Notes du cours d'Analyse Numérique

Sup Galilée - MACS 1 - année 2008-2009

Benoît Merlet

Sources bibliographiques :

Ces notes de cours reprennent le plan et les développements des premiers chapitres du polycopié de Ernst Hairer (professeur à Genève) qui est disponible en ligne. Cependant, j'ai laissé de coté certaines notions. Je recommande donc la lecture du polycopié du professeur Hairer.

Le lecteur qui souhaite en apprendre plus sur l'approximation polynomiale, pourra consulter le polycopié du cours 2007-2008 de Martin Campos Pinto, lui même inspiré des notes manuscrites de Laurence Halpern. Ce polycopié est disponible sur le site de la Macs.

Pour la partie Algèbre linéaire du cours je recommande le livre *Algèbre linéaire numérique : Cours et exercices* de Grégoire Allaire et Mahmoud Kaber Sidi.

Table des matières

I Introduction

1 Qu'est-ce que le calcul scientifique ?

Le calcul scientifique se propose de prendre les modèles venus de la physique, la biologie, l'économie, $\cdots$ et de trouver les solutions de ces modèles à l'aide d'un *ordinateur*.

Les modèles considérés seront donc composés d'ensemble d'équations que dans la plupart des cas on ne saura pas résoudre *explicitement*. Si on ne sait pas les résoudre *à la main*, si on ne trouve pas de solution *analytique*, c'est souvent parce qu'il n'y en a pas. La solution existe bel et bien mais elle ne peut pas être décrite à l'aide de fonctions classiques.

À la place, on utilisera un ordinateur pour résoudre ces systèmes d'équations de manière *approchée* (avec une erreur petite).

Notre tâche est alors de trouver des méthodes *constructives* (qui permettent de construire effectivement une solution approchée en un nombre fini d'étapes de calcul) et efficaces.

On aura les deux exigences contradictoires suivantes :

- Le nombre de calculs élémentaires pour avoir la solution doit être le plus petit possible.
- La solution approchée construite devra être la plus proche possible de la solution exacte.

Définition 1.1 *Le nombre de calculs élémentaires effectués (par le processeur) lors de l'application d'une méthode numérique (additions, multiplications de réels ou d'entiers) est appelé* ***coût*** *de la méthode.*

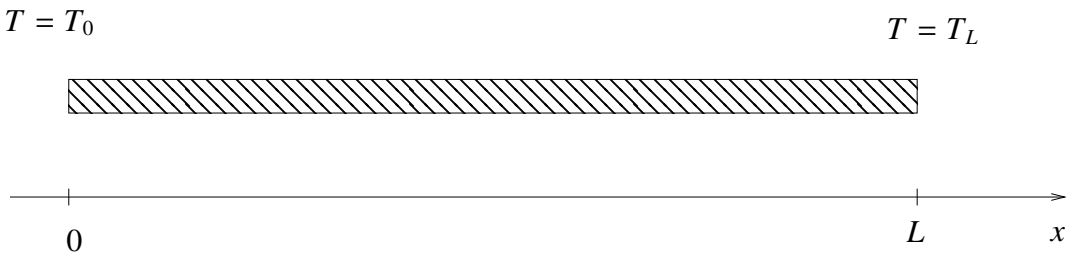
La différence entre la solution exacte et la solution approchée est appelée ***erreur*** *de la méthode numérique.*

La tâche de l'analyse numérique est de

- Développer des méthodes concrètes, des algorithmes de recherche de solutions efficaces.
- Montrer la convergence de ces méthodes : on montre quand on le peut que la solution dite approchée est effectivement proche de la solution recherchée. On évalue l'ordre de grandeur de l'erreur. C'est la partie la plus mathématique de l'analyse numérique, on fait des démonstrations.
- Évaluer le coût numérique des méthodes, comparer les méthodes entre elles.

2 Un exemple tiré de la Thermostatique

Soit une barre métallique de longueur L isolée dans sa longueur et dont on impose la température à ces deux extrémités.



La quantité de chaleur par unité de longueur $Q(x)$ vérifie l'équation

$$\frac{\partial Q}{\partial t} + \frac{\partial \phi}{\partial x} = 0 \qquad \Longrightarrow \qquad \frac{\partial \phi}{\partial x} = 0.$$

où ϕ est le flux de chaleur dans la direction x. La loi de Fick s'écrit

$$\phi = -k(x)\frac{dT}{dx} = 0,$$

où $k = k(x)$ est un coefficient qui dépend du matériau. On a donc à résoudre le *problème aux limites*

$$\text{(2.1)} \qquad \begin{cases} \dfrac{d}{dx}\left(-k(x)\dfrac{dT}{dx}\right) = 0, & \text{pour } 0 < x < L, \\ \\ T(0) = T_0, \qquad T(L) = T_L. \end{cases}$$

2.1 Cas d'une barre homogène : $k(x) = k_0 = Cste$.

Dans ce cas, le système (2.1) devient

$$\text{(2.2)} \qquad \begin{cases} \dfrac{d^2T}{dx^2} = 0, & \text{pour } 0 < x < L, \\ \\ T(0) = T_0, \qquad T(L) = T_L. \end{cases}$$

L'équation différentielle donne $T(x) = ax + b$ et on détermine a et b avec les conditions aux limites.

$$T(x) = T_0 + \frac{T_L - T_0}{L}x.$$

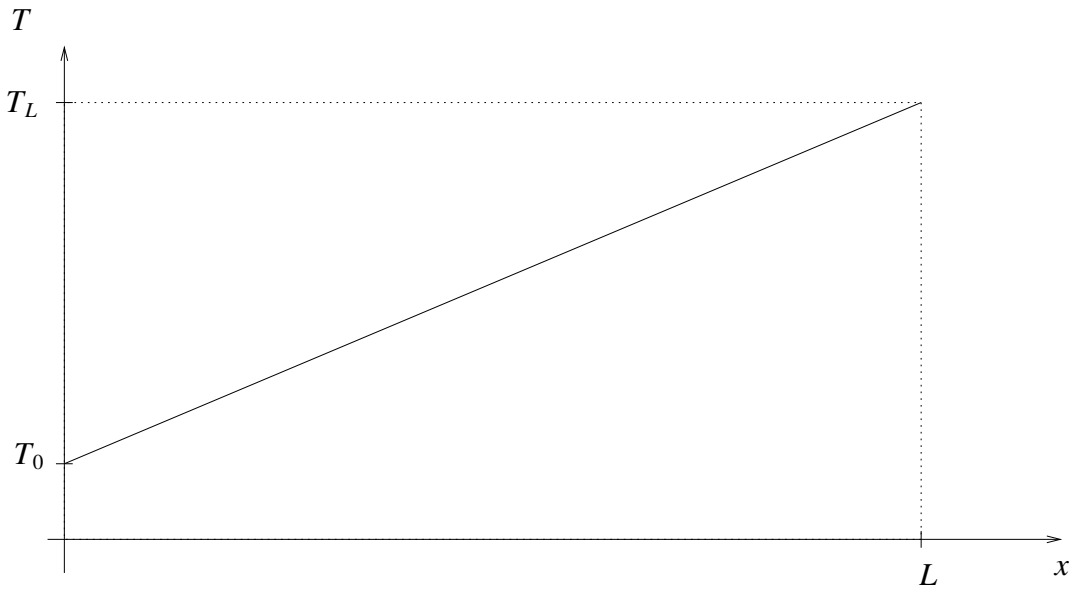


FIG. I.1 – solution

2.2 Cas où $k(x)$ est polynomiale.

On prend cette fois

$$k(x) \;=\; 1 + \frac{x}{L},$$

soit

$$\text{(2.3)} \qquad \begin{cases} \dfrac{d}{dx}\left\{\left(1+\dfrac{x}{L}\right)\dfrac{dT}{dx}\right\} \;=\; 0, & \text{pour } \; 0 < x < L, \\ \\ T(0) \;=\; T_0, \qquad T(L) = T_L. \end{cases}$$

L'équation différentielle donne

$$\left(1+\frac{x}{L}\right)\frac{dT}{dx} \;=\; c,$$

pour une constante $c \in \mathbf{R}$ à déterminer. Donc

$$T(x) \;=\; T_0 + c\int_0^x \frac{1}{1+y/L}\,dy \;=\; T_0 + cL\int_0^{x/L} \frac{1}{1+z}\,dz \;=\; T_0 + cL\ln\left(1+\frac{x}{L}\right).$$

En faisant $x = L$, on a $T_L = T_0 + cL\ln 2$ d'où $c = (T_L - T_0)/(L\ln 2)$. Finalement

$$T(x) \;=\; T_0 + \frac{T_L - T_0}{\ln 2}\ln\left(1+\frac{x}{L}\right) \qquad \text{pour } \; 0 < x < L.$$

2.3 Cas où $k(x)$ est un sinus cardinal

Cette fois k est de la forme

$$k(x) \;=\; \frac{\sin(\pi x/(2L))}{\pi x/(2L)}.$$

D'où

$$\frac{\sin(\pi x/(2L))}{\pi x/(2L)}\,\frac{dT}{dx} \;=\; c,$$

et

$$T(x) \;=\; T_0 + \int_0^x T'(y)\,dy \;=\; T_0 + c\int_0^x \frac{\pi y/(2L)}{\sin(\pi y/(2L))}\,dy$$

Pour déterminer c, on utilise $T(L) = T_L$, on obtient

$$T_L - T_0 \;=\; c\int_0^L \frac{\pi y}{2L\sin((\pi y)/(2L))}\,dy \;=\; c\frac{2L}{\pi}\int_0^{\pi/2} \frac{y}{\sin y}\,dy.$$

Malheureusement, il n'y a pas de formule analytique pour exprimer la valeur de la dernière intégrale.

Nous verrons des méthodes numériques pour approcher les intégrales dans un prochain chapitre.

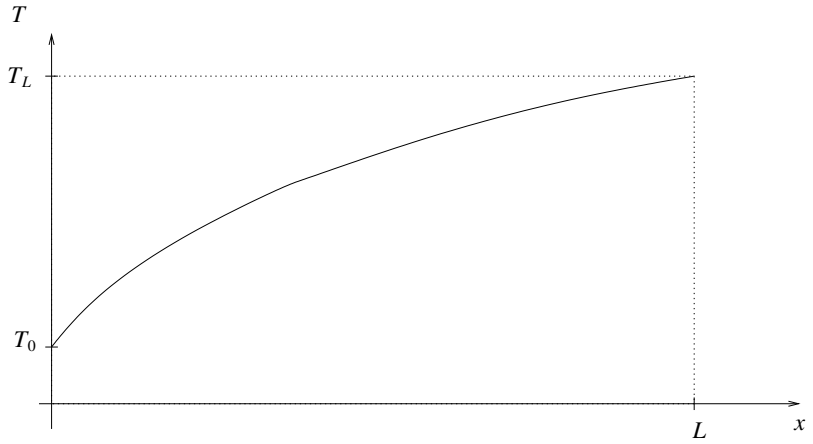
Par exemple, on peut utiliser une somme de Riemann

$$\int_a^b f(s)\,ds \;=\; \lim_{n\uparrow\infty} \frac{b-a}{n+1}\sum_{i=0}^{n} f\left(a + \frac{i}{n}(b-a)\right).$$

Et pour n assez grand on pourra approcher la quantité $I := \displaystyle\int_a^b f(s)\,ds$ par l'expression calculable en un nombre fini d'opérations.

$$S_n \;:=\; \frac{b-a}{n+1}\sum_{i=0}^{n} f\left(a + \frac{i}{n}(b-a)\right).$$

Plus n sera grand plus l'erreur $e_n := S_n - I$ sera petite, par contre plus le coût de calcul sera grand, ici $\hat{c} \sim n$.



Exercice 2.1

- *Donner un équivalent du nombre d'opérations élémentaires nécessaires pour calculer* S_n.
- *Supposons* f *de classe* C^1. *Montrer qu'il existe* $C > 0$ *tel que* $\sup_{[a,b]} |e_n| \leq C/n$ *pour* $n \geq 1$.

2.4 Cas où $k(x)$ n'est connu expérimentalement qu'en certains points

On ne dispose que de quelques mesures expérimentales donnant une valeur approchée k_i de $k(x)$ pour une suite de points x notés $0 \leq x_0 < x_1 < \cdots < x_n \leq L$.

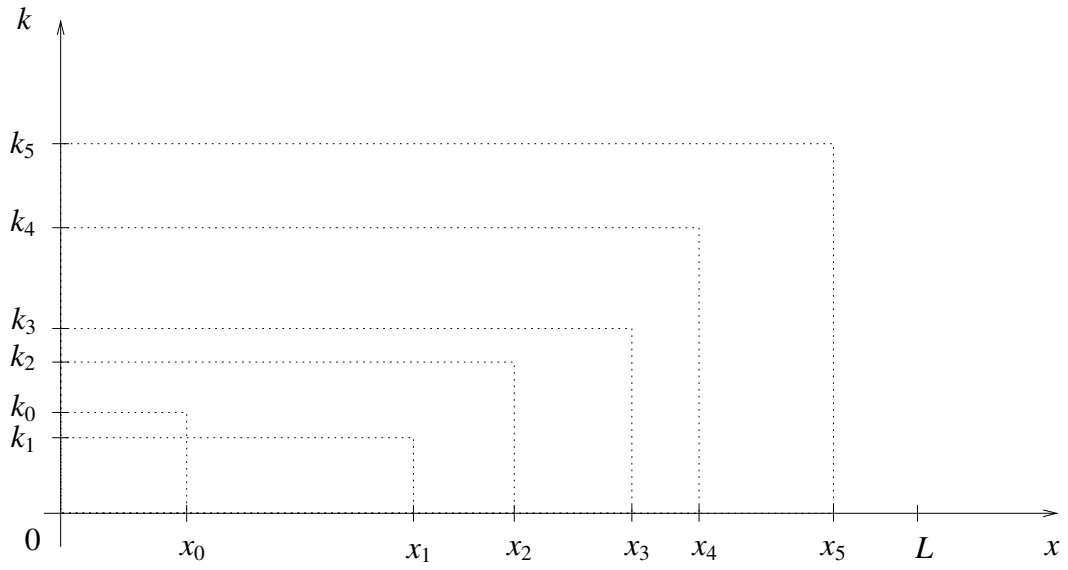


FIG. I.2 – points expérimentaux

Pour résoudre le probleme (2.1), nous avons besoin de connaître k sur tout l'intervalle $(0, L)$ et pas seulement en quelques points. Il y a plusieurs méthodes pour *reconstruire* k sur l'intervalle $[0, L]$, nous en verrons deux dans ce cours.

1. L'interpolation polynomiale. On approche k par un polynôme P qui interpole k aux points $x_1, \cdots, x_n$, i.e :

$$P(x_i) = k_i, \qquad \text{pour } i = 1, \cdots, n.$$

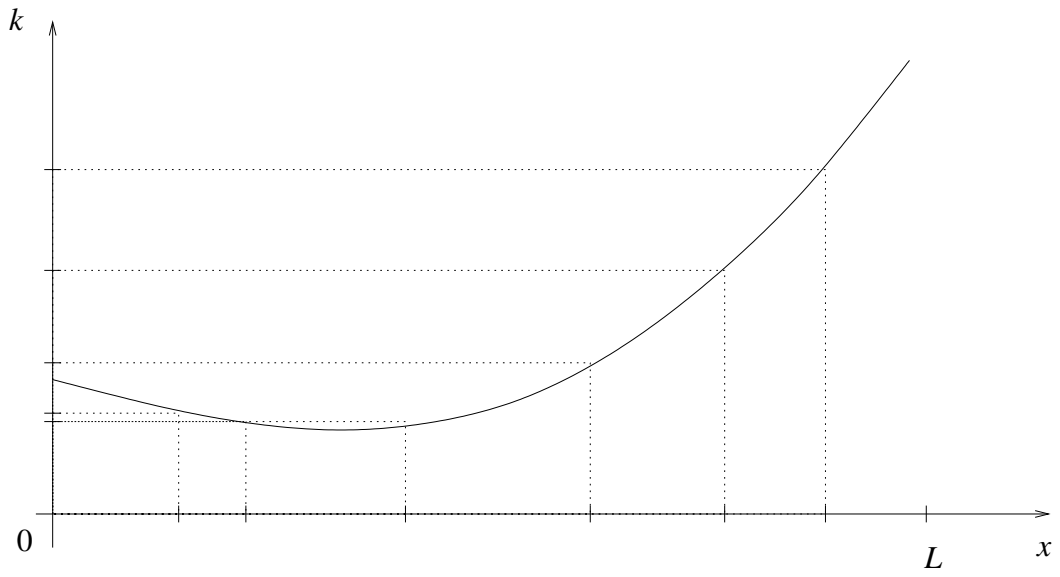


FIG. I.3 – interpolation polynomiale

2. L'approximation polynomiale. On approche k par un polynôme P en un certain sens. Par exemple, on va chercher le polynôme de degré infèrieur à N tel que la quantité

$$\sum_{i=1}^{n} |P(x_i) - k_i|^2$$

soit la plus petite possible. On parlera alors d'approximation au sens des moindres carrés.

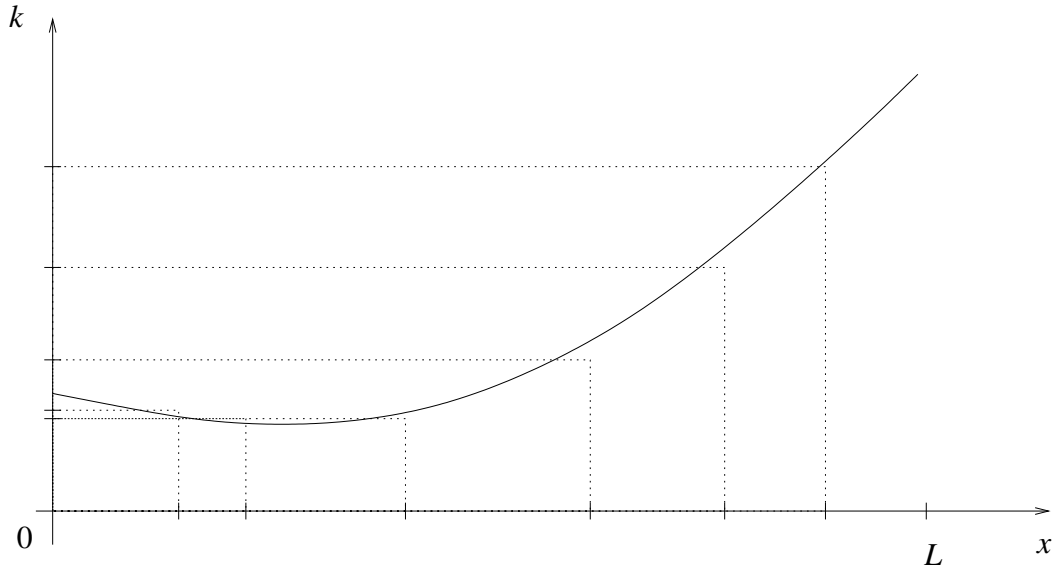


FIG. I.4 – approximation polynomiale

Ensuite on applique la méthode du paragraphe suivant.

2.5 Cas général, méthode des différences finies

On revient sur le problème

$$\begin{cases} \dfrac{d}{dx}\left(-k(x)\dfrac{dT}{dx}\right) = 0, & \text{pour } 0 < x < L, \\ \\ T(0) = T_0, \qquad T(L) = T_L. \end{cases}$$

Étape 1 : discrétisation en espace

Soit $N > 0$, on considère une discrétisation uniforme de l'intervalle $[0, L]$ de pas $h = 1/N$.

$$0 = x_0^h < x_1^h < x_2^h < \cdots < x_N^h = L, \qquad \text{avec} \quad x_i^h := ih, \quad i = 0, \cdots, N.$$

FIG. I.5 – subdivision uniforme

Vocabulaire : h est appelé le pas de la discrétisation. On dit que $x_0^h, x_1^h, \cdots, x_N^h$ est une subdivision uniforme de pas h. "Uniforme" signifiant que $x_{i+1}^h - x_i^h = h$ est constant.

Les inconnues discrètes seront $T_0^h, T_1^h, \cdots, T_N^h$ où on souhaite

$$T_i^h \simeq T(x_i^h).$$

Étape 2 : discrétisation des équations

Les conditions aux limites se discrétisent par

- $T(0) = T_0 \Longrightarrow T_0^h = T_0$,
- $T(L) = T_L \Longrightarrow T_N^h = T_L$.

Comment discrétiser l'équation différentielle $(k(x)T'(x))' = 0$?

Il y a plusieurs méthodes qui utilisent toutes les développements de Taylor. Ici on utilise que si U est de classe C^2 alors

$$U'(x_i^h) = \frac{U(x_i^h + \frac{h}{2}) - U(x_i^h - \frac{h}{2})}{h} + O(h^2).$$

On applique ce principe à $U(x) := k(x)T'(x)$ d'où

$$(kT')'(x_i^h) \;=\; \frac{(kT')(x_i^h + \frac{h}{2}) - (kT')(x_i^h - \frac{h}{2})}{h} + O(h^2).$$

Les quantités $k(x_i^h \pm \frac{h}{2})$ sont connues, il reste à approcher $T'(x_i^h \pm \frac{h}{2})$. On utilise la même méthode.

$$T'(x_i^h + \frac{h}{2}) \;=\; \frac{T(x_i^h + h) - T(x_i^h)}{h} + O(h^2) \;=\; \frac{T(x_{i+1}^h) - T(x_i^h)}{h} + O(h^2).$$

D'où pour $i = 1, \cdots, N-1$,

$$(kT')'(x_i^h) \;=\; \frac{k(x_i^h + \frac{h}{2})}{h^2} T(x_{i+1}^h) - \frac{k(x_i^h + \frac{h}{2}) + k(x_i^h - \frac{h}{2})}{h^2} T(x_i^h) + \frac{k(x_i^h - \frac{h}{2})}{h^2} T(x_{i-1}^h) + O(h).$$

On obtient le problème discret en remplaçant $T(x_i^h)$ par l'inconnue discrète T_i^h et en négligeant le terme $O(h)$. Soit

$$\frac{k(x_i^h - \frac{h}{2})}{h^2} T_{i-1}^h - \frac{k(x_i^h - \frac{h}{2}) + k(x_i^h + \frac{h}{2})}{h^2} T_i^h + \frac{k(x_i^h + \frac{h}{2})}{h^2} T_{i+1}^h \;=\; 0, \qquad i = 1, \cdots, N-1.$$

On obtient donc le système linéaire de taille $(N+1) \times (N+1)$ pour les inconnues discrètes

$$\left\{ \begin{array}{rcl} T_0^h &=& T_0 \\ \frac{k(x_1^h - \frac{h}{2})}{h^2} T_0^h - \frac{k(x_1^h - \frac{h}{2}) + k(x_1^h + \frac{h}{2})}{h^2} T_1^h + \frac{k(x_1^h + \frac{h}{2})}{h^2} T_2^h &=& 0, \\ \vdots \quad \vdots \quad \vdots \quad \vdots & & \vdots \\ \frac{k(x_i^h - \frac{h}{2})}{h^2} T_{i-1}^h - \frac{k(x_i^h - \frac{h}{2}) + k(x_i^h + \frac{h}{2})}{h^2} T_i^h + \frac{k(x_i^h + \frac{h}{2})}{h^2} T_{i+1}^h &=& 0, \qquad i = 1, \cdots, N-1, \\ \vdots \quad \vdots \quad \vdots \quad \vdots & & \vdots \\ \frac{k(x_{N-1}^h - \frac{h}{2})}{h^2} T_{N-2}^h - \frac{k(x_{N-1}^h - \frac{h}{2}) + k(x_{N-1}^h + \frac{h}{2})}{h^2} T_{N-1}^h + \frac{k(x_{N-1}^h + \frac{h}{2})}{h^2} T_N^h &=& 0, \\ T_N^h &=& T_L. \end{array} \right.$$

Soit en notant T^h le vecteur inconnu $T^h := {}^t(T_0^h, \cdots, T_N^h)$ et $b^h := {}^t(T_0, 0, \cdots, 0, T_L)$, on a à résoudre

$$A^h T^h \;=\; b^h,$$

où A^h est la matrice

$$\begin{pmatrix}
1 & 0 & 0 & \cdots & \cdots & 0 \\
-\frac{k(x^h_{1/2})}{h^2} & \frac{k(x^h_{1/2})+k(x^h_{3/2})}{h^2} & -\frac{k(x^h_{3/2})}{h^2} & \ddots & & \vdots \\
0 & -\frac{k(x^h_{3/2})}{h^2} & \frac{k(x^h_{3/2})+k(x^h_{5/2})}{h^2} & -\frac{k(x^h_{5/2})}{h^2} & & \vdots \\
\vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\
\vdots & & -\frac{k(x^h_{N-5/2})}{h^2} & \frac{k(x^h_{N-5/2})+k(x^h_{N-3/2})}{h^2} & -\frac{k(x^h_{N-3/2})}{h^2} & 0 \\
\vdots & & & -\frac{k(x^h_{N-3/2})}{h^2} & \frac{k(x^h_{N-3/2})+k(x^h_{N-1/2})}{h^2} & -\frac{k(x^h_{N-1/2})}{h^2} \\
0 & \cdots & \cdots & 0 & 0 & 1
\end{pmatrix}$$

Il faut

1. S'assurer que la soultion trouvée T^h_i est bien proche de la solution exacte $T(x^h_i)$. Notamment vous verrez (2^{nd} semestre) que sous certaines hypothèses, l'erreur $\max_i |T(x^h_i) - T^h_i|$ tend vers 0 quand h tend vers 0 (N tend vers l'infini).

2. Résoudre le système linéaire. Remarquez que pour avoir une erreur petite il faudra prendre N grand. Remarquez aussi que le système est *creux* : les coefficients de la matrice sont presques tous nuls. Un bon tiers de ce cours (E. Audusse) sera consacré à la résolution de systèmes linéaires : trouver x tel que $Ax = b$. En particulier quand la taille du système est grande et qu'il est creux.

Plan du cours

Chapitre III. Méthodes de résolution numérique de $f(x) = 0$

Chapitre IV. Méthodes de quadratures, polynômes orthogonaux

Chapitre V. Interpolation polynomiale, approximation polynomiale

Chapitre VI. Résolution numérique des Equations Différentielles Ordinaires

- Algèbre linéaire numérique

II Représentation des nombres en machine, erreur de troncature

Cette partie est trés fortement inspirée du premier chapitre du livre " Analyse Numérique, une approche Mathématique" de *Michelle Schatzman*.

1 Écriture des entiers en base b.

On souhaite faire des calculs sur des nombres réels avec un ordinateur. C'est impossible car un ordinateur est fini et l'ensemble des nombres réels ne l'est pas.

Définition 1.1 *L'***erreur de troncature***, c'est l'erreur qu'on fait en arrondissant. Par exemple si on écrit* $\pi \simeq 3.1416$*, on fait l'erreur de troncature* $e_{tr} := \pi - 3.1416 \simeq -7.10^{-6}$.

Dans la machine, les nombres sont stockés en base 2. Par exemple,

$$1 = 1\times 2^0 = \overline{1}^2, \qquad 2 = 1\times 2^1 = \overline{10}^2, \qquad 7 = 1\times 2^2 + 1\times 2^1 + 1\times 2^0 = \overline{111}^2,$$

$$\begin{aligned} 287 &= 256 + 16 + 8 + 4 + 2 + 1, \\ &= 1\times 2^8 + 0\times 2^7 + 0\times 2^6 + 0\times 2^5 + 1\times 2^4 + 1\times 2^3 + 1\times 2^2 + 1\times 2^1 + 1\times 2^0 \\ &= \overline{100011111}^2 \end{aligned}$$

Plus généralement :

Proposition 1.2 *Soit* $b \geq 2$ *un entier. Tout nombre entier positif n admet une unique écriture en base b,*

$$n = \sum_{i=0}^{r} a_i b^i = \sum_{i=0}^{+\infty} a_i b^i, \qquad \text{avec } 0 \leq a_i < b \text{ entiers, } a_r \neq 0 \text{ et } a_i = 0 \text{ pour } i > r.$$

Preuve
Unicité. Si $n = \sum_{i\geq 0} a_i b^i = \sum_{i\geq 0} a'_i b^i$, on a

$$\sum_{i\geq 0}(a_i - a'_i)b^i \;=\; 0, \qquad \text{avec} \quad -b < a_i - a'_i < b.$$

Supposons par contradiction qu'il existe i tel que $a_i \neq a'_i$. Soit q le plus grand entier tel que $a_q \neq a'_q$, on a

$$b^q \;\leq\; \left|(a_q - a'_q)b^q\right| \;=\; \left|\sum_{0\leq i<q}(a_i - a'_i)b^i\right| \;\leq\; \sum_{0\leq i<q}(b-1)b^i \;=\; b^q - 1. \qquad \text{Contradiction.}$$

Existence. Par récurrence sur n. C'est vrai pour $n = 0$. Supposons que jusqu'au rang $n-1$, on ait une écriture en base b. Au rang n, il existe r tel que

$$b^r \;\leq\; n \;<\; b^{r+1}. \tag{1.1}$$

On effectue la division euclidienne de n par b^r, on a

$$n \;=\; a_r b^r + q \qquad \text{avec } 0 \leq q < b^r \leq n \quad \text{et } a_r < b \text{ (par (1.1))}.$$

On applique l'hypothèse de récurrence à $q < n$, on a $q = \sum_{i=0}^{r'} a_i b^i$. On a $r' < r$ car $b^r > q \geq a_{r'} b^{r'} \geq b^{r'}$. Donc

$$n \;=\; a_r b^r + \sum_{i=0}^{r'} a_i b^i \;=\; \sum_{i=0}^{r} a_i b^i.$$

□

2 Représentation des nombres réels en machine

Dans l'ordinateur les nombres réels sont stockés en base 2. Sous la forme

$$n \;=\; \pm\overline{0, a_{-1}a_{-2}\cdots a_{-r}}^{\,2} \times 2^p, \qquad \text{avec } a_{-i} = 0 \text{ ou } 1 \qquad \text{et } a_{-1} = 1. \tag{2.1}$$

Définition 2.1 *le nombre r est le nombre de chiffres significatifs. La partie* $\overline{0, a_{-1}a_{-2}\cdots a_{-r}}^{\,2}$ *est appelée* mantisse *et p est l'*exposant.

Exemples :

$$8 \;=\; \overline{1000}^{\,2} \;=\; \overline{0,100000}^{\,2} \times 2^4, \qquad 0,25 \;=\; 1\times 2^{-2} \;=\; \overline{0,100000}^{\,2} \times 2^{-1},$$
$$1/3 \;=\; \overline{0,010101010\cdots}^{\,2} \;\simeq\; \overline{0,101010}^{\,2} \times 2^{-1}.$$

Définition 2.2 *L'exposant p est borné par $-m \leq p \leq M$. Les entiers m, M et r dépendent de la machine. Ces entiers étant fixés, on appelle ensemble des flottants l'ensemble $F(r, M, m)$ des nombres pouvant s'écrire sous la forme* (2.1).

Exercice 2.1 *Déterminer $F(3, 1, 2)$. Pensez-vous que cet ensemble est stable par addition, multiplication, ... ?*

On se donne une application arrondi

$$A \ : \ \mathbf{R} \to F(r, M, m)$$

telle que pour $x \in F(r, m, M)$, on ait $A(x) = x$ et telle que sinon l'erreur de troncature $x - A(x)$ soit la plus petite possible. L'addition, la multiplication, ... sont définies par

$$x \oplus y \ := \ A(x + y), \qquad x \otimes y \ := \ A(x \times y), \qquad \cdots$$

Exercice 2.2 *Ces opérations sont elles associatives ?* Indication : *on pourra travailler avec $F(3, 1, 2)$.*

Vocabulaire L'unité de mesure de vitesse d'un ordinateur est le flops : *Floating point Operation Per Second* (le nombre d'opération à virgule flottante par seconde) autrement dit on compte le nombre d'opérations $\oplus$, $\otimes$, ... entre flottants réalisables en une seconde.

3 Problèmes d'instabilité numérique : deux exemples

Premier exemple. On utilise un ordinateur pour calculer la suite récurrente définie par

$$u_{n+1} \ = \ (q + 1)u_n - p, \tag{3.1}$$

avec $p, q \geq 1$ entiers.

Commençons par étudier les suites définies par la relation de récurrence (3.1). On voit facilement qu'on a

$$\left(u_{n+1} - \frac{p}{q}\right) \ = \ (q + 1)\left(u_n - \frac{p}{q}\right).$$

Donc en posant $y_n := u_n - p/q$, la suite (y_n) est géométrique de raison $q + 1$ et $y_n = y_0(q + 1)^n$ pour $n \geq 0$. D'où

$$u_n = \left(u_0 - \frac{p}{q}\right)(q + 1)^n, \qquad n \geq 0.$$

En particulier, si $u_0 = p/q$, on a $u_n = 0$ pour tout n.

Regardons ce qui se passe numériquement. A chaque étape de (3.1), on peut s'attendre à une erreur d'arrondi. En fait, on construit une suite vérifiant

$$\tilde{u}_{n+1} = (q + 1)\tilde{u}_n - p + e_n,$$

où e_n est l'erreur de troncature à l'étape n. En posant $\tilde{y}_n := \tilde{u}_n - p/q$, on obtient

$$\tilde{y}_{n+1} = (q+1)\tilde{y}_n + e_n,$$

d'où

$$\tilde{y}_n = (q+1)^n y_0 + (q+1)^{n-1} e_1 + (q+1)^{n-2} e_2 + \cdots + (q+1) e_{n-1} + e_n \;=\; (q+1)^n y_0 + \sum_{i=1}^{n} (q+1)^{n-i} e_i.$$

Si $y_0 = 0$, au lieu d'avoir $\tilde{y}_n = 0$ pour tout n, on a

$$\tilde{y}_n = \sum_{i=1}^{n} (q+1)^{n-i} e_i.$$

Imaginons qu'il y ait une erreur $e_i \neq 0$ seulement à la première étape, on a alors

$$\tilde{y}_n = (q+1)^{n-1} e_1 \quad \overset{n\uparrow\infty}{\longrightarrow} \quad \infty.$$

Voici les valeurs calculées à l'aide du logiciel MATLAB avec $p = 1$, $q = 3$, $u_0 = p/q$ (donc $u_n = p/q$ pour $n \geq 0$).

$$\begin{array}{ccccccc} \tilde{u}_0 \simeq 0.3333 & & \tilde{u}_1 \simeq 0.3333 & & \cdots\cdots & \tilde{u}_{21} \simeq 0.3333 & \tilde{u}_{22} \simeq 0.3333 \\ \tilde{u}_{23} \simeq 0.3330 & \tilde{u}_{24} \simeq 0.3320 & & \tilde{u}_{25} \simeq 0.3281 & & \tilde{u}_{26} \simeq 0.3125 & \tilde{u}_{27} \simeq 0.2500 \\ \tilde{u}_{28} \simeq 0 & \tilde{u}_{29} \simeq -1.0000 & \cdots\cdots & & \tilde{u}_{50} \simeq -5.8641\,10^{12} & & \tilde{u}_{100} \simeq -7.4336\,10^{42}. \end{array}$$

Exercice 3.1 *Que se passe-t-il si on prend $p = 3$, $q = 4$?*

Exemple 2. Résolution de l'équation

$$x^2 + 2x + c \;=\; 0, \qquad\qquad \text{avec } \; c = -\varepsilon \; \text{ réel donné.}$$

On a les solutions

$$\begin{aligned} x_1 &= -1 + \sqrt{1-c} \\ x_2 &= -1 - \sqrt{1-c}. \end{aligned} \tag{3.2}$$

supposons que $c = -\varepsilon$ soit petit, on a à l'ordre principal,

$$x_1 \;=\; -1 + \sqrt{1+\varepsilon} \;=\; -1 + \left(1 + \frac{\varepsilon}{2} + O(\varepsilon^2)\right) \;=\; \frac{\varepsilon}{2} + O(\varepsilon^2).$$

Si on utilise la formule (3.2), l'ordinateur effectue successivement les opérations :

$$\begin{aligned} y &\leftarrow 1 + \varepsilon, \\ z &\leftarrow \sqrt{y} \\ x &\leftarrow -1 + z \end{aligned}$$

Dans le cas $\varepsilon = 2^{-q}$ où $q > r$ (r est la taille de la mantisse), l'ordinateur calcule

$$\begin{aligned} y &\leftarrow 1+\varepsilon, && \text{soit } y = A(1+2^{-q}) = A(\overline{0.10\cdots001}^{\,2}\,2^1) = \overline{0.10\cdots0}^{\,2}\,2^1 = 1.\\ z &\leftarrow \sqrt{y}, && \text{soit } z = A(\sqrt{1}) = 1.\\ x &\leftarrow -1+z, && \text{soit } x = 0 \neq \frac{\varepsilon}{2}. \end{aligned}$$

Méthode alternative, au lieu de (3.2), on utilise

$$\begin{aligned} x_1 &= -1+\sqrt{1-c} = -\sqrt{1}+\sqrt{1-c} = \left(-\sqrt{1}+\sqrt{1-c}\right)\frac{\sqrt{1}+\sqrt{1-c}}{\sqrt{1}+\sqrt{1-c}}\\ &= -\frac{c}{\sqrt{1}+\sqrt{1-c}}. \end{aligned}$$

Qui cette fois conduit à $x \simeq \varepsilon/2$.

III Résolution numérique du problème non-linéaire : trouver $x \in \mathbf{R}^d$ tel que $f(x) = 0$.

1 Introduction du problème

On se donne une application

$$f \ : \ \mathbf{R}^d \ \longrightarrow \mathbf{R}^d,$$

qu'on suppose au moins continue. On cherche x tel que

$$f(x) \ = \ 0. \tag{1.1}$$

Remarque 1.1 *Le problème* (1.1) *peut venir d'un problème d'extrema : si on cherche les extrema d'une fonction régulière* $F \ : \ \mathbf{R}^d \ \longrightarrow \ \mathbf{R}$*, ces extrema vérifient* $\nabla F(x) \ = \ 0$ *où* $\nabla F \ : \ \mathbf{R}^d \ \longrightarrow \ \mathbf{R}^d$.

Le problème (1.1) n'a pas nécessairement de solution. Par exemple $f \ : \ \mathbf{R} \to \mathbf{R}, \ x \mapsto 1 + x^2$ ne s'annule pas. Une première étape est de s'assurer de l'existence d'au moins une solution $x^\star$.Une fois qu'on est certaine de l'existence d'une solution $x^\star$, les tâches du numéricien sont :

- Trouver une méthode qui permette de construire *x proche* d'une racine $x^\star$ de f en un nombre fini d'opérations élémentaires.
- Evaluer l'erreur commise : la distance $|x - x^\star|$.

2 La méthode de bissection

Dans cette partie, la dimension d est égale à 1. On suppose f continue. On rappelle un premier résultat bien connu d'existence

Théorème 2.1 (des valeurs intermédiaires)

Soit $f \ : \ [a, b] \ \to \ \mathbf{R}$ *continue telle que* $f(a)f(b) < 0$. *Alors, il existe* $c \in (a, b)$ *tel que* $f(c) = 0$.

Preuve
Quite à remplacer f par $-f$, on peut supposer $f(a) < 0$ et $f(b) > 0$. Soit

$$I^- \;=\; \{x \in [a,b] \;:\; f(x) < 0\}\,,$$

Cet ensemble est non vide car il contient a. Donc I^- a un plus petit majorant $c = \sup\{x \in I^-\}$. Supposons par contradiction que $f(c) \neq 0$, on a deux cas :

cas 1. Si $f(c) < 0$ alors par continuité de f il existe $\eta > 0$ tel que $f < 0$ sur $[c, c+\eta]$ et donc $c + \eta \in I^-$ donc $c + \eta \leq \sup\{x \in I^-\} = c$. Contradiction.

cas 2. Si $f(c) > 0$ alors il existe η tel que $f > 0$ sur $[c-\eta, c]$ et donc $I^- \subset [a, c-\eta]$ et $c - \eta \geq c$. Contradiction.

On conclut donc que $f(c) = 0$. □

Cette preuve ne permet pas directement de construire c. On va voir une autre preuve qui utilise le principe de bissection (*Dichotomie* en Grec).

Seconde Preuve On pose $a_0 = a$ et $b_0 = b$. On construit ensuite de manière récursive les suites (a_n) et (b_n) par

$$c_n \;:=\; \frac{a_n + b_n}{2}.$$

Puis

$$\text{Si } \; f(a_n)f(c_n) \leq 0, \qquad \text{alors} \quad \begin{cases} a_{n+1} := a_n, \\ b_{n+1} := c_n, \end{cases} \qquad \text{Sinon} \quad \begin{cases} a_{n+1} := c_n, \\ b_{n+1} := b_n. \end{cases}$$

On a par construction

$$a_0 \leq a_1 \leq \cdots \leq a_n \leq a_{n+1} \leq \cdots \leq b_{n+1} \leq b_n \leq \cdots \leq b_0,$$

$$b_n - a_n \;=\; \frac{b_{n-1} - a_{n-1}}{2} \;=\; \cdots \;=\; \frac{b-a}{2^n} \quad \overset{n\uparrow\infty}{\longrightarrow} \quad 0, \qquad \text{et} \qquad f(a_n)f(b_n) \leq 0.$$

Les deux suites adjacentes (a_n), (b_n) convergent donc vers une même limite c. En passant à la limite dans $f(a_n)f(b_n) \leq 0$, on obtient par continuité de f, l'inégalité $(f(c))^2 \leq 0$ et donc $f(c) = 0$. □

On déduit de cette méthode l'algorithme de bissection (Dichotomie)

- Données : $a < b$, f fonction continue sur $[a,b]$ telle que $f(a)f(b) < 0$, N nombre d'itérations
- Résultat c tel que $|c - c^\star| \leq (b-a)/2^{N+1}$ où $f(c^\star) = 0$.
- Algorithme :

```
Pour      i ← 1 à N faire
          c ← (a + b)/2
          Si f(a)f(c) ≤ 0   faire
                        b ← c
          Sinon faire
                        a ← c
          Fin Si
Fin Pour  i
Renvoyer  c.
```

Proposition 2.2 *L'algorithme de Dichotomie est tel que*

$$|c - c^\star| \ \leq \ \frac{b-a}{2^{N+1}}.$$

L'algorithme de bissection a plusieurs avantages. Tout d'abord, il donne un résultat satisfaisant de manière sûre. On a une majoration explicite de l'erreur commise. D'autre part c'est pour l'instant le seul que nous connaissions. Il a un défaut principal : on ne voit pas bien comment le généraliser à la dimension supérieure.

3 Méthodes de point fixe

On revient au cas général. $d \geq 1$. On transforme le pb

(3.1) Trouver x tel que $f(x) = 0$.

en

(3.2) Trouver x tel que $g(x) = x$.

Il y a une infinité de choix pour g. Par exemple : si

$$g(x) \ := \ x + Af(x)$$

où A est une matrice $d \times d$ inversible, alors (3.1) $\Longleftrightarrow$ (3.2).

3.1 Rappels

Définition 3.1 *Dans* $\mathbf{R}^d$*, l'application* $\mathbf{R}^d \longrightarrow \mathbf{R}_+,\ x \longmapsto ||x||$ *est une norme ssi*

- $||x|| = 0 \iff x = 0 \qquad \forall\, x \in \mathbf{R}^\cdot$
- $||\lambda x|| = |\lambda|\,||x|| \qquad \forall\, x \in \mathbf{R}^d,\ \lambda \in \mathbf{R}.$
- $||x + y|| \leq ||x|| + ||y|| \qquad \forall\, x, y \in \mathbf{R}^d.$

On dit que $(\mathbf{R}^d, \|\cdot\|)$ *a la structure d'espace vectoriel normé.*

Proposition 3.2 *Dans* $\mathbf{R}^d$*, toutes les normes sont équivalentes i.e. si* $\|\cdot\|_1$ *et* $\|\cdot\|_2$ *sont deux normes sur* $\mathbf{R}^d$ *alors* $\exists \kappa \in (0, 1]$ *tel que*

$$\kappa \|x\|_1 \ \leq \ \|x\|_2 \ \leq \kappa^{-1} \|x\|_1 \qquad \forall\, x \in \mathbf{R}^d.$$

En particulier les notions de convergences de suites ou bien de voisinage d'un point de $\mathbf{R}^d$ ne dépendent pas de la norme choisie.

Théorème 3.3 *L'espace vectoriel normé* $(\mathbf{R}^d, \|\cdot\|)$ *est complet. On dit que c'est un espace de Banach.*

Cela signifie que toute suite de Cauchy de $(\mathbf{R}^d, \|\cdot\|)$ *converge. Plus précisément, soit* $(x_n) \subset \mathbf{R}^d$ *telle que pour tout* $\varepsilon > 0$ *il existe* $N \geq 0$ *tel que*

$$\forall\, n, p \geq N, \qquad \|x_n - x_p\| \;<\; \varepsilon.$$

Alors il existe $x^\star \in \mathbf{R}^d$ *tel que* $x_n \overset{n\uparrow\infty}{\longrightarrow} x^\star$ *dans* $\mathbf{R}^d$. $\Longleftrightarrow$ $(\|x_n - x^\star\| \overset{n\uparrow\infty}{\longrightarrow} 0)$.

3.2 Le théorème de Point Fixe de Picard

Définition 3.4 *On dit que* $g : \mathbf{R}^d \longrightarrow \mathbf{R}^d$ *est strictement contractante pour la norme* $\|\cdot\|$ *si il existe* $\lambda \in (0,1)$ *tel que pour tout* $x, y \in \mathbf{R}^d$

$$\|g(x) - g(y)\| \;\leq\; \lambda \|x - y\|. \tag{3.3}$$

Exercice 3.1 *Les applications suivantes de* $\mathbf{R}$ *à valeurs dans* $\mathbf{R}$ *sont-elles contractantes ?*

$$g_1(x) := 1 + \frac{1}{2}x, \qquad g_2(x) := \sin x, \qquad g_3(x) := -5 + 2x.$$

Exercice 3.2 *Dans cet exercice,* $d = 2$. *Montrer que*

$$\|(x,y)\| = \sqrt{x_1^2 + x_2^2} \qquad \textit{et} \qquad \|(x,y)\|' = \sqrt{2x_1^2 + x_2^2}$$

définissent deux normes sur $\mathbf{R}^2$.

Donner une application $g : \mathbf{R}^2 \longrightarrow \mathbf{R}^2$ *qui soit contractante dans* $(\mathbf{R}^2, \|\cdot\|')$ *et pas dans* $(\mathbf{R}^2, \|\cdot\|)$.

Exercice 3.3 *On suppose g de classe* C^1. *A quelle condition sur sa différentielle, l'application g est-elle contractante ?*

Théorème 3.5 (point fixe de Picard) *Soit* $g : \mathbf{R}^d \longrightarrow \mathbf{R}^d$ *une application strictement contractante. Alors g admet un unique point fixe.*

Preuve
Unicité. Soient $x^\star$ et $y^\star$ deux points fixes, on a $g(x^\star) - g(y^\star) = x^\star - y^\star$, d'où (par (3.3))

$$\|x^\star - y^\star\| \;\leq\; \lambda\, \|x^\star - y^\star\|,$$

et comme $\lambda < 1$, on a $x^\star = y^\star$.

Existence. On se donne $x_0 \in \mathbf{R}^d$ quelconque et on définit de manière récursive la suite $(x_n)_{n\geq 0}$ par

$$x_{n+1} \;:=\; g(x_n), \qquad \forall\, n \geq 0. \tag{3.4}$$

On a, par (3.3), pour $n \geq 1$

$$\|x_{n+1} - x_n\| \;=\; \|g(x_n) - g(x_{n-1})\| \;\leq\; \lambda\, \|x_n - x_{n-1}\| \;\leq\; \cdots \;\leq\; \lambda^n\, \|x_1 - x_0\|\,. \tag{3.5}$$

En particulier $\|x_{n+1} - x_n\| \overset{n\uparrow\infty}{\longrightarrow} 0$. Avec la même idée, on calcule pour $n \geq 0$ et $p \geq 1$,

$$\begin{aligned}
\big\|x_{n+p} - x_n\big\| &= \big\|(x_{n+p} - x_{n+p-1}) + (x_{n+p-1} - x_{n+p-2}) + \cdots + (x_{n+p+1} - x_{n+p})\big\| \\
&\leq \big\|x_{n+p} - x_{n+p-1}\big\| + \big\|x_{n+p-1} - x_{n+p-2}\big\| + \cdots + \|x_{n+1} - x_n\| \\
&\leq \big\|x_{n+p} - x_{n+p-1}\big\| + \big\|x_{n+p-1} - x_{n+p-2}\big\| + \cdots + \|x_{n+1} - x_n\| \\
&\leq \lambda^{p-1}\, \|x_{n+1} - x_n\| + \lambda^{p-2}\, \|x_{n+1} - x_n\| + \cdots + \|x_{n+1} - x_n\| \\
&\leq \underbrace{\big(1 + \lambda + \cdots + \lambda^{p-1}\big)}_{\leq\, 1/(1-\lambda)} \|x_{n+1} - x_n\| \;\leq\; \frac{1}{1-\lambda}\, \|x_{n+1} - x_n\|\,. \qquad (3.6)
\end{aligned}$$

Et comme $\lim\limits_{n\uparrow\infty} \|x_{n+1} - x_n\| \;=\; 0$, la suite (x_n) est de Cauchy et converge donc vers une limite $x^\star$.

Finalement, par continuité de g, on a

$$g(x^\star) \;=\; \lim_{n\uparrow\infty} g(x_n) \;=\; \lim_{n\uparrow\infty} x_{n+1} \;=\; x^\star.$$

□

Cette preuve fournit une méthode pratique pour construire des approximations x_n de $x^\star$. De plus, on a une estimation d'erreur. En effet passant à la limite $p \to \infty$ dans (3.6), on obtient

$$\|x^\star - x_n\| \;\leq\; \frac{1}{1-\lambda} \|x_{n+1} - x_n\|. \tag{3.7}$$

Et par (3.5),

$$\|x^\star - x_n\| \;\leq\; \frac{\lambda^n}{1-\lambda} \|x_1 - x_0\|.$$

Si on connaît λ, on peut obtenir une approximation de $x^\star$ avec la précision souhaitée ε. Pour cela, on se donne x_0, on calcule x_1 puis $R = \|x_1 - x_0\|$. Pour avoir $\|x^\star - x_n\| < \varepsilon$, il suffit d'avoir

$$\varepsilon > \frac{\lambda^n}{1-\lambda} R \quad\Longleftrightarrow\quad \ln \varepsilon > n \ln \lambda + \ln\left(\frac{R}{1-\lambda}\right) \quad\Longleftrightarrow\quad n > \frac{\ln 1/\varepsilon + \ln(R/(1+\lambda))}{\ln 1/\lambda} \sim \frac{\ln 1/\varepsilon}{\ln 1/\lambda}.$$

Notant e_n l'erreur $\|x_n - x^\star\|$, on a

$$\ln e_n \;\leq\; n \underbrace{\ln \lambda}_{<\, 0.} + \ln \ln(R/(1+\lambda))$$

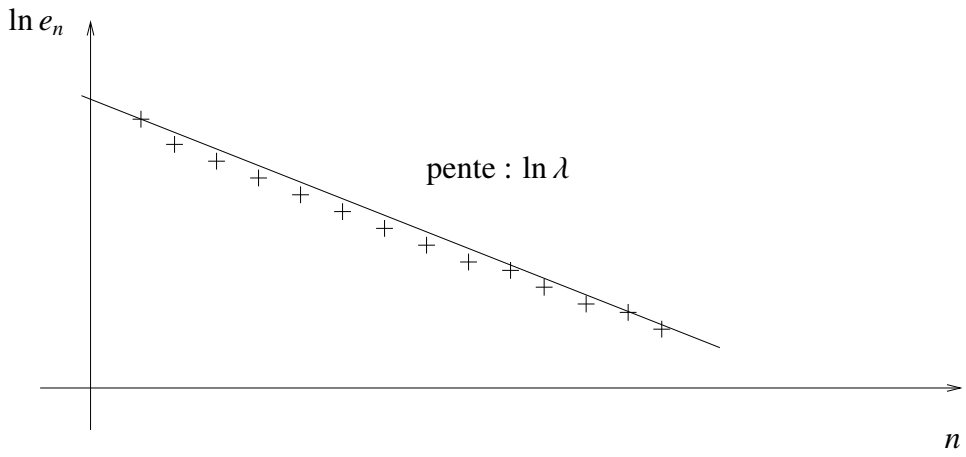


FIG. III.1 – Forme typique du graphe de convergence

On a même une estimation d'erreur à chaque itération par (3.7). Cette inégalité permet d'obtenir un résultat avec une précision fixée.

Algorithme :

- Données : $\lambda \in (0,1)$, g application λ-contractante $\mathbf{R}^d \longrightarrow \mathbf{R}^d$, $x_0 \in \mathbf{R}^d$, $tol > 0$.
- Résultat x tel que $|x - x^\star| < tol$ où $g(x^\star) = x^\star$.
- Algorithme :

$x \leftarrow x_0$
$y \leftarrow g(x)$
$err \leftarrow \|x - y\|$
Tant que $err > tol(1-\lambda)$
$x \leftarrow y$
$y \leftarrow g(x)$
$err \leftarrow \|x - y\|$
Fin Tant que

En fait pour caractériser ce type de méthode, on regarde plutôt le rapport entre deux erreurs successives.

$$e_n := x_n - x^\star.$$

Comme $g(x^\star) = x^\star$, on a $e_n + 1 = g(x_n) - g(x^\star)$ et donc

$$\|e_{n+1}\| \leq \lambda \|e_n\|. \tag{3.8}$$

Définition 3.6 *On dit qu'une suite (x_n) construite par une méthode numérique converge vers $x^\star$ avec un ordre $p \geq 1$ si il existe $C > 0$ tel que*

$$\|e_{n+1}\| \leq C\|e_n\|^p, \qquad \forall\, n \geq n_0;$$

Si $p = 1$, il faut que $\lambda < 1$ pour pouvoir en déduire que la suite converge ; dans ce cas on dit que la convergence est linéaire. *On dit que la convergence est* quadratique *si $p = 2$ et* cubique *si $p = 3$. La constante C est appelée facteur de convergence de la méthode.*

Exemples

$$p = 1,\ C = 1/2\ : \qquad e = (1, 1/2, 1/4, 1/8, 1/16, \cdots)$$
$$p = 2,\ C = 1\ : \qquad e = (10^{-1}, 10^{-2}, 10^{-4}, 10^{-8}, 10^{-16}, \cdots).$$

Remarque 3.7 *Plus p est grand, plus la méthode converge rapidement.*

3.3 Méthode d'accélération d'Aitken

On souhaite améliorer la convergence de la méthode du point fixe. Pour cela, nous allons utiliser l'information qu'on a sur la manière dont la suite converge vers $x^\star$. On se place en dimension $d = 1$ et on suppose g de classe C^2.

Exercice 3.4 *On considère une suite construite par* (3.4). *On note* $\rho = g'(x^\star)$ *On suppose* $\rho \neq 0$. *Montrer que*

$$x_{n+1} - x^\star \ = \ \rho(x_n - x^\star) + O(|x_n - x^\star|^2).$$

En déduire $\rho \ = \ \rho_n + o(1)$ *où* $\rho_n := \dfrac{x_{n+2} - x_n}{x_{n+1} - x_n}$.

Finalement, on pose $\tilde{x}_n \ := \ \dfrac{x_{n+1} - \rho_n x_n}{1 - \rho_n}$. *Donner un équivalent de* $\dfrac{\tilde{x}_{n+1} - x^\star}{\tilde{x}_n - x^\star}$. *Conclure.*

4 La Méthode de Newton

Si on sait que f est différentiable et qu'on sait calculer sa différentielle, on peut utiliser la méthode de Newton. En voici l'idée dans le cas de la dimension $d = 1$.

On se donne $x_0 \in \mathbf{R}$. Au voisinage de x_0, on a

$$f(x) \ = \ f(x_0) + f'(x_0)(x - x_0) + o(|x - x_0|).$$

En particulier

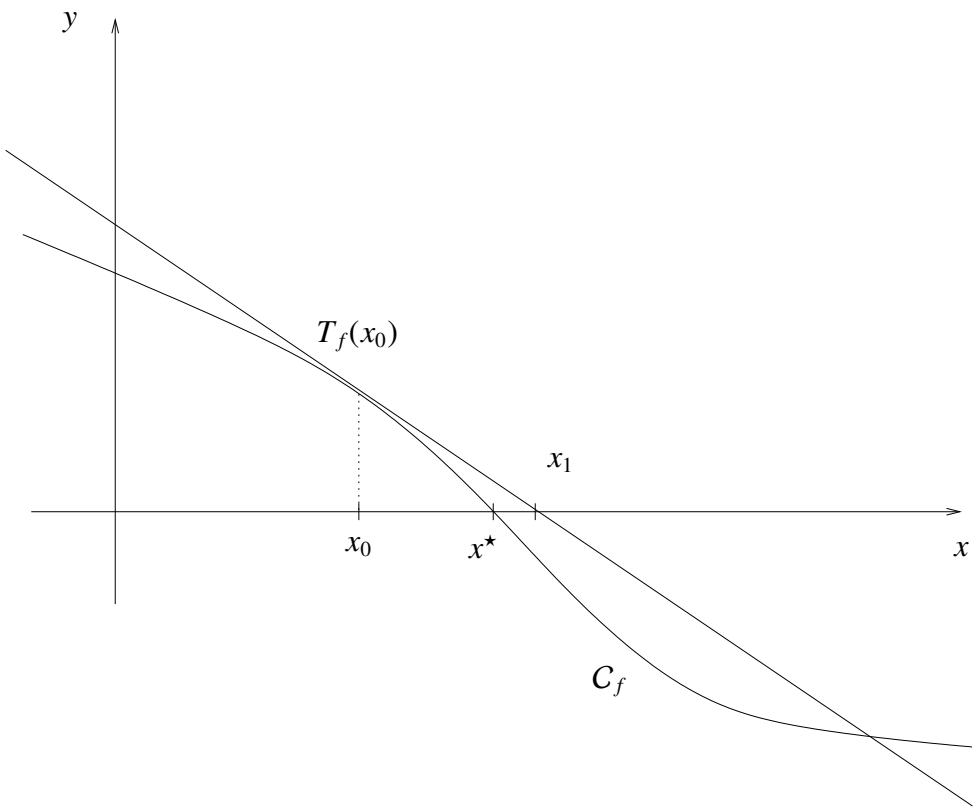
$$0 \ = \ f(x^\star) \ = \ f(x_0) + f'(x_0)(x^\star - x_0) + o(|x - x^\star{}_0|).$$

Si x_0 est proche de $x^\star$, il est raisonnable de négliger le terme en $o(|x^\star - x_0|)$.

$$0 \ \simeq \ f(x_0) + f'(x_0)(x^\star - x_0).$$

On définit alors le terme suivant de la suite x_1 comme étant solution de

$$0 \ = \ f(x_0) + f'(x_0)(x_1 - x_0) \qquad \Longleftrightarrow \qquad x_1 \ := \ x_0 - \big(f'(x_0)\big)^{-1} f(x_0).$$



Définition 4.1 *Soit $f \in C^1(\mathbf{R}^d, \mathbf{R}^d)$. La méthode de Newton est définie par*

- *Prendre $x_0 \in \mathbf{R}$*
- *Pour $n \geq 0$, si $f'(x_n) \neq 0$, on pose*

$$x_{n+1} := x_n - \left(f'(x_n)\right)^{-1} f(x_n). \tag{4.1}$$

Remarque 4.2 *Il est impératif de savoir calculer f et f' pour mettre en œuvre la méthode de Newton.*

Exercice 4.1 *Que se passe-t-il si f est une fonction affine ?*

La méthode est identique dans le cas des dimensions supérieures $d > 1$. En effet si $f \in C^1(\mathbf{R}^d, \mathbf{R}^d)$, pour $x_0 \in \mathbf{R}^d$, on a

$$f(x) = f(x_0) + Df(x_0)(x - x_0) + o(|x - x_0|),$$

où $Df(x_0)$ est la matrice de coefficients $\left(\dfrac{\partial f_i}{\partial x_j}(x_0)\right)_{1 \leq i,j \leq d}$.

On a alors

$$0 = f(x^\star) = f(x_0) + Df(x_0)(x^\star - x_0) + o(|x^\star - x_0|),$$

En négligeant le terme $o(|x^\star - x_0|)$, on déduit l'algorithme :

Définition 4.3 *Soit $f \in C^1(\mathbf{R}^d, \mathbf{R}^d)$. La méthode de Newton est définie par*

- *Prendre $x_0 \in \mathbf{R}^d$*
- *Pour $n \geq 0$, si $Df(x_n) \neq 0$, on pose*

$$x_{n+1} := x_n - (Df(x_n))^{-1} f(x_n). \tag{4.2}$$

On a le résultat de convergence suivant.

Théorème 4.4 *Si $f \in C^2(\mathbf{R}^d, \mathbf{R}^d)$ et $x^\star \in \mathbf{R}^d$ sont tels que $f(x^\star) = 0$ et $Df(x^\star)$ inversible. Alors il existe $\eta > 0$ tel que $\forall\, x_0 \in B(x^\star, \eta)$, on a $x_n \overset{n\uparrow\infty}{\longrightarrow} x^\star$. De plus la méthode est d'ordre 2, i.e. il existe $C_1, n_1 > 0$ tels que*

$$\|x_{n+1} - x^\star\| \leq C_1 \|x_n - x^\star\|^2, \qquad \forall\, n \geq n_1 \tag{4.3}$$

Cette inégalité entraînant l'estimation d'erreur : il existe $C_2, n_2 > 0$ et $0 \leq \rho < 1$ telles que

$$\|x_n - x^\star\| \leq C_2 \rho^{2^n}, \qquad \forall\, n \geq n_2. \tag{4.4}$$

Preuve
On écrit la formule de Taylor avec reste intégrale à l'ordre 2 pour la fonction $g(t) = f(x_n + t(x^\star - x_n))$ entre les points 0 et 1 . On a

$$g(1) = g(0) + g'(0) + \int_0^1 (1-t) g''(t)\, dt$$

Soit

$$0 = f(x^\star) = f(x_n) + Df(x_n) \cdot (x^\star - x_n) + \int_0^1 D^2 f(x_n + t(x^\star - x_n))(x^\star - x_n, x^\star - x_n)\, dt.$$

Or par définition, on a

$$0 = f(x_n) + Df(x_n) \cdot (x_{n+1} - x_n).$$

D'où, en soustrayant,

$$Df(x_n) \cdot (x_{n+1} - x^\star) = \int_0^1 D^2 f(x_n + t(x^\star - x_n))(x^\star - x_n, x^\star - x_n)\, dt. \tag{4.5}$$

Maintenant, comme $Df(x^\star)$ est inversible, il existe $\rho > 0$ tel que si $\|x - x^\star\| \leq \rho_1$ alors $Df(x)$ est inversible. De plus, il existe $\kappa_1 > 0$ tel que

$$\|Df^{-1}(x)\| \leq \kappa_1 \qquad \text{si} \qquad \|x - x^\star\| \leq \rho_1.$$

De même par continuité de $D^2 f$, il existe $\kappa_2 > 0$ telle que

$$\|D^2 f(x)(z_1, z_2)\| \leq \kappa_2 \|z_1\| \|z_2\|, \qquad \text{pour} \qquad \|x - x^\star\| < \rho_1,\ z_1, z_2 \in \mathbf{R}^d.$$

Supposons que $\|x_n - x^\star\| < \rho_1$, alors (4.5) implique

$$\|x_{n+1} - x^\star\| \;=\; \kappa_1\kappa_2\|x_n - x^\star\|^2 \tag{4.6}$$

Et si on choisit $\rho \;<\; \min\Big((\kappa_1\kappa_2)^{-1}, \rho_1\Big)$, on déduit de (4.6),

$$\|x_n - x^\star\| < \rho \quad \Longrightarrow \quad \|x_{n+1} - x^\star\| \;\leq\; \rho\kappa_1\kappa_2\,\|x_n - x^\star\| \;<\; \|x_n - x^\star\| \;<\; \rho.$$

On en déduit par récurrence que si $\|x_0 - x^\star\| < \rho$ alors pour $n \geq 0$, $\|x_n - x^\star\| < \rho$. De plus on a

$$\|x_{n+1} - x^\star\| \;<\; \rho\kappa_1\kappa_2\,\|x_n - x^\star\| \;<\; \cdots \;<\; (\rho\kappa_1\kappa_2)^n\,\|x_0 - x^\star\|,$$

et comme, $(\rho\kappa_1\kappa_2) < 1$, on a $x_n \overset{n\uparrow\infty}{\longrightarrow} x^\star$.

Pour finir, la méthode est d'ordre 2 d'après (4.6) et on en déduit (4.4) par récurrence. □

Remarque 4.5 *La méthode de Newton est très efficace mais elle est instable : elle peut ne pas converger si le premier itéré est trop loin de la solution exacte $x^\star$.*

En pratique : soit on connaît déjà une valeur approchée de $x^\star$, soit on commence par utiliser une autre méthode, plus robuste que la méthode de Newton. On obtient ainsi une première approximation x_0 qu'on utilise ensuite comme point de départ de la méthode de Newton.

Exercice 4.2 *On suppose $d = 1$. Sous les hypothèses du théorème, calculer la limite de la suite*

$$\frac{x_{n+1} - x^\star}{(x_n - x^\star)^2}.$$

Exercice 4.3 *Quel est alors l'ordre de la méthode dans le cas $f(x) = x^p$, $p > 1$?*

IV Intégration numérique

1 Rappels

1.1 Formule de Taylor avec reste intégral

Proposition 1.1 *Soit $I = [a, b]$ un intervalle fermé de* $\mathbf{R}$. *Si $f : I \longrightarrow \mathbf{R}$ de classe C^{k+1}, alors on a*

$$\begin{aligned} f(b) &= f(a) + f'(a)(b-a) + \cdots + f^{(k)}(a)\frac{(b-a)^k}{k!} + \left(\int_0^1 \frac{(1-t)^k}{k!} f^{(k+1)}(a + t(b-a))\, dt\right)(b-a)^{k+1} \\ &= \sum_{j=0^k} f^{(j)}(a)\frac{(b-a)^j}{j!} + \left(\int_0^1 \frac{(1-t)^k}{k!} f^{(k+1)}(a + t(b-a))\, dt\,(b-a)^{k+1}\right) \\ &= \sum_{j=0^k} f^{(j)}(a)\frac{(b-a)^j}{j!} + \int_a^b \frac{(b-x)^k}{k!} f^{(k+1)}(x)\, dx. \end{aligned}$$

Preuve
On prouve le résultat par récurrence sur k. Pour $k = 0$, on considère f de classe C^1 sur $[a, b]$. On étudie la fonction h définie par

$$h(x) = f(x) - f(a) - \int_a^x f'(y)\, dy$$

En revenant à la définition de la dérivée et en utilisant la ccontinuité de f', on déduit que h est dérivable de dérivée identiquement nulle sur $[a, b]$. Par Rolle, il existe $\theta \in (a, b)$ tel que $h(b) = h(a) + (b-a)h'\theta) = h(a) = 0$ donc $h(b) = 0$. Ce qui montre la formule de Taylor avec reste intégral pour $k = 0$.

Supposons maintenant la formule vraie jusqu'au rang k. Soit f de classe C^{k+2} sur $[a, b]$. On pose

$$h(x) = f(x) - f(a) - f'(a)(x-a) - \cdots - f^{(k+1)}(a)\frac{(x-a)^k}{(k+1)!} - \int_a^x \frac{(x-a)^{k+1}}{(k+1)!} f^{(k+2)}(y)\, dy.$$

La fonction h est dérivable sur $[a, b]$ et on calcule

$$h'(x) = f'(x) - f'(a) - f''(a)(x-a) - \cdots - f^{(k+1)}(a)\frac{(x-a)^k}{k!} - \int_a^x \frac{(x-a)^k}{k!} f^{(k+2)}(y)\, dy.$$

En appliquant la formule de Taylor avec reste intégral au rang k à la fonction f' (qui est bien de classe C^k) sur l'intervalle $[a, x]$ on déduit $h' = 0$ sur $[a, b]$ et comme $h(a) = 0$, on obtient $h(b) = 0$ et la formule est vraie au rang $k + 1$. □

1.2 Polynômes

Soit **K** le corps **R** des réels ou **C** des complexes. On note $\mathcal{P}_n(\mathbf{K})$ l'ensemble des polynômes de degrés inférieurs à n à coefficients dans **K**. On note $\mathcal{P}(\mathbf{K})$ l'ensemble des polynômes à coefficients dans **K**.

L'ensemble $\mathcal{P}_n(\mathbf{K})$ est un **K**-espace vectoriel de dimension $n + 1$ admettant pour base les polynômes $(1, X, X^2, \cdots, X^n)$. L'ensemble $\mathcal{P}(\mathbf{K})$ a la structure de **K**-algèbre. Si un polynôme P s'écrit

$$P(X) = \sum_{i=0}^{n} a_i X^i, \qquad \text{avec} \quad a_n \neq 0,$$

on dit que n est le degré de P. On définit le degré du polynôme nul par $d^o(0) = -\infty$. Si P et Q sont deux polynômes de degrés m et n alors PQ a pour degrés $m + n$.

Proposition 1.2 (Division euclidienne des polynômes) *Soient P et Q deux polynômes de $\mathcal{P}(\mathbf{K})$ tels que $d = d^oQ \geq 0$. Alors il existe $M, R \in \mathcal{P}(\mathbf{K})$ uniques tels que*

$$P = MQ + R, \qquad \textit{avec} \;\; d^o(R) < d.$$

Preuve
La preuve de l'existence se fait par récurrence sur le degré de P.

• si $d^oP < d$ alors $M = 0$ et $R = P$ conviennent.

• Soit $n \geq d$ et supposons qu'on ait l'existence tant que le degré de P est strictement plus petit que n. Pour P de degré n on écrit $P = b_nX^n + P_1$, avec P_1 de degré strictement inférieur à n. De même, on écrit $Q = a_dX^d + Q_1$ avec Q_1 de degré strictement inférieur à d. On a alors

$$P = \frac{b_n}{a_d} X^{n-d} Q + P_2$$

avec $d^oP_2 < n$. On applique alors l'hypothèse de récurrence au polynôme P_2 : on a $P_2 = QM_2+R$ avec $d^oR < d$. On en déduit

$$P = \left(\frac{b_n}{a_d} X^{n-d} + M_2 \right) Q + R,$$

qui a la forme souhaitée.

L'existence d'une division euclidienne vient d'être établie par récurrence. Il reste à montrer l'unicité. Supposons qu'on ait

$$P = M_1Q + R_1 = M_2Q + R_2, \qquad \textit{avec} \;\; d^oR_1 < d \;\; \text{et} \;\; d^oR_2 < d.$$

En soustrayant on a

$$(M_1-M_2)Q = R_2-R_1 \quad \Rightarrow \quad d^o(M_1-M_2)+d^oQ = d^o(R_1-R_2) < d^oQ \quad \Rightarrow \quad M_1 = M_2,\ R_1 = R_2.$$

Et l'unicité est établie. □

Proposition 1.3 *Soit P un polynôme à coefficients complexes de degré $n \geq 1$ et de coefficient dominant a_n. Alors il existe $(x_1, \cdots, x_n) \in C$ uniques à permutation près tels que*

$$P = a_n(X - x_1)\cdots(X - x_n).$$

On écrit aussi

$$P = a_n(X - y_1)^{r_1}\cdots(X - y_m)^{r_m},$$

avec $(y_1, \cdots, y_m)$ deux à deux distincts et $r_1 + \cdots + r_m = n$. Les y_i sont les racines de P et r_i est appelée multiplicité de la racine y_i.

1.3 Espaces euclidiens

Soit E un **R**-espace vectoriel, on dit que $(\cdot;\cdot)$ est un produit scalaire sur E si

1) l'application $E \times E \longrightarrow \mathbf{R},\ (x, y) \longmapsto (x; y)$ est bilinéaire ;

2) $\forall x \in E \quad (x; x) \geq 0$;

3) $\forall x \in E \quad (x; x) = 0 \Longleftrightarrow x = 0$;

4) $\forall x, y \in E \quad (x; y) = (y; x)$.

Tout **R** e.v. munit d'un produit scalaire est normé par $\|x\| = \sqrt{(x; x)}$.

On a l'inégalité de Cauchy-Schwarz :

$$\forall x, y \in E, \qquad (x; y) \leq \|x\|\,\|y\|.$$

Si $(e_1, \cdot, e_n)$ est une base de E, alors il existe $(f_1, \cdots, f_n)$ base orthogonale de E telle que

$$\text{vect}\,\{e_1, \cdots, e_i\} = \text{vect}\,\{f_1, \cdots, f_i\}, \qquad \text{pour } i \leq n.$$

Cette base se construit de manière récursive par $f_1 = e_1$; puis pour $i \geq 2$, $f_i = e_i - \sum_{j<i} \dfrac{(e_i; f_j)}{\|f_j\|^2} f_j$.
Cette base est appelée orthonormalisation de Schmidt de la base $(e_1, \cdots, e_n)$.

1.4 Algèbre linéaire

Soit $L : E \longrightarrow F$ un endomorphisme d'espaces vectoriels. Si E et F sont de dimensions finies n et p et admettent respectivement pour base $(e_1, \cdots, e_n)$ et $(f_1, \cdots, f_p)$, on peut représenter L

par la matrice $(M_{i,j})_{1\leq i\leq p,\, 1\leq j\leq n}$ définie par

$$Le_j = \sum_{i=1}^{p} M_{i,j} f_i.$$

Si $n = p$, l'endomorphisme L est inversible si et seulement si le déterminant de M est non nul.

1.5 Déterminant de Van der Monde

Proposition 1.4 *Soient* $(x_0, \cdots, x_n) \in \mathbf{R}^{n+1}$, *alors*

$$\det \begin{pmatrix} 1 & \cdots & \cdots & 1 \\ x_0 & \cdots & \cdots & x_n \\ \vdots & & & \vdots \\ x_0^n & \cdots & \cdots & x_n^n \end{pmatrix} = \prod_{i<j} (x_j - x_i).$$

Preuve
On fait la preuve par récurrence sur n.

- pour $n = 0$, la formule est vraie.

- Supposons la formule vraie pour un $n \geq 0$ et soit $x_0, \cdots, x_{n+1} \in \mathbf{R}$, on note $L_0, \cdots, L_{n+1}$ les matrices lignes extraites de la matrice de Van der Monde :

$$\begin{pmatrix} 1 & \cdots & \cdots & 1 \\ x_0 & \cdots & \cdots & x_{n+1} \\ \vdots & & & \vdots \\ x_0^{n+1} & \cdots & \cdots & x_{n+1}^{n+1} \end{pmatrix} = \begin{pmatrix} L_0 \\ \vdots \\ \vdots \\ L_{n+1} \end{pmatrix}$$

Soit $P(X)$ le polynôme $P = \prod_{0\leq i\leq n}(X - x_i) = X^{n+1} + \sum_{i=0}^{n} a_i X^i$. On a, en utilisant les propriétés élémentaires du déterminant,

$$\begin{aligned} \det \begin{pmatrix} L_0 \\ \vdots \\ L_{n+1} \end{pmatrix} &= \det \begin{pmatrix} L_0 \\ \vdots \\ L_n \\ L_{n+1} + \sum_{i=0}^{n} a_i L_i \end{pmatrix} = \det \begin{pmatrix} L_0 \\ \vdots \\ L_n \\ P(x_0) \cdots P(x_n)\, P(x_{n+1}) \end{pmatrix} = \det \begin{pmatrix} L_0 \\ \vdots \\ L_n \\ 0 \cdots 0\, P(x_{n+1}) \end{pmatrix} \\ &= P(x_{n+1}) \det \begin{pmatrix} 1 & \cdots & \cdots & 1 \\ x_0 & \cdots & \cdots & x_n \\ \vdots & & & \vdots \\ x_0^n & \cdots & \cdots & x_n^n \end{pmatrix} = P(x_{n+1}) \prod_{i<j\leq n} (x_j - x_i) = \prod_{i<j\leq n+1} (x_j - x_i). \end{aligned}$$

Et la formule est vraie au rang $n + 1$. □

2 Formules de quadrature

Soit $f : [a,b] \longrightarrow \mathbf{R}$. On veut calculer l'intégrale $\int_a^b f(x)\,dx$ ou au moins une approximation de cette intégrale.

2.1 Somme de Riemann

Exercice 2.1 *Calculer*

$$\lim_{n\to\infty} \sum_{i=0}^{n} \frac{1}{n+1+i}.$$

Théorème 2.1 *Si f est continue, alors*

$$\int_a^b f(x)\,dx = \lim_{n\to\infty} \underbrace{\frac{b-a}{n+1} \sum_{i=0}^{n} f\left(a + \frac{i}{n+1}(b-a)\right)}_{I_n(a,b;f)}$$

La formule $I_n(a,b;f)$ est appelée formule des rectangles à gauche.

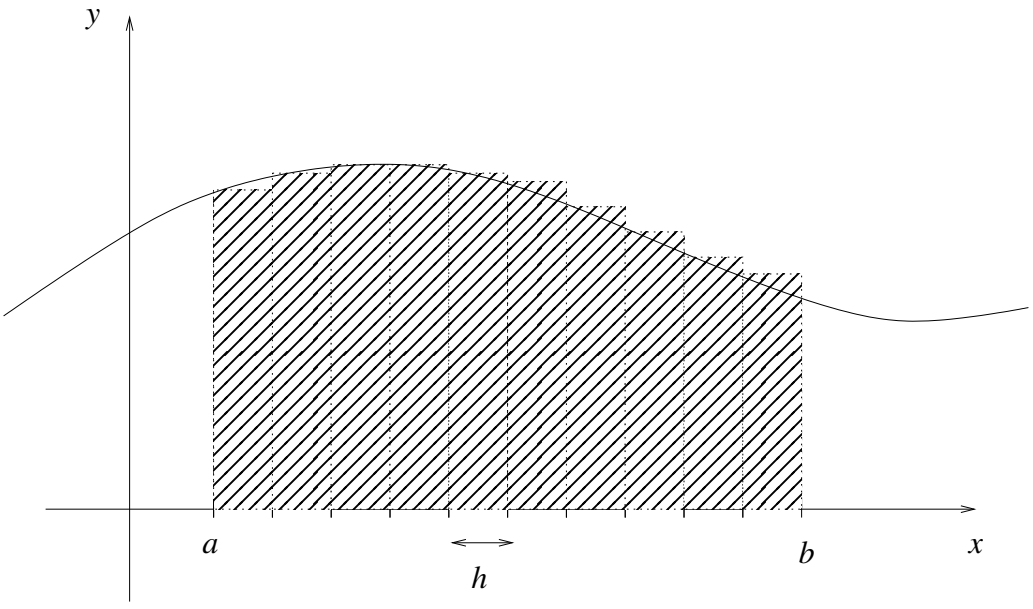


FIG. IV.1 – Dans la formule des rectangles à gauche $\int_a^b f$ est approchée par l'aire hachurée.

On notera $h = \dfrac{b-a}{n+1}$ et $x_i = a + \dfrac{i}{n+1}(b-a), \quad i = 0, \cdots, n+1$. l'ensemble $\{x_0, \cdots, x_{n+1}\}$ sera appelé sudivision uniforme de l'intervalle $[a,b]$. Le réel h sera appelé pas de la subdivision.

Plus généralement $x_0 = a < x_1 < \cdots < x_{n+1} = b$ sera une subdivision de $[a, b]$ et on notera $h_j = x_{j+1} - x_j$.

2.2 Méthode générale

On construit une subdivison $a = x_0 < x_1 < \cdots < x_n = b$ de l'intervalle $[a, b]$. On décompose $\int_a^b f$ en utilisant la relation de Chasles :

$$\int_a^b f(x)\,dx = \sum_{j=0}^{n-1} \int_{x_j}^{x_{j+1}} f(x)\,dx.$$

On est ramené au calcul de plusieurs intégrales pour lesquelles la longueur de l'intervalle d'intégration est relativement petite. Posant $h_j = x_{j+1} - x_j$, on a :

$$\int_{x_j}^{x_{j+1}} f(x)\,dx \;=\; h_j \int_0^1 \underbrace{f(x_j + th_j)}_{g(t)}\,dt.$$

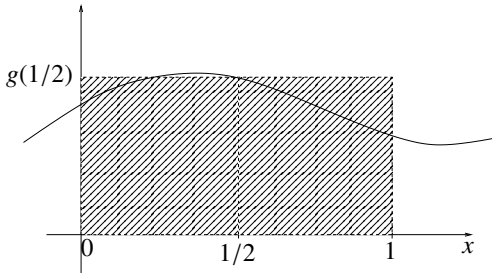
On s'est ramené au calcul de quantités de la forme $\int_0^1 g(t)\,dt$.

Remarque 2.2 *Si f est de classe C^1, g l'est aussi et $g'(t) = h_j f'(x_j + th_j) = O(h_j)$. De même, si f est de classe C^2, g aussi et $g''(t) = h_j^2 f''(x_j + th_j) = O(h_j^2)$. On doit approcher l'intégrale de fonctions qui varient lentement.*

Exemples

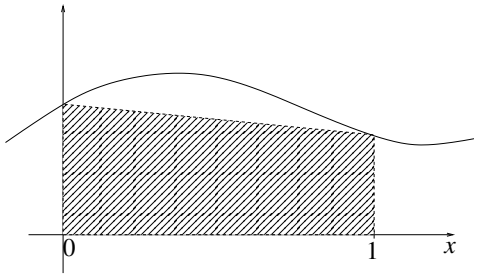
1) La formule du point milieu :

$$\int_0^1 g(t)\,dt \approx g\left(\frac{1}{2}\right)$$



2) La formule du trapèze :

$$\int_0^1 g(t)\,dt \approx \frac{1}{2}(g(0)+g(1))$$



Dans le cas 1) la méthode revient à remplacer g par la fonction constante égale à $g(1/2)$ sur $[0,1]$. Cette formule est exacte si g est une fonction constante (un polynôme de degré 0).

Dans le cas 2) on remplace g par la fonction affine P telle que $P(0) = g(0)$ et $P(1) = g(1)$. On voit que cette formule est exacte si g est une fonction polynomiale de degré 1.

3) En poursuivant la même idée. Si on considère le polynôme de degré deux satisfaisant $P(x) = g(x)$ pour $x = 0, 1/2, 1$ et qu'on approche l'intégrale de g par celle de P, on obtient la formule de Simpson.

$$\int_0^1 g(t)\,dt \;\approx\; \frac{1}{6}\left(g(0)+4g(1/2)+g(1)\right),$$

qui est exacte pour les polynômes de degré inférieur à 2.

4) Généralisation. On fixe $p \geq 1$ et on considère la subdivision uniforme $0 = x_0 < x_1, \cdot < x_p = b$ de l'intervalle $[0,1]$. Si on considère le polynôme P de degré inférieur à $p-1$ tel que $P(x_i) = g(x_i)$ pour $i = 0, \cdots, p$ et qu'on approche l'intégrale $\int_0^1 g$ par $\int_0^1 P$, on obtient les formules de Newton-Cotes.

2.3 Ordre

Définition 2.3 *Une formule de quadrature à N étages est donnée par*

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

où les (c_i) *sont des réels distincts deux à deux appelés* nœuds *d'intégration de la formule de quadrature. Les* (b_i) *sont appelés les* poids *de la formule de quadrature.*

Définition 2.4 *On dit que l'ordre de la formule de quadrature* (2.1) *est p si cette formule est* exacte *pour les polynômes de degré* $d \leq p-1$ *:*

$$\int_0^1 g(t)\,dt = \sum_{i=1}^{N} b_i g(c_i) \qquad si \quad g \quad est\ un\ polynôme\ tel\ que\ \ d^o g \leq p-1. \tag{2.2}$$

Théorème 2.5 *La formule* (2.1) *est d'ordre p si et seulement si*

$$\sum_{i=1}^{N} b_i c_i^{q-1} = \frac{1}{q} \qquad pour\ \ q = 1, \cdots, p. \tag{2.3}$$

Preuve
Il suffit de vérifier (2.2) pour la base canonique $1, X, X^2, \cdot, X^{p-1}$ de $\mathcal{P}_n(\mathbf{R})$. Ce qui donne (2.3).
□

Dans le cas $N = p$, le système linéaire (2.3) s'écrit

$$\begin{pmatrix} 1 & \cdots & \cdots & 1 \\ c_1 & \cdots & \cdots & c_N \\ c_1^2 & \cdots & \cdots & c_N^2 \\ \vdots & & & \vdots \\ c_1^{N-1} & \cdots & \cdots & c_N^{N-1} \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ \vdots \\ b_N \end{pmatrix} = \begin{pmatrix} 1 \\ 1/2 \\ \vdots \\ \vdots \\ 1/N \end{pmatrix}. \tag{2.4}$$

On reconnaît une matrice de Van der Monde qui est inversible dès que les (c_i) sont distincts deux à deux. On peut donc résoudre (2.4) en $b_1, \cdots, b_N$ de manière unique. Autrement dit :

Proposition 2.6 *Étant donnés N noeuds* $c_1, \cdots, c_N \subset [0,1]$*, 2 à 2 distincts, il existe une unique suite de poids* $b_1, \cdots, b_N$ *tel que la formule de quadrature correspondante soit d'ordre N.*

Exercice 2.2 *De quel ordre est la méthode du point milieu ? de Simpson ?*

Définition 2.7 *On dit qu'une formule de quadrature est symétrique si*

$$c_i = 1 - c_{N+1-i}, \qquad b_i = b_{N+1-i}, \qquad pour\ \ i = 1, \cdots, N.$$

Proposition 2.8 *Une formule de quadrature symétrique a toujours un ordre pair. C'est à dire que si elle est exacte pour les polynômes de degré inférieur à $2m-2$, alors elle est aussi exacte pour les polynômes de degré $2m-1$.*

Preuve
Supposons qu'on a une formule de quadrature symétrique d'ordre $2m-2$. Soit g une fonction polynomiale de degré $2m-1$, on a

$$g(t) \;=\; \alpha\left(t-\frac{1}{2}\right)^{2m-1} + g_1(t),$$

où $\alpha \in \mathbf{R}$ et g_1 est une fonction polynomiale de degré inférieur à $2m-2$. La formule de quadrature est exacte pour g_1 par hypothèse. D'autre part, par symétrie on a

$$\sum_{i=1}^{N} b_i\left(c_i-\frac{1}{2}\right)^{2m-1} \;=\; 0 \;=\; \int_0^1 \left(t-\frac{1}{2}\right)^{2m-1} dt.$$

Donc la formule est aussi exacte pour la fonction $t \mapsto \alpha\left(t-\frac{1}{2}\right)^{2m-1}$. □

3 Étude de l'erreur

On considère une subdivision unforme de $[a,b]$ de pas $h = \dfrac{b-a}{n}$. On veut étudier la différence

$$err \;=\; \int_a^b f(x)\,dx - \sum_{j=0}^{n-1}\sum_{i=1}^{N} b_i f(x_j + c_i h)$$

en fonction du coût numérique $\hat{c}$. Ce coût sera de l'ordre du nombre d'évaluations de la fonction f, i.e : $\hat{c} \;\propto\; Nn$.

On commence par étudier l'erreur sur un sous-intervalle de longueur h. On définit

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

On suppose que la formule de quadrature utilisée est d'ordre p mais pas $p+1$ et on suppose que f est de classe C^{p+1}. On a le développememnt limité pour $t \in [0,1]$,

$$f(x^\star+th) \;=\; \sum_{j=0}^{p-1} \frac{f^{(j)}(x^\star)}{j!} h^j t^j + \frac{f^{(p)}(x^\star)}{p!} h^p t^p + O(h^{p+1}).$$

De même

$$f(x^\star + c_i h) \;=\; \sum_{j=0}^{p-1} \frac{f^{(j)}(x^\star)}{j!} h^j c_i^j + \frac{f^{(p)}(x^\star)}{p!} h^p c_i^p + O(h^{p+1}).$$

D'où

$$E(f, x^\star, h) \;=\; \sum_{j=0}^{p-1} \frac{f^{(j)}(x^\star)}{j!} h^{j+1} \underbrace{\left\{ \int_0^1 t^j \, dt - \sum_{i=1}^{N} b_i c_i^j \right\}}_{=0} + \frac{f^{(p)}(x^\star)}{p!} h^{p+1} \underbrace{\left\{ \int_0^1 t^p \, dt - \sum_{i=1}^{N} b_i c_i^p \right\}}_{= 1/(p+1) - \sum_{i=1}^{N} b_i c_i^p} + O(h^{p+2}).$$

car la méthode est d'ordre p. $\qquad =: C_p \; (\neq 0).$

Donc

$$err \;=\; \frac{C_p}{p!} \left(\sum_{i=0}^{n-1} \int_{x_i}^{x_i+h} f^{(p)}(y) \, dy \right) h^p + O(h^{p+1}) \;=\; \frac{C_p}{p!} \Big[f^{(p-1)}(b) - f^{(p-1)}(a) \Big] h^p + O(h^{p+1}).$$

On a donc si $f^{(p-1)}(b) \neq f^{(p-1)}(a)$, on a

$$\ln err \overset{h \to 0^+}{\sim} p \ln h \qquad \text{(Voir figure IV.2.)}$$

On a démontré

Théorème 3.1 *Soit une formule de quadrature d'ordre p et f de classe C^{p+1}. On a :*

$$err \;=\; \frac{C_p}{p!} \Big(f^{(p-1)}(b) - f^{(p-1)}(a) \Big) h^p + O(h^{p+1}),$$

avec

$$C_p \;=\; \frac{1}{p+1} - \sum_{i=1}^{N} b_i c_i^p.$$

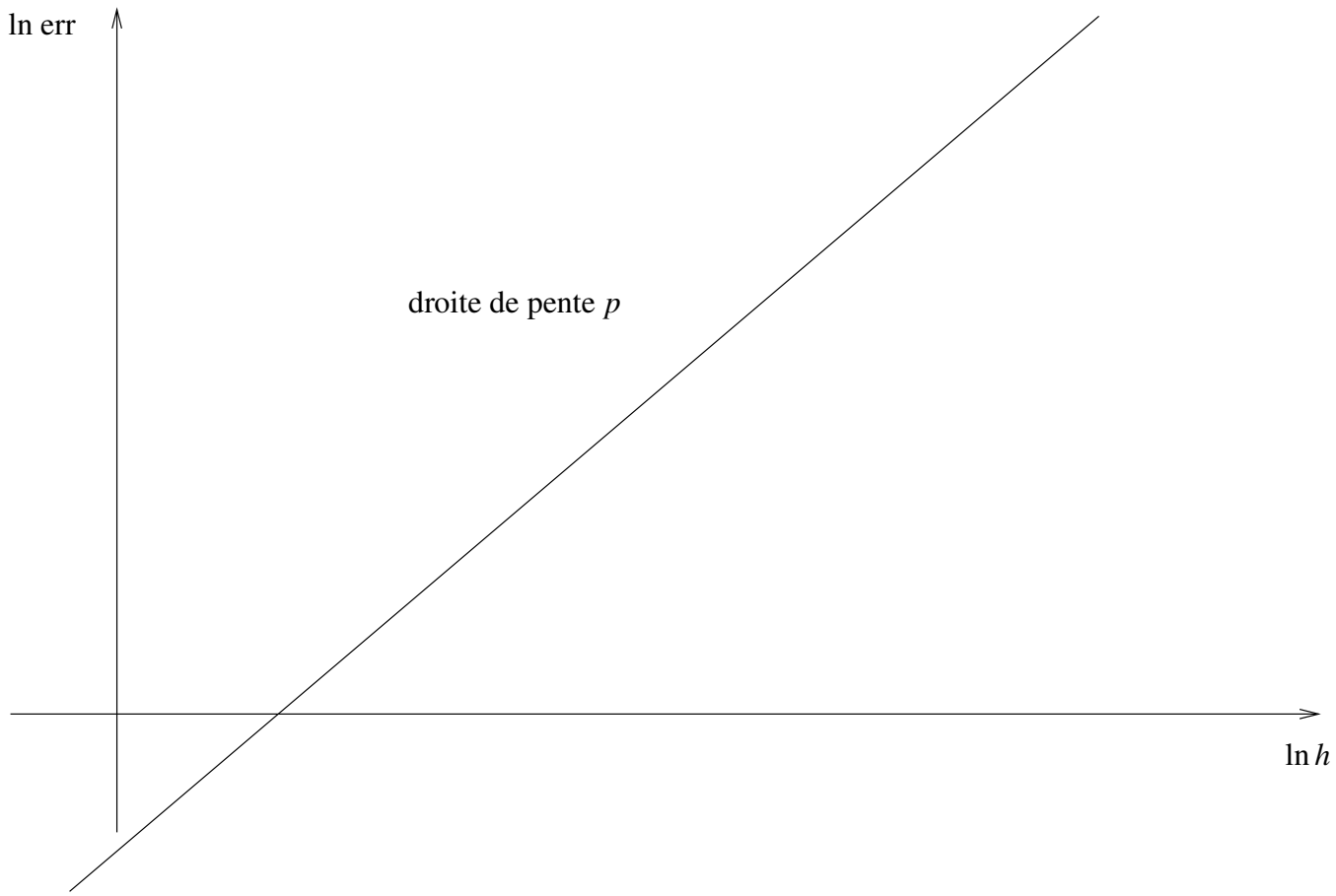


FIG. IV.2 – forme typique du graphe d'erreur

4 Étude approfondie de l'erreur

Théorème 4.1 *Soit une formule de quadrature d'ordre p et un entier $k \leq p$. Si $f : [x^\star, x^\star + h] \to \mathbf{R}$ est k fois continûment dérivable. L'erreur vérifie*

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

où $N_k(\tau)$ est le $k^{ième}$ noyau de Péano associé à la formule de quadrature

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

où pour $\sigma \in \mathbf{R}$, on a posé $\sigma_+ := \max(\sigma, 0)$ et où, par convention on a posé $0^0 = 0$.

Remarque 4.2 *On a ainsi une formule exacte pour l'erreur sans $o(\cdot)$ ou $O(\cdot)$.*

Preuve
On utilise la formule de Taylor avec reste intégral

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

On utilise l'identité

$$\int_0^t (t-\tau)^{k-1} g(\tau)\, d\tau \;=\; \int_0^1 (t-\tau)_+^{k-1} g(\tau)\, d\tau.$$

La méthode étant d'ordre k, l'erreur ne fait pas intervenir la partie de degré inférieur à $k-1$ du développpement de Taylor. En effet, on a

$$f(x^\star + th) \;=\; P_{k-1}(t) + h^k \int_0^t \frac{(t-\tau)^{k-1}}{(k-1)!} f^{(k)}(x^\star + \tau h)\, d\tau,$$

avec $E(P_{k-1}, 0, 1) = 0$ car $d^o P_{k-1} \leq k-1$. Donc

$$\begin{aligned}
E(f, x^\star, h) &= h^{k+1} \left(\int_0^1 \int_0^1 \frac{(t-\tau)_+^{k-1}}{(k-1)!} f^{(k)}(x^\star + \tau h)\, d\tau\, dt - \sum_{i=1}^N b_i \int_0^1 \frac{(c_i-\tau)_+^{k-1}}{(k-1)!} f^{(k)}(x^\star + \tau h)\, d\tau \right) \\
&\overset{\text{Fubini}}{=} h^{k+1} \int_0^1 \left\{ \int_0^1 \frac{(t-\tau)_+^{k-1}}{(k-1)!} dt - \sum_{i=1}^N b_i \frac{(c_i-\tau)_+^{k-1}}{(k-1)!} \right\} f^{(k)}(x^\star + \tau h)\, d\tau \\
&= h^{k+1} \int_0^1 \left\{ \frac{(1-\tau)^k}{k!} - \sum_{i=1}^N b_i \frac{(c_i-\tau)_+^{k-1}}{(k-1)!} \right\} f^{(k)}(x^\star + \tau h)\, d\tau.
\end{aligned}$$

□

On déduit du résultat précédent des estimations d'erreur

Proposition 4.3 *Si f est k fois continûment dérivable et que la formule est d'ordre $p \geq k$, on a :*

1) $\quad |err| \;\leq\; \|N_k\|_{L^\infty([0,1])} \left\|f^{(k)}\right\|_{L^1([a,b])} h^k$;

2) $\quad |err| \;\leq\; (b-a)\, \|N_k\|_{L^1([0,1])} \left\|f^{(k)}\right\|_{L^\infty([a,b])} h^k.$

Exercice 4.1 *Prouver la Proposition précédente.*

Exemples de noyaux de Péano

Formule du point milieu : $N = 1$, $b_1 = 1$, $c_1 = 1/2$.

1) on calcule

$$N_1(\tau) \;=\; \frac{(1-\tau)^1}{1!} - \frac{(\frac{1}{2}-\tau)_+^0}{0!} \;=\; 1 - \tau - \mathbf{1}_{[0,\frac{1}{2}]}(\tau) \;=\; \begin{cases} -\tau & \text{si } \tau \leq \frac{1}{2}, \\ 1-\tau & \text{si } \tau > \frac{1}{2}. \end{cases}$$

2) on calcule

$$N_2(\tau) \;=\; \frac{(1-\tau)^2}{2!} - \frac{(\frac{1}{2}-\tau)_+^1}{1!} \;=\; \frac{(1-\tau)^2}{2} - (\frac{1}{2}-\tau)_+ \;=\; \begin{cases} \dfrac{\tau^2}{2} & \text{si } \tau \le \frac{1}{2}, \\ \dfrac{(1-\tau)^2}{2} & \text{si } \tau > \frac{1}{2}. \end{cases}$$

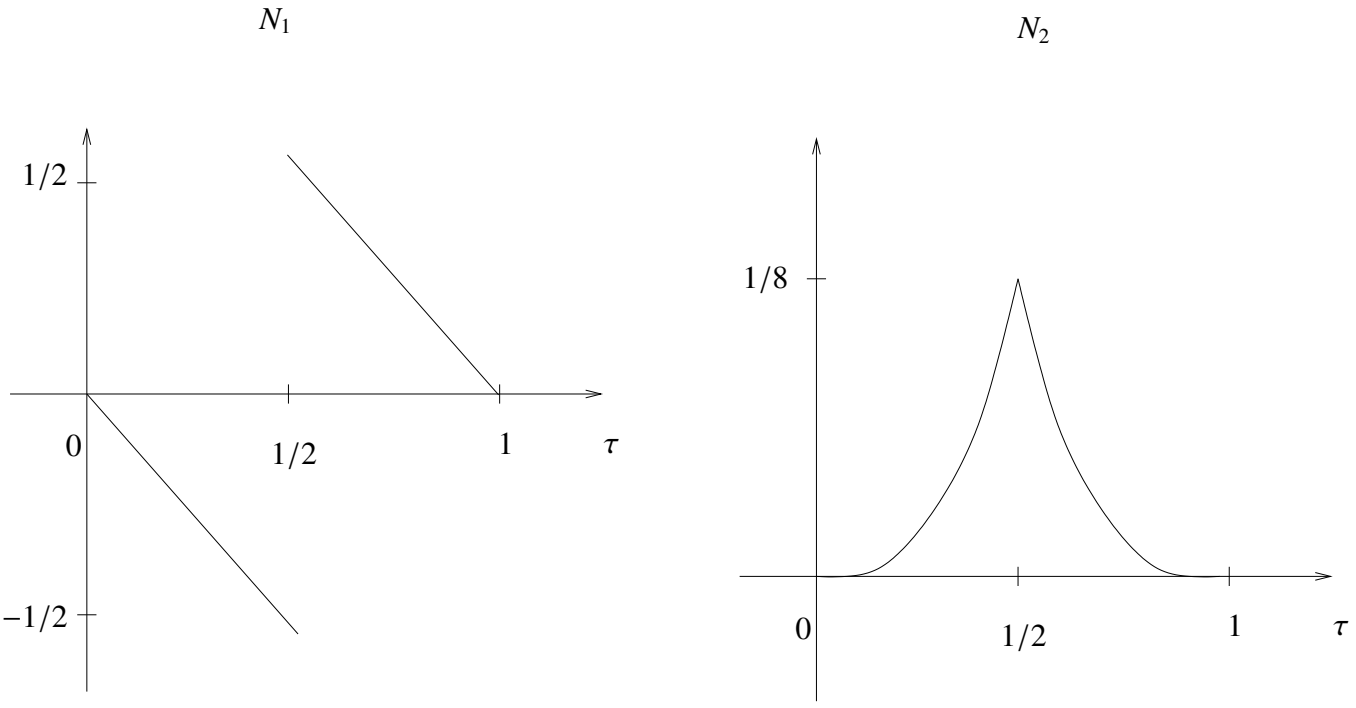


FIG. IV.3 – Repésentation graphique des noyaux N_1 et N_2 pour la formule du point milieu.

5 Choisir les nœuds d'intégration

On a vu que, une fois que les p nœuds d'intégration $c_1, \cdots, c_p$ étaient fixés, pour obtenir une méthode d'ordre p (au moins) on avait une et une seule façon de choisir les poids $b_1, \cdots, b_p$. Ces poids sont solution de (2.4), i.e :

$$\begin{pmatrix} 1 & \cdots & \cdots & 1 \\ c_1 & \cdots & \cdots & c_p \\ c_1^2 & \cdots & \cdots & c_p^2 \\ \vdots & & & \vdots \\ c_1^{p-1} & \cdots & \cdots & c_p^{p-1} \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ \vdots \\ b_p \end{pmatrix} = \begin{pmatrix} 1 \\ 1/2 \\ \vdots \\ \vdots \\ 1/p \end{pmatrix}.$$

Question : Y a-t-il un moyen de choisir les points $c_1, \cdots c_p$ de sorte que la méthode soit d'ordre $q > p$?

Remarque 5.1 *Si une méthode est d'ordre p_1 alors elle est d'ordre p_2 dès que $p_2 \leq p_1$.*

Théorème 5.2 (Caractérisation de l'ordre d'une méthode) *Soient $(b_i, c_i)_{i=1}^p$ les poids et nœuds d'une formule d'ordre p. Soit*

$$M(t) := (t - c_1) \cdots (t - c_p) = \prod_{i=1}^{p} (t - c_i).$$

Alors la méthode est d'ordre $p + m$ si et seulement si

$$\int_0^1 M(t) g(t)\, dt = 0 \qquad \text{pour tout polynôme } g \text{ tel que } d^o g \leq m - 1. \tag{5.1}$$

Preuve
Soit f un polynôme de degré p+m-1. On fait la division euclidienne de f par M :

$$f(t) = M(t)g(t) + r(t) \qquad \text{où } d^o r < d^o M = N \quad \text{et} \quad d^o g \leq d^o f - d^o M = m - 1.$$

On a

- $$\int_0^1 f(t)\, dt = \int_0^1 M(t) g(t)\, dt + \int_0^1 r(t)\, dt.$$
- $$\sum_{i=1}^{N} b_i f(c_i) = \sum_{i=1}^{N} b_i \underbrace{M(c_i)}\, g(c_i) + \underbrace{\sum_{i=1}^{N} b_i r(c_i)}.$$
  $$= 0 \text{ par déf de } M \qquad = \int_0^1 r(t)\, dt \quad \text{car } d^o r < p.$$

Donc on a

$$\int_0^1 f(t)\, dt = \sum_{i=1}^{N} b_i f(c_i) \qquad \text{pour tout polynôme } f \text{ tel que } d^o f \leq p + m - 1,$$

si et seulement si (5.1) est vrai. □

Corollaire 5.3 *Si q est l'ordre d'une formule de quadrature à N étages alors $q \leq 2N$.*

Preuve
Si elle est d'ordre $2N + 1$ alors

$$\int_0^1 M(t) g(t)\, dt = 0 \qquad \text{pour} \qquad d^o g = N.$$

En particulier pour $g = M$, on a $\int_0^1 M^2(t)\, dt = 0.$ *Ce qui est faux.* □

Conclusion : Pour construire une formule de quadrature d'ordre maximal $2N$ il faut (et il suffit) que le polynôme

$$M(t) \;=\; (t-c_1)\cdots(t-c_N)$$

soit orthogonal aux polynômes de degré strictement inférieur à N pour le produit scalaire

$$(f;g) \;=\; \int_0^1 f(t)g(t)\,dt.$$

6 Polynômes orthogonaux

On cherche donc M de degré N qui soit orthogonal aux polynômes de degré $< N$ pour le produit scalaire

$$(f;g) \;=\; \int_0^1 f(t)g(t)\,dt.$$

Remarque 6.1 *En fait on s'interesse aux racines $(c_1,\cdots,c_N)$ du polynôme M (encore faut-il qu'il ait N racines distinctes). Nous devrons donc : prouver l'existence de M, démontrer si possible qu'il a bien N racines distinctes, trouver une méthode de calcul de ces racines.*

Définition 6.2 *Plus généralement. Soit $\omega : (a,b) \to (0,+\infty)$ une fonction de poids intégrable telle que*

$$\int_a^b t^k\omega(t)\,dt \;<\; \infty, \qquad k=1,2,\cdots$$

Eventuellement, a peut être $-\infty$ ou (/ et) b peut être égal à $+\infty$.

Définition 6.3 *On définit le produit scalaire*

$$\langle f\,;\,g\rangle \;=\; \int_a^b f(t)g(t)\omega(t)\,dt$$

sur l'espace des polynômes : $f,g \in \mathcal{P}(\mathbf{R})$.

Définition 6.4 *On dit que les fonctions polynomiales f et g sont orthogonales si et seulement si*

$$\langle f\,;\,g\rangle \;=\; 0.$$

Théorème 6.5 *Il existe une suite de polynômes $p_0,p_1,\cdots,p_k,\cdots$ telle que*

1) $p_0 = 1$.

2) Pour $k \geq 1$, $p_k(t) \;=\; t^k + q_k(t)$, avec $d^o q_k \leq k-1$.

3) Pour $k \geq 1$, on a $\langle p_k\,;\,g\rangle \;=\; 0$ pour tout polynôme g tel que $d^o g \leq k-1$.

En particulier,

4) pour $k \geq 0$, $(p_0, \cdots, p_k)$ est une base $\mathcal{P}_k(\mathbf{R})$.

Cette suite de polynômes est unique et satisfait la relation de récurrence :

5) $p_{-1}(t) = 0$, $p_0(t) = 1$, et pour $k \geq 0$,

$$p_{k+1}(t) \;=\; (t - \beta_{k+1})p_k(t) - \gamma_{k+1}^2 p_{k-1}(t),$$

avec

$$\beta_{k+1} \;=\; \frac{\langle tp_k \,;\, p_k\rangle}{\langle p_k \,;\, p_k\rangle}, \qquad\qquad \gamma_{k+1}^2 \;=\; \frac{\langle p_k \,;\, p_k\rangle}{\langle p_{k-1} \,;\, p_{k-1}\rangle}.$$

Preuve
On construit $p_0, p_1, p_2, \cdots$ comme l'orthonormalisation de Schmidt de la base canonique $1, t, t^2, \cdots$. C'est-à-dire qu'on pose $p_0(t) = 1$ puis pour $k \geq 0$,

$$p_{k+1}(t) \;=\; t^{k+1} - \sum_{i=0}^{k} \frac{\langle t^{k+1} \,;\, p_i\rangle}{\langle p_i \,;\, p_i\rangle}\, p_i(t). \tag{6.1}$$

On montre par récurrence que 1), 2) et 3) sont vérifiées.

- Pour $k = 0$, 1) est vrai par construction et pour 2) et 3), il n'y a rien à vérifier.

- Supposons que 1), 2) et 3) soient vérifiées jusqu'au rang k. Au rang $k + 1$, par construction on a (6.1) et comme par hypothèse de récurrence les polynômes p_i sont de degré i pour $i \leq k$, on en déduit que 2) est vérifiée au rang $k + 1$.

Pour 3), on sait par hypothèse de récurrence que la famille $(p_0, \cdots, p_k)$ est une base $\mathcal{P}_k(\mathbf{R})$, il suffit donc de vérifier 3) pour $g = p_0, \cdots, p_k$. Or pour $n = 0, \cdots, k$, on a

$$\begin{aligned}
\langle p_{k+1} \,;\, p_n\rangle \;&=\; \langle t^{k+1} \,;\, p_n\rangle - \sum_{i=0}^{k} \frac{\langle t^{k+1} \,;\, p_i\rangle}{\langle p_i \,;\, p_i\rangle} \underbrace{\langle p_i \,;\, p_n\rangle}_{= 0 \text{ pour } i \neq n} \\
&=\; \langle t^{k+1} \,;\, p_n\rangle - \langle t^{k+1} \,;\, p_n\rangle \;=\; 0.
\end{aligned}$$

Donc 2) et 3) sont vérifiées au rang $k + 1$.

Nous avons donc établi 2) et 3) par récurrence.

Le point 4) est clair : par construction, on a vect $(p_0, \cdots, p_k) = \mathcal{P}_k(\mathbf{R})$.

Pour montrer la relation de récurrence 5) on remarque que comme p_{k+1} est de degré $k + 1$ avec pour coefficient dominant 1 et que p_k est de degré k avec pour coefficient dominant 1, on peut écrire

$$p_{k+1}(t) \;=\; tp_k(t) + h_{k+1}(t),$$

avec $d^o h_{k+1} \leq k$. Comme $(p_0, \cdots, p_k)$ est une base de $\mathcal{P}_k(\mathbf{R})$, on a le développement

$$p_{k+1}(t) \;=\; tp_k(t) + \sum_{i=0}^{k} \alpha_i p_i(t).$$

Pour déterminer les α_i, on utilise l'orthogonalité de la suite $(p_0, \cdots, p_{k+1})$: prenant le produit scalaire avec les p_i, on a

$$0 \;=\; \langle p_i\,;\, p_{k+1}\rangle \;=\; \langle p_i\,;\, tp_k\rangle + \sum_{j=0}^{k} \alpha_j \underbrace{\langle p_j\,;\, p_i\rangle}_{=0 \text{ pour } j\neq i} \;=\; \langle p_i\,;\, tp_k\rangle + \alpha_i \langle p_i\,;\, p_i\rangle$$

D'où, pour $i = 0, \cdots, k$,

$$\alpha_i \;=\; -\frac{\langle p_i\,;\, tp_k\rangle}{\langle p_i\,;\, p_i\rangle} \;=\; -\frac{\langle tp_i\,;\, p_k\rangle}{\langle p_i\,;\, p_i\rangle}.$$

Si $i < k-1$, alors $tp_i \in \mathcal{P}_{k-1}(\mathbf{R})$ et $\langle tp_i\,;\, p_k\rangle = 0$ ainsi $\alpha_i = 0$ pour $i = 0, \cdots, k-2$.

Pour $i = k-1$, on écrit $tp_{k-1} = p_k + h_k$ avec $h_k \in \mathcal{P}_{k-1}(\mathbf{R})$, d'où

$$\alpha_{k-1} \;=\; -\frac{\langle p_k\,;\, p_k\rangle}{\langle p_{k-1}\,;\, p_{k-1}\rangle} \;=:\; -\gamma_{k+1}^2.$$

Finalement, pour $i = k$, on a

$$\alpha_k \;=\; -\frac{\langle tp_k\,;\, p_k\rangle}{\langle p_k\,;\, p_k\rangle} \;=:\; -\beta_{k+1}.$$

Et 5) est établie. □

Exemples : Les polynômes orthogonaux peuvent être normalisés autrement que par la condition $p_k(t) - t^k \in \mathcal{P}_{k-1}(\mathbf{R})$. Les polynômes orthogonaux les plus connus ont une normalisation traditionnelle, nous en énumérons quelques uns en précisant leurs noms, l'intervalle (a, b) et le poids ω :

- les polynômes de Legendre, $P_k(t)$: $(a, b) = (-1, 1)$, $\omega(t) = 1$; ils sont normalisés par la condition $P_k(1) = 1$.
- les polynômes de Chebychev, $T_k(t)$: $(a, b) = (-1, 1)$, $\omega(t) = \dfrac{1}{\sqrt{1-t^2}}$;
- les polynômes de Laguerre, $L_k^{\alpha}(t)$: $(a, b) = (0, \infty)$, $\omega(t) = t^{\alpha}e^{-t}$, $\alpha > -1$;
- les polynômes de Hermite, $H_k(t)$: $(a, b) = (-\infty, +\infty)$, $\omega(t) = e^{-t^2}$;
- les polynômes de Jacobi, $P_k^{\alpha,\beta}$: $(a, b) = (-1, 1)$, $\omega(t) = (1-t)^{\alpha}(1+t)^{\beta}$.

Théorème 6.6 *Soit $p_k(t)$ un polynôme orthogonal à tous les polynômes de degré inférieur à $k-1$ alors toutes les racines de $p_k(t)$ sont réelles, simples et dans l'intervalle (a, b).*

Preuve
Soient $t_1, \cdots, t_r$ les racines de $p_k(t)$ qui sont telles que $t_i \in (a, b)$ et p_k change de signe en t_i.

On pose $g(t) = (t - t_1)\cdots(t - t_r)$ alors si $r < k$, on a $d^o g < k$ et $\langle g\,;\, p_k\rangle = 0$, i.e :

$$\int_a^b \underbrace{p_k(t)g(t)}_{\text{de signe constant}} \underbrace{\omega(t)}_{>0}\, dt \;=\; 0.$$

Donc $p_k(t)g(t) = 0$ pour $t \in (a,b)$, ce qui est absurde. Donc $r = k$ et comme p_k est de degré k, toutes ces racines sont simples. □

Théorème 6.7 (Formule de Rodrigues) *Considérons*

$$\tilde{p}_k(t) \;=\; \frac{1}{\omega(t)}\frac{d^k}{dt^k}\Big\{\omega(t)(t-a)^k(b-t)^b\Big\}, \tag{6.2}$$

Dès que le membre de droite de (6.2) *est un polynôme de degré k, le polynôme orthogonal p_k s'écrit $C_k\tilde{p}_k$ où la constante $C_k \neq 0$ dépend de la normalisation choisie.*

Preuve
Soit $g(t)$ un polynôme de degré $q \leq k-1$, on a en intégrant par partie successivement

$$\begin{aligned}
\langle \tilde{p}_k(t)\,;\, g(t)\rangle \;&=\; \int_a^b \frac{d^k}{dt^k}\Big(\omega(t)(t-a)^k(b-t)^k\Big)\,g(t)\,dt \\
&\overset{ipp}{=} -\int \frac{d^{k-1}}{dt^{k-1}}\Big(\omega(t)(t-a)^k(b-t)^k\Big)\,g'(t) + \Bigg[\omega(t)\,\underbrace{\frac{d^{k-1}}{dt^{k-1}}\Big((t-a)^k(b-t)^k\Big)}_{=\,0 \text{ car } a \text{ et } b \text{ racines de multiplicité } k}\,g(t)\Bigg]_a^b \\
&= \;\cdots\; = 0.
\end{aligned}$$

□

Exemple : les polynôme de Legendre : $\omega(t) = 1$ sur $(-1,1)$. Ces polynômes sont normalisés par la condition $P_k(1) = 1$. Ils sont donnés par

$$P_k(t) \;=\; \frac{(-1)^k}{k!2^k}\frac{d^k}{dt^k}\Big(\omega(t)(1+t)^k(1-t)^k\Big)$$

On a

$$P_0(t) = 1\,; \qquad P_1(t) = t\,; \qquad P_2(t) = \frac{3}{2}t^2 - \frac{1}{2}; \qquad P_3(t) = \frac{5}{2}t^3 - \frac{3}{2}t.$$

7 Formules de quadrature de Gauss

On construit ici une formule de quadrature avec un ordre $p = 2N$. On a vu qu'il existait un polynôme $M(t)$ de degré N orthogonal à tous les polynômes de degré inférieurs à $N-1$: i.e.

$$\int_0^1 M(t)g(t)\,dt \;=\; 0 \qquad \text{si } d^o g \leq N-1.$$

Ce polynôme est donné par

$$M(t) \;=\; CP_N(2t-1)$$

où P_N est le $N^{\text{ième}}$ polynôme de Legendre car

$$\int_0^1 P_N(2t-1)g(t)\,dt \;=\; \frac{1}{2}\int_{-1}^1 P_N(x)g((x+1)/2)\,dx \;=\; 0.$$

Toutes les racines de M sont situées dans l'intervalle $(0,1)$.

Théorème 7.1 *Il existe une unique formule de quadrature à N étages d'ordre $2N$. Elle est donnée par $c_1,\cdots,c_N$ racines de $P_N(2t-1)$ et $b_1,\cdots,b_N$ donnés comme l'unique solution du système linéaire* (2.4)*.*

Calcul des nœuds :

$N=1$: $\quad \int_0^1 g(t)\,dt \simeq g(\frac{1}{2})$;

$N=2$: $\quad \int_0^1 g(t)\,dt \simeq \frac{1}{2}g\left(\frac{1}{2}-\frac{\sqrt{3}}{6}\right)+\frac{1}{2}g\left(\frac{1}{2}+\frac{\sqrt{3}}{6}\right)$;

$N=3$: $\quad \int_0^1 g(t)\,dt \simeq \frac{5}{18}g\left(\frac{1}{2}-\frac{\sqrt{15}}{10}\right)+\frac{4}{9}g\left(\frac{1}{2}\right)+\frac{5}{18}g\left(\frac{1}{2}+\frac{\sqrt{15}}{10}\right)$.

Au delà de $N=10$, on ne sait plus calculer exactement les racines. En effet on doit résoudre des équations de degré 5 et plus et il n'existe pas de formules par radicaux pour calculer les racines du polynôme de Legendre. Comment fait-on ?

On utilise la formule de récurrence

$$(k+1)P_{k+1}(x) \;=\; (2k+1)xP_k(x) - kP_{k-1}(x).$$

pour calculer les coefficients du polynômes P_{k+1}. Pour le calcul des racines de P_{k+1} on utilise le fait que les racines sont encadrées par celles de P_k : on a

$$-1 < x_1^{k+1} < x_1^k < x_2^{k+1} < x_2^k < \cdots < x_k^k < x_{k+1}^{k+1} < 1.$$

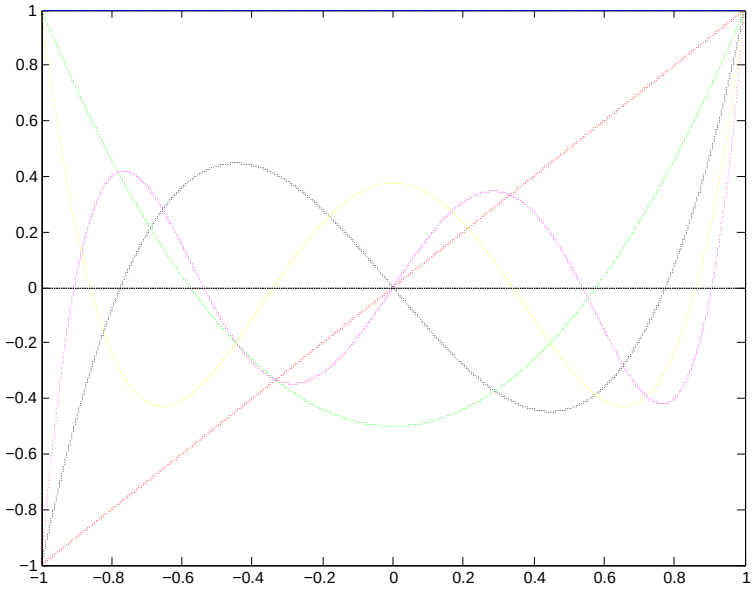
On applique une méthode de Dichotomie (puis Newton) pour déterminer ces racines.

Pour le calcul des poids (b_i), on a la formule explicite :

$$b_i \;=\; \frac{1-x_i^2}{N^2(P_{N-1}(x_i))^2}.$$

Les premiers polynômes de Legendre

$$\begin{aligned}
P_0(x) &= 1\\
P_1(x) &= x\\
P_2(x) &= \frac{1}{2}(3x^2-1)\\
P_3(x) &= \frac{1}{2}(5x^3-3x)\\
P_4(x) &= \frac{1}{8}(35x^4-30x^2+3)\\
P_5(x) &= \frac{1}{8}(63x^5-70x^3+15x)
\end{aligned}$$


1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1

V Interpolation polynomiale

Le problème de l'interpolation consiste à chercher une fonction simple (polynôme, polynôme trigonométrique, polynôme par morceaux) prenant en des points donnés $x_0, \cdots, x_n$ les valeurs prescrites $y_0, \cdots, y_n$.

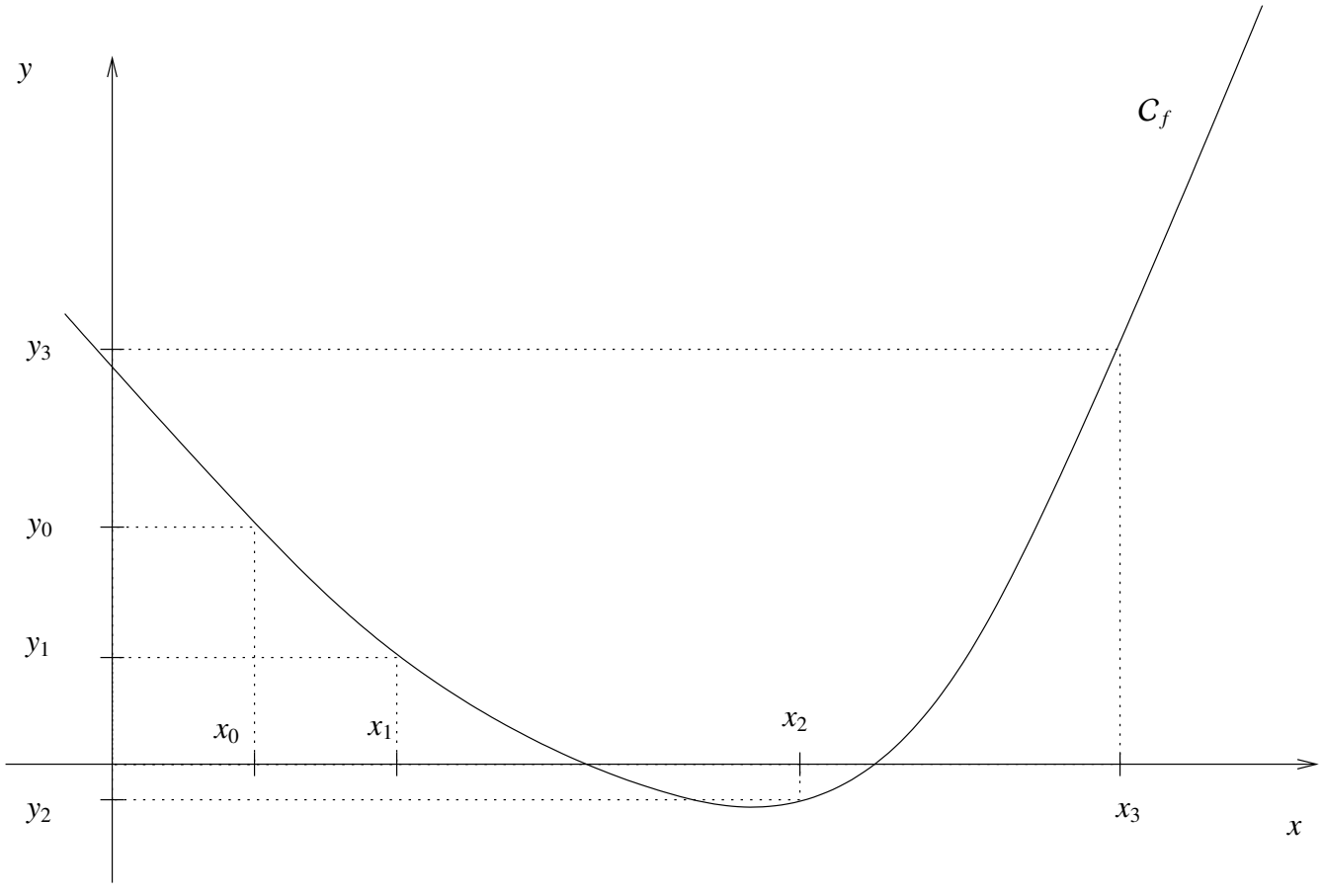


FIG. V.1 – Graphe d'une fonction interpolant les points $(x_0, y_0), \cdots, (x_3, y_3)$

Autrement dit on cherche une fonction f (parmi une certaine famille de fonctions) dont le graphe passe par les points $(x_0, y_0), \cdots, (x_n, y_n)$. On dit que la fonction f interpole les points $(x_0, y_0), \cdots, (x_n, y_n)$.

1 Différences divisées et formule de Newton

On se donne les réels $(x_0, \cdots, x_n)$ et $(y_0, \cdots, y_n)$. On souhaite trouver une fonction polynomiale de degré n interpolant les points $(x_0, y_0), \cdots, (x_n, y_n)$. On a à résoudre $(n+1)$ équations

$$P(x_0) = y_0, \quad \cdots \quad , P(x_n) = y_n,$$

et on a $(n+1)$ inconnues : les coefficients $a_{n+1}, a_n, \cdots, a_0$ du polynôme P.

On aboutit à la résolution du système linéaire

$$\begin{pmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & & & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

On reconnaît une matrice de Van der Monde et on en déduit que le système admet une solution unique dès que les (x_i) sont deux à deux distincts.

Pour construire le polynôme P, plutôt que de chercher à inverser la matrice, nous utilisons la méthode suivante : on cherche directement une base $(l_0, \cdots, l_n)$ de polynômes de degrés inférieurs à n satisfaisant

$$l_i(x_j) \; = \; \delta_{ij}, \qquad \text{pour } \; 0 \le i, j \le n.$$

(On rappelle que le symbole de Kronecker δ_{ij} vaut 1 si $i = j$ et 0 si $i \neq j$.)

Ensuite on écrira

$$P(x) \; = \; \sum_{i=0}^{n} y_i l_i(x). \tag{1.1}$$

Pour construire les (l_i), on remarque que si $j \neq i$, on a $l_i(x_j) = 0$ donc x_j est racine de l_i. Le polynôme l_i s'écrit donc

$$l_i(x) \; = \; a(x) \prod_{j=0,\cdots,n,\, j\neq i} (x - x_j)\,.$$

Comme le produit est de degré n est que l_i doit être au plus de degré n on en déduit que a est constant. Pour déterminer a, on écrit $l_i(x_i) = 1$. On a finalement

$$l_i(x) \; = \; \prod_{j=0,\cdots,n,\, j\neq i} \frac{x - x_j}{x_i - x_j}, \qquad i = 0, \cdots, n. \tag{1.2}$$

On déduit le polynôme interpolateur $P(x)$ des identités (1.1) et (1.2).

Remarque 1.1 (historique) *Ce type de problème est apparu en astronomie. La situation est la suivante : des astronomes ont fait plusieurs observations d'un astéroïde , ils ont mesuré sa*

position (angle y) à certaines dates et heures (x) ; ils souhaitent savoir dans quelle partie du ciel il va apparaître à une date ultérieure.

A chaque nouvelle observation, les astronomes ont une information en plus et il leur faut recalculer tous le polynômes (l_i) depuis le début. La méthode qui suit due à Newton permet de réduire la quantité de calculs à effectuer.

On part du polynôme P_0 interpolant (x_0, y_0) puis on ajoute successivement des points en augmentant le degré du polynôme interpolant.

Pour $n = 0$, on a $P_0(x) = y_0$.

Pour $n = 1$, on a clairement $P_1(x) = y_0 + \alpha(x - x_0)$. Et on détermine α en faisant $P_1(x_1) = y_1$. On obtient

$$P_1(x) \;=\; y_0 + \frac{y_1 - y_0}{x_1 - x_0}(x - x_0).$$

Pour $n = 2$, on écrit $P_2(x) = P_1(x) + \beta(x - x_1)(x - x_0)$ et on détermine β en faisant $P_2(x_2) = y_2$. On calcule

$$\beta \;=\; \frac{1}{x_2 - x_1}\left(\frac{y_2 - y_0}{x_2 - x_0} - \frac{y_1 - y_0}{x_1 - x_0}\right).$$

En poursuivant, on est amené à introduire les quantités ci-après.

Définition 1.2 *Pour une famille $(x_i, y_i)_{0\le i\le n}$, on définit récursivement*

$$\begin{aligned}
y[x_i] &= y_i,\\
\Delta y[x_i, x_j] &= \frac{y[x_j] - y[x_i]}{x_j - x_i},\\
&\vdots\\
\Delta^n y[x_{i_0}, x_{i_1}, \cdots, x_{i_n}] &= \frac{\Delta^{n-1} y[x_{i_1}, \cdots, x_{i_n}] - \Delta^{n-1} y[x_{i_0}, \cdots, x_{i_{n-1}}]}{x_{i_n} - x_{i_0}}.
\end{aligned}$$

Théorème 1.3 (formule de Newton)

Le polynôme d'interpolation de degré n qui passe par $(x_0, y_0), \cdots, (x_n, y_n)$ (les (x_i) étant deux à deux distincts) est unique et donné par

$$p(x) \;=\; y[x_0] + \Delta y[x_0, x_1]\,(x - x_0) + \cdots + \Delta^n y[x_0, \cdots, x_n]\,(x - x_0)\cdots(x - x_{n-1}).$$

Preuve
Pour $n = 0, 1, 2$ le résultat correspond aux calculs ci-dessus.

Supposons que la formule de Newton soit vraie au rang $n - 1$. On se donne les $(n + 1)$ paires de points $(x_0, y_0), \cdots, (x_n, y_n)$ et on cherche le polynôme p de degré au plus n interpolant ces

points. Par hypothèse de récurrence, on a nécessairement

$$p(x) = \underbrace{y[x_0] + \Delta y[x_0, x_1]\,(x - x_0) + \cdots + \Delta^{n-1} y[x_0, \cdots, x_{n-1}]\,(x - x_0)\cdots(x - x_{n-2})}_{=:\ p_1(x)} + a(x - x_0)\cdots(x - x_{n-1}). \tag{1.3}$$

où $a \in \mathbf{R}$ se détermine en écrivant $p(x_n) = y_n$. On note aussi

$$p_2(x) \ := \ y[x_1] + \Delta y[x_1, x_2]\,(x - x_1) + \cdots + \Delta^{n-1} y[x_1, \cdots, x_n]\,(x - x_1)\cdots(x - x_{n-1}).$$

En fait on remarque qu'on a

$$p(x) \ = \ \frac{1}{x_n - x_0}\left((x_n - x)p_1(x) + (x - x_0)p_2(x)\right) \ =: \ q(x). \tag{1.4}$$

En effet pour $1 \le i \le n$, on a $p_1(x_i) = p_2(x_i) = y_i$ donc $q(x_i) = y_i$. Pour $i = 0$, on a $p_1(x_0) = y_0$ d'où $q(x_0) = y_0$ et pour $i = n$, on a $p_2(x_n) = y_n$ d'où $q(x_n) = y_n$. Finalement, $q(x_i) = p(x_i)$ pour $i = 0, \cdots, n$ et comme le polynôme $p - q$ est de degré au plus n, on a $p = q$. On a prouvé la validité de (1.4).

En comparant (1.3) et (1.4) et en identifiant les coefficients de plus haut degré, on obtient

$$a \ = \ \frac{\Delta^{n-1} y[x_1, \cdots, x_n] - \Delta^{n-1} y[x_0, \cdots, x_{n-1}]}{x_n - x_0}.$$

Ce qui établit le résultat par récurrence sur n. □

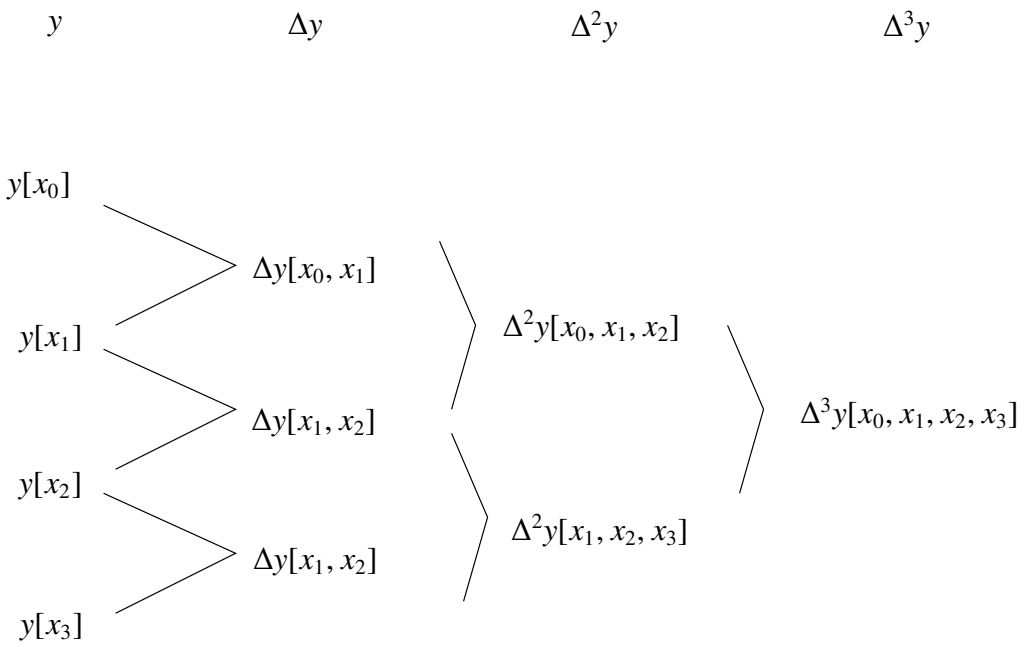


FIG. V.2 – Diagramme de calcul des différences divisées de Newton ($n = 3$).

2 Estimation d'erreur pour l'interpolation et polynômes de Chebyshev

Soit $f : [a,b] \longrightarrow \mathbf{R}$. Soient $x_0, \cdots x_n \in [a,b]$ des points distincts deux à deux. On pose $y_i = f(x_i)$ pour $i = 0, \cdots, n$. Dans toute cette section p désignera le polynôme d'interpolation des points (x_i, y_i). On s'interesse à la distance

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

Lemme 2.1 *Soit f une fonction n fois différentiable et $y_i = f(x_i)$ pour $i = 0, \cdots, n$ distincts deux à deux. Alors il existe $\xi \in (\min x_i, \max x_i)$ tel que*

$$\Delta^n y[x_0, \cdots, x_n] = \frac{f^{(n)}(\xi)}{n!}.$$

Preuve
Soit p le polynôme d'interpolation de degré n passant par (x_i, y_i). La fonction $d = f - p$ s'annule en $x_0, \cdots, x_n$: en appliquant successivement le Théorème de Rolle, on en déduit que d' s'annule n fois puis d'' s'annule $(n-1)$ fois, ... Au final, la fonction $d^{(n)}$ s'annule au moins une fois en un point ξ de l'intervalle $(\min x_i, \max x_i)$:

$$0 = d^{(n)}(\xi) = f^{(n)}(\xi) - n!\Delta^n y[x_0, \cdots, x_n].$$

□

Proposition 2.2 (Erreur d'interpolation) *Soit $f : [a,b] \longrightarrow \mathbf{R}$ une fonction $n+1$ fois différentiable. Il existe $\xi \in (\min\{x_i, x\} \max\{x_i, x\})$ tel que*

$$f(x) - p(x) = (x - x_0) \cdots (x - x_n) \frac{f^{(n+1)}(\xi)}{(n+1)!}.$$

Preuve
On pose $x_{n+1} = x$, $y_{n+1} = f(x)$ et on considère le polynôme p_{n+1}de degré $n+1$ interpolant les points $(x_0, y_0), \cdots, (x_{n+1}, y_{n+1})$. On a

$$f(x) = p_{n+1}(x) = p(x) + \Delta^{n+1} y[x_0, \cdots, x_n, x](x - x_0) \cdots (x - x_n).$$

On conclut en appliquant le Lemme précédent. □

Pour avoir une erreur $\|f - p\|_\infty$ la plus petite possible, on va chercher des nœuds d'interpolation $(x_0, \cdots, x_n)$ tels que

$$\sup_{x \in [a,b]} |(x - x_0) \cdots (x - x_n)|$$

soit le plus petit possible. Pour cela nous allons introduire les polynômes de Chebyshev.

Définition 2.3 *Pour $n = 0, 1, \cdots$ et pour $x \in [-1, 1]$, on pose*

$$T_n(x) \;=\; \cos(n \arccos x).$$

Proposition 2.4 *Les fonctions (T_n) satisfont*

a) $T_0(x) = 1$, $T_1(x) = x$. *et pour* $n \geq 1$, $\quad T_{n+1}(x) \;=\; 2xT_n(x) - T_{n-1}(x)$.

En particulier, T_n est une fonction polynomiale de degré n et de coefficient dominant 2^{n-1} (pour $n \geq 1$).

b) $|T_n(x)| \leq 1$ *pour* $x \in [-1, 1]$ *et* $n \geq 1$.

c) Pour $n \geq 0$ *et* $k = 0, \cdots, n$, *on a* $T_n\left(\cos\left(\dfrac{k\pi}{n}\right)\right) \;=\; (-1)^k$.

d) Pour $n \geq 0$ *et* $k = 0, \cdots, n-1$, *on a* $T_n\left(\cos\left(\dfrac{(2k+1)\pi}{2n}\right)\right) \;=\; 0$.

e) La suite (T_n) forme une suite de polynômes orthogonaux pour le poids $\omega(x) \;=\; \dfrac{1}{\sqrt{1-x^2}}$, $x \in (-1, 1)$, *i.e.*

$$\textit{pour} \quad m, n \geq 0, \qquad \int_{-1}^{1} T_n(x) T_m(x) \frac{1}{\sqrt{1-x^2}}\, dx \;=\; \begin{cases} \pi & \textit{si } n = m = 0, \\ \pi/2 & \textit{si } n = m \neq 0, \\ 0 & \textit{si } n \neq m. \end{cases}$$

Preuve

a) On a $T_0(x) = 1$ et $T_1(x) = x$ pour $x \in [-1, 1]$ par définition. Soit $n \geq 1$ et $y \in \mathbf{R}$, on a la relation trigonométrique

$$\cos(n+1)y + \cos(n-1)y \;=\; 2 \cos y \cos ny.$$

En posant $y = \arccos x$, on obtient la relation souhaitée.

b) Par définition de T_n, on a $|T_n(x)| = \cos(n \arccos x)| \leq 1$ pour tout $|x| \geq 1$.

c) et d) proviennent respectivement des identités

$$\cos(k\pi) \;=\; (-1)^k, \qquad \cos(k\pi + \pi/2) \;=\; 0, \qquad \text{pour } \; k \in \mathbf{Z}.$$

e) Pour évaluer l'intégrale, on procède au changement de variable $x = \cos y$, $0 < y < \pi$, on obtient

$$\int_{-1}^{1} T_n(x) T_m(x) \frac{1}{\sqrt{1-x^2}}\, dx \;=\; \int_0^{\pi} \cos ny \cos my\, dy \;=\; \frac{1}{2}\int_0^{\pi} \cos(n+m)y\, dy + \frac{1}{2}\int_0^{\pi} \cos(n-m)y\, dy.$$

On conclut en calculant explicitement ces intégrales. □

Revenons au problème de minimisation de $|(x - x_0) \cdots (x - x_n)|$.

Lemme 2.5 *Soit g un polynôme de degré $n \geq 1$ dont le coefficient dominant est 2^{n-1} (le même que T_n). Alors*

$$g \not\equiv T_n \qquad \Longrightarrow \qquad \max_{[-1,1]} |g(x)| > 1 \quad (= \max_{[-1,1]} |T_n(x)|).$$

Preuve
Soit g un polynôme de degré n dont le coefficient devant x^n est 2^{n-1}. Si $\max |g| \leq 1$, alors $p = T_n - g$ est un polynôme de degré inférieur à $n-1$. Par la propriété c) de la Proposition précédente, pour $\varepsilon > 0$, le polynome $p_\varepsilon = T_n - (1-\varepsilon)g$ prend une valeur non nulle du même signe que $(-1)^k$ au point $x_k := \cos k\pi/n$ pour $k = 0, \cdots n$. Donc p_ε admet n racines distinctes $y_{\varepsilon,1}, \cdots, y_{\varepsilon,n}$ avec $y_{\varepsilon,i} \in (x_{i-1}, x_i)$. Comme p_ε est de degré n de coefficient dominant $\varepsilon 2^{n-1}$, on a

$$p_\varepsilon(x) = \varepsilon 2^{n-1} \prod_{i=1}^{n} (x - y_{\varepsilon,i}).$$

Donc pour $x \in [-1, 1]$,

$$|p_\varepsilon(x)| \;\leq\; 2^{2n-1}\varepsilon \;\stackrel{\varepsilon\to 0}{\longrightarrow}\; 0.$$

Et comme $p_\varepsilon(x) \stackrel{\varepsilon\to 0}{\longrightarrow} p(x)$, on en déduit $p(x) = 0$ pour $x \in [-1, 1]$. Et finalement $g = T_n$. □

On déduit du Lemme 2.5 la réponse à notre problème de minimisation.

Corollaire 2.6 *L'expression*

$$\sup_{[a,b]} |(x - x_0) \cdots (x - x_n)|$$

est minimale pour

$$x_k \;=\; \frac{a+b}{2} + \frac{b-a}{2} \cos\left(\frac{(2k+1)\pi}{2n+2}\right), \qquad k = 0, \cdots, n.$$

Citons cette autre version du Lemme 2.5 qui pourra être utile lors de l'étude de la méthode du gradient conjugué.

Lemme 2.7 *Soit $y \in \mathbf{R} \setminus [-1, 1]$. Si g de degré inférieur à n est tel que $g(y) = T_n(y)$ alors*

$$g = T_n \qquad \textit{ou} \qquad \sup_{[-1,1]} |g| > 1.$$

Preuve
Supposons que $\sup_{[-1,1]} |g| \leq 1$ et montrons que dans ce cas $g = T_n$. On considère à nouveau les polynômes $p = T_n - g$ et $p_\varepsilon = T_n - (1-\varepsilon)g$. Soit a le coefficient de degré n de g, comme plus haut, pour $0 < \varepsilon < 1$, il existe $y_{\varepsilon,1}, \cdots, y_{\varepsilon,n}$ satisfaisant $y_{\varepsilon,i} \in (x_{i-1}, x_i)$ tels que

$$p_\varepsilon(x) \;=\; (2^{n-1} - (1-\varepsilon)a) \prod_{i=1}^{n} (x - y_{\varepsilon,i}). \tag{2.1}$$

Les $y_{\varepsilon,i}$ étant borné, on peut extraire une suite (ε_n) convergent vers 0 telle que pour $i = 1, \cdots, n$, la suite $(y_{\varepsilon_n,i})$ converge vers $y_i \in [-1, 1]$. Passant à la limite dans (2.1), on obtient

$$p(x) \;=\; (2^{n-1} - a) \prod_{i=1}^{n} (x - y_i).$$

Comme on a aussi $p(y) = 0$ avec $y \in \mathbf{R} \setminus [-1, 1]$, on en déduit $p = 0$. □

3 Phénomène de Runge

Proposition 3.1 *Soit $(x_0^{(n)}, \cdots, x_n^{(n)})$ une suite de points d'interpolation de $[a, b]$. Alors il existe f continue telle que le polynôme d'interpolation p_n aux points*

$$(x_0^{(n)}, f(x_0^{(n)})), \cdots, (x_n^{(n)}, f(x_n^{(n)})).$$

soit tel que

$$\max_{[-1,1]} |f - p_n| \quad \overset{n\to\infty}{\longrightarrow} \quad +\infty.$$

Par contre le choix des zéros des polynômes de Chebychev comme noeuds d'interpolation a de bonnes propriétés.

Théorème 3.2 *Soit f de classe C^1 sur $[a, b]$ et p_n son polynôme d'interpolation aux points de Chebychev. Alors*

$$\max_{[-1,1]} |f - p_n| \quad \overset{n\to\infty}{\longrightarrow} \quad 0.$$

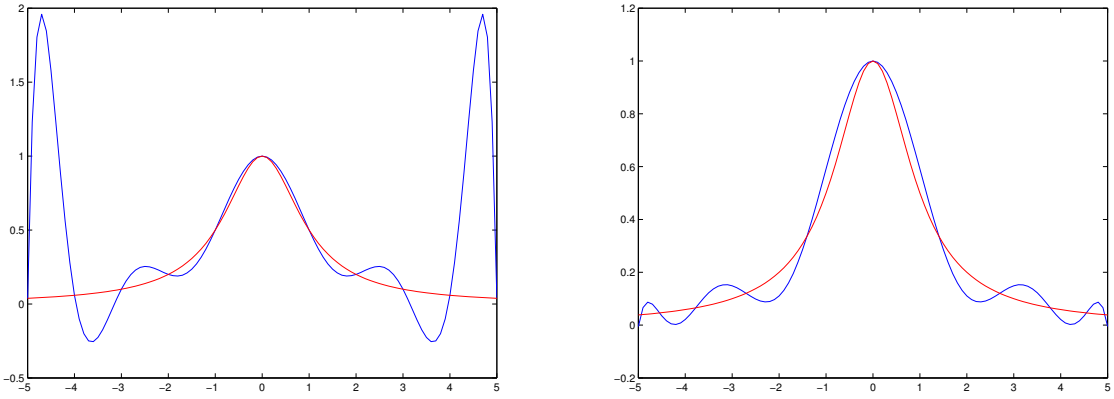


FIG. V.3 – Interpolation de Lagrange de degré 10 de $x \mapsto 1/(1 + x^2)$ avec des nœuds equirépartis (à gauche) et avec les nœuds de Chebyshev (à droite).

Le résultat de la Proposition 3.1 concerne l'approximation des fonctions à l'aide de polynômes *d'interpolation*. Il reste vrai qu'on peut approcher une fonction continue sur un compact par une fonction polynomiale avec une erreur aussi petite que l'on souhaite (voir Section 6).

Théorème 3.3 (Weierstraß) *L'algèbre des fonctions polynomiales est dense dans l'espace* $C([a,b],\mathbf{R})$ *munit de la norme* $\|\cdot\|_\infty$.

4 Interpolation polynomiale par morceaux

A cause du phénomène de Runge, plutôt que d'appocher f sur $[a,b]$ par une fonction polynomiale de degré élevé, on découpe l'intervalle $[a,b]$ en sous intervalles $I_1,\cdots,I_m$ avec $I_k=[a_{k-1},a_k]$ sur lesquels on interpole f à l'aide de fonctions polynomiales de degré relativement faibles.

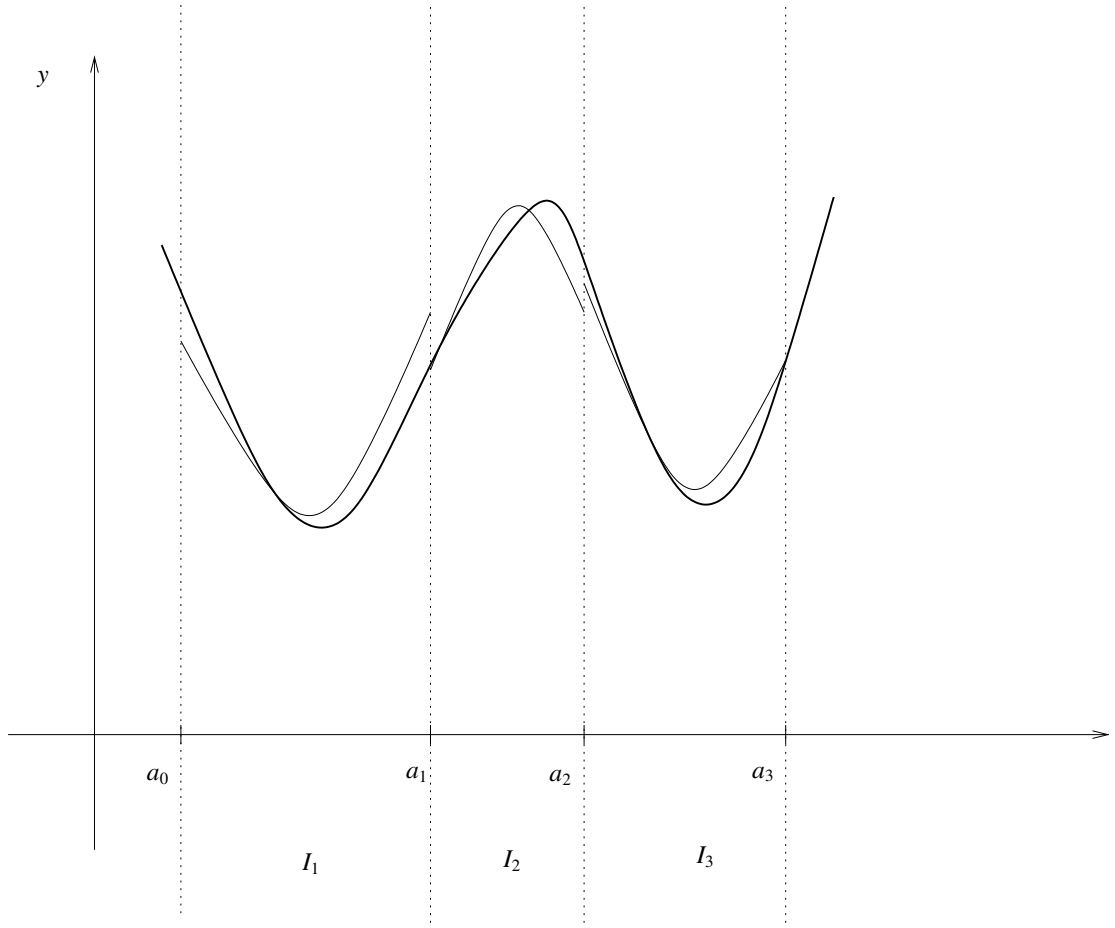


FIG. V.4 – Approximation polynomiale par morceaux.

Exemples :

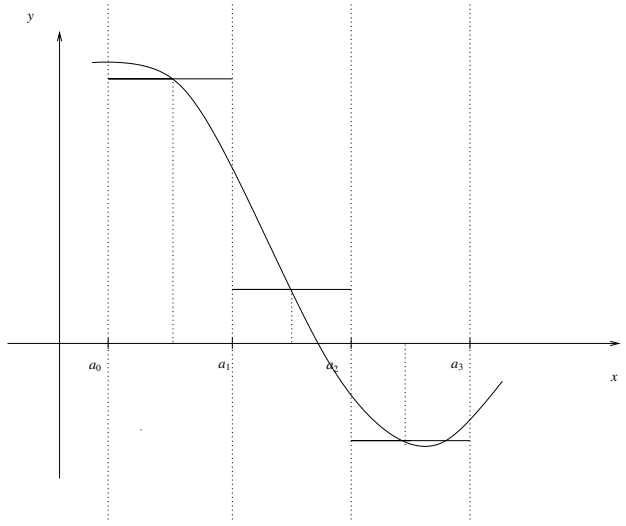


FIG. V.5 – Approximation de degré 0, interpolation au point milieu. L'approximation est discontinue.

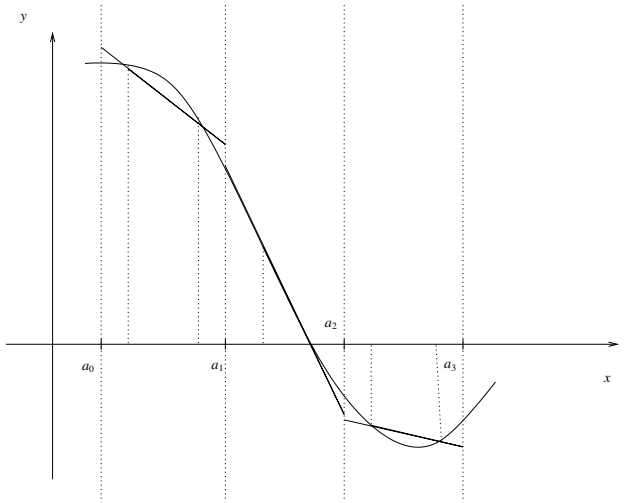


FIG. V.6 – Approximation de degré 1, aux points de Chebychev. L'approximation est discontinue.

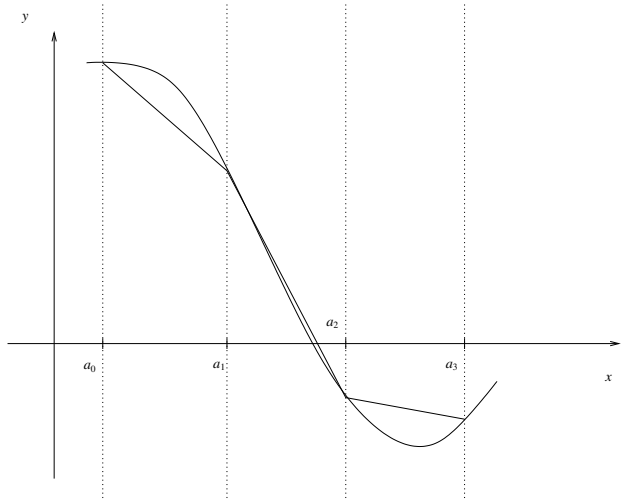


FIG. V.7 – Approximation de degré 1, interpolation aux points (a_{k-1}, a_k). L'approximation est continue.

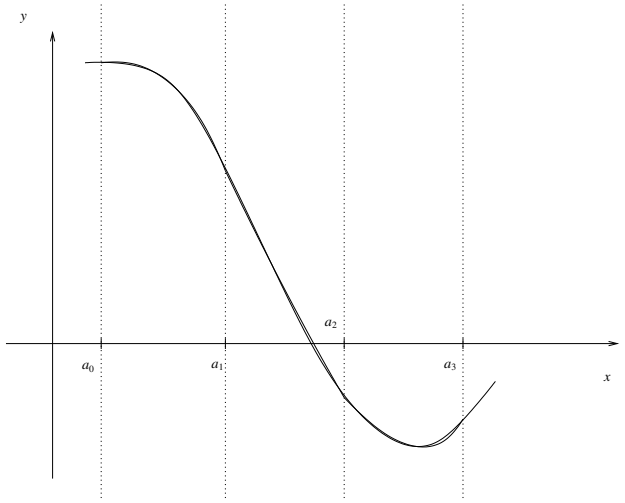


FIG. V.8 – Approximation de degré 2, interpolation aux points $(a_{k-1}, \frac{a_{k-1}+a_k}{2}, a_k)$. L'approximation est continue.

5 Stabilité numérique de l'interpolation

Lors du calcul du polynôme interpolant, l'ordinateur fait des erreurs d'arrondis (de troncature). Dans cette section, on s'interesse à l'influence de ces erreurs sur le polynôme d'interpolation. En fait on fait des erreurs sur

a) les x_i,

b) les y_i,

c) les calculs de différences divisées.

Ici, on ne s'interesse qu'au point b). Si les y_i sont des mesures effectuées par un astronome aux dates x_i, alors on s'interesse à l'erreur de mesure entre les valeurs y_i notées par l'astronomes et la valeur réelle $\tilde{y}_i$.

Comme plus haut on note p le polynôme d'interpolation aux points $(x_0, y_0), \cdots, (x_n, y_n)$, on a

$$p(x) \;=\; \sum_{i=0}^{n} y_i l_i(x) \qquad \text{où pour } \; i = 0, \cdots, n, \quad l_i(x) = \prod_{0 \le j \le n,\ j \neq i} \frac{x - x_j}{x_i - x_j}.$$

On introduit les valeurs erronées

$$\tilde{y}_i \;=\; y_i + \varepsilon_i,$$

et le polynôme d'interpolation correspondant

$$\tilde{p}(x) \;=\; \sum_{i=0}^{n} \tilde{y}_i l_i(x).$$

L'erreur commise au point x est clairement

$$err(x) \;=\; \sum_{i=0}^{n} \varepsilon_i l_i(x),$$

pour laquelle la meilleure majoration disponible est

$$\|err\|_{L^\infty([a,b])} \;\le\; \max|\varepsilon_i| \underbrace{\sup_{x \in [a,b]} \sum_{i=0}^{n} |l_i(x)|}_{=:\,\Lambda_n}.$$

Là encore on peut observer que l'interpolation aux points de Chebyshev a de meilleure propriétés que l'interpolation sur la subdivision uniforme. En effet pour une subdivision de uniforme, on a

$$\Lambda_n \;\sim\; \frac{2^{n+1}}{en \ln n}.$$

Et pour les noeuds de Chebychev, on a la croissance bien plus lente suivante

$$\Lambda_n \;\sim\; \frac{2}{\pi} \ln n.$$

6 Polynômes de Bernstein

L'objet de cette section est de donner une construction explicite pour le Théorème de Weierstraß.

Théorème 6.1 (Weierstraß) *Soit* $f : [a,b] \longrightarrow \mathbf{R}$ *une fonction continue. Pour tout* $\varepsilon > 0$*, il existe une fonction polynomiale* p *telle que*

$$\max_{x\in[a,b]} |f(x) - p(x)| < \varepsilon.$$

L'erreur $\|f - p\|_\infty$ sera évaluée à l'aide du module de continuité de f dont nous rappelons la définition.

Définition 6.2 (Module de continuité) *Si* f *est continue sur* $[a,b]$*, pour* $0 \leq h < b - a$*, on pose*

$$\omega(h) = \max\{|f(y) - f(x)| ; a \leq x \leq y \leq x + h \leq b\}.$$

La fonction ω *est continue, croissante et telle que* $\omega(0) = 0$*. De plus elle sous-linéaire (i.e.* $\omega(h_1 + h_2) \leq \omega(h_1) + \omega(h_2)$*). Elle satisfait clairement*

$$\forall x, y \in [a,b] \qquad |f(x) - f(y)| \leq \omega(|x - y|)$$

La fonction ω *est appelée module de continuité de* f*.*

Définition 6.3 *On travaille sur l'intervalle* $[a,b] = [0,1]$*. Pour* $0 \leq j \leq n$*, posons*

$$\beta_n^j(x) = \binom{n}{j} x^j (1-x)^{n-j},$$

et pour $f \in C([0,1], \mathbf{R})$ *et* $n \geq 0$*,*

$$B_n(f,x) = \sum_{j=0}^{n} f\left(\frac{j}{n}\right) \beta_n^j(x)$$

Le Théorème 6.1 est une conséquence de l'estimation explicite qui suit.

Théorème 6.4 *Soit* f *continue sur* $[0,1]$ *et* ω *son module de continuité. On a*

$$\sup_{x\in[0,1]} |f(x) - B_n(f,x)| \leq 2\omega\left(\frac{1}{2\sqrt{n}}\right).$$

Preuve
On rappelle la formule du binôme de Newton

$$(a+b)^n = \sum_{j=0}^{n} \binom{n}{j} a^j b^{n-j}.$$

C'est un exercice classique d'obtenir (en étudiant la fonction $f(y) = (1 - x + y)^n$ et ses dérivées au voisinage de $y = x$) les formules suivantes

$$(6.1)\quad \text{a)}\ \sum_{j=0}^{n} \beta_n^j(x) = 1, \qquad \text{b)}\ \sum_{j=0}^{n} \frac{j}{n}\beta_n^j(x) = x, \qquad \text{c)}\ \sum_{j=0}^{n} \frac{j^2}{n^2}\beta_n^j(x) = \left(1 - \frac{1}{n}\right)x^2 + \frac{1}{n}x.$$

Ces préliminaires étant faits, on définit l'erreur au point x pour la $n^{\text{ième}}$ approximation de Bernstein par

$$e_n(f,x) = f(x) - B_n(f,x) \overset{(6.1-a)}{=} \sum_{j=0}^{n} \beta_n^j(x)\left(f(x) - f\left(\frac{j}{n}\right)\right).$$

D'où

$$|e_n(f,x)| \leq \sum_{j=0}^{n} \beta_n^j(x)\left|f(x) - f\left(\frac{j}{n}\right)\right|.$$

On se donne $\delta > 0$ qui sera fixé à la fin de la preuve et on considère alors deux cas pour $0 \leq x \leq 1$ et $0 \leq j \leq n$:

- $\left|x - \frac{j}{n}\right| \leq \delta$; dans ce cas on utilisera l'inégalité $\left|f(x) - f\left(\frac{j}{n}\right)\right| \leq \omega(\delta)$.
- $\left|x - \frac{j}{n}\right| > \delta$; dans ce cas on note p l'unique entier $p \geq 1$ tel que $p\delta \leq \left|x - \frac{j}{n}\right| < (p+1)\delta$.

On utilisera la majoration

$$\left|f(x) - f\left(\frac{j}{n}\right)\right| \leq \omega((p+1)\delta) \leq (p+1)\omega(\delta) \leq \left(1 + \frac{1}{\delta}\left|x - \frac{j}{n}\right|\right)\omega(\delta) \leq \left(1 + \frac{1}{\delta^2}\left(x - \frac{j}{n}\right)^2\right)\omega(\delta).$$

On a utilisé pour la dernière inégalité le fait que $\frac{1}{\delta}\left|x - \frac{j}{n}\right| \geq 1$ entraîne $\frac{1}{\delta}\left|x - \frac{j}{n}\right| \leq \frac{1}{\delta^2}\left(x - \frac{j}{n}\right)^2$.

Finalement dans les deux cas, on a

$$\left|f(x) - f\left(\frac{j}{n}\right)\right| \leq \left(1 + \frac{1}{\delta^2}\left(x - \frac{j}{n}\right)^2\right)\omega(\delta).$$

On en déduit

$$|e_n(f,x)| \leq \left\{\sum_{j=0}^{n} \beta_n^j(x)\left(1 + \frac{1}{\delta^2}\left(x - \frac{j}{n}\right)^2\right)\right\}\omega(\delta) \overset{(6.1)}{=} \left(1 + \frac{x(1-x)}{n\delta^2}\right)\omega(\delta) \leq \left(1 + \frac{1}{4n\delta^2}\right)\omega(\delta).$$

Choisissant $\delta = 1/(2\sqrt{n})$, on obtient le résultat annoncé. □

7 Interpolation de Hermite

On décrit ici l'interpolation de Hermite. L'interpolation par splines cubiques et d'autres applications de l'interpolation de Lagrange seront vues en exercice.

Pour améliorer l'approximation de la fonction f par le polynôme interpolateur, on va utiliser les valeurs de f et de ses dérivées aux nœuds d'interpolation. Plus précisément on a le résultat général qui suit.

Théorème 7.1 *Soient $x_0, \cdots, x_k \in [a,b]$ des points deux à deux distincts. Soients $\alpha_0, \cdots, \alpha_k$ des entiers positifs. On pose $\beta = \max \alpha_i$ et $n = -1 + \sum_{j=0}^{k}(\alpha_i + 1)$. Alors soit $f \in C^\beta([a,b], \mathbf{R})$ il existe un unique polynôme $p_n \in \mathcal{P}_n(\mathbf{R})$ tel que*

$$\textit{pour } \; j = 0, \cdots, k, \quad \textit{pour } \; i = 0, \cdots, \alpha_j, \qquad p_n^{(i)}(x_j) \; = \; f^{(i)}(x_j).$$

Preuve
On a à résoudre un système linéaire $(n+1) \times (n+1)$. Il suffit donc de montrer l'injectivité. Soit $p \in \mathcal{P}_n(\mathbf{R})$ tel que

$$\text{pour } \; j = 0, \cdots, k, \quad \text{pour } \; i = 0, \cdots, \alpha_j, \qquad p_n^{(i)}(x_j) \; = \; 0.$$

Pour $j = 0, \cdots, k$, le point x_j est donc une racine de multiplicité $\alpha_j + 1$ de p_n et on a

$$p_n(x) \; = \; q(x) \prod_{j=0}^{k} (x - x_j)^{\alpha_j + 1}.$$

Comme le degré de p_n est au plus $n \; < \; \sum_{j=0}^{k}(\alpha_j + 1)$, on en déduit $q = 0$. □

8 Approximation polynomiale

On change de problème. On se donne toujours des points $x_0, \cdots, x_n$ et des valeurs $y_0, \cdots, y_n$ et on cherche un polynôme $p_N \in \mathcal{P}_N(\mathbf{R})$ tel que

$$\Big\| (y_i - p_N(x_i))_{i=0,\cdots,n} \Big\|$$

soit minimal. Évidemment, le résultat dépend de la norme $\| \cdot \|$ choisie sur l'espace $\mathbf{R}^{n+1}$. Ici on prendra l'une des deux normes suivantes :

$$\|z\| \; = \; \begin{cases} \|z\|_\infty \; = \; \max\limits_{0 \le i \le n} |z_i| \\ \|z\|_2 \; = \; \left(\sum\limits_{0 \le i \le n} |z_i|^2 \right)^{1/2} \end{cases}$$

On n'impose pas que le polynôme passe exactement par les points (x_i, y_i) mais que le graphe de p_N passe globalement proche de ces points. Cela permet de ne tenir compte que faiblement de valeurs absurdes qui seraient dues à une erreur de mesure grossière et de lisser les erreurs de mesures. Contrairement au cas de l'interpolation, le degré maximal du polynôme N peut être beaucoup plus petit que le nombre de nœuds $n+1$.

Le choix de la norme euclidienne donne la méthode d'*approximation par moindres carrés*. C'est la méthode utilisée par exemple quand on effectue une régression linéaire : quand on cherche la droite qui passe au plus près d'un nuage de points. L'approximation par moindres carrés a été inventée par Gauss quand celui-ci cherchait à prédire la position de l'astroïde Cérès. Elle sera étudiée dans la partie Algèbre linéaire du cours.

Nous nous concentrons donc ici sur l'approximation en norme uniforme.

8.1 Meilleure approximation en norme L^∞

On a un résultat d'existence et d'unicité et une caractérisation du polynôme dans le cas où le degré maximal du polynôme approchant recherché est égal au nombre de nœuds moins 2.

Proposition 8.1 *Soit $f \in C([a,b], \mathbf{R})$. Soit $n \geq 1$, Soient $x_0 < \cdots < x_n$ un ensemble de $n+1$ points de $[a,b]$ donnés. Il existe un unique polynôme $p \in \mathcal{P}_{n-1}(\mathbf{R})$ tel que*

$$(f-p)(x_i) \;=\; (-1)^i (f-p)(x_0), \qquad \textit{pour } i = 0, \cdots, n. \tag{8.1}$$

De plus si $q \in \mathcal{P}_{n-1}(\mathbf{R}) \setminus \{p\}$, alors

$$\max_{0\leq i\leq n} |(f-p)(x_i)| \;<\; \max_{0\leq i\leq n} |(f-q)(x_i)|.$$

La propriété (8.1) *est appelée propriété d'équi-oscillation.*

Remarque 8.2 *Dans le cas $N < n$, on a existence mais pas toujours unicité. Dans le cas des $N = n+1$, on retrouve le polynôme interpolateur de Lagrange p_n. Dans le cas $N > n+1$, on n'a jamais unicité, en fait tous les polynômes de la forme $p_n(x) + \lambda \prod_{i=0}^{n}(x - x_i)$ sont solutions avec une erreur nulle aux nœuds $x_0, \cdots, x_n$.*

Preuve
Si on part de la propriété d'equi-oscillation, on a un système linéaire de n équations à n inconnues (les coefficients de p) à résoudre.

$$p(x_i) + (-1)^{i+1} p(x_0) \;=\; f(x_i) + (-1)^{i+1} f(x_0), \qquad i = 1, \cdots, n.$$

L'existence est donc équivalente à l'unicité.

Unicité : Supposons qu'on ait $p(x_i) = (-1)^i p(x_0)$ pour $i = 1, \cdots, n$. Si $p(x_0) \neq 0$, alors p change de signe n fois et on en déduit $p = 0$.

Le polynôme trouvé est bien un minimiseur.

En effet soit $q \in \mathcal{P}_{n-1}(\mathbf{R})$ tel que $\max_{0\leq i\leq n} |(f-p)(x_i)| = \max_{0\leq i\leq n} |(f-q)(x_i)|$, alors $(p-q)(x_i)$ a le même signe que $(p-f)(x_i)$ et par théorème des valeurs intermédiaires, on en déduit que $p-q$ admet au moins n racines en tenant compte de leur multiplicité. Finalement comme $d^o(p-q) < n$, on conclut que $p = q$. □

On peut également rechercher la meilleure approximation de f dans $\mathcal{P}_n(\mathbf{R})$ au sens de la norme uniforme sur $[a,b]$. On a le résultat d'existence et d'unicité suivant.

Théorème 8.3 *Soit $f \in C([a,b],\mathbf{R})$, il existe une unique fonction polynomiale $p \in \mathcal{P}_n(\mathbf{R})$ tel que*

$$\|f-p\|_{L^\infty([a,b])}$$

soit minimale.

Ce polynôme est caractérisé par l'equi-oscillation entre $n+2$ points

Preuve

Existence d'un minimiseur. Fixons $x_0,\cdots,x_n \in (a,b)$ et

$$l_i(x) \;=\; \prod_{0\le j\le n,\, j\neq i} \frac{x-x_j}{x_i-x_j}.$$

On a vu que l'application

$$\begin{aligned} \pi \;:\; \mathbf{R}^{n+1} &\longrightarrow \mathcal{P}_n(\mathbf{R}),\\ (y_0,\cdots,y_n) &\longmapsto \sum_{i=0}^{n} y_i l_i \end{aligned}$$

est un isomorphisme. On considère la fonction

$$\begin{aligned} F \;:\; \mathbf{R}^{n+1} &\longrightarrow \mathbf{R},\\ y &\longmapsto \|\pi(y)-f\|_{L^\infty([a,b])}. \end{aligned}$$

On laisse au lecteur le soin de vérifier que cette fonction est continue et qu'elle vérifie l'inégalité $F(y) \ge \max|y_i| - \|f\|_\infty$. On en déduit qu'elle admet un minimiseur y. Le polynôme $p=\pi(y)$ minimise bien la distance

$$\|f-q\|_{L^\infty([a,b])}$$

parmi les polynômes $q \in \mathcal{P}_n(\mathbf{R})$.

Equi-oscillation d'un minimiseur. On considère l'ensemble des points de $[a,b]$ où le maximum de $|f(x)-p(x)|$ est atteint. Supposons qu'on ne puisse trouver dans S qu'au maximum $k \le n+1$ points $\xi_1 < \cdots < \xi_k$ tels que $(f-p)(\xi_i)$ et $(f-p)(\xi_{i+1})$ aient des signes opposés. On choisit alors des points $\zeta_i \in (\xi_i,\xi_{i+1})$ tels que $(f-p)(\zeta_i)=0$ pour $i=1,\cdots k-1$. On peut alors vérifier que pour ε assez petit ayant le bon signe, on a

$$\left\| f-p-\varepsilon\prod_{i=1}^{k-1}(x-\zeta_i)\right\|_{L^\infty([a,b])} \;<\; \|f-p\|_{L^\infty([a,b])}.$$

Ce qui est faux. On a donc l'existence de $n+2$ points $a \le \xi_1 < \cdots < \xi_{n+2} \le b$ pour lesquels

$$(f-p)(\xi_i) \;=\; s\,(-1)^i\,\|f-p\|_{L^\infty([a,b])}, \qquad i=1,\cdots,n+2.$$

où $s = \pm 1$.

Unicité. On utilise la propriété d'équi-oscillation. Supposons que p soit un minimiseur equi-oscillant en $\xi_1 < \cdots < \xi_{n+2}$. Soit $q \in \mathcal{P}_n(\mathbf{R})$. Supposons que $\|q - f\|_{L^\infty([a,b])} = \|p - f\|_{L^\infty([a,b])}$. Dans ce cas pour $i = 1, \cdots, n+2$,

$$(p-q)(\xi_i) \;=\; (p-f)(\xi_i) + (f-q)(\xi_i) \qquad \text{s'annule ou a le même signe que } \; (p-f)(\xi_i).$$

On en déduit (par théorème des valeurs intermédiaires) l'existence de $n+1$ racines de $p - q$ en tenant compte de la multiplicité. Et comme $p - q \in \mathcal{P}_n(\mathbf{R})$, on a finalement $p = q$. □

VI Résolution numérique des Equations Différentielles Ordinaires

1 Généralités

On se donne une application $f : \mathbf{R}^d \times \mathbf{R} \longrightarrow \mathbf{R}^d$ continue et on s'interesse au système différentiel

$$y'(t) = f(y,t), \qquad t \geq 0 \tag{1.1}$$

On se fixe aussi une donnée initiale $y_0 \in \mathbf{R}^d$ et on impose

$$y(0) = y_0 \tag{1.2}$$

L'ensemble

$$\begin{cases} y'(t) = f(y,t), & t \geq 0, \\ y(0) = y_0 \end{cases} \tag{1.3}$$

est appelé problème de Cauchy.

Exemple : On considère un ressort de raideur k, de masse m se déplaçant selon un axe horizontal. Sa position est repérée par l'abscisse x le point d'équilibre est en $x = x_0$. La relation fondamentale de la dynamique donne

$$mx''(t) = -k(x - x_0). \tag{1.4}$$

Cette équation est une équation différentielle linéaire du second ordre pourtant elle peut s'écrire sous la forme d'un système différentielle de degré un. Pour cela on remarque que si on pose $y = (x, x')$ alors (1.4) est équivalent à

$$\begin{cases} y_1'(t) = y_2(t) \\ y_2'(t) = -\frac{k}{m}(y_1(t) - x_0). \end{cases} \tag{1.5}$$

Si on ajoute des condition initiale $y(t^0) = (y_1^0, y_2^0)$, le problème (1.4) rentre bien dans le cadre de (1.3).

Nous commençons par montrer un résultat classique d'existence et d'unicité des solutions du problème de Cauchy. Nous montrons aussi que la solution dépend continûment de la donnée initiale. La preuve de ce dernier résultat est instructive. Les idées qui y sont développées seront utiles dans la section suivante : quand nous nous voudrons établir la convergence de nos approximations numériques vers la solution de (1.3).

Rappel L'ensemble $C([a,b],\mathbf{R}^d)$ est un espace vectoriel normé si on le muni de la norme $\|\cdot\|_{L^\infty([a,b])}$ où

$$\|y\|_{L^\infty([a,b])} := \sup_{t\in(a,b)} \|y(t)\|_{\mathbf{R}^d}.$$

Quand il n'y aura pas d'ambiguïté sur les bornes, nous noterons cette norme $\|\cdot\|_\infty$.

De plus cet espace est complet (c'est un espace de Banach) : si $(y_n) \subset C([a,b],\mathbf{R}^d)$ est telle que

$$\forall \varepsilon > 0\ \exists N \text{ tel que } \forall n,p > N, \text{ on a } \|y_n - y_p\|_\infty < \varepsilon$$

alors il existe $y_\star \in C([a,b],\mathbf{R}^d)$ telle que

$$\|y_n - y_\star\|_\infty \overset{n\uparrow\infty}{\longrightarrow} 0.$$

Un résultat fondamental

Lemme 1.1 (de Grönwall) *Soit* $u : [0,T] \longrightarrow \mathbf{R}$ *une fonction continue positive satisfaisant*

$$u(t) \le A + B\int_0^t u(s)\,ds, \qquad \forall t \in [0,T]. \tag{1.6}$$

Alors on a

$$u(t) \le A\exp(Bt), \qquad \forall t \in [0,T].$$

Preuve
On pose

$$\Phi(t) := \exp(-Bt)\int_0^t u(s)\,ds.$$

Cette fonction est dérivable, on calcule

$$\Phi'(t) = \left(u(t) - B\int_0^t u(s)\,ds\right) exp(-Bt).$$

Et par hypothèse $\Phi'(t) \le A\exp(-Bt)$ sur $[0,T]$. Comme $\Phi(0) = 0$, on en déduit

$$\Phi(t) \le \frac{A}{B}(1-\exp(-Bt)), \qquad \forall t \in [0,T],$$

d'où par (1.6) et par définition de Φ : on obtient $u(t) \le A\exp(Bt), \qquad \forall t \in [0,T].$ □

Théorème 1.2 (de Cauchy Lipschitz) *Soit* $f : \mathbf{R}^d \times \mathbf{R}_+ \longrightarrow \mathbf{R}^d$ *une fonction continue localement Lipschitienne par rapport à la première variable : c'est-à-dire : quel que soit R et* $T > 0$*, il existe K tel que*

$$\forall 0 \leq t \leq T,\ \forall y_1, y_2 \in \mathbf{R}^d |y_1|, ||y_2| \leq R, \qquad |f(t, y_1) - f(t, y_2)| \leq K|y_1 - y_2|. \tag{1.7}$$

I] Alors pour tout $R > 0$ *il existe un temps* $T = T(R) > 0$ *tel que pour tout* $y_0 \in \mathbf{R}^d$*, le problème* (1.3) *admet une unique solution* $y := S_T y_0$ *définie sur un intervalle* $[0, T]$.

II] De plus l'application

$$\begin{array}{rccc} S_T\ :\ B(0,R) & \longrightarrow & C([0,T], \mathbf{R}^d), \\ y_0 & \longmapsto & S_T y_0 \end{array}$$

est Lipschitzienne.

Le point II] est important pour l'analyse numérique. Il sera utile de savoir que l'erreur faite à un instant (icit l'instant initial) n'est pas trop amplifiée.

Pour démontrer le Théorème, on va plutôt voir (1.3) sous sa forme intégrée :

$$y(t) \ = \ y_0 + \int_0^t f(y(s), s)\, ds \tag{1.8}$$

qui est équivalente à (1.3).

Preuve
Unicité. Si on a deux solutions y_1, y_2 ayant la même donnée initiale y_0 alors par continuité de ces solutions on sait qu'elles sont bornées sur $[0, T]$. Soit R la borne et $K = K(R)$ la constante de Lipshitz correspondante. On écrit

$$y_1(t) \ = \ y_0 + \int_0^t f(y_1(s), s)\, ds, y_2(t) \ = \ y_0 + \int_0^t f(y_2(s), s)\, ds,$$

donc par (1.7), on a

$$|y_1 - y_2|(t) \ \leq \ \int_0^t |f((y_1(s), s) - f(y_2(s), s)|\, ds, \ \leq \ K \int_0^t |y_1 - y_2|(s)\, ds$$

On pose $u(t) := \sup_{s \in [0,t]} |y_1 - y_2|(s)$, on a donc

$$0 \leq u(t) \leq K \int_0^t u(s)\, ds,$$

et par le Lemme de Grönwall, on a $u \equiv 0$ donc $y_1 = y_2$.

Existence. On construit une suite de fonctions (y^n) par récurrence par :

$$y^0(t) \ := t, \qquad \forall t \geq 0.$$

Et pour $n \geq 0$,

$$y^{n+1}(t) := y_0 + \int_0^t f(y^n(s), s)\, ds, \qquad \forall t \geq 0. \tag{1.9}$$

Supposons que $|y_0| < R/4$. Soit $n \geq 0$. Soit $T_0 > 0$ quelconque et $C := \sup_{0 \leq t \leq T_0, y \in B(0,R)} |f(t,y)|$. Supposons que $|y_n| \leq R$ sur $[0, T_0]$. On a alors pour $t \leq T_0$,

$$|y^n|(t) = |y_0 + \int_0^t f(y^n(s), s)\, ds| \ \leq \ R/2 + Ct.$$

Donc pour $0 \leq t \leq T$, où $T := \min((R/(2C), T_0)$, on a $|y^n(t)| \leq R$. On peut donc établir par récurrence que pour $n \geq 0$,

$$\forall t \in [0, T], \qquad |y^n|(t) \ \leq \ R. \tag{1.10}$$

Ceci étant établi, posons $K = K(R, T)$, on a pour $n \geq 1$ t $0 \leq t \leq T$,

$$\begin{aligned} |y^{n+1}(t) - y^n(t)| &= \left| \int_0^t \{f(s, y^n(s)) - f(s, y^{n-1}(s))\}\, ds \right| \ \leq \ \int_0^t \left| f(s, y^n(s)) - f(s, y^{n-1}(s)) \right| ds, \\ &\leq K \int_0^t |y^{n+1}(s) - y^n(s)|\, ds. \end{aligned}$$

Posons maintenant pour $n \geq 1$, $A_n(t) := \sup_{[0,t]} |y^n - y^{n-1}|$. On vient de montrer que pour $0 \leq t \leq T$ et $n \geq 1$, on avait

$$A_n(t) \ \leq \ K \int_0^t A_n(s)\, ds.$$

On montre facilement par récurrence que

$$A_n(t) \ \leq \ A_0(T) \frac{K^n t^n}{n!}.$$

Donc $\sum_n A_n(t) \leq A_0(T) \exp(KT)$ converge. La suite (y^n) vérifie donc le critère de Cauchy dans $C([0, T])$. Elle converge donc en norme uniforme vers une fonction $y \in C([0, T])$. En passant à la limite dans (1.9), on en déduit que y est la solution cherchée.

Continuité. C'est une simple application du Lemme de Grönwall comme pour l'unicité. □

Dans le cas où f est uniformément Lipschitzienne, i.e : il existe $K > 0$ tel que

$$\forall t \geq 0 \forall y_1, y_2 \in \mathbf{R}^d |f(t, y_1) - f(t, y_2)| \leq K|y_1 - y_2|, \tag{1.11}$$

on a un résultat plus fort :

Théorème 1.3 (de Cauchy Lipschitz) *Soit $f : \mathbf{R}^d \times \mathbf{R}_+ \longrightarrow \mathbf{R}^d$ une fonction continue uniformément Lipschitienne par rapport à la première variable I] Alors pour tout $T > 0$ tel que pour tout $y_0 \in \mathbf{R}^d$, le problème* (1.3) *admet une unique solution $y := S_T y_0$ définie sur un intervalle $[0, T]$.*

II] De plus l'application

$$\begin{array}{rcl} S_T \ : \ \mathbf{R}^d & \longrightarrow & C([0,T],\mathbf{R}^d), \\ y_0 & \longmapsto & S_T y_0 \end{array}$$

est Lipschitzienne.

2 Définition des Méthodes à un pas

Dans toute la suite pour simplifier on suppose que f vérifie (1.11). On est donc dans le cadre du Théorème 1.3. On s'interesse à des solutions de (1.3) sur un intervalle $[0,T]$ donné.

On va étudier une famille de méthodes de résolution des EDO : les méthodes à un pas. En voici le principe. On commence par découper l'intervalle de temps $[0,T]$ en n sous intervalles. On note $h := T/n$ et $t_i^h := ih$, pour $i = 0, \cdots, n$. On note y_i^h l'approximation de la solution exacte y de (1.3) au temps t_i^h. Dans la suite pour éviter une surcharge d'écriture et quand il n'y aura pas d'ambiguïté, on notera t_i et y_i pour t_i^h et y_i^h. On reprendra la notation rigoureuse quand h sera variable.

La solution exacte vérifie

$$y(t_{i+1}) \ = \ y(t_i) + \int_{t_i}^{t_{i+1}} f(y(s),s)\,ds,$$

qu'on réécrit

$$y(t_{i+1}) \ = \ y(t_i) + h \underbrace{\int_0^1 f(y(t_i+\theta h), t_i+\theta h)\,d\theta}_{=:\ \Phi_{ex}(h,t_i).} \tag{2.1}$$

On a bien sûr

$$\lim_{h\downarrow 0} \Phi_{ex}(h,t_i) \ = \ f(y(t_i),t_i).$$

Il paraît donc raisonnable d'approcher $\Phi_{ex}(h,t_i)$ par $f(y(t_i),t_i)$ dans (2.1). On en déduit la **méthode d'Euler explicite :**

$$y_{i+1} \ := \ y_i + hf(y_i,t_i). \tag{2.2}$$

Une autre façon d'écrire cette m'éthode est

$$\frac{y_{i+1}-y_i}{h} = f(y_i,t_i) \tag{2.3}$$

qui ressemble plus à (1.1) alors que (2.2) a plutôt la forme (1.8).

2.1 Les méthodes à un pas

D'une manière générale les méthodes à un pas s'écrivent

$$\left\{ \begin{array}{rcl} y_0^h & = & y_0 \\ y_{i+1}^h & := & y_i^h + h\Phi(h, t_i^h, y_i^h). \end{array} \right. \tag{2.4}$$

Ce sont des méthodes à un pas car y_{i+1} ne dépend que de y_i et pas de $y_{i-1}, y_{i-2}, \cdots$.

Exemple : la méthode de Runge. On approche l'intégrale définissant Φ_{ex} par la formule du point milieu :

$$\Phi_{ex} := \int_0^1 f(y(t_i + \theta h), t_i + \theta h)\, d\theta \simeq f\left(\underbrace{y\left(t_i + \frac{h}{2}\right)}, t_i + \frac{h}{2}\right),$$

approché par la méthode d'Euler explicite.

D'où la méthode :

$$y_{i+1} := y_i + hf\left(y_i + \frac{h}{2} f(y_i, t_i), t_i + \frac{h}{2}\right).$$

On a donc une méthode à un pas caractérisée par

$$\Phi_{Runge}(h, t, y) := f\left(y + \frac{h}{2} f(y, t), t + \frac{h}{2}\right).$$

3 Convergence

Définition 3.1 *On dira que l'approximation définie par* (2.4) *est convergente si pour tout* $y_0 \in \mathbf{R}^d$*, si* $y_0^h \overset{h\downarrow 0}{\longrightarrow} y_0$ *alors*

$$\sup_{0 \le j \le T/h} |y_j^h - y(t_j^h)| \overset{h\downarrow 0}{\longrightarrow} 0.$$

On dit que le schéma est convergent d'ordre p si

$$\sup_{0 \le j \le T/h} |y_j^h - y(t_j^h)| = O(h^p).$$

Remarque 3.2 *Cette définition impose que la solution exacte et la solution numérique approchée soient stables par rapport aux perturbations de la donnée initiale.*

Définition 3.3 *On dira que le schéma défini par* (2.4) *est consistante si*

$$\lim_{h\downarrow 0} \Phi(h, \bar{y}, \bar{t}) = f(\bar{y}, \bar{t}), \qquad \forall \bar{y}, \bar{t}.$$

On dit que le schéma est consistant d'ordre p si

$$\Phi(h, \bar{y}, \bar{t}) = \int_0^1 f(z(\bar{t} + \theta h), \bar{t} + \theta h)\, d\theta + O(h^p),$$

où z est la solution de

$$\left\{ \begin{array}{rcl} z' & = & f(z,t) \\ z(\bar{t}) & = & \bar{y}. \end{array} \right.$$

Définition 3.4 *On dira que le schéma défini par* (2.4) *est stable si pour tout T, il existe des constantes $M(T) > 0$ et $\bar{h} > 0$ telles que*

$$\forall 0 < h \leq \bar{h}, \qquad \forall (\varepsilon_j)_{j=0,\cdots} \subset \mathbf{R},$$

si on définit

$$\begin{array}{rcl} y_{j+1} & := & y_j + h\Phi(h, y_j, t_j), \\ \tilde{y}_{j+1} & := & \tilde{y}_j + h\Phi(h, \tilde{y}_j, \tilde{t}_j) + \varepsilon_{j+1}, \\ \tilde{y}_0 & = & y_0 + \varepsilon_0. \end{array}$$

Alors

$$|y_N - \tilde{y}_N| \ \leq \ M(T)\left(\sum_{j=0}^{N} |\varepsilon_j|\right), \qquad \forall t_N \leq T.$$

Théorème 3.5 *Soit f satisfaisant les hypothèses du Théorème 1.3. Si le schéma défini par* (2.4) *est consistant (d'ordre p) et stable alors il est convergent (d'ordre p).*

Preuve
On pose

$$\tilde{y}_j \ = \ y(t_j)$$

où y est la solution, exacte. On a

$$\tilde{y}_{j+1} \ = \ \tilde{y}_j + h\Phi(h, \tilde{y}_j, t_j) + he_j.$$

Par consistance du schéma, on a $e_j = O(h^p)$ et par stabilité, (en posant $\varepsilon_j = he_j$, on déduit

$$|y_N - \tilde{y}_N| \ \leq \ M(T) \sum_{k=0}^{N} |\varepsilon_j| = O(h^p).$$

□

La consistance est facile à vérifier. La stabilité est plus délicate. Nous allons présenter une condition suffisante plus facile à vérifier pour la consistance. Pour cela nous aurons besoin d'une version discrète du Lemme de e Grönwall.

Lemme 3.6 ([de Grönwall discret) *Soient $(a_j)_{j\geq 0}$ et $(b_j)_{j\geq 0}$ deux suites de nombres positifs et soit $\Lambda > 0$. Si on a pour $j \geq 0$,*

$$a_{j+1} \leq (1 + \lambda)a_j + b_j.$$

Alors pour $j \geq 0$,

$$a_j \ \leq \ a_0 \exp(\Lambda j) + \sum_{k=0}^{j-1} b_k \exp(\Lambda(j - k - 1)).$$

Preuve
De l'inégalité $1 + x \leq \exp(x)$, on tire

$$a_{j+1} \ \leq \ \exp(\Lambda) a_j + b_j.$$

Posons

$$\alpha_j := a_j \exp(-\Lambda j),$$

d'où

$$\alpha_{j+1} \leq \alpha_j + b_j \exp(-\Lambda(j+1)) \ \leq \cdots \leq \alpha_0 + b_0 \exp(-\Lambda) + b_1 \exp(-2\Lambda) + \cdots + b_{j+1} \exp(-\Lambda(j+1)).$$

D'où le résultat. □

Théorème 3.7 (condition suffisante de stabilité) *Pour que le schéma* (2.4) *soit stable, il suffit que pour tout $T > 0$, il existe $\lambda \geq 0$ et $\bar{h} > 0$ telle que*

$$|\Phi(h, y, t) - \Phi(h, \tilde{y}, t)| \leq \lambda |y - \tilde{y}|, \qquad \forall 0 < h < \bar{h}, \ \forall y, \tilde{y} \in \mathbf{R}^d, \forall t \in [0, T]. \tag{3.1}$$

Preuve
On reprend les notations de la définition de la stabilité. En utilisant (3.1), on a

$$|\tilde{y}_{j+1} - y_j| \ \leq \ (1 + \lambda h)|\tilde{y}_j - y_j| + |\varepsilon_{j+1}|$$

On applique alors le Lemme de Grönwall discret avec $\Lambda = \lambda h$, $a_j = |\tilde{y}_j - y_j|$ et $b_j = |\varepsilon_{j+1}|$:

$$|\tilde{y}_j - y_j| \ \leq \ \exp(\lambda j h) \sum_{k=0}^{j-1} |\varepsilon_k|.$$

Et comme $\exp(\lambda j h) \leq \exp(\lambda T)$ pour $jh \leq T$ on a bien la stabilité du schéma. □

Exercice Le schéma d'Euler explicite (2.2) est-il consistant ? stable ?

4 Schémas d'Euler, θ-schémas

Rappelons le schéma d'Euler explicite :

$$y_{i+1} \ := \ y_i + h f(y_i, t_i). \tag{4.1}$$

Il existe aussi une version implicite appelée aussi schéma d'Euler rétrograde :

$$y_{i+1} \ := \ y_i + h f(y_{i+1}, t_i) \tag{4.2}$$

Cette fois pour y_{i+1} apparaît dans le membre de droite. Pour le déterminer il faut résoudre un système non-linéaire. On pourra utiliser pour cela une méthode de Newton. En effet on dispose

comme point départ de la méthode de Newton la valeur y_i qui est une assez bonne approximation de y_{i+1} (on s'attend à avoir $y_{i+1} = y_i + O(h)$).

Une généralisation des schémas d"Euler explicite et implicite (4.1)(4.2) sont les θ-schémas. On se donne un paramètre $\theta \in [0, 1]$ le θ-schéma est défini par

$$y_{i+1} = y_i + h\{\theta f(y_{i+1}, t_{i+1}) + (1-\theta) f(y_i, t_i)\} \tag{4.3}$$

Exercice : Dans l'écriture (4.3) du θ-schéma, la vecteur inconnue y_{i+1} apparaît dans le membre de droite. Ce schéma peut-il se mettre sous la forme (2.4) des schémas à un pas ?

Remarque 4.1 *Pour $\theta = 0$, on retrouve le schéma d'Euler explicite et pour $\theta = 1$, le schéma d'Euler implicite. Pour $\theta = 1/2$, le schéma s'appelle schéma de Crank Nicolson.*

On commence par étudier l'ordre de consistance de ces schémas. Nous laissons la preuve du résultat suivant en exercice.

Proposition 4.2 *Le θ-schéma est d'ordre 1 si $\theta \neq 1/2$ et d'ordre 2 pour $\theta = 1/2$.*

Pourquoi utiliser des méthodes implicites qui sont plus coûteuses puisqu'il y a un système linéaire à résoudre ? L'intérêt des schémas implicites réside dans leur stabilité. La suite de cette section est consacrée à un exemple qui met en évidence les problèmes d'amplification d'erreur qui peuvent conduire à préférer une méthode implicite à une méthode explicite.

4.1 coefficient d'amplification

On s'interesse à l'équation en dimension d

$$y' = -Ay, \tag{4.4}$$

où A est une matrice symétrique définie positive de valeurs propres $0 < \lambda_1 \leq \cdots \leq \lambda_d$. On note $e_1, \cdots, e_d$ une base orthonormée de vecteurs propres associés. On utilise la décomposition

$$y(t) = \sum \alpha_i(t) e_i.$$

En prenant le produit scalaire de (4.4) avec e_i, on déduit que le coefficient α_i évolue selon l'équation différentielle

$$\alpha_i'(t) = -\lambda_i \alpha_i$$

Donc

$$\alpha_i(t) = \alpha_i^0 \exp(-\lambda_i t), \tag{4.5}$$

où

$$y_0 = \sum \alpha_i^0 e_i.$$

La quantité

$$|y(t)| = \sqrt{\sum |\alpha_i(t)|^2} = \sqrt{\sum |\alpha_i^0|^2 \exp(-2\lambda_i t)}$$

est donc décroissante du temps. On souhaite que cette propriété soit respectée par le schéma.

Proposition 4.3 *Le θ schéma appliqué au problème* (4.4)*, s'écrit*

$$y_{i+1} \;=\; C(\theta, h) y_i.$$

où $C(\theta, h)$ est une matrice symétrique. On a $|y_{i+1}| \leq |y_i|$ pour tout y_i si

(4.6) *les valeurs propres de $C(\theta, h)$ sont de norme plus petite que 1.*

Pour $\theta \geq 1/2$, c'est toujours vrai.

Pour $\theta > 1/2$, c'est vrai ssi $h \leq 2/(\lambda_d(1 - 2\theta))$.

Le choix d'un θ-schéma à dominante explicite impose donc une restriction sur la taille des pas de temps.

Exercice : Prouver la Proposition.

Execrcice : D'après (4.5) le temps caractéristique d'évolution du coefficient α_i est $\tau_i = 1/\lambda_i$. L'évolution la plus lente ayant lieu pour $i = 1$. Donner, dans le cas $\theta > 1/2$, le nombre minimal de pas de temps nécessaires pour aller jusqu'au temps caractéristique τ_1. Que se passe-t-il quand la matrice A est mal conditionnée ?