

Chapitre 3

Dérivation et intégration numérique

3.1 Introduction

Le but de ce chapitre est double : indiquer comment calculer de façon approchée une dérivée et décrire les méthodes de base pour le calcul numérique d'intégrales. Un point commun de ces deux thèmes qui vont nous occuper est l'utilisation de polynômes d'interpolation. L'idée sous-jacente est simple : la fonction avec laquelle on travaille est trop compliquée ou pas assez connue pour qu'on puisse faire des opérations simples comme la dérivation ou l'intégration, on choisit alors de la remplacer par un polynôme et on dérive ou on intègre celui-ci. On présentera assez rapidement les formules de dérivation approchée qui en résultent avec un aperçu sur les méthodes de différences finies. On passera plus de temps sur le calcul d'intégrales du type

$$\int_a^b f(x)w(x)dx \quad (3.1)$$

où w est une fonction poids (continue strictement positive sur $]a; b[$) et $f(x)w(x)$ est intégrable sur $[a; b]$. Les méthodes d'intégration numérique que nous décrirons consistent toutes à remplacer l'intégrale (3.1) par une expression le plus souvent de la forme :

$$\sum_{i=0}^N A_i f(x_i) \quad (3.2)$$

où les x_i sont des points distincts de $[a; b]$ et A_i des coefficients réels, le tout choisi pour que la différence

$$E = \int_a^b f(x)w(x)dx - \sum_{i=0}^N A_i f(x_i)$$

soit petite.

3.2 Dérivation numérique

Dans ce paragraphe, la fonction f n'est bien sûr pas connue par une formule explicite mais

- ou bien par ses valeurs sur un ensemble discret (en supposant les points assez proches pour que la notion de dérivée ait un sens)
- ou bien, le plus souvent, par un algorithme de calcul ou une formule compliquée qui permet, au moins en théorie, de la calculer en tout point. On suppose bien sûr que la dérivée n'est pas accessible par un procédé analogue.

Dans toute la suite, on supposera f connue ou calculable aux points $\dots, x_{i-2}, x_{i-1}, x_i, x_{i+1}, x_{i+2}, \dots$ qu'on supposera proches. On notera $h_i = x_{i+1} - x_i$.

3.2.1 Dérivée première

Supposons qu'on veuille calculer une valeur approchée de $f'(x_i)$. Une première idée, déjà évoquée dans le chapitre précédent, consiste à remplacer f par une fonction spline passant par les mêmes points et à prendre la dérivée de la fonction spline. D'un point de vue numérique, cette idée est très bonne, assez stable et donne de bons résultats en terme d'erreur. Malheureusement, elle est assez lourde à mettre en place, nécessite beaucoup de calculs. Elle a de plus un petit côté moralement désagréable : la dérivée est une notion purement locale, en ce sens que la dérivée de f au point x_i ne dépend que des valeurs prises par f au voisinage de x_i , alors que la fonction spline dépend globalement de f par l'intermédiaire d'un système linéaire qui fait intervenir toutes les valeurs de f .

On préfère donc le plus souvent utiliser une idée plus simple : on écrit un polynôme d'interpolation au voisinage du point x_i et on dérive celui-ci. Les formules vont varier en fonction du nombre de points qu'on choisit pour écrire le polynôme d'interpolation (en général 2 ou 3, plus rarement 4 ou 5).

Formules à deux points

Le polynôme d'interpolation sur les deux points x_i et x_{i+1} s'écrit :

$$P(x) = f(x_i) + f[x_i, x_{i+1}](x - x_i).$$

On a donc $P'(x) = f[x_i, x_{i+1}]$, ce qui fournit la

1 Formule décentrée à droite

$$f'(x_i) \simeq \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i}. \quad (3.3)$$

On a aussi la

2 Formule décentrée à gauche

$$f'(x_i) \simeq \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}}. \quad (3.4)$$

Formules à trois points

On choisit d'interpoler sur les points x_{i-1}, x_i, x_{i+1} (ce qui est moralement plus satisfaisant). Dans ce cas on a :

$$P(x) = f(x_{i-1}) + f[x_{i-1}, x_i](x - x_{i-1}) + f[x_{i-1}, x_i, x_{i+1}](x - x_{i-1})(x - x_i).$$

On a donc

$$P'(x) = f[x_{i-1}, x_i] + f[x_{i-1}, x_i, x_{i+1}](x_i - x_{i-1}),$$

ce qui fournit, après simplification la

Formule centrée

$$f'(x_i) \simeq f(x_{i+1}) \frac{h_{i-1}}{h_i(h_{i-1} + h_i)} + f(x_i) \left(\frac{1}{h_{i-1}} - \frac{1}{h_i} \right) - f(x_{i-1}) \frac{h_i}{h_{i-1}(h_{i-1} + h_i)} \quad (3.5)$$

La formule (3.5) se simplifie notablement dans le cas de points équidistants ($h_{i-1} = h_i = h$) pour donner la

Formule centrée-points équidistants

$$f'(x_i) \simeq \frac{f(x_{i+1}) - f(x_{i-1}))}{2h}. \quad (3.6)$$

Remarquons que la formule ci-dessus n'est autre que la moyenne des deux formules décentrées dans le cas de points équidistants.

Erreur

Pour le calcul théorique de l'erreur commise quand on remplace $f'(x_i)$ par l'une des formules approchées ci-dessus, on revient à la démonstration effectuée dans le cadre de l'interpolation. On obtient :

$$\left| f'(x_i) - \frac{f(x_i) - f(x_{i-1}))}{x_i - x_{i-1}} \right| \leq \frac{M_2}{2} h_{i-1}. \quad (3.7)$$

On a de même

$$\left| f'(x_i) - \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} \right| \leq \frac{M_2}{2} h_i \quad (3.8)$$

et dans le cas centré, avec les points équidistants

$$\left| f'(x_i) - \frac{f(x_{i+1}) - f(x_{i-1}))}{2h} \right| \leq \frac{M_3}{6} h^2 \quad (3.9)$$

3.2.2 Dérivée d'ordre supérieur

Le principe est exactement le même, il faut simplement prendre garde que le degré du polynôme d'interpolation soit suffisant pour que sa dérivée n -ième soit non nulle ! Par exemple, pour la dérivée seconde, on choisit en général d'interpoler sur 3 points, ce qui donne (dans le cas de points équidistants)

Dérivée seconde - points équidistants

$$f''(x_i) \simeq \frac{f(x_{i+1}) + f(x_{i-1}) - 2f(x_i)}{h^2}. \quad (3.10)$$

avec une erreur majorée par

$$\left| f''(x_i) - \frac{f(x_{i+1}) + f(x_{i-1}) - 2f(x_i)}{h^2} \right| \leq \frac{M_4}{12} h^2 \quad (3.11)$$

Remarque 3.1. *Application : les méthodes de différences finies. Les formules ci-dessus sont à la base des méthodes de différences finies pour résoudre de façon approchée des équations différentielles ou aux dérivées partielles.*

3.3 Intégration numérique : méthodes composites

3.3.1 Principe

On veut calculer $\int_a^b f(x)dx$. On décompose l'intervalle $[a; b]$ en $a = x_0 < x_1 < \dots < x_n = b$. On a alors :

$$\int_a^b f(x)dx = \sum_{i=1}^n \int_{x_{i-1}}^{x_i} f(x)dx.$$

Sur chaque $[x_{i-1}; x_i]$, on applique une méthode d'intégration élémentaire consistant à remplacer $f(x)$ par

$$P_i = f(x_{i,0}) + f[x_{i,0}, x_{i,1}](x - x_{i,0}) + \dots + f[x_{i,0}, x_{i,1}, \dots, x_{i,l}](x - x_{i,0}) \dots (x - x_{i,l}) \quad (3.12)$$

le polynôme d'interpolation de f en des points $x_{i,0}, \dots, x_{i,l}$ de l'intervalle $[a; b]$ (qui peuvent être ou non dans $[x_{i-1}, x_i]$).

3.3.2 Méthode des rectangles

On prend $l = 0$: on approche f par une constante $f(x_{i,0})$ où $x_{i,0} \in [x_{i-1}; x_i]$. D'où la formule de quadrature (i.e. intégration numérique) élémentaire :

$$\int_{x_i}^{x_{i+1}} f(x)dx \simeq h_i f(\xi_i), \quad \xi_i \in [x_{i-1}; x_i], \quad h_i = x_i - x_{i-1} \quad (3.13)$$

et la formule de quadrature composée :

$$\int_a^b f(x)dx \simeq \sum_{i=1}^n h_i f(\xi_i), \quad (3.14)$$

On reconnait là une somme de Riemann dont on sait qu'elle converge vers $\int_a^b f(x)dx$ quand $\max_{1 \leq i \leq n} h_i \rightarrow 0$ si f est Riemann-intégrable. Les choix courants pour ξ_i sont :

— rectangles à gauche

$$\xi_i = x_{i-1} \Rightarrow \int_a^b f(x)dx \simeq \sum_{i=1}^n h_i f(x_{i-1})$$

— rectangles à droite

$$\xi_i = x_i \Rightarrow \int_a^b f(x)dx \simeq \sum_{i=1}^n h_i f(x_i)$$

— formule du point milieu

$$\xi_i = \frac{x_{i-1} + x_i}{2} = x_{i-\frac{1}{2}} \Rightarrow \int_a^b f(x)dx \simeq \sum_{i=1}^n h_i f(x_{i-\frac{1}{2}})$$

3.3.3 Méthode des trapèzes

On prend $l = 1$ et on remplace f par son interpolé linéaire aux points $[x_{i-1}; x_i]$

$$\int_{x_i}^{x_{i-1}} f(x)dx \simeq \frac{1}{2}(x_i - x_{i-1})(f(x_i) + f(x_{i-1}))$$

$$\int_a^b f(x)dx \simeq \sum_{i=1}^{n-1} \frac{h_i + h_{i+1}}{2} f(x_i) + \frac{1}{2}(h_1 f(x_0) + h_n f(x_n))$$

et dans le cas d'une subdivision régulière ($h_i = h, \forall i = 1, \dots, n$)

$$\int_a^b f(x)dx \simeq h \left[\sum_{i=1}^{n-1} f(x_i) + \frac{1}{2}(f(x_0) + f(x_n)) \right]$$

3.3.4 Méthode de Simpson

Cette fois on interpole f aux points x_{i-1}, x_i et $x_{i-\frac{1}{2}} = \frac{x_{i-1} + x_i}{2}$ soit, en utilisant la formule d'interpolation de Newton (2.8)

$$\begin{aligned} \int_{x_i}^{x_{i-1}} f(x) dx &\simeq \int_{x_i}^{x_{i-1}} f(x_{i-1}) + f[x_{i-1}, x_i](x - x_{i-1}) + f[x_{i-1}, x_i, x_{i-\frac{1}{2}}](x - x_{i-1})(x - x_i) dx \\ &= h_i f(x_{i-1}) + \frac{1}{2} h_i^2 f[x_{i-1}, x_i] + f[x_{i-1}, x_i, x_{i-\frac{1}{2}}] \int_{x_i}^{x_{i-1}} (x - x_{i-1})(x - x_i) dx \end{aligned}$$

or

$$\int_{x_i}^{x_{i-1}} (x - x_{i-1})(x - x_i) dx = \frac{1}{3} h_i^3 - \frac{1}{2} h_i^3$$

(en écrivant $x - x_i = x - x_{i-1} + x_{i-1} - x_i$)

et

$$\begin{aligned} h_i^2 f[x_{i-1}, x_i] &= h_i (f(x_i) - f(x_{i-1})) \\ h_i^3 f[x_{i-1}, x_{i-\frac{1}{2}}, x_i] &= h_i^2 \left(\frac{f(x_i) - f(x_{i-\frac{1}{2}})}{\frac{h_i}{2}} - \frac{f(x_{i-\frac{1}{2}}) - f(x_{i-1})}{\frac{h_i}{2}} \right). \end{aligned}$$

On obtient donc

$$\int_{x_i}^{x_{i-1}} f(x) dx = \frac{h_i}{6} (f(x_{i-1}) + 4f(x_{i-\frac{1}{2}}) + f(x_i))$$

La formule composée dans le cas de pas constants devient

$$\int_a^b f(x) dx = \frac{h}{6} \left[f_0 + f_n + 2 \sum_{i=1}^{n-1} f_i + 4 \sum_{i=1}^n f_{i-\frac{1}{2}} \right]$$

où on note

$$f_i = f(x_i), \quad f_{i-\frac{1}{2}} = f(x_{i-\frac{1}{2}}) = f\left(\frac{x_i + x_{i-1}}{2}\right)$$

Remarque 3.2. Pour cette méthode, il est nécessaire de connaître f aux points x_i et en leurs milieux, soit en un nombre impair de points. De manière générale, on appelle Formules de Newton-Cotes (fermées) les formules composées lorsque, sur chaque intervalle $[x_{i-1}, x_i]$, f est interpolée en l points équidistants. Les cas $l = 1$ et $l = 2$ correspondent donc respectivement aux formules des trapèzes et de Simpson, le cas $l = 4$ est connu dans la littérature sous le nom de Boole-Villarceau.

3.4 Analyse de l'erreur dans les méthodes d'intégration

3.4.1 Théorie

Nous évaluons ici l'erreur faite lorsqu'on remplace $\int_a^b f(x)dx$ par $\int_a^b P(x)dx$ où P est le polynôme d'interpolation de f aux points $x_0, \dots, x_N$. On a vu au chapitre précédent que si $P_N(x)$ est le polynôme d'interpolation de f aux points $x_0, \dots, x_N$ alors :

$$f(x) - P_N(x) = f[x_0, x_1, \dots, x_N, x] \Psi_N \quad (3.15)$$

avec $\Psi_N = (x - x_0)(x - x_1) \dots (x - x_N)$.

Ainsi

$$E(f) = \int_a^b f(x)dx - \int_a^b P_N(x)dx = \int_a^b f[x_0, x_1, \dots, x_N, x] \Psi_N(x)dx \quad (3.16)$$

L'expression de $E(f)$ se simplifie dans deux cas particuliers.

Premier cas Ψ_N est de signe constant sur $[a, b]$, on peut alors utiliser le théorème de la moyenne pour les intégrales qui affirme que, si g et h sont continues sur $[a, b]$ et si $h(x) \geq 0$ sur $[a, b]$, il existe $\xi \in [a, b]$ tel que

$$\int_a^b g(x)h(x)dx = g(\xi) \int_a^b h(x)dx$$

Appliquer à (3.16), ceci montre l'existence de $\xi \in [a, b]$ tel que

$$E(f) = f[x_0, x_1, \dots, x_N, \xi] \int_a^b \Psi_N(x)dx$$

D'après le lemme 2.2 du chapitre précédent, on en déduit encore qu'il existe $\eta \in [c, d]$ ($[c, d]$ est un intervalle contenant les x_i , a et b) tel que

$$E(f) = \frac{1}{(N+1)!} f^{N+1}(\eta) \int_a^b \Psi_N(x)dx \quad (3.17)$$

(On supposera ici et dans toute la suite que f est assez régulière).

Deuxième cas $\int_a^b \Psi_N(x)dx = 0$, dans ce cas, on utilise

$$f[x_0, x_1, \dots, x_N, x] = f[x_0, x_1, \dots, x_N, x_{N+1}] + (x_{N+1} - x)f[x_0, x_1, \dots, x_N, x_{N+1}, x] \quad (3.18)$$

valable pour x_{N+1} arbitraire. Alors (3.18) devient

$$E(f) = \int_a^b f[x_0, x_1, \dots, x_N, x_{N+1}] \Psi_N(x)dx + \int_a^b f[x_0, x_1, \dots, x_N, x_{N+1}, x] \Psi_{N+1}(x)dx$$

$$E(f) = \int_a^b f[x_0, x_1, \dots, x_N, x_{N+1}, x] \Psi_{N+1}(x) dx. \quad (3.19)$$

Si on peut choisir x_{N+1} de telle façon que $\Psi_{N+1}(x)$ reste de signe constant sur $[a, b]$, on aura d'après (3.17)

$$E(f) = \frac{1}{(N+2)!} f^{N+2}(\eta) \int_a^b \Psi_{N+1}(x) dx \quad (3.20)$$

pour un $\eta \in [c, d]$ ($[c, d]$ est un intervalle contenant les x_i , a et b)

Définition 3.1. On dit qu'une méthode d'intégration numérique est d'ordre k si elle est exacte pour tous les polynômes de degré inférieur ou égal à k .

Remarque 3.3. Lorsque $\Psi_N(x)$ est de signe constant sur $[a, b]$, d'après (3.17), on a une méthode d'ordre N puisque $P^{(N+1)}(\eta) = 0$ pour tout polynôme de degré inférieur ou égal à N . Dans la deuxième situation, si (3.17) est vraie, on aura même une méthode d'ordre $N+1$ (avec un même nombre de points).

3.4.2 Application aux méthodes usuelles

Méthode des rectangles à gauche ou à droite

$N = 0$, $x_0 = a$, $\Psi_0(x) = x - a$. On peut appliquer (3.17) qui donne

$$E(f) = f'(\eta) \int_a^b (x - a) dx = \frac{(b - a)^2}{2} f'(\eta)$$

La méthode est d'ordre 0 et exacte seulement pour les constantes.

$$\left| \int_{x_{i-1}}^{x_i} f(x) dx - h_i f(x_{i-1}) \right| \leq \frac{h_i^2}{2} f'(\eta_i) \leq h_i^2 2M_1$$

où $M_1 = \max_{a \leq x \leq b}$. Pour la formule composée, il faut ajouter les erreurs provenant de chaque intégration élémentaire. Dans le cas d'un pas constant, on obtient :

$$\left| \int_a^b f(x) dx - h \sum_{i=1}^N f(x_{i-1}) \right| \leq \frac{M_1}{2} (b - a) h$$

Formule du point milieu

$$N = 0, x_0 = \frac{a+b}{2}, \Psi_0(x) = x - \frac{a+b}{2}.$$

Ici, $\int_a^b \Psi_0(x) dx = 0$. On va donc obtenir une méthode d'ordre 1 (exacte pour toute fonction linéaire) bien qu'on interpole par une constante. En effet, d'après (3.20), appliqué

$$\text{avec } x_{N+1} = x_0 = x_1 = \frac{a+b}{2}$$

$$\left| \int_a^b f(x)dx - \frac{b-a}{2}(f(a) + f(b)) \right| \leq \frac{M_2}{2} \int_a^b \left(x - \frac{b-a}{2}\right)^2 dx = \frac{M_2}{24}(b-a)^3$$

avec $M_2 = \max_{a \leq x \leq b} |f''(x)|$. Pour la formule composée associée, on obtient

$$\left| \int_a^b f(x)dx - h \sum_{i=1}^N f(x_{i-\frac{1}{2}}) \right| \leq \frac{M_2}{24} h^2 (b-a)$$

Méthode des trapèzes

$N = 1, x_0 = a, x_1 = b, \Psi_1(x) = (x-a)(x-b)$.

Comme Ψ_1 est de signe constant sur $[a; b]$ on peut appliquer (3.17) pour trouver

$$\left| \int_a^b f(x)dx - \frac{b-a}{2}(f(a) + f(b)) \right| \leq \frac{|f''(\eta)|}{2} \int_a^b (x-a)(x-b)dx$$

soit

$$\left| \int_a^b f(x)dx - \frac{b-a}{2}(f(a) + f(b)) \right| \leq \frac{M_2}{12}(b-a)^3$$

Pour la formule composée de la méthode des trapèzes, on en déduit que l'erreur est majorée par

$$E(f) \leq \frac{M_2}{12} h^3 (b-a)$$

Méthode de Simpson

$N = 2, x_0 = a, x_1 = \frac{a+b}{2}, x_2 = b, \Psi_2(x) = (x-a)(x - \frac{a+b}{2})(x-b)$.

Par raison de symétrie, on vérifie que

$$\int_a^b \Psi_2(x)dx = 0$$

comme $\Psi_3(x) = (x-a)(x - \frac{a+b}{2})^2(x-b)$ est de signe constant sur $[a; b]$ on peut appliquer (3.20) pour obtenir

$$\left| \int_a^b f(x)dx - \frac{b-a}{6} \left(f(a) + 4f\left(\frac{b-a}{2}\right) + f(b) \right) \right| \leq \frac{M_4}{24} \int_a^b \Psi_3(x)dx$$

soit tous calculs faits

$$|E(f)| \leq \frac{M_4}{90} \left(\frac{b-a}{2}\right)^5$$

$$M_4 = \max_{a \leq x \leq b} |f^{(4)}(x)|.$$

La formule composée correspondante avec pas constant h donnera une erreur majorée par

$$|E(f)| \leq \frac{M_4}{2880} h^4 (b - a)$$

Par ailleurs, la méthode de Simpson (bien que reposant sur une interpolation à trois points) est exacte pour tout polynôme de degré inférieur ou égal à 3.

3.5 Exemple d'application

Exemple 3.1. Calcul de $\int_0^1 e^{-x^2} dx$.

Utilisons les méthodes élémentaires précédentes à l'aide des valeurs de $f(x) = e^{-x^2}$ aux points $0, \frac{1}{2}, 1$. soit

$$f(0) = 1, \quad f\left(\frac{1}{2}\right) = 0,77880, \quad f(1) = e - 1 = 0,36788.$$

On a les résultats

	rectangle	trapèzes	Simpson
$\int_0^1 e^{-x^2} dx$	1	0,68394	0,74718
erreur	0,25318	-0,06288	0,00036

La valeur exacte est 0,74682. Ces résultats sont tout à fait en accord avec les considérations théoriques précédentes. Noter comment la méthode de Simpson donne une approximation à $4 \cdot 10^{-4}$ avec seulement 3 valeurs de f . Ceci est bien sûr dû au fait que les dérivées d'ordre supérieur ne varient pas trop sur l'intervalle $[0, 1]$.

Exemple 3.2. Calcul d'une valeur approchée de $\ln 2 = \int_1^2 \frac{1}{t} dt$.

Si l'on pose, pour tout x de $[1, 2]$, $f(x) = \frac{1}{x}$, on a pour tout x de $[1, 2]$, $f'(x) = -\frac{1}{x^2}$ et $f''(x) = \frac{2}{x^3}$ et donc $\sup_{x \in [1, 2]} |f'(x)| = 1$ et $\sup_{x \in [1, 2]} |f''(x)| = 2$.

(1) Utilisation de la Méthode des rectangles

Avec $n = 10$, on obtient

$$\left| \ln 2 - \frac{1}{10} \left(1 + \frac{10}{11} + \frac{10}{12} + \frac{10}{13} + \frac{10}{14} + \frac{10}{15} + \frac{10}{16} + \frac{10}{17} + \frac{10}{18} + \frac{10}{19} \right) \right| \leq \frac{1}{20}.$$

Ce qui donne

$$\left| \ln 2 - \frac{33464927}{46558512} \right| \leq \frac{1}{20},$$

$$\text{puis} \quad \frac{33464927}{46558512} - \frac{1}{20} \leq \ln 2 \leq \frac{33464927}{46558512} + \frac{1}{20}.$$

On en déduit que

$$0.66877 \leq \ln 2 \leq 0.76877.$$

(2) Utilisation de la Méthode des trapèzes

Avec $n = 10$, on obtient Posons

$$A = \frac{1}{10} \left(1 + \frac{10}{11} + \frac{10}{12} + \frac{10}{13} + \frac{10}{14} + \frac{10}{15} + \frac{10}{16} + \frac{10}{17} + \frac{10}{18} + \frac{10}{19} \right).$$

$$\left| \ln 2 - \frac{1}{20} \left(\frac{1}{2} - 1 \right) - A \right| \leq \frac{1}{600}.$$

Ce qui donne

$$\left| \ln 2 - \frac{161504821}{232792560} \right| \leq \frac{1}{600},$$

$$\text{puis} \quad \frac{161504821}{232792560} - \frac{1}{600} \leq \ln 2 \leq \frac{161504821}{232792560} + \frac{1}{600}.$$

On en déduit que

$$0.6921 \leq \ln 2 \leq 0.69544.$$

Chapitre 4

Résolution numérique des équations différentielles

Les équations différentielles décrivent l'évolution de nombreux phénomènes dans des domaines variés. Une équation différentielle est une équation impliquant une ou plusieurs dérivées d'une fonction inconnue. Si toutes les dérivées sont prises par rapport à une seule variable, on parle d'équation différentielle ordinaire. Une équation mettant en jeu des dérivées partielles est appelée équation aux dérivées partielles.

4.1 Généralité

On appelle équation différentielle une équation reliant une fonction et ses dérivées successives. Si l'équation ne fait intervenir que la fonction et sa dérivée, on parle d'équation du premier ordre.

Soit Ω un ouvert de $\mathbb{R} \times \mathbb{R}^n$. Etant données $f : \Omega \subset \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ qui à tout couple $(t, y) \in \Omega$ associe $f(t, y)$ continu, un point $t_0 \in \mathbb{R}$, un point quelconque $Y_0 \in \mathbb{R}^n$, nous cherchons une fonction $y : I \rightarrow \mathbb{R}^n$, où I est un voisinage de t_0 dans $\mathbb{R}$, qui à associe $y(t)$, continûment dérivable, telle que y vérifie **le problème de Cauchy** suivant :

$$\begin{cases} y'(t) = f(t, y(t)) \\ y(t_0) = Y_0 \end{cases} \quad \text{condition initiale ou condition de Cauchy} \quad (4.1)$$

On dit qu'il y a unicité au problème de Cauchy en (t_0, Y_0) s'il existe au moins une solution à ce problème et si pour toutes solutions $\varphi : I \rightarrow \mathbb{R}^n$ et $\psi : J \rightarrow \mathbb{R}^n$, les fonctions φ et ψ coïncident sur $I \cap J$.

L'étude mathématique de l'existence et de l'unicité d'une solution y est délicate et constitue une branche entière des mathématiques. Nous nous contenterons de donner une condition suffisante à l'existence et l'unicité d'une solution au problème de Cauchy. Nous ne nous intéresserons ensuite qu'à la résolution numérique de ces équations différentielles.

CHAPITRE 4. RÉOLUTION NUMÉRIQUE DES ÉQUATIONS DIFFÉRENTIELLES

Théorème 4.1. *Théorème de Cauchy-Lipschitz (Admis)*

Si f est continue dans Ω et si f est localement lipschitzienne par rapport à sa deuxième variable, ie si pour tout $(t_0, Y_0) \in \Omega$ il existe un voisinage V de (t_0, Y_0) et $L > 0$ tels que

$$\forall (t, y), (t, z) \in V, |f(t, y) - f(t, z)| \leq |y - z|,$$

alors pour tout $(t_0, Y_0) \in \Omega$, il existe I voisinage de t_0 dans $\mathbb{R}$ et une application y continûment dérivable qe I dans $\mathbb{R}^n$ solution unique de

$$\begin{cases} y'(t) = f(t, y(t)) \\ y(t_0) = Y_0 \end{cases}$$

Remarque 4.1. Si f est de classe C^1 alors f est localement lipschitzienne par rapport à sa deuxième variable, et le théorème de Cauchy-Lipschitz s'applique.

4.2 Résolution numérique : méthode d'Euler

On suppose que les hypothèses du théorème (4.1) sont vérifiées. Nous cherchons à approcher les valeurs de la solution y de l'équation (4.1) pour différentes valeurs de $t \in I = [0, T]$. Soit h la longueur du pas, on définit $t_j = jh, j = 0, 1, 2, \dots, n$:

$$\begin{aligned} \int_{t_j}^{t_{j+1}} y'(t) dt &= \int_{t_j}^{t_{j+1}} f(t, y(t)) dt \\ y(t_{j+1}) - y(t_j) &= \underbrace{\int_{t_j}^{t_{j+1}} f(t, y(t)) dt}_{\text{intégration numérique}} \end{aligned}$$

On pose $y_j = y(t_j)$ et $y_{j+1} = y(t_{j+1})$

En utilisant la formule des rectangles à gauche, on obtient :

$$\begin{aligned} y(t_{j+1}) - y(t_j) &= (t_{j+1} - t_j)(f(t_j, y_j)) \\ y(t_{j+1}) &= y(t_j) + (t_{j+1} - t_j)(f(t_j, y_j)) \end{aligned}$$

On obtient le schéma d'Euler explicite (explicite car on a directement y_{j+1} en fonction de y_j) :

$$\begin{aligned} y_{j+1} &= y_j + hf(t_j, y_j) \\ y(0) &= Y_0 \end{aligned}$$

En utilisant la formule des rectangles à droite, on obtient :

$$\begin{aligned} y(t_{j+1}) - y(t_j) &= (t_{j+1} - t_j)(f(t_{j+1}, y_{j+1})) \\ y(t_{j+1}) &= y(t_j) + (t_{j+1} - t_j)(f(t_{j+1}, y_{j+1})) \end{aligned}$$

CHAPITRE 4. RÉOLUTION NUMÉRIQUE DES ÉQUATIONS DIFFÉRENTIELLES

On obtient le schéma d'Euler implicite (implicite car y_{j+1} est des deux côtés de l'équation) :

$$\begin{aligned}y_{j+1} &= y_j + hf(t_j, y_{j+1}) \\ y(0) &= Y_0\end{aligned}$$

Remarque 4.2. *Le schéma d'Euler explicite est plus simple à utiliser mais moins précis que le schéma d'Euler implicite. De plus, le schéma d'Euler implicite est plus stable que le schéma d'Euler explicite.*

Exemple 4.1. Soit $\alpha > 0$ et l'équation

$$\begin{aligned}y'(t) &= -\alpha y(t) \\ y(0) &= 1\end{aligned}$$

La fonction $y(t) = \exp(-\alpha t)$ est solution.

Par le schéma d'Euler explicite :

$$y_{j+1} = (1 - \alpha h)y_j \iff y_j = (1 - \alpha h)^n$$

Donc comme $y(t)$ tend vers 0 quand t tend vers l'infini, $|1 - \alpha h| < 1$, soit la condition de stabilité $h < \frac{2}{\alpha}$.

Par le schéma d'Euler implicite :

$$(1 + \alpha h)y_{j+1} = y_j \iff y_j = \frac{1}{(1 + \alpha h)^n}$$

Alors y_j tend vers 0 pour tout $h > 0$. Il n'y a pas de condition de stabilité.

Interprétation graphique du schéma d'Euler explicite

Comme $y(0) = f(0, y(0))$, on confond la courbe de $f : t \rightarrow f(t, y(t))$ et la tangente en 0 pour trouver le point suivant, A_1 . Puis on construit la tangente à la courbe au nouveau point et on trouve un point A_2 , etc...

Le θ -schéma

En combinant les schémas d'Euler explicite et implicite, on obtient le θ -schéma :

$$\begin{aligned}y_{j+1} &= y_j + h[\theta f(t_{j+1}, y_{j+1}) + (1 - \theta)f(t_j, y_j)], 0 \leq \theta \leq 1 \\ y(0) &= Y_0\end{aligned}$$

On retrouve le schéma d'Euler explicite pour $\theta = 0$, et l'implicite pour $\theta = 1$.

Remarque 4.3. *Pour les schémas implicites, la valeur de départ est en général donnée par un schéma explicite. Les schémas implicites sont à considérer comme des équations de point fixe.*

4.3 Méthode à un pas

Les méthodes à un pas s'écrivent

$$\begin{aligned}y_{j+1} &= y_j + h\phi(t_j, y_j, h), \\ y(0) &= Y_0\end{aligned}$$

avec ϕ continue par rapport aux trois variables. Choisir une méthode revient à choisir la fonction ϕ ! Quelles conditions imposer à ϕ pour que la méthode fonctionne ?

Exemple 4.2. Schéma d'Euler : $\phi(t, y, h) = f(t, y)$

On pose $e_j(h) = y(t_j) - y_j, j = 0, 1, \dots, n$: l'erreur à l'étape j .

Objectif 1 Quand h tend vers 0, $y(t_j)$ doit tendre vers y_j ie on a convergence de la solution approchée trouvée y_j vers la solution exacte au point $t_j, y(t_j)$.

Définition 4.1. Une méthode à un pas est dite **convergente** sur $[0, T]$ si quelle que soit Y_0 condition initiale

$$\begin{aligned}\lim_{h \rightarrow 0} \max_{0 \leq j \leq n} |e_j(h)| &= 0, \\ \lim_{h \rightarrow 0} \max_{0 \leq j \leq n} |y(t_j) - y_j| &= 0\end{aligned}$$

Objectif 2 Il faut que la méthode approche bien l'équation différentielle.

Définition 4.2. Une méthode à un pas est dite **consistante** si pour tout u solution de $y' = f(t, u)$

$$\lim_{h \rightarrow 0} \max_{0 \leq j \leq n} \left| \frac{1}{h} (y(t_j) - y_j) - \phi(t, y, h) \right| = 0$$

Objectif 3 Connaître et contrôler la répercussion les erreurs d'arrondi

Définition 4.3. Une méthode à un pas est dite **stable** si il existe K indépendant de h tel que , pour tous $Y_0, \tilde{Y}_0$ et $y_{j+1} = y_j + h\phi(t_j, y(t_j), h), y_{j+1} = \tilde{y}_j + h\phi(t_j, y(t_j), h) + \varepsilon_j$ vérifiant $|y_j - \tilde{y}_j| \leq K(|Y_0 - \tilde{Y}_0| + \sum_{i=0}^{j-1} |\varepsilon_i|) \forall 0 \leq j \leq n$.

Ces trois notions sont liées pour les schémas à un pas. En effet

Proposition 4.1. Pour une méthode à un pas, la convergence implique la consistance.

Théorème 4.2. Si une méthode à un pas est consistante et stable alors elle est convergente.

Proposition 4.2. Une condition nécessaire et suffisante pour qu'une méthode à un pas soit consistante est $\phi(t, y, h) = f(t, y, 0), \forall t \in [0, T], y \in \mathbb{R}^n$.

Justification

$$\lim_{h \rightarrow 0} \frac{y(t+h) - y(t)}{h} = y'(t) = f(t, y(t))$$

Comme ϕ est continue,

$$\lim_{h \rightarrow 0} \phi(t, y(t), h) = \phi(t, y(t), 0)$$

soit

$$\phi(t, y, h) = f(t, y, 0)$$

Proposition 4.3. *Si ϕ est localement lipschitzienne par rapport à sa deuxième variable, alors la méthode est stable*

Définition 4.4. *On dit qu'une méthode est d'ordre p si*

$$\max_j \left| \frac{1}{h} (y(t_{j+1}) - y(t_j)) - \phi(t_j, y(t_j), h) \right| = O(h^p)$$

4.4 Méthodes de Runge-Kutta

Les méthodes de Runge-Kutta visent à étendre à la résolution d'équations différentielles ordinaires l'usage des techniques de calcul approché d'intégrales que sont les formules de quadrature interpolatoires. Elles font pour cela appel à de multiples évaluations de la fonction f , en des points obtenus par substitutions successives (pour les méthodes explicites), sur chaque sous-intervalle de la grille de discrétisation, cet « échantillonnage » de la dérivée de la courbe intégrale recherchée permettant de réaliser l'intégration numérique approchée de cette dernière au moyen d'une somme pondérée des valeurs recueillies. Pour illustrer cette idée, donnons un premier exemple.

Exemple de méthode de Runge-Kutta explicite

On considère le problème

$$\begin{cases} y'(t) = f(t, y(t)), & t \in]t_0, t_f] \\ y(t_0) = Y_0, \end{cases}$$

et une subdivision $t_0 \leq t_1 \dots \leq t_N = t_0 + T$.

Sur l'intervalle $[t_n, t_{n+1}]$, on a $y(t_{n+1}) = y(t_n) + \int_{t_n}^{t_{n+1}} f(s, y(s)) ds$ soit encore

$$y(t_{n+1}) = y(t_n) + h_n \int_0^1 f(t_n + \sigma h_n, y(t_n + \sigma h_n)) d\sigma$$

ou bien

$$y(t_{n+1}) = y(t_n) + h_n \int_0^1 g(\sigma) d\sigma, \quad g(\sigma) = f(t_n + \sigma h_n, y(t_n + \sigma h_n)).$$

CHAPITRE 4. RÉOLUTION NUMÉRIQUE DES ÉQUATIONS DIFFÉRENTIELLES

On peut approcher le calcul de $\int_0^1 g(\sigma) d\sigma$ par la formule de quadrature (rectangle à gauche, rectangle à droite ou trapèze).

On veut maintenant utiliser une quadrature générale.

$$\int_0^1 g(\sigma) d\sigma \approx \sum_{i=1}^s b_i g(c_i), \quad g(c_i) = f(t_n + c_i h_n, y(t_n + c_i h_n))$$

Le problème est que l'on ne connaît pas $y(t_n + c_i h_n) := y(t_{n,i})$. On évalue la fonction y au point $t_{n,i} = t_n + c_i h_n$ par une quadrature

$$y(t_{n,i}) = y(t_n) + h_n \int_0^{c_i} g(\sigma) d\sigma$$

Pour simplifier la présentation, on choisit ici des méthodes explicites où on approche $\int_0^{c_i} g(\sigma) d\sigma$ avec des valeurs de $g(c_j) = f(t_{n,j}, y(t_{n,j}))$, $j = 1, \dots, i-1$

$$\int_0^{c_i} g(\sigma) d\sigma \approx \sum_{j=1}^{i-1} a_{ij} g(c_j)$$

et on a donc

$$y(t_{n,i}) \approx y(t_n) + h_n \sum_{j=1}^{i-1} a_{ij} f(t_{n,j}, y(t_{n,j}))$$

$$y(t_{n+1}) \approx y(t_n) + h_n \sum_{i=1}^s b_i f(t_{n,i}, y(t_{n,i}))$$

avec $t_{n,i} = t_n + c_i h_n$.

La méthode de Runge-Kutta à s-étages explicite est donnée par

$$\begin{cases} k_i = f(t_{n,i}, y_n + h_n \sum_{j=1}^{i-1} a_{ij} k_j), & i = 1, \dots, s \\ y_{n+1} = y_n + h_n \sum_{i=1}^s b_i k_i \end{cases}$$

la valeur x_0 étant donnée. Une telle méthode est donc entièrement caractérisée par la donnée des coefficients $(a_{ij})_{1 \leq i, j \leq s}$, $(b_i)_{1 \leq i \leq s}$ et $(c_i)_{1 \leq i \leq s}$, que l'on a coutume de présenter dans le tableau de Butcher

c_1	a_{11}	a_{12}	$\cdots$	a_{1s}
c_2	a_{21}	a_{22}	$\cdots$	a_{2s}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
c_s	a_{s1}	a_{ss}	$\cdots$	a_{ss}
	b_1	b_2	$\cdots$	b_s

Tableau de Butcher

CHAPITRE 4. RÉOLUTION NUMÉRIQUE DES ÉQUATIONS DIFFÉRENTIELLES

Remarque 4.4. *En se référant au tableau de Butcher, on voit qu'une méthode de Runge-Kutta est explicite si et seulement si la matrice A est strictement triangulaire inférieure. Si A est seulement triangulaire inférieure, c'est-à-dire si $a_{ij} = 0$ pour tout couple (i, j) de $1, \dots, s^2$ tel que $i < j$ mais qu'au moins l'un des coefficients a_{ii} , $1 \leq i \leq s$, est non nul, la méthode est semi-implicite.*

Bibliographie

- [1] **Gloria Faccanoni** : *Analyse numérique*. Année 2013 - 2014.