepsilon_n| \le C(\Delta T)^p$$

où

$$\varepsilon_n = y(t_{n+1}) - y(t_n) - \Delta T\Phi(t_n, y(t_n), \Delta T) \quad (0 \le n \le N-1).$$

Remarque : La notion de consistance revient en fait à s'assurer que la méthode considérée est cohérente (ou consistante) avec le problème initial. La stabilité signifie que les éventuelles erreurs numériques commises dans l'évaluation des termes de la suite d'approximations sont contrôlables. Enfin, la notion d'ordre est reliée avec la vitesse de convergence d'une méthode si celle-ci est stable (voir la démonstration du Théorème 5). Les sous-paragraphes suivants vont s'efforcer de donner des conditions

nécessaires et/ou suffisantes sur Φ permettant de vérifier ces différentes propriétés. Auparavant, deux lemmes souvent utilisés dans la suite de la leçon sont démontrés.

1.6.2 Lemmes techniques

Lemme 2. *soit $(z_n)_{n\in\mathbf{N}}$ une suite réelle telle que*

$$\begin{cases} z_0 \geq 0, \\ z_{n+1} \leq Az_n + B \quad (n \in \mathbf{N}) \end{cases}$$

où A et B sont deux réels tels que $A \geq 1$ et $B \geq 0$. On a l'inégalité suivante valable pour tout $n \in \mathbf{N}$:

$$z_n \leq e^{nQ} z_0 + \frac{e^{nQ}-1}{Q} B$$

avec $Q = A-1$ et la convention

$$\frac{e^{nQ}-1}{Q} = n \quad \text{lorsque } Q = 0.$$

Démonstration. On démontre ce résultat en appliquant n fois la propriété de la suite $(z_n)_{n\in\mathbf{N}}$:

$$\begin{cases} z_n \leq A^n z_0 + \dfrac{A^n-1}{A-1} B \quad \text{lorsque } A > 1, \\ z_n \leq z_0 + nB \quad \text{lorsque } A = 1 \end{cases}$$

et en remarquant que

$$A = Q + 1 \leq e^Q \quad \square$$

Lemme 3. *soit $(z_n)_{n\in\mathbf{N}}$ une suite réelle et $(h_n, \alpha_n)_{n\in\mathbf{N}} \in (\mathbf{R}_+^2)^{\mathbf{N}}$ tels que :*

$$\begin{cases} z_0 \geq 0, \\ z_{n+1} \leq (1+h_n) z_n + \alpha_n. \end{cases}$$

En notant

$$t_0 = 0 \quad \text{et} \quad t_n = \sum_{i=0}^{n-1} h_i \quad (n \in \mathbf{N}^*),$$

on a l'inégalité suivante valable pour tout $n \in \mathbf{N}$ *:*

$$z_n \leq e^{t_n} z_0 + \sum_{i=0}^{n-1} e^{(t_n - t_{i+1})} \alpha_i.$$

Démonstration. : Il suffit d'adapter la démonstration précédente □

1.6.3 Consistance d'une méthode explicite à un pas

Théorème 9. *: une méthode explicite à un pas de type (2) est consistante si et seulement si*

$$\forall (t,y) \in [t_0, t_0 + T] \times \mathbf{R}, \quad \Phi(t,y,0) = f(t,y).$$

Démonstration. : Soit $\varepsilon_n = y(t_{n+1}) - y(t_n) - \Delta T \Phi(t_n, y(t_n), \Delta T) \quad (0 \leq n \leq N-1)$. Par le théorème des acroissements finis, on sait qu'il existe $c_n \in]t_n, t_{n+1}[$ tel que

$$\varepsilon_n = \Delta T f(c_n, y(c_n)) - \Delta T \Phi(t_n, y(t_n), \Delta T) = \Delta T (\alpha_n + \beta_n)$$

avec

$$\begin{cases} \alpha_n = f(c_n, y(c_n)) - \Phi(c_n, y(c_n), 0), \\ \beta_n = \Phi(c_n, y(c_n), 0) - \Phi(t_n, y(t_n), \Delta T). \end{cases}$$

Comme la fonction $\tilde{\Phi} : \begin{pmatrix} [t_0, t_0 + T] \times [0, T] \to \mathbf{R} \\ (t,h) \mapsto \Phi(t, y(t), h) \end{pmatrix}$ est continue sur l'ensemble compact $[t_0, t_0 + T] \times [0, T]$, donc uniformément continue, on peut affirmer que

$$\forall \varepsilon > 0, \exists N_0 \in \mathbf{N}^*, \forall N \geq N_0, \forall n \in \{0, \ldots, N-1\}, |\beta_n| \leq \varepsilon.$$

Ainsi, pour $N \geq N_0$,

$$\left| \sum_{n=0}^{N-1} |\varepsilon_n| - \sum_{n=0}^{N-1} \Delta T |\alpha_n| \right| \leq \sum_{n=0}^{N-1} \Delta T |\beta_n| \leq \varepsilon T.$$

De plus, par définition de l'intégrale de Rieman,

$$\lim_{N \to +\infty} \sum_{n=0}^{N-1} \Delta T |\alpha_n| = \int_{t_0}^{t_0+T} |f(t, y(t)) - \Phi(t, y(t), 0)| dt.$$

Au vu de la définition et des remarques précédentes, une condition nécessaire et suffisante de consistance de la méthode étudiée est donc que pour toute solution y de (1) on ait

$$\forall t \in [t_0, t_0+T], \quad f(t, y(t)) = \Phi(t, y(t), 0). \tag{3}$$

Or, pour tout couple $(t^*, y^*) \in [t_0, t_0+T] \times \mathbf{R}$, il existe par le théorème de Cauchy-Lipschitz une (unique) solution y définie sur $[t_0, t_0+T]$ du problème de Cauchy suivant :

$$\begin{cases} y'(t) = f(t, y(t)), & t \in [t_0, t_0+T], \\ y(t^*) = y^*. \end{cases}$$

En écrivant la relation (3) en t^* pour cette fonction, également solution d'un problème de type (1), on obtient :

$$\Phi(t^*, y^*, 0) = f(t^*, y^*)$$

ce qui est bien le résultat recherché □

1.6.4 Stabilité d'une méthode explicite à un pas

Théorème 10. *: pour qu'une méthode explicite à un pas de type (2) soit stable, il suffit que la fonction Φ soit Lipschitzienne en y, à savoir qu'il existe $\Lambda > 0$ tel que*

$$\forall (t, y_1, y_2, h) \in [t_0, t_0+T] \times \mathbf{R}^2 \times [0, T], \quad |\Phi(t, y_2, h) - \phi(t, y_1, h)| \leq \Lambda |y_2 - y_1|$$

On peut alors prendre comme constantes de stabilité $M = M' = e^{\Lambda T}$.

Démonstration. : En conservant les notations de la définition de la stabilité, on peut écrire sous les hypothèses du théorème :

$$\begin{aligned} |y_{n+1} - z_{n+1}| &= |y_n - z_n + \Delta T\,(\Phi(t_n, y_n, \Delta T) - \Phi(t_n, z_n, \Delta T)) - \varepsilon_n| \\ &\leq (1 + \Lambda \Delta T)|y_n - z_n| + |\varepsilon_n| \end{aligned}$$

Grâce au Lemme 2, on a alors

$$\begin{aligned} |y_n - z_n| &\leq e^{n\Lambda\Delta T}|y_0 - z_0| + \sum_{i=0}^{n-1} e^{(n-1-i)\Lambda\Delta T}|\varepsilon_i| \\ &\leq e^{\Lambda T}|y_0 - z_0| + e^{\Lambda T}\sum_{i=0}^{N-1}|\varepsilon_i| \end{aligned}$$

ce qui clôt la démonstration □

Remarque : Étant sous forme d'exponentielles, les constantes de stabilité peuvent malheureusement être très grandes (voir exemples ultérieurs). Cette estimation ne peut en fait être améliorée dans le cas général (voir l'exemple trivial où $y' = \lambda y$ avec $\lambda > 0$).

1.6.5 Convergence d'une méthode explicite à un pas

Théorème 11. *: si une méthode explicite à un pas est stable et consistante, alors elle est convergente.*

Démonstration. : On pose $\varepsilon_n = y(t_{n+1}) - y(t_n) - \Delta T \Phi(t_n, y(t_n), \Delta T) \quad (0 \leq n \leq N-1)$. La famille $(z_n)_{0 \leq n \leq N-1}$ solution de la relation de récurrence perturbée

$$\begin{cases} z_0 = y(t_0), \\ z_{n+1} = z_n + \Delta T \Phi(t_n, z_n, \Delta T) + \varepsilon_n \quad (0 \leq n \leq N-1) \end{cases}$$

vérifie facilement par construction même :

$$\forall n \in \{0, ..., N\}, \quad z_n = y(t_n).$$

La méthode considérée étant stable et consistante, il vient respectivement

$$\max_{0 \leq n \leq N-1} |y_n - y(t_n)| \leq M |y_0 - y(t_0)| + M' \sum_{n=0}^{N-1} |\varepsilon_n|$$

et

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

En regroupant ces deux résultats, on a bien démontré la convergence de la méthode, à savoir :

$$\lim_{\substack{N \to +\infty \\ y_0 \to y(t_0)}} \max_{0 \leq n \leq N} |y(t_n) - y_n| = 0$$

□

Le corollaire suivant permet de déterminer une condition suffisante de convergence d'une méthode simplement à partir d'informations sur Φ :

Corollaire 7. *si la fonction Φ associée à une méthode explicite à un pas est Lispchitzienne par rapport à la variable y et si*

$$\forall (t, y) \in [t_0, t_0 + T] \times \mathbf{R} \quad \Phi(t, y, 0) = f(t, y),$$

alors la méthode est convergente.

Démonstration. : Il suffit d'utiliser les Théorèmes 3, 4 et 5 □

1.6.6 Ordre d'une méthode explicite à un pas

Théorème 12. *: une méthode explicite à un pas de type (2) est d'ordre $p \in \mathbf{N}^*$ si et seulement si Φ et f sont p-fois continûment différentiables sur leur ensemble de définition et si*

$$\forall l \in \{0,...,p-1\}, \quad \forall (t,y) \in [t_0, t_0+T] \times \mathbf{R}, \ \frac{\partial^l}{\partial h^l}\Phi(t,y,0) = \frac{1}{l+1} f^{[l]}(t,y), \quad (4)$$

où $f^{[l]}(t,y)$ désigne la l-ième dérivée totale de f suivant les caractéristiques du problème (1) :

$$\begin{cases} f^{[0]}(t,y) = f(t,y), \\ f^{[k+1]}(t,y) = \partial_t f^{[k]}(t,y) + \nabla_y f^{[k]}(t,y) \cdot f(t,y) \quad (0 \le k \le p-1). \end{cases}$$

Démonstration. : Condition suffisante : on remarque tout d'abord que si $f \in C^p([t_0, t_0+T] \times \mathbf{R}, \mathbf{R})$, alors toute solution y de (1) appartient à $C^{p+1}([t_0, t_0+T], \mathbf{R})$ et vérifie

$$\forall l \in \{0,...,p\}, \quad \forall t \in [t_0, t_0+T], \quad y^{(l+1)}(t) = f^{[l]}(t, y(t)).$$

On écrit alors la relation de Taylor-Lagrange à l'ordre p pour la fonction $h \mapsto \Phi(t_n, y(t_n), h)$ entre 0 et ΔT :

$$\Phi(t_n, y(t_n), \Delta T) = \sum_{l=0}^{p-1} \frac{(\Delta T)^l}{l!} \frac{\partial^l}{\partial h^l}\Phi(t_n, y(t_n), 0) + \frac{(\Delta T)^p}{p!} \frac{\partial^p}{\partial h^p}\Phi(t_n, y(t_n), \lambda_n)$$

($\lambda_n \in]0, \Delta T[$) et à l'ordre $p+1$ pour la fonction y entre t_n et t_{n+1} :

$$\begin{aligned} y(t_{n+1}) - y(t_n) &= \sum_{k=1}^{p} \frac{(\Delta T)^k}{k!} y^{(k)}(t_n) + \frac{(\Delta T)^{p+1}}{(p+1)!} y^{(p+1)}(c_n) \qquad (c_n \in]t_n, t_{n+1}[) \\ &= \sum_{k=1}^{p} \frac{(\Delta T)^k}{k!} f^{[k-1]}(t_n, y(t_n)) + \frac{(\Delta T)^{p+1}}{(p+1)!} y^{(p+1)}(c_n). \end{aligned}$$

En soustrayant ces deux égalités après un changement d'indices ($k = l+1$), il vient

$$\begin{aligned} \varepsilon_n &= y(t_{n+1}) - y(t_n) - \Delta T \Phi(t_n, y(t_n), \Delta T) \\ &= \sum_{l=0}^{p-1} \frac{(\Delta T)^{l+1}}{l!} \left[\frac{f^{[l]}(t_n, y(t_n))}{l+1} - \frac{\partial^l \Phi(t_n, y(t_n), 0)}{\partial h^l} \right] \\ &+ \frac{(\Delta T)^{p+1}}{p!} \left[\frac{y^{(p+1)}(c_n)}{p+1} - \frac{\partial^p}{\partial h^p}\Phi(t_n, y(t_n), \lambda_n) \right] \\ &= \frac{(\Delta T)^{p+1}}{p!} \left[\frac{y^{(p+1)}(c_n)}{p+1} - \frac{\partial^p}{\partial h^p}\Phi(t_n, y(t_n), \lambda_n) \right] \end{aligned}$$

grâce à la relation (4). Ainsi, avec la régularité supposée de Φ et f

$$\exists C > 0, \quad \forall n \in \{0, ..., N-1\}, \quad |\varepsilon_n| \leq C(\Delta T)^{p+1}.$$

En sommant ces N inégalités, on trouve bien que la méthode est d'ordre p. Condition nécessaire : on raisonne par l'absurde en supposant que la méthode est d'ordre p et que la relation (4) est violée à partir d'un certain $l \in \{0, ..., p-1\}$. En écrivant les mêmes égalités de Taylor-Lagrange que précédemment mais cette fois à l'ordre $l+1$ et $l+2$ respectivement, il vient

$$\varepsilon_n = \frac{(\Delta T)^{l+1}}{l!}\Big[\frac{f^{[l]}(t_n, y(t_n))}{l+1} - \frac{\partial^l}{\partial h^l}\Phi(t_n, y(t_n), 0)\Big] + O((\Delta T)^{l+2})$$

puis en sommant

$$\sum_{n=0}^{N-1}\Big|\frac{\varepsilon_n}{(\Delta T)^l}\Big| = \frac{1}{l!}\sum_{n=0}^{N-1}\Delta T|\psi(t_n)| + O(\Delta T)$$

où ψ désigne la fonction continue $\psi : \begin{pmatrix} [t_0, t_0+T] \to \mathbf{R} \\ t \mapsto \dfrac{f^{[l]}(t, y(t))}{l+1} - \dfrac{\partial^l}{\partial h^l}\Phi(t, y(t), 0) \end{pmatrix}$.

On déduit alors par passage à la limite dans les deux membres de la dernière égalité lorsque ΔT tend vers 0 que

$$0 = \frac{1}{l!}\int_{t_0}^{t_0+T} |\psi(s)|ds$$

ce qui implique immédiatement que pour toute fonction y solution de (1)

$$\forall t \in [t_0, t_0+T], \quad \psi(t) = \frac{f^{[l]}(t, y(t))}{l+1} - \frac{\partial^l}{\partial h^l}\Phi(t, y(t), 0) = 0.$$

En raisonnant comme dans la fin de la démonstration du Théorème 3, on en déduit facilement que

$$\forall (t^*, y^*) \in [t_0, t_0+T] \times \mathbf{R} \quad \frac{\partial^l}{\partial h^l}\Phi(t^*, y^*, 0) = \frac{f^{[l]}(t^*, y^*)}{l+1}$$

ce qui achève la démonstration par l'absurde □

1.7 Résolution approchée : les méthodes d'Euler

1.7.1 Méthode d'Euler explicite

La méthode d'Euler explicite est un cas particulier de méthode explicite à un pas correspondant au choix de Φ suivant :

$$\forall (t,y,h) \in [t_0, t_0+T] \times \mathbf{R}^m \times [0,T], \quad \Phi(t,y,h) = f(t,y).$$

Définition : On appelle méthode de résolution approchée d'Euler explicite (à pas de temps constant) du problème (1), la construction pour tout $N \in \mathbf{N}^*$ de la suite finie $(y_n)_{0\leq n\leq N}$ telle que

$$\begin{cases} y_0 \in \mathbf{R}^m \text{ fixé,} \\ y_{n+1} = y_n + \Delta T f(t_n, y_n) \quad 0 \leq n \leq N-1. \end{cases} \tag{5}$$

Dans toute la suite, on désignera par e_n l'erreur commise en remplaçant $y(t_n)$ par y_n :

$$e_n = y_n - y(t_n).$$

Remarque : La méthode d'Euler revient en fait à approcher l'intégrale $\displaystyle\int_{t_n}^{t_{n+1}} y'(t)dt$ par la valeur $(t_{n+1}-t_n)f(t_n,y_n)$, c'est à dire à utiliser la méthode de quadrature élémentaire des rectangles à gauche.

1.7.2 Convergence de la méthode d'Euler explicite

Théorème 13. *: la méthode d'Euler explicite définie par la relation (5) est consistante, stable, convergente et d'ordre 1 (si f est C^1). De plus, on a l'inégalité suivante valable pour tout $n \in \{0,...,N\}$ et toute solution y du problème (1) :*

$$||e_n|| \leq \frac{e^{nL\Delta T}-1}{L}\omega(\Delta T, y') + e^{nL\Delta T}||e_0||$$

où ω désigne le module de continuité d'une fonction :

$$\omega(\delta, g) = \max_{\{(t,t')\in[t_0,t_0+T]^2 \ / \ |t'-t|\leq\delta\}} ||g(t')-g(t)||$$

et L la constante de Lipschitz de f par rapport à la seconde variable.

Démonstration. : Les propriétés de consistance, de stabilité, de convergence et l'ordre de la méthode d'Euler se déduisent immédiatement de l'étude générale du paragraphe précédent.

Pour démontrer l'estimation de l'erreur dans la seconde partie du théorème, on considère une fonction y solution du problème (1). On a :

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

en notant $\varepsilon_n = \int_{t_n}^{t_{n+1}} (y'(t) - y'(t_n))dt$. On peut rapprocher cette égalité de la définition (5) de la suite des approximations :

$$y_{n+1} = y_n + \Delta T f(t_n, y_n)$$

En soustrayant ces deux égalités et en utilisant la définition du module de continuité, on obtient

$$||y(t_{n+1}) - y_{n+1}|| \leq ||y(t_n) - y_n|| + L\Delta T||y(t_n) - y_n|| + \Delta T\,\omega(\Delta T, y').$$

On se trouve alors dans les conditions d'application du Lemme 1. Il vient :

$$||y(t_n) - y_n|| = ||e_n|| \leq e^{nL\Delta T}||e_0|| + \frac{e^{nL\Delta T} - 1}{L\Delta T}\Delta T\,\omega(\Delta T, y')$$

et le résultat annoncé est bien démontré □

Le corollaire suivant peut être déduit de la convergence de la méthode d'Euler :

Corollaire 8. *soit la suite de fonctions $(u_N)_{N\in\mathbf{N}}$ définies sur $[t_0, t_0+T]$ par interpolation affine entre les points $(t_n)_{0\leq n\leq N}$ pour lesquels $u_N(t_n) = y_n$ où la famille $(y_n)_{0\leq n\leq N}$ est construite suivant la méthode d'Euler (5) avec $y_0 \in \mathbf{R}^m$ fixé. Alors u_N converge dans $L^\infty([t_0, t_0+T], \mathbf{R}^m)$ vers l'unique solution y de (1) telle que $y(t_0) = y_0$.*

Démonstration. : Soit N fixé et $t \in [t_0, t_0 + T[$. On note n l'unique entier compris entre 0 et $N-1$ tel que $t \in [t_n, t_{n+1}[$.On a alors l'estimation suivante :

$$\begin{aligned} ||u_N(t) - y(t)|| &\leq ||u_N(t) - y_n|| + ||y_n - y(t_n)|| + ||y(t_n) - y(t)|| \\ &\leq ||u_N(t) - y_n|| + ||e_n|| + \omega(\Delta T, y). \end{aligned}$$

u_N étant affine sur $[t_n, t_{n+1}]$, on a

$$||u_N(t) - y_n|| \leq ||y_{n+1} - y_n|| = \Delta T ||f(t_n, y_n)||$$

Il suffit donc pour conclure à l'uniforme convergence de u_N vers y de montrer que y_n reste borné indépendamment de n et N. Pour cela, on remarque d'abord que

$$\begin{aligned} ||y_{n+1}|| \leq ||y_n|| + \Delta T ||f(t_n, y_n)|| &\leq ||y_n|| + \Delta T (||f(t_n, 0)|| + L||y_n||) \\ &\leq (1 + \Delta T L)||y_n|| + \Delta T \max_{t \in [t_0, t_0+T]} ||f(t, 0)|| \end{aligned}$$

et on utilise ensuite le Lemme 1 :

$$||y_n|| \leq e^{TL}||y_0|| + \frac{e^{TL} - 1}{L} \max_{t \in [t_0, t_0+T]} ||f(t, 0)||$$

□

L'estimation de l'erreur donnée dans le Théorème 8, peu utilisable en pratique (car dépendant de la fonction y' a priori inconnue) peut être reformulée ou améliorée avec des hypothèses supplémentaires sur f :

Proposition 2. *Les notations et hypothèses du Théorème 8 sont conservées. On note K un compact de* $\mathbf{R}^m$ *tel que* $y([t_0, t_0 + T]) \subset K$ *et* $Q = [t_0, t_0 + T] \times K$. *Alors :*

$$||e_n|| \leq C \frac{e^{nL\Delta T} - 1}{L} + e^{nL\Delta T}||e_0||$$

avec

$$C = L\Delta T \sup_{(t,z) \in Q} ||f(t, z)|| + \sup_{z \in K} \omega(\Delta T, t \mapsto f(t, z))$$

Si de plus $f \in C^1(Q, \mathbf{R}^m)$, *on a pour tout* $n \in \{0, ..., N\}$

$$||e_n|| \leq \Delta T \int_{t_0}^{t_n} e^{L(t_n - z)} ||y''(z)|| dz + e^{nL\Delta T}||e_0||$$

et

$$||e_n|| \leq \frac{\Delta T}{2} \max_{(t,z) \in Q} ||f^{[1]}(t, z)|| \frac{e^{nL\Delta T} - 1}{L} + e^{nL\Delta T}||e_0||.$$

Démonstration. : La première inégalité revient seulement à déterminer l'expression de $\omega(\Delta T, y')$ en fonction de f. On a pour tout $(s, s') \in [t_0, t_0 + T]^2$:

$$\begin{aligned} y'(s') - y'(s) &= f(s', y(s')) - f(s, y(s)) \\ &= f(s', y(s')) - f(s', y(s)) + f(s', y(s)) - f(s, y(s)) \end{aligned}$$

et ainsi

$$\begin{aligned}|s'-s| \leq \Delta T \Rightarrow ||y'(s')-y'(s)|| &\leq L||y(s')-y(s)|| + \omega(\Delta T, t \mapsto f(t,y(s))) \\ &\leq L\Delta T \sup_{\xi\in[s,s']} ||y'(\xi)|| + \omega(\Delta T, t \mapsto f(t,y(s))) \\ &\leq L\Delta T \sup_{(t,z)\in Q} ||f(t,z)|| + \sup_{z\in K} \omega(\Delta T, t \mapsto f(t,z))\end{aligned}$$

Pour démontrer la deuxième inégalité, on remarque tout d'abord que

$$\forall n \in \{0,...,N-1\}, \quad y(t_{n+1}) = y(t_n) + \int_{t_n}^{t_{n+1}} y'(s)ds = y(t_n) + \Delta T f(t_n, y(t_n)) + \alpha_n$$

avec

$$\alpha_n = \int_{t_n}^{t_{n+1}} (\int_{t_n}^{s} y''(z)dz)ds$$

En raisonnant alors exactement comme dans la démonstration du Théorème 8 (avec α_n à la place de ε_n et en utilisant cette fois le Lemme 2), il vient

$$\begin{aligned}||e_n|| - e^{nL\Delta T}||e_0|| \leq \sum_{i=0}^{n-1} e^{L(t_n-t_{i+1})}|\alpha_i| &\leq \sum_{i=0}^{n-1} e^{L(t_n-t_{i+1})} \int_{t_i}^{t_{i+1}} (\int_{t_i}^{s} ||y''(z)||dz)ds \\ &\leq \sum_{i=0}^{n-1} \int_{t_i}^{t_{i+1}} (\int_{t_i}^{t_{i+1}} e^{L(t_n-z)}||y''(z)||dz)ds \\ &\leq \Delta T \int_{t_0}^{t_n} e^{L(t_n-z)}||y''(z)||dz.\end{aligned}$$

La dernière inégalité se déduit de la même façon que la précédente en utilisant la majoration de α_n suivante :

$$||\alpha_n|| \leq \frac{\Delta T^2}{2} \max_{s\in[t_0,t_0+T]} ||y''(s)|| \leq \frac{\Delta T^2}{2} \max_{(t,z)\in Q} ||f^{[1]}(t,z)||$$

□

1.7.3 Généralisations de la méthode d'Euler explicite

1.7.3.1 Pas de temps variable

La première généralisation possible consiste à étendre la méthode d'Euler à des pas de temps non constants

$$t_0 < t_1 < ... < t_N = t_0 + T.$$

Dans ce cas, pour tout $n \in \{0, ..., N-1\}$, on note $h_n = t_{n+1} - t_n$ et

$$\delta_N = \max_{0 \leq n \leq N-1} h_n.$$

En gardant la même définition (5) pour la suite $(y_n)_{0 \leq n \leq N}$ (avec h_n à la place de ΔT), on montre un résultat équivalent à celui du Théorème 8 (en utilisant le Lemme 2 à la place du Lemme 1), à savoir :

$$\forall n \in \{0, ..., N\} \quad ||e_n|| \leq \frac{e^{L(t_n - t_0)} - 1}{L} \omega(\delta_N, y') + e^{L(t_n - t_0)} ||e_0||$$

et le même corollaire que précédemment se déduit en remplaçant simplement ($\lim_{N \to +\infty}$) par ($\lim_{\substack{N \to +\infty \\ \delta_N \to 0}}$).

Remarque : La possibilité de pouvoir utiliser un pas de temps variable est très utile pour simuler certains problèmes raides (voir exemples ultérieurs). Il est alors possible de réduire localement le pas de temps près des singularités de façon à conserver une erreur de méthode de l'ordre de la précision exigée avec un temps de calcul raisonnable.

1.7.3.2 Affaiblissement des hypothèses sur f

On peut également généraliser la définition et la convergence de la méthode d'Euler à partir d'hypothèses plus faibles sur f :

Théorème 14. *: on suppose qu'il existe une fonction $y \in C^0([t_0, t_0 + T], \mathbf{R}^m)$ et une fonction f définie et continue au voisinage de chaque point $(t, y(t))$ telle que y soit l'unique solution du problème (1). Alors, la méthode d'Euler est encore correctement définie pour N assez grand par la relation (5) avec $y_0 = y(t_0)$ et demeure convergente.*

Démonstration. : On suppose pour commencer que f est définie, continue et bornée sur $[t_0, t_0 + T] \times \mathbf{R}^m$. On construit les deux fonctions f_N et u_N de $[t_0, t_0 + T]$ dans $\mathbf{R}^m$, respectivement constante par morceaux et affine par morceaux, telles que

$$\begin{cases} f_N(t_n) = f(t_n, y_n) \quad \text{lorsque } n \in \{0, ..., N-1\}, \\ u_N(t_n) = y_n \quad \text{lorsque } n \in \{0, ..., N\}, \end{cases}$$

où la suite $(y_n)_{0 \leq n \leq N}$ est construite à partir de (5) avec $y_0 = y(t_0)$. Par construction, on a la relation suivante :

$$\forall t \in [t_0, t_0 + T], \quad u_N(t) = y(t_0) + \int_{t_0}^{t} f_N(s) ds. \tag{6}$$

On en déduit que la famille de fonctions $(u_N)_{N\in\mathbf{N}}$ est bornée et équicontinue dans $C^0([t_0,t_0+T],\mathbf{R})$. Par le théorème d'Ascoli (voir [Br]), soit $u\in C^0([t_0,t_0+T],\mathbf{R}^m)$ la limite (au sens de la norme infinie) d'une suite extraite $(u_{\alpha(N)})_{N\in\mathbf{N}}$. On montre par une méthode de découpe en trois analogue à celle utilisée précédemment, que la famille de fonctions $(f_{\alpha(N)})_{N\in\mathbf{N}}$ converge uniformément vers la fonction $t\mapsto f(t,u(t))$ lorsque $N\to+\infty$. Par passage à la limite dans l'égalité (6), on obtient alors

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

et par unicité de la solution y du problème (1), il vient $u\equiv y$ puis par unicité de la limite extraite, $y=\lim\limits_{N\to+\infty} u_N$. La convergence de la méthode d'Euler à savoir

$$\lim_{N\to+\infty}(\max_{0\le n\le N}|y_n-y(t_n)|)=0$$

est en particulier démontrée dans ce cas.

Dans le cas général, il existe $\varepsilon>0$ tel que f soit définie, continue et bornée sur l'ensemble $K=\{(t,z)\in[t_0,t_0+T]\times\mathbf{R}^m\ /\ |z-y(t)|\le\varepsilon\}$. Par un théorème de prolongement, il existe une fonction $\tilde{f}$ continue et bornée sur $[t_0,t_0+T]\times\mathbf{R}^m$ coincidant avec f sur K. Soit alors pour tout $N\in\mathbf{N}^*$, la suite $(\tilde{y}_n)_{0\le n\le N}$ telle que

$$\begin{cases}\tilde{y}_0=y(t_0),\\ \tilde{y}_{n+1}=\tilde{y}_n+\Delta T\tilde{f}(t_n,\tilde{y}_n)\quad 0\le n\le N-1.\end{cases}$$

Étant ici ramené au cas précédent, on en déduit qu'il existe un entier N_0 tel que

$$\forall N\in\mathbf{N},\quad N\ge N_0\Rightarrow\forall n\in\{0,\dots,N\},\quad |y(t_n)-\tilde{y}_n|\le\varepsilon$$

Ceci implique en particulier que pour $N\ge N_0$, la suite $(y_k)_{0\le k\le N}$ est correctement définie par la relation (5) et coincide avec la suite $(\tilde{y}_k)_{0\le k\le N}$. L'existence et la convergence de la méthode d'Euler sont donc aussi démontrées dans le cas général □

1.7.4 Méthode d'Euler implicite

Il existe une version implicite de la méthode d'Euler qui peut s'avérer à l'usage mieux conditionnée que la version explicite (voir exemple ultérieur). La justification de sa construction peut être trouvée dans [CrMi]. Définition : On appelle méthode

de résolution approchée d'Euler implicite (à pas de temps variable) de l'équation (1),

la construction valable pour tout $N \in \mathbf{N}^*$ tel que $\delta_N L < 1$ d'une suite finie $(y_n)_{0\leq n\leq N}$ vérifiant

$$\begin{cases} y_0 \in \mathbf{R}^m \text{ fixé,} \\ y_{n+1} = y_n + h_n f(t_n, y_{n+1}) \quad 0 \leq n \leq N-1. \end{cases} \tag{7}$$

Théorème 15. *: la méthode d'Euler implicite est un cas particulier de méthode explicite à un pas. Elle est consistante, stable, convergente et d'ordre 1 (si f est C^1). De plus, en notant $L_1 = \dfrac{L}{1-L\delta_N}$, on a une estimation de l'erreur commise :*

$$\forall n \in \{0,...,N\}, \quad ||e_n|| \leq \frac{e^{L_1(t_n-t_0)}-1}{L_1}\omega(\delta_N, y') + e^{L_1(t_n-t_0)}||e_0||$$

si f est seulement continue. Si f est continûment dérivable, on a de plus

$$||e_n|| \leq \frac{\delta_N}{1-L\delta_N}\int_{t_0}^{t_n} e^{L_1(t_n-s)}||y''(s)||ds + e^{L_1(t_n-t_0)}||e_0||.$$

Démonstration. On se limite ici à la démonstration des deux estimations d'erreur (voir par exemple [Sch] pour une preuve complète). Pour cela, on écrit

$$e_{n+1} = e_n + h_n\big(f(t_{n+1}, y_{n+1}) - f(t_{n+1}, y(t_{n+1})\big) + \varepsilon_n, \quad n \in \{0,...,N-1\}$$

avec

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

En particulier

$$||e_{n+1}|| \leq ||e_n|| + h_n L||e_{n+1}|| + ||\varepsilon_n||.$$

On remarque alors que

$$0 < (1-h_nL)^{-1} \leq 1 + h_n(\frac{L}{1-L\delta_N}) = 1 + h_nL_1$$

et en utilisant à nouveau le Lemme 2, il vient

$$||e_n|| \leq e^{L_1(t_n-t_0)}||e_0|| + \sum_{i=0}^{n-1} e^{L_1(t_n-t_{i+1})}||\varepsilon_i||$$

ce qui permet ensuite de conclure en raisonnant comme dans la Proposition 10 □

Remarque : Il est également possible de construire la méthode d'Euler implicite à partir d'hypothèses plus faibles sur f. On peut seulement se contenter de supposer que $f \in C^0([t_0, t_0+T]\times \mathbf{R}^m, \mathbf{R}^m)$ et qu'il existe $l \in \mathbf{R}$ tel que

$$\forall (y,z) \in (\mathbf{R}^m)^2, \quad \forall t \in [t_0, t_0+T], \quad < f(t,y) - f(t,z), y-z > \leq l||y-z||_2^2$$

pour la norme $||.||_2$ associé au produit scalaire euclidien $<.,.>$ sur $\mathbf{R}^m$. On montre alors que l'existence et l'unicité de y solution de (1) est toujours assurée et de plus que la méthode d'Euler implicite est bien définie sous la condition $l\delta_N < 1$. Enfin, les estimations du Théorème 12 sont encore valables avec L remplacé par l (voir [CrMi]). Cette extension est en particulier intéressante lorsque $l < 0$ car dans ce cas l'erreur de la méthode d'Euler implicite décroit lorsque n croit.

1.7.5 Autres méthodes d'Euler

Il existe d'autres méthodes d'Euler (Euler Cauchy, Euler modifiée) dont le principe consiste toujours à approcher $\int_{t_n}^{t_{n+1}} y'(t)dt$ par une méthode de quadrature élémentaire, simplement différente de la méthode des rectangles à gauche ou à droite. Elles seront vues au paragraphe suivant dans le cadre général des méthodes de Runge-Kutta.

1.8 Résolution approchée : les méthodes de Runge-Kutta

La famille des méthodes de Runge-Kutta explicites construites dans ce paragraphe (et à laquelle appartient la méthode d'Euler explicite) est un exemple important de méthodes explicites à un pas.

1.8.1 Définition

On suppose pour simplifier que $m = 1$ et que le pas de temps est constant :

$$\begin{cases} \Delta T = \dfrac{T}{N}, \\ t_n = t_0 + n\Delta T \quad (0 \leq n \leq N). \end{cases}$$

Définition : Soit $q \in \mathbf{N}^*$ et $(a_{i,j})_{1\leq j<i\leq q}$, $(b_i)_{1\leq i\leq q}$ et $(c_i)_{1\leq i\leq q}$ trois familles de coefficients réels (avec $c_i \in [0,1]$). On appelle méthode de résolution approchée de Runge-Kutta explicite à un pas (et à pas de temps constant) du problème (1), la construction pour tout $N \in \mathbf{N}^*$ de la suite finie $(y_n)_{0\leq n\leq N}$ telle que

$$\begin{cases} y_0 \in \mathbf{R}^m \text{ fixé}, \\ t_{n,j} = t_n + c_j \Delta T, \quad 1 \leq j \leq q, \\ y_{n,i} = y_n + \Delta T \sum_{j=1}^{i-1} a_{i,j} f(t_{n,j}, y_{n,j}), \quad 1 \leq i \leq q, \\ y_{n+1} = y_n + \Delta T \sum_{j=1}^{q} b_j f(t_{n,j}, y_{n,j}), \quad 0 \leq n \leq N-1. \end{cases} \tag{8}$$

On représente conventionnellement une méthode de Runge-Kutta par le tableau de ses coefficients rangés de la manière suivante :

$$\begin{array}{c|cccc} c_1 & 0 & & & \\ c_2 & a_{2,1} & 0 & & \\ c_3 & a_{3,1} & a_{3,2} & 0 & \\ \vdots & \vdots & \vdots & \ddots & \ddots \\ c_q & a_{q,1} & a_{q,2} & \dots & a_{q-1,q}\ 0 \\ \hline & b_1 & b_2 & \dots & b_{q-1}\ b_q \end{array}$$

Remarque : Il est également possible de définir des méthodes de Runge-Kutta implicites. Dans ce cas, le tableau des coefficients n'est plus triangulaire inférieur strict mais rectangulaire.

1.8.2 Interprétation

Pour mieux comprendre la définition des méthodes de Runge-Kutta, les formules (8) sont à rapprocher de celles valables pour toute solution exacte y du problème (1) :

$$\begin{cases} y(t_{n,i}) = y(t_n) + \Delta T \int_0^{c_i} f(t_n + u\Delta T, y(t_n + u\Delta T))du, \quad 1 \leq i \leq q, \\ y(t_{n+1}) = y(t_n) + \Delta T \int_0^{1} f(t_n + u\Delta T, y(t_n + u\Delta T))du \quad 0 \leq n \leq N-1. \end{cases}$$

Le principe des méthodes de Runge-Kutta consiste donc à approcher successivement par une méthode de quadrature, $y(t_{n,i})$ pour tout $i \in \{1,...,q\}$, puis $y(t_{n+1})$ à l'aide des précédentes valeurs calculées.

Ainsi, les familles de coefficients $(a_{i,j})_{1\leq j<i\leq q}$ et $(b_i)_{1\leq i\leq q}$ sont associées aux méthodes de quadrature :

$$\begin{cases} \displaystyle\int_0^{c_i} g(t)dt \simeq \sum_{j=1}^{i-1} a_{i,j} g(c_j), \quad 1 \leq i \leq q, \\ \displaystyle\int_0^{1} g(t)dt \simeq \sum_{j=1}^{q} b_j g(c_j). \end{cases}$$

On obtient ainsi deux premières conditions (qui seront supposées vérifiées dans toutes la suite) sur les familles de coefficients pour que les méthodes de quadrature soient au moins d'ordre 0 (c'est à dire exactes sur les constantes) :

$$\forall i \in \{1,...,q\}, \quad c_i = \sum_{j=1}^{i-1} a_{i,j} \tag{9}$$

et

$$1 = \sum_{j=1}^{q} b_j. \tag{10}$$

En particulier, ceci impose que $c_1 = 0$ (et aussi par conséquent $t_{n,1} = t_n$ et $y_{n,1} = y_n$).

1.8.3 Stabilité des méthodes de Runge-Kutta explicites

Les méthodes de Runge-Kutta explicites sont un cas particulier de méthodes explicites à un pas correspondant à une fonction Φ égale à :

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

où la famille $(y_i)_{1\leq i\leq q}$ est définie pout tout (t,y,h) par

$$y_i = y + h \sum_{j=1}^{i-1} a_{i,j} f(t + c_j h, y_j), \quad 1 \leq i \leq q. \tag{11}$$

On a alors le résultat suivant :

Proposition 3. *Les méthodes de Runge-Kutta explicites sont stables et la constante de stabilité $M = e^{\Lambda T}$ est donnée par les relations (12) et (13).*

Démonstration. : Grâce au Théorème 4, il suffit de montrer que Φ est Lipschitzienne par rapport à la variable y. Pour cela, soit

$$\alpha = \max_{1\leq i\leq q} \sum_{j=1}^{i-1} |a_{i,j}|. \tag{12}$$

Si $(y_i)_{1\leq i\leq q}$ et $(z_i)_{1\leq i\leq q}$ sont deux familles construites suivant la formule (11) (à partir de y et z respectivement), on montre aisément par récurrence l'inégalité

$$|y_i - z_i| \leq \left(1 + (\alpha L h) + \ldots + (\alpha L h)^{i-1}\right)|y - z|$$

où L désigne la constante de Lispschitz de f par rapport à sa deuxième variable. On peut alors écrire pour tout $(t,h) \in [t_0, t_0 + T] \times [0, \Delta T]$:

$$|\Phi(t,y,h) - \Phi(t,z,h)| \leq \sum_{j=1}^{q} |b_j| L |y_j - z_j| \leq \Lambda |y - z|$$

où

$$\Lambda = L \sum_{j=1}^{q} |b_j| \left(1 + (\alpha L \Delta T) + \ldots + (\alpha L \Delta T)^{j-1}\right), \tag{13}$$

ce qui achève la démonstration □

Remarque : On peut estimer le coefficient Λ lorsque le pas de temps ΔT est petit et lorsque les coefficients b_j sont tous positifs : en effet, dans ce cas et grâce à la relation (10), on a

$$\Lambda \leq L(1 + (\alpha L \Delta T) + \ldots + (\alpha L \Delta T)^{q-1}) = L \frac{1 - (\alpha L \Delta T)^q}{1 - \alpha L \Delta T} \simeq L$$

si $\Delta T \ll \dfrac{1}{\alpha L}$.

1.8.4 Convergence et ordre des méthodes de Runge-Kutta explicites

Théorème 16. *: toute méthode de Runge-Kutta explicite vérifiant (9) et (10) est convergente et d'ordre* 1 *(si f est C^1). Une condition nécessaire et suffisante pour qu'elle soit d'ordre* 2 *(si f est C^2) est que ses coefficients vérifient en outre :*

$$\sum_{j=1}^{q} b_j c_j = \frac{1}{2}. \tag{14}$$

Démonstration. En reprenant l'expression de la fonction Φ donnée au sous-paragraphe précédent, on s'aperçoit aisément que

$$\Phi(t,y,0) = \sum_{j=1}^{q} b_j f(t,y) = f(t,y)$$

en utilisant la relation (10). Grâce à l'étude générale précédemment effectuée, on en déduit que toute méthode de Runge-Kutta vérifiant (9) et (10) est convergente car consistante et stable. De plus son ordre est au moins égal à 1 (si f est C^1) grâce au Théorème 7. Afin d'améliorer éventuellement cette valeur (si f est C^2), on calcule (toujours en se référant au Théorème 7)

$$\frac{\partial \Phi}{\partial h}(t,y,h) = \sum_{j=1}^{q} b_j \left(c_j \partial_t f(t + c_j h, y_j) + \partial_y f(t + c_j h, y_j) \frac{\partial y_i}{\partial h} \right)$$

avec

$$\frac{\partial y_i}{\partial h} = \sum_{k=1}^{j-1} a_{j,k} f(t + c_k h, y_k) + h \sum_{k=1}^{j-1} a_{j,k} \left(c_k \partial_t f(t + c_k h, y_k) + \partial_y f(t + c_k h, y_k) \frac{\partial y_k}{\partial h} \right).$$

En particulier,

$$\frac{\partial y_i}{\partial h}(t,y,0) = \sum_{k=1}^{j-1} a_{j,k} f(t,y) = c_j f(t,y)$$

et

$$\frac{\partial \Phi}{\partial h}(t,y,0) = \sum_{j=1}^{q} b_j c_j \left(\partial_t f(t,y) + \partial_y f(t,y) f(t,y) \right) = \left(\sum_{j=1}^{q} b_j c_j \right) f^{[1]}(t,y).$$

On obtient donc bien la condition nécessaire et suffisante (14) pour qu'une méthode de Runge-Kutta soit d'ordre 2 □

Remarque : Les conditions nécessaires et suffisantes sur les coefficients pour obtenir des méthodes de Runge-Kutta d'ordre supérieur à 2 deviennent rapidement très

complexes à exprimer. Elles sont données par les solutions d'un système polynomial à plusieurs indéterminées. Le recours à un logiciel de calcul formel comme Maple s'avère alors très utile : voir [GH] pour un exemple de détermination de méthodes de Runge-Kutta d'ordre 3 par résolution du système associé avec la méthode du résultant.

Exemples

(i) $q = 1$: on a nécessairement $b_1 = 1$ et la méthode se réduit à :

$$y_{n+1} = y_n + (t_{n+1} - t_n) f(t_n, y_n).$$

On retouve la méthode d'Euler explicite étudiée au paragraphe précédent. (ii) $q = 2$: pour tout $\alpha \in [0, 1]$, on construit des méthodes de Runge-Kutta d'ordre 2 avec le tableau

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

En particulier, lorsque $\alpha = 1$, la méthode (dite de Heun ou d'Euler Cauchy) s'écrit

$$y_{n+1} = y_n + \frac{\Delta T}{2} \left[f(t_n, y_n) + f\left(t_n + \Delta T, y_n + \Delta T f(t_n, y_n)\right) \right]$$

et lorsque $\alpha = \frac{1}{2}$, la méthode (dite d'Euler modifié ou du point milieu) s'écrit

$$y_{n+1} = y_n + \Delta T f\left(t_n + \frac{\Delta T}{2}, y_n + \frac{\Delta T}{2} f(t_n, y_n)\right).$$

(iii) $q = 4$: un exemple de méthode de Runge-Kutta fréquemment utilisé dans la pratique est le suivant :

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

On peut montrer que cette méthode est d'ordre 4 (voir par exemple [CrMi]).

Remarque : Lorsque $5 \leq q \leq 7$ (respectivement $8 \leq q$), on montre que les méthodes de Runge-Kutta explicites sont forcément d'ordre inférieur à $q - 1$ (respectivement $q - 2$).

1.9 Résolution approchée : mise en œuvre et difficultés

Beaucoup de problèmes applicatifs se ramènent au terme de leur modélisation à la résolution d'un système d'équations différentielles. Si on désire simuler celui-ci numériquement avec les méthodes exposées précédemment, il est indispensable au préalable de s'assurer que le problème est bien posé.

1.9.1 Problème bien posé mathématiquement et numériquement

La notion de problème bien (ou mal) posé peut s'entendre à deux niveaux. Le premier niveau est purement théorique : on dit qu'un problème de Cauchy est mathématiquement bien posé si la solution est unique et dépend continûment de la donnée initiale. Par exemple, le problème

$$\begin{cases} y'(t) = 2\sqrt{|y(t)|}, \quad t \in \mathbf{R}_+, \\ y(0) = 0 \end{cases}$$

est mathématiquement mal posé car il existe une infinité de solutions différentes de la solution nulle : en effet, pour tout $a \geq 0$

$$y(t) = \begin{cases} 0 & \text{si } t \in [0,a], \\ (t-a)^2 & \text{si } t > a \end{cases}$$

est solution. Le second niveau est d'ordre numérique : on dit qu'un problème de

Cauchy est numériquement bien posé s'il est bien posé mathématiquement et si la dépendance en la donnée initiale est numériquement contrôlable : par exemple, le problème

$$\begin{cases} y'(t) = 3y(t) - 1, \quad t \in [0,30], \\ y(0) = \dfrac{1}{3} \end{cases}$$

est mathématiquement bien posé mais numériquement mal posé : en effet, il possède bien une unique solution $y(t) = \frac{1}{3}$, mais la solution $\tilde{y}(t)$ obtenue en perturbant légèrement la donnée initiale, $\tilde{y}(0) = \frac{1}{3} + \varepsilon$, est égale à

$$\tilde{y}(t) = \frac{1}{3} + \varepsilon e^{3t}.$$

Par exemple, si $\varepsilon = 10^{-17}$ (précision machine de Scilab), on a $\tilde{y}(30) - y(30) \simeq 10^{22}$.

De manière manifeste, aucune simulation numérique sur cette équation ne pourra donc être accomplie. Une façon de s'assurer qu'un problème est numériquement bien posé pourra consister à déterminer une estimation a priori de la solution en fonction des données initiales.

1.9.2 Conditionnement

Lorsqu'un problème de Cauchy est numériquement bien posé, il est encore essentiel avant toute simulation de s'assurer qu'il est bien conditionné pour la méthode numérique envisagée, c'està dire que celle-ci peut approcher de manière satisfaisante la solution exacte avec un pas de temps raisonnable. Soit par exemple (voir [Dem]) le problème de Cauchy

$$\begin{cases} y'(t) = -150y(t) - 30, \quad t \in [0,1], \\ y(0) = \dfrac{1}{5} \end{cases}$$

Il est mathématiquement bien posé (solution $y(t) = \frac{1}{5}$) et aussi numériquement car la solution $\tilde{y}(t)$ correspondant à une donnée initiale $\tilde{y}(0) = \frac{1}{5} + \varepsilon$ est égale à :

$$\tilde{y}(t) = \frac{1}{5} + \varepsilon e^{-150t}.$$

La méthode d'Euler explicite à pas constant appliquée à ce problème donne :

$$y_{n+1} = y_n + \frac{1}{N}(-150y_n + 30) = (1 - \frac{150}{N})y_n + \frac{30}{N}, \quad n \in \{0, ..., N-1\}$$

soit

$$y_{n+1} - \frac{1}{5} = (1 - \frac{150}{N})(y_n - \frac{1}{5})$$

puis enfin

$$y_n - \frac{1}{5} = (1 - \frac{150}{N})^n (y_0 - \frac{1}{5}).$$

En prenant $N = 50$ et $y_0 = \frac{1}{5} + \varepsilon$, on obtient ainsi $y_n = \dfrac{1}{5} + (-2)^n \varepsilon$, approximation totalement aberrante (par exemple, $y_{50} \simeq \dfrac{1}{5} + 10^{15}\varepsilon$).

La condition sur le pas de temps pour obtenir une bonne approximation de la solution exacte est en fait $N > 75$. On peut donc légitimement considérer que le problème est mal conditionné pour la méthode d'Euler explicite. Par contre, le lecteur pourra vérifier qu'il est bien conditionné pour la méthode d'Euler implicite.

Pour s'assurer qu'un problème de Cauchy est bien conditionné pour la méthode d'Euler explicite, on pourra par exemple vérifier que e^{LT} n'est pas trop grand (il existe d'autres critères plus précis : cf [CrMi]). Un autre exemple de comparaison du conditionnement des méthodes usuelles est traité dans [Fer] sur l'exemple de l'équation

$$\begin{cases} y'(t) = 3\dfrac{y(t)}{t} - \dfrac{5}{t^3}, \quad t \in [1,5], \\ y(1) = 1. \end{cases}$$

1.9.3 Problèmes raides

Il existe certains problèmes numériquement bien posés mais mal conditionnés pour toutes les méthodes usuelles. On parle alors de problèmes raides. Ceux- ci interviennent dans de nombreux domaines applicatifs et doivent faire l'objet d'études préalables complexes : mécanique (problème à trois corps : voir [GH] pour une étude théorique), cinétique chimique (réactions très rapides, réactions nucléaires, photodissociation), etc... Une simulation numérique peut éventuellement s'adapter à ces circonstances particulières : localement, un pas de temps très faible devra être adopté. Les solveurs incorporés dans les logiciels Scilab (fonction `ode`) ou Matlab ont recours à des stratégies d'adaptation de pas de temps permettant, dans une certaine mesure, de parvenir à un résultat satisfaisant.

1.9.4 Notion d'erreurs

Il existe différents niveaux d'erreurs dans le traitement d'un problème applicatif complet. En imaginant que celui-ci est modélisé par une équation différentielle, elle même approchée à l'aide de la méthode d'Euler, on peut distinguer les différentes erreurs suivantes : tout d'abord, une erreur de modélisation peut être commise (par exemple, en négligeant dans un problème de mécanique le frottement dans l'air). Ensuite, la méthode d'Euler appliquée à l'équation différentielle obtenue rajoute une erreur pouvant être estimée par exemple par le Théorème 8 :

$$||e_n|| \leq \frac{e^{nL\Delta T} - 1}{L} \omega(\Delta T, y') + e^{nL\Delta T} ||y_0 - y(t_0)||$$

Le premier terme du second membre correspond à l'erreur de la méthode proprement dite et le second terme à l'effet d'une erreur dans l'estimation de $y(t_0)$. De plus,

d'autres erreurs d'arrondi (pouvant être quantifiées en fonction de la précision machine) se rajoutent à chaque évaluation de f et à chaque nouvelle estimation de y_n. Afin d'effectuer des simulations cohérentes, il est nécessaire de s'assurer que la somme de celles-ci reste inférieure à l'erreur de méthode. Enfin, si les résultats de la simulation sont confrontés à des résultats expérimentaux, une erreur de type expérimentale est à prendre en considération (par exemple en définissant des intervalles de confiance).