Interpolation polynomiale, intégration numérique, résolution numérique d'équations différentielles

Gilles Leborgne

17 avril 2021

Table des matières

(Paragraphes commençant par une étoile (*) : hors examen.)

1 Interpolation polynomiale

L'interpolation polynomiale consiste à faire passer une fonction localement polynomiale par des points.

L'approximation polynomiale consiste à approcher une fonction f par un polynôme p (au moins localement) de telle sorte que f et p soient "proches".

1.1 Introduction à l'interpolation de Lagrange

1.1.1 Rappel

Le développement de Taylor à l'ordre n d'une fonction f C^{n+1} consiste, au voisinage d'un point x_0, à approcher f à l'aide d'un polynôme p_n de degré n, polynôme donné uniquement à l'aide des valeurs ponctuelles $f^{(k)}(x_0)$ pour $k = 0, ..., n$. Il s'écrit :

$$f(x) = p_n(x) + o((x-x_0)^n), \qquad \text{au voisinage de } x_0,$$

où p_n est le polynôme de degré n donné par :

$$p_n(x) = f(x_0) + (x-x_0)f'(x_0) + ... + \frac{(x-x_0)^n}{n!} f^{(n)}(x_0).$$

Nous n'étudierons pas cette approximation polynomiale supposée connue, et qui s'applique aux x proches de x_0.

1.1.2 Présentation de l'interpolation de Lagrange

Un autre type de développement polynomial de f, dit de Lagrange, consiste à, sur un intervalle $[a, b]$, considérer $n+1$ points x_i deux à deux distincts et les valeurs $y_i = f(x_i)$, pour $i = 0, 1, ..., n$, et à construire le polynôme p de degré n t.q. $p(x_i) = y_i$ pour $i = 0, 1, ..., n$.

Définition 1.1 Les points x_i, pour $i = 0, 1, ..., n$, sont appelés les points de collocation, et pour des valeurs y_i, pour $i = 0, 1, ..., n$, le polynôme p vérifiant $p(x_i) = y_i$ pour $i = 0, 1, ..., n$ est appelé le polynôme d'interpolation de Lagrange relatif aux points (x_i, y_i).

Les polynômes p_n cherchés peuvent être décomposés sur la base $(1, x, ..., x^n)$, i.e. mis sous la forme $p_n(x) = a_0 + a_1 x + ... + a_n x^n$, et les inconnues sont alors les $n+1$ composantes a_i qui sont solutions du système de Vandermonde formé des $n+1$ équations $p(x_i) = y_i$, i.e. :

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

Usuellement la matrice de Vandermonde est très mal conditionnée, et on évite de résoudre ce système.

Ici, on préfèrera développer les polynômes de degré n sur des bases plus adaptées, en commençant par les bases :

1- donnée au § 1.3 par Lagrange : $(L_i(x) = \prod_{j \neq i} \frac{(x - x_j)}{(x_i - x_j)})_{0 \leq i \leq n}$, (tous de degré n), ou bien

2- donnée au § 1.5 par Newton : $(1, (x-x_0), (x-x_0)(x-x_1), ..., (x-x_0)...(x-x_{n-1}))$.

1.1.3 * Applications : maillage

Pour une fonction f et les valeurs $y_i = f(x_i)$, le polynôme p d'interpolation de Lagrange relatif aux points (x_i, y_i) donnera une approximation de f sur $[a, b]$; mais cette approximation en génral sera "mauvaise" si $[a, b]$ est "grand".

Pour avoir "une bonne approximation de f" sur un "grand" intervalle $[a, b]$, on commence par "mailler $[a, b]$", i.e. on décompose $[a, b]$ en N_{int} intervalles :

$$[a, b] = \bigcup_{k=1}^{N_{\text{int}}} [a_{k-1}, a_k], \tag{1.1}$$

où $a_0 = a$, $a_{N_{\text{int}}} = b$ et $a_{k-1} < a_k$ pour tout $k = 1, ..., N_{\text{int}}$. Et on cherche un polynôme d'interpolation de Lagrange sur chaque $[a_{k-1}, a_k]$.

1.2 * Rappels généraux

On note $\mathcal{P}_n$ l'ensemble des polynômes réels de degré au plus n.

(Lors de la recherche de racines, on considère en général les polynômes complexes. Ici on propose un cadre réel pour simplifier la présentation.)

1.2.1 Pour les racines d'un polynôme

Définition 1.2 Un polynôme p est une fonction $\mathbb{R} \to \mathbb{R}$ telle que :

$$\exists n \in \mathbb{N}, \quad \exists (a_i)_{i=0,...,n} \in \mathbb{R}^{n+1}, \quad \forall x \in \mathbb{R}, \quad p(x) = a_0 + a_1 x + ... + a_n x^n \qquad (= \sum_{i=0}^{n} a_i x^i). \tag{1.2}$$

p est polynôme de degré n si $a_n \neq 0$, et p est dit normalisé si $a_n = 1$. Un polynôme de type $p(x) = a_n x^n$ est appelé un monôme (ne comporte qu'un seul terme).

Proposition 1.3 *Un polynôme de degré n est en tout point égal à son développement limité d'ordre n : pour tout $x \in \mathbb{R}$, pour tout $h \in \mathbb{R}$:*

$$p(x+h) = p(x) + h\,p'(x) + ... + \frac{h^n}{n!}\,p^{(n)}(x) \qquad (= \sum_{k=0}^{n} \frac{h^k}{k!} p^{(k)}(x)). \tag{1.3}$$

Preuve. Il suffit de le vérifier pour les $p(x) = x^n$ pour tout $n \in \mathbb{N}$, la dérivation étant linéaire (si f_1 et f_2 sont deux fonctions qui ont un développement limité à l'ordre n au voisinage de x, $f_i(x+h) = \sum_{k=0}^{n} \frac{h^k}{k!} f_i^{(k)}(x) + o(h^n)$ alors la fonction $f_1 + \alpha f_2$ a pour développement limité $(f_1+\alpha f_2)(x+h) = \sum_{k=0}^{n} \frac{h^k}{k!} (f_1+\alpha f_2)^{(k)}(x) + o(h^n) = \sum_{k=0}^{n} \frac{h^k}{k!} f_1^{(k)}(x) + o(h^n) + \sum_{k=0}^{n} \frac{h^k}{k!} f_2^{(k)}(x) + o(h^n)$).

On a $(x+h)^n = \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} h^k x^{n-k}$ avec $(x^n)^{(k)} = n(n-1)...(n-k+1)x^{n-k} = \frac{n!}{(n-k)!} x^{n-k}$, d'où $(x+h)^n = \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} h^k \frac{(n-k)!}{n!} (x^n)^{(k)} = \sum_{k=0}^{n} \frac{1}{k!} h^k (x^n)^{(k)}$: c'est (1.3) pour $p(x) = x^n$. ▪

Proposition 1.4 *L'ensemble $\mathcal{P}_n$ des polynômes de degré inférieur ou égal à n est un espace vectoriel, sous-espace vectoriel de $C^\infty(\mathbb{R};\mathbb{R})$. L'ensemble de tous les polynômes est un espace vectoriel, sous-espace vectoriel de $C^\infty(\mathbb{R};\mathbb{R})$.*

Preuve. Le caractère C^∞ des polynômes sur $\mathbb{R}$ est connu (basé sur la relation de Neumann $\frac{x^{n+1}-x_0^{n+1}}{x-x_0} = x^n + x^{n-1}x_0 + ... + x x_0^{n-1} + x_0^n$ immédiate à vérifier, et qui donne la valeur de la dérivée $(n+1)x_0^n$). Et on a la stabilité par addition et multiplication par un scalaire (immédiat). ▪

Définition 1.5 On dit que $x_0 \in \mathbb{R}$ est racine du polynôme p ssi $p(x_0)=0$.

x_0 est racine simple ssi $p(x_0) = 0$ et $p'(x_0) \neq 0$.

$x_0 \in \mathbb{R}$ est racine multiple d'ordre m du polynôme p ssi $p(x_0)=0$, $p'(x_0)=0$,..., $p^{(m-1)}(x_0)=0$ et $p^{(m)}(x_0)\neq 0$. Si $m=2,3,4,...$, on dit racine double, triple, quadruple...

Exemple 1.6 Si $p(x) = (x+1)(x-2)^2$, alors -1 est racine simple et $+2$ est racine double. ▪

Proposition 1.7 *Soit p un polynôme de degré n.*

1- Si x_0 est racine de p, alors il existe un polynôme q de degré $n-1$ tel que $p(x) = (x-x_0)q(x)$. Et si p a n racines simples $(x_i)_{i=0,...,n-1}$, il est de la forme $p(x) = c \prod_{i=0}^{n-1}(x-x_i)$, où c est une constante.

2- Si x_0 est racine multiple d'ordre m de p, pour $m \leq n$, alors il existe un polynôme q de degré $n-m$ tel que $p(x) = (x-x_0)^m q(x)$, et x_0 n'est pas racine de q.

3- Un polynôme p de degré n qui a $n+1$ racines est le polynôme nul.

4- Il existe un unique polynôme p de degré n qui en $n+1$ points distincts x_i prend $n+1$ valeurs $y_i = p(x_i)$.

Preuve. 1- Un polynôme est égal à son développement limité : $p(x) = p(x_0) + (x-x_0)(q(x))$ avec q polynôme de degré $n-1$. Et $p(x_0)=0$ donne $p(x) = (x-x_0)q(x)$. Et si $x_1 \neq x_0$ et $p(x_1)=0$, alors x_1 annule q et q est de la forme $q(x) = (x-x_1)r(x)$ avec r polynôme de degré $n-2$. La suite de la récurrence est laissée en exercice.

2- C'est vrai pour $m = 1$. Pour $m = 2$, comme x_0 est racine, on a $p(x) = q(x)(x-x_0)$. Puis $p'(x) = q(x) + (x-x_0)q'(x)$ et $p'(x_0) = 0$ donne $q(x_0) = 0$, d'où q est de la forme $q(x) = (x-x_0)r(x)$ avec r polynôme de degré $n-2$. La suite de la récurrence est laissée en exercice.

Puis si x_0 était racine de q, alors q serait de la forme $q(x) = (x-x_0)r(x)$, et x_0 serait une racine d'ordre $m+1$. Donc x_0 n'est pas racine de q.

3- On a $p(x) = q(x)(x-x_0)...(x-x_n)$ avec $q(x)$ polynôme, avec les $n+1$ x_i répétés autant de fois que leur multiplicité. Donc p est de degré $\geq n+1$ avec p de degré $\leq n$, ce qui est absurde, à moins que $q = 0$, i.e. à moins que $p = 0$.

4- C'est vrai lorsque tous les y_i sont nuls d'après 3- : c'est le polynôme nul. Cas général, on cherche p sous la forme $p(x) = a_0 + a_1x + ... + a_nx^n$, les inconnues étant les coefficients a_i, et ces coefficients sont solutions du système $V.\vec{a} = \vec{y}$ où $\vec{a} = (a_0, ..., a_n) \in \mathbb{R}^{n+1}$, $\vec{y} = (y_0, ..., y_n)$, et V est la matrice de Vandermonde (de coefficient $[V_{ij}] = [x_i^j]_{i,j=0,...,n}$). Les x_i étant distincts 2 à 2, cette matrice est inversible, et l'unique solution est donc $\vec{a} = \vec{0}$.

Cependant une matrice de Vandermonde est mal conditionnée (pour n grand), et on évite de s'en servir pour calculer les a_j ; les polynômes de Lagrange, voir paragraphe 1.3, donnent une démonstration immédiate ($p(x) = \sum_{i=0}^{n} y_i L_i(x)$). ∎

1.2.2 Division euclidienne

Théorème 1.8 *Soit p_n un polynôme de degré $n \geq 1$, soit d_m un polynôme non nul de degré $m \geq 1$ (le diviseur) avec $m \leq n$. Alors il existe un unique polynôme q_{n-m} de degré $n-m$ (le quotient) et un unique polynôme r_{m-1} de degré $m-1$ (le reste) tels que :*

$$p_n = d_m q_{n-m} + r_{m-1}. \tag{1.4}$$

Preuve. Notons $p_n(x) = a_nx^n + a_{n-1}x^{n-1} + ... + a_0$, où $a_n \neq 0$, et $d_m(x) = b_mx^m + b_{m-1}x^{m-1} + ... + b_0$, où avec $b_m \neq 0$. Donc il existe une unique constante c_{n-m} t.q. $a_nx^n = b_mx^m x^{n-m}c_{n-m}$, à savoir $c_{n-m} = \frac{a_n}{b_m}$. D'où $p_n = d_m(x)c_{n-m}x^{n-m} + p_{n-1}(x)$ où p_{n-1} est un polynôme de degré $\leq n-1$.

Si $m = n$, c'est terminé. Si $m \leq n-1$, on procède de même avec p_{n-1}... ∎

Remarque 1.9 On peut remarquer qu'en écrivant l'égalité de tous les monômes dans (1.4), on a $n+1$ équations (une pour chaque monôme de p_n) et $(n-m+1) + (m-1+1) = n+1$ inconnues (les coefficients monomiaux de q_{n-m} et r_{m-1}). ∎

Exemple 1.10 $P(x) = x^3 + x^2 - 1$ divisé par $D(x) = x - 1$: on identifie les termes de plus haut degré, d'où $x^3 + x^2 - 1 = x^2(x-1) + 2x^2 - 1$, puis $2x^2 - 1 = 2x(x-1) + 2x - 1$, puis $2x - 1 = 2(x-1) + 1$, et finalement on a trouvé $x^3 + x^2 - 1 = (x-1)(x^2 + 2x + 2) + 1$. Les étapes présentés sous forme de division euclidienne :

$$\begin{array}{rrrr|l} x^3 & x^2 & 0x & -1 & x \quad -1 \\ \cline{5-5} -x^3 & +x^2 & & & x^2 \ +2x \ +2 \\ \cline{1-2} & 2x^2 & 0x & -1 & \\ & -2x^2 & +2x & & \\ \cline{2-3} & & 2x & -1 & \\ & & -2x & +2 & \\ \cline{3-4} & & & +1 & \end{array}$$

∎

1.2.3 Théorème de Weierstrass

Soient deux réels a, b tels que $-\infty < a < b < \infty$.

Pour une fonction $f : [a,b] \to \mathbb{R}$ bornée, on note $||f||_\infty = \sup_{x\in[a,b]} |f(x)|$.

On considère l'espace vectoriel (normé complet) $(C^0([a,b];\mathbb{R}), ||.||_\infty)$ des fonctions continues sur le compact (fermé borné) $[a,b]$, à valeurs réelles.

Théorème 1.11 *Sur l'intervalle compact $[a,b]$ de $\mathbb{R}$, toute fonction continue $f \in C^0([a,b];\mathbb{R})$, peut être approchée uniformément par des fonctions polynomiales :*

$$\forall \varepsilon > 0, \quad \exists p \text{ polynôme}, \quad ||f(x) - p(x)||_\infty < \varepsilon.$$

Autrement dit, l'espace vectoriel des polynômes est dense dans $(C^0([a,b];\mathbb{R}), ||.||_\infty)$.

Preuve. Voir par exemple Arnaudiès et Fraysse [2]. Idée :

1- les fonctions continues sur $[a,b]$ et affines par morceaux forment un espace vectoriel dense dans $C^0([a,b];\mathbb{R})$.

(Une fonction continue f est affine par morceaux s'il existe un nombre fini de points x_i, $0 \leq i \leq n$, tels que $x_0=a$, $x_n=b$ et $x_{i-1}<x_i$ pour tout $i=1,...,n$, pour lesquelles la restriction $f_{|[x_{i-1},x_i]}$ de f à $[x_{i-1},x_i]$ est une fonction affine.)

2- Les fonctions polynômes approchent uniformément toute fonction affine par morceaux.

3- Les points 1- et 2- impliquent le résultat. ∎

Remarque 1.12 Une autre démonstration, à l'aide des polynômes de Bernstein, est donnée dans Schatzman [13]. ∎

1.3 Polynômes de Lagrange

On se place sur un seul intervalle $[a,b]$. (Si $[a,b]$ est "grand", on "maille" $[a,b]$, cf. (1.1), et on applique la procédure pour chaque $[a_{k-1},a_k]$.)

Soit $n+1$ points (de collocation) $(x_i)_{i=0,...,n}$ dans $[a,b]$, distincts 2 à 2, et soit $n+1$ réels $(y_i)_{i=0,...,n}$. On cherche un polynôme p de degré n tel que $p(x_i)=y_i$ pour tout $i=0,...,n$.

Exemple 1.13 Cas $n=0$, donc un seul point x_0 dans $[a,b]$ et une seule valeur y_0, et donc le polynôme cherché est le polynôme constant sur $[a,b]$ donné par $p(x)=y_0$ pour tout $x \in [a,b]$. (Dans la pratique on prend $x_0 \in]a,b[$ pour ne pas avoir de problème pour la généralisation à un maillage.) ∎

Exemple 1.14 Cas $n=1$, donc deux points x_0,x_1 dans $[a,b]$ et deux valeurs y_0,y_1, et donc une seul polynôme de degré 1 t.q. $p(x_0)=y_0$ et $p(x_1)=y_1$, à savoir $p(x)=y_0\frac{x-x_1}{x_0-x_1}+y_1\frac{x-x_0}{x_1-x_0}$, vérification immédiate (le membre de droite est bien un polynôme de degré 1 et donne trivialement le résultat). ∎

Exemple 1.15 Cas $n=2$, donc trois points x_0,x_1,x_2 dans $[a,b]$ et trois valeurs y_0,y_1,y_2, et donc une seul polynôme de degré 2 t.q. $p(x_0)=y_0$, $p(x_1)=y_1$ et $p(x_2)=y_2$, à savoir $p(x)=y_0\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}+y_1\frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}+y_2\frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-1)}$, vérification immédiate (le membre de droite est de degré 2 et donne trivialement le résultat). ∎

Les trois exemples précédents donnent une base de $\mathcal{P}_n$ adaptée à la recherche du polynôme d'interpolation :

• Pour $n=0$, la base $\mathcal{P}_0$ est la fonction $L_0=1_{\mathbb{R}}$ (fonction de Lagrange pour du $\mathcal{P}_0$).

• pour $n=1$, la base $\mathcal{P}_1$ est constituée des deux fonctions L_0 et L_1 (fonction de Lagrange pour du $\mathcal{P}_1$) données par $L_0(x)=\frac{x-x_1}{x_0-x_1}$ et $L_1(x)=\frac{x-x_0}{x_1-x_0}$, fonctions qui vérifient $L_i(x_j)=\delta^i_j$.

• pour $n=2$, la base $\mathcal{P}_2$ est constituée des trois fonctions L_0, L_1 et L_2 (fonction de Lagrange pour du $\mathcal{P}_2$) données par $L_0(x)=\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}$, $L_1(x)=\frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}$, et $L_2(x)=\frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-1)}$, fonctions qui vérifient $L_i(x_j)=\delta^i_j$.

Et on généralise à tout n :

Définition 1.16 Pour $n \in \mathbb{N}$, pour $i \in [0,n]_{\mathbb{N}}$, le polynôme L_i de degré n qui vérifie

$$\forall j=0,...,n, \quad L_i(x_j)=\delta_{ij} \overset{\text{déf}}{=} \begin{cases} =1 & \text{si } i=j, \\ =0 & \text{si } i \neq j \end{cases} \tag{1.5}$$

est appelé le i-ème polynôme de base de Lagrange de degré n.

Exemple 1.17 Cas $n=2$:

$$\begin{cases} L_0(x_0)=1 \\ L_0(x_1)=0 \\ L_0(x_2)=0 \end{cases} \quad \begin{cases} L_1(x_0)=0 \\ L_1(x_1)=1 \\ L_1(x_2)=0 \end{cases} \quad \begin{cases} L_2(x_0)=0 \\ L_2(x_1)=0 \\ L_2(x_2)=1 \end{cases}$$

qui ressemble à la base canonique dans $\mathbb{R}^3$ (même ressemblance quelque soit n). ∎

Donc pour $n \geq 1$, pour $0 \leq i \leq n$, les L_i ont pour racines les n points x_j pour $j \neq i$: ils sont donc donnés par (après normalisation pour avoir $L_i(x_i)=1$) :

$$L_i(x)=\prod_{j \neq i} \frac{(x-x_j)}{(x_i-x_j)} \tag{1.6}$$

Proposition 1.18 *Les $(L_i)_{i=0,n}$ donnés en (1.6) forment une base de l'espace $\mathcal{P}_n$ des polynômes de degré $\leq n$. Et tout polynôme $p_n \in \mathcal{P}_n$ vérifiant $p_n(x_j) = y_j$, pour tout $j = 0, ..., n$, est donné par $p_n = \sum_{i=0}^n y_i L_i$, où donc les y_i sont les composantes de p_n sur la base $(L_i)_{i=0,n}$, i.e. :*

$$\forall x \in [a,b], \quad p_n(x) = \sum_{i=0}^{n} y_i L_i(x). \tag{1.7}$$

(p_n est le polynôme dont le graphe passe par les points (x_i, y_i) pour $i = 0, ..., n$.)

Preuve. $\mathcal{P}_n$ étant un espace vectoriel, ayant $L_i \in \mathcal{P}_n$ pour tout i, toute combinaison linéaire des L_i est dans $\mathcal{P}_n$ (est un polynôme de degré $\leq n$).

1- Les L_i sont indépendants dans $\mathcal{P}_n$: étant donnés $n+1$ scalaires $(\alpha_i)_{i=0,n}$, si $\sum_{i=0}^n \alpha_i L_i = 0$, i.e. si pour tout x on a $\sum_{i=0}^n \alpha_i L_i(x) = 0$, alors en particulier $\sum_{i=0}^n \alpha_i L_i(x_j) = 0$, pour tout $j \in [0,n]$, i.e. $\sum_{i=0}^n \alpha_i \delta_{ij} = 0 = \alpha_j$ pour tout j. Donc les L_i forment une famille libre.

2- Les L_i sont générateurs de $\mathcal{P}_n$: ils sont indépendants en nombre $n+1$ dans $\mathcal{P}_n$ qui est de dimension $n+1$. Ou encore : soit p_n est un polynôme de degré n. On note $y_j =^{\text{déf}} p_n(x_j)$ pour tout j. On pose $q(x) = \sum_{i=0}^n y_i L_i(x)$ qui est un polynôme de degré n (combinaison linéaire de polynômes de degré n). Montrons que $p = q$: on a $q(x_j) = \sum_{i=0}^n y_i L_i(x_j) = \sum_{i=0}^n \delta_{ij} = p(x_j)$ pour tout j, donc p et q, polynômes de degrés n, sont égaux pour $n+1$ points distincts, donc $p_n = q$. ▪

Exemple 1.19 Avec x_0=0 et x_1=1, les polynômes L_0 et L_1 définis par $L_0(x) = 1-x$ et $L_1(x) = x$ forment une base de P_1. Et tout polynôme de la forme $p_1(x) = ax + b$ tel que $p_1(x_0) = y_0$ et $p_1(x_1) = y_1$ s'écrit $p_1(x) = y_0 L_0(x) + y_1 L_1(x)$. En effet, on a $y_0 L_0(x) + y_1 L_1(x) = y_0 + (y_1 - y_0)x$ vaut bien y_0 en $x = 0$ et y_1 en $x = 1$. ▪

Exemple 1.20 Par 2 points (x_0, y_0) et (x_1, y_1) passe l'unique polynôme

$$\begin{aligned} p_1(x) &= y_0 \frac{x - x_1}{x_0 - x_1} + y_1 \frac{x - x_0}{x_1 - x_0} \\ &= \frac{y_1 - y_0}{x_1 - x_0} x - \frac{y_1 x_0 - y_0 x_1}{x_1 - x_0}, \end{aligned} \tag{1.8}$$

dont le graphe est une droite (la pente est bien $\frac{y_1-y_0}{x_1-x_0} = \frac{\text{côté opposé}}{\text{côté adjacent}}$). ▪

Exemple 1.21 Par 3 points (x_0, y_0), (x_1, y_1) et (x_2, y_2) passe l'unique polynôme

$$\begin{aligned} p_2(x) = y_0 &\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} + y_1 \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} + y_2 \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} \\ = & - \frac{y_0(x_1-x_2) + y_1(x_2-x_0) + y_2(x_0-x_1)}{(x_0-x_1)(x_1-x_2)(x_2-x_0)} x^2 \\ & + \frac{y_0(x_1^2-x_2^2) + y_1(x_2^2-x_0^2) + y_2(x_0^2-x_1^2)}{(x_0-x_1)(x_1-x_2)(x_2-x_0)} x \\ & - \frac{y_0(x_1-x_2)x_1x_2 + y_1(x_2-x_0)x_2x_0 + y_2(x_0-x_1)x_0x_1}{(x_0-x_1)(x_1-x_2)(x_2-x_0)} \end{aligned} \tag{1.9}$$

dont le graphe est une parabole. ▪

Remarque 1.22 L'approximation par les polynômes de Lagrange est utilisée lors de la résolution des équations aux dérivées partielles par la technique des éléments finis. Ils sont alors appelés "polynômes de base".

Par exemple, pour les éléments finis dits "P_1" sur l'intervalle $[a_{k-1}, a_k]$, les deux polynômes de base sont souvent notés $\varphi_1(x) = \frac{x-a_k}{a_{k-1}-a_k}$ $(= L_0(x))$ et $\varphi_2(x) = \frac{x-a_{k-1}}{a_k-a_{k-1}}$ $(= L_1(x))$: ils vérifient $\varphi_1(a_{k-1}) = 1$, $\varphi_1(a_k) = 0$ et $\varphi_2(a_{k-1}) = 0$, $\varphi_2(a_k) = 1$. ▪

Exercice 1.23 Soit le polynôme de degré $n+1$:

$$\pi_{n+1}(x) = \prod_{i=0}^{n} (x - x_i). \tag{1.10}$$

Soit f une fonction et p_n son polynôme d'interpolation de Lagrange, cf. (1.7). Montrer :

$$p_n(x) = \pi_{n+1}(x) \sum_{i=0}^{n} \frac{y_i}{\pi'_{n+1}(x_i)(x - x_i)}. \tag{1.11}$$

Ecriture abusive de $p_n(x) = \sum_{i=0}^{n} \frac{y_i}{\pi'_{n+1}(x_i)} \frac{\pi_{n+1}(x)}{(x - x_i)}$, où on a noté $\frac{\pi_{n+1}(x)}{(x - x_i)} \overset{\text{déf}}{=} \prod_{\substack{j=1,...,n \\ j \neq i}} (x - x_j)$.

Réponse. $\pi_{n+1}(x) = \prod_{k=0}^{n}(x-x_k)$ donne $\pi'_{n+1}(x) = \sum_{i=0}^{n} \prod_{\substack{k=0,\dots,n \\ k\neq i}} (x-x_k)$, donc $\pi'_{n+1}(x_i) = \prod_{\substack{k=0,\dots,n \\ k\neq i}} (x_i - x_k)$. ∎

1.4 * Polynômes de Lagrange sur un maillage : spline C^0

On considère le maillage de $[a,b]$ donné par (1.1), i.e. $[a,b] = \bigcup_{k=1}^{N_{\rm int}} [a_{k-1}, a_k]$, où a_0=a, $a_{N_{\rm int}}$=b et a_{k-1}<a_k pour tout $k = 1, ..., N_{\rm int}$.

Et on applique la procédure de Lagrange sur chaque $[a_{k-1}, a_k]$. Pour simplifier la présentation, dans chaque intervalle $[a_{k-1}, a_k]$ on prend n+1 points de collocation, points qu'on notera x_{ki} pour $i = 1, ..., n+1$. Et on associe n+1 valeurs y_{ki} pour $i = 1, ..., n+1$. On obtient ainsi $N_{\rm int}$ polynôme de Lagrange $p_{(k)}$ vérifiant $p_{(k)}(x_{ki}) = y_{ki}$.

On définit alors $f : [a,b] \to \mathbb{R}$ la fonction définit par $f_{|]a_{k-1},a_k[} = p_{(k)}$, avec $f(a_k)$ quelconque, par exemple $f(a_k) = 0$ pour tout k. En général f n'est pas continue sur $[a,b]$.

Cas particulier : si $f(a_k-) = f(a_k+)$, i.e. si $p_{(k)}(a_k) = p_{(k+1)}(a_k)$, ce pour tout $k = 1, ..., n-1$, alors la fonction f est continue sur tout $[a,b]$. On dit alors que f est une spline C^0 sur $[a,b]$.

Ces splines C^0 sont à la base de la méthode des éléments finis P_k (approximation C^0 localement de degré k).

1.5 * Polynômes de Newton

L'interpolation de Lagrange présente un inconvénient : si on dispose d'un point supplémentaire (x_{n+1}, y_{n+1}) en plus des n+1 points précédents $(x_i, y_i)_{i=0,\dots,n}$, alors le nouveau polynôme d'interpolation de Lagrange p_{n+1} de degré n+1 sous la forme classique $\sum_{i=0}^{n+1} a_i x^i$ ne se déduit pas facilement du polynôme de Lagrange p_n de degré n précédent. Comparer par exemple (1.8) et (1.9).

L'idée de Newton est d'écrire les polynômes de manière différente : à partir du polynôme constant (de degré 0) :

$$p_0(x) = y_0, \tag{1.12}$$

où donc $p_0(x_0) = y_0$, on ajoute un point $x_1 \neq x_0$ et la valeur y_1, et on veut obtenir le polynôme p_1 de degré 1 t.q. $p_1(x_0) = y_0$ et $p_1(x_1) = y_1$ sous la forme :

$$p_1(x) = p_0(x) + a_1(x - x_0) = a_0 + a_1(x - x_0). \tag{1.13}$$

On a ainsi ajouté $a_1(x-x_0)$ (polynôme de degré 1) qui s'annule en x_0 : la valeur en x_0 est donc inchangée. Donc en particulier $a_0 = y_0$. Comme on doit avoir $p_1(x_1) = y_1$, il vient :

$$a_1 = \frac{y_1 - y_0}{x_1 - x_0}, \tag{1.14}$$

c'est la pente. Puis on poursuit le processus : on ajoute un point x_2, distinct de x_0 et de x_1, et une valeur y_2 pour obtenir :

$$p_2(x) = p_1(x) + a_2(x - x_0)(x - x_1) \qquad (= a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1)),$$

où on a ajouté à p_1 un polynôme de degré 2 qui s'annule en x_0 et x_1 : la valeur en x_0 et x_1 est donc inchangée. Comme on doit avoir $p_2(x_2) = y_2$, il vient $y_2 = y_0 + \frac{y_1-y_0}{x_1-x_0}(x_2-x_0) + a_2(x_2-x_0)(x_2-x_1)$ soit :

$$a_2 = \frac{(y_2 - y_0)(x_1 - x_0) - (y_1 - y_0)(x_2 - x_0)}{(x_2 - x_0)(x_2 - x_1)(x_1 - x_0)} = \frac{1}{(x_2 - x_0)}\left(\frac{y_2 - y_1}{x_2 - x_1} - \frac{y_1 - y_0}{x_1 - x_0}\right). \tag{1.15}$$

On poursuit le processus : disposant des points $(x_i, y_i) \in \mathbb{R}^2$ pour $i = 0, ..., k-1$, où les x_i sont distincts 2 à 2, on ajoute un point (x_k, y_k) et le polynôme obtenu s'écrit :

$$p_k(x) = p_{k-1}(x) + a_k(x - x_0)(x - x_1)...(x - x_{k-1}),$$

et donc $p_k - p_{k-1}$ est le polynôme de degré k qui s'annule aux k points x_0,...,x_{k-1} et qui vaut y_k au point x_k, ce qui détermine a_k.

Définition 1.24 Pour une fonction f définie en deux points x_0 et x_1 distincts, on appelle première différence divisée de f la valeur :

$$f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0} \tag{1.16}$$

(pente moyenne entre x_0 et x_1.) Si f est définie en 3 points distincts x_0, x_1 et x_2, on appelle deuxième différence divisée de f la valeur :

$$f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0}. \tag{1.17}$$

(C'est la valeur a_2 donnée en (1.15).) Et de manière générale, on définit la n-ième différence divisée en n+1

points distincts par :

$$f[x_0, ..., x_n] = \frac{f[x_1, ..., x_n] - f[x_0, ..., x_{n-1}]}{x_n - x_0}. \tag{1.18}$$

(Noter qu'on n'a pas besoin de supposer que $x_n > x_{n-1}$.)

Pour le polynôme de Newton p_1 on a donc :

$$p_1(x) = f(x_0) + f[x_0, x_1](x - x_0),$$

et que pour le polynôme de Newton p_2 on a donc :

$$p_2(x) = f(x_0) + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1).$$

Proposition 1.25 *Étant donnés $n+1$ points $(x_i, f(x_i))_{i=0,...,n}$ l'unique polynôme de degré n qui passe par ces points est donné par :*

$$p_n(x) = f(x_0) + f[x_0, x_1](x - x_0) + ... + f[x_0, ..., x_n](x - x_0)...(x - x_{n-1}). \tag{1.19}$$

Preuve. Démonstration par récurrence. C'est vrai pour p_1.

Supposons que ce le soit pour p_{n-1}, $n \geq 2$, i.e. qu'étant donnés n points, $(x_i, y_i)_{i=0,...,n-1}$ où $y_i = f(x_i)$, le polynôme p_{n-1} de Newton est donné par :

$$p_{n-1}(x) = f(x_0) + f[x_0, x_1](x - x_0) + f[x_0, ..., x_{n-1}](x - x_0)...(x - x_{n-2}).$$

Et pour les points $(x_i, f(x_i))_{i=1,...,n}$, le polynôme q_{n-1} de Newton est donné par :

$$q_{n-1}(x) = f(x_1) + f[x_1, x_2](x - x_1) + f[x_1, ..., x_n](x - x_1)...(x - x_{n-1}).$$

Et par hypothèse de récurrence :

$$p_{n-1}(x_i) = q_{n-1}(x_i) \quad (= y_i = f(x_i)) \quad \text{pour } i = 1, ..., n-1.$$

D'où immédiatement, en posant :

$$p_n(x) = \frac{(x_n - x)p_{n-1}(x) + (x - x_0)q_{n-1}(x)}{(x_n - x_0)}, \tag{1.20}$$

on a $p_n(x_i) = f(x_i)$ pour tout $i = 0, ..., n$. Donc p_n est le polynôme cherché. Il reste à montrer que :

$$p_n(x) - p_{n-1}(x) = f[x_0, ..., x_n](x - x_0)...(x - x_{n-1}).$$

Mais posant :

$$s_n(x) = p_n(x) - p_{n-1}(x) = \frac{(x - x_0)}{(x_n - x_0)}(q_{n-1}(x) - p_{n-1}(x)),$$

le polynôme s_n de degré n s'annule en les n points x_i, $0 \leq i \leq n-1$: c'est trivial pour $x = x_0$ et on a $p_{n-1}(x_i) = q_{n-1}(x_i)$ pour $1 \leq i \leq n-1$. Donc s_n est de la forme $s_n(x) = \alpha(x - x_0)...(x - x_{n-1})$ pour un scalaire α donné. Et il reste à montrer que $\alpha = f[x_0, ..., x_n]$.

Mais le coefficient de x^n dans p_n est donné à l'aide de (1.20) par :

$$\frac{-f[x_0, ..., x_{n-1}] + f[x_1, ..., x_n]}{(x_n - x_0)} \quad \text{donc } = f[x_0, ..., x_n],$$

par définition de $f[x_0, ..., x_n]$. ∎

Remarque 1.26 On présente généralement les coefficients du polynôme de Newton sous la forme d'une table "de différences divisées", par exemple pour les polynômes de degré 3 :

x_i	$f(x_i)$	$f[x_i, x_{i+1}]$	$f[x_i, x_{i+1}, x_{i+2}]$	$f[x_i, x_{i+1}, x_{i+2}, x_{i+3}]$
x_0	$f(x_0) = a_0$			
		$f[x_0, x_1] = a_1$		
x_1	$f(x_1)$		$f[x_0, x_1, x_2] = a_2$	
		$f[x_1, x_2]$		$f[x_0, x_1, x_2, x_3] = a_3$
x_2	$f(x_2)$		$f[x_1, x_2, x_3]$	
		$f[x_2, x_3]$		
x_3	$f(x_3)$			

dans lesquelles on remplace les valeurs génériques $f[...]$ par les valeurs calculées. Autre présentation :

x_i	$f(x_i)$		$f[x_i, x_{i+1}]$		$f[x_i, x_{i+1}, x_{i+2}]$		$f[x_i, x_{i+1}, x_{i+2}, x_{i+3}]$
x_0	$f(x_0) = a_0$	$\to$	$f[x_0, x_1] = a_1$	$\to$	$f[x_0, x_1, x_2] = a_2$	$\to$	$f[x_0, x_1, x_2, x_3] = a_3$
		$\nearrow$		$\nearrow$		$\nearrow$	
x_1	$f(x_1)$	$\to$	$f[x_1, x_2]$	$\to$	$f[x_1, x_2, x_3]$		
		$\nearrow$		$\nearrow$			
x_2	$f(x_2)$	$\to$	$f[x_2, x_3]$				
		$\nearrow$					
x_3	$f(x_3)$						

Ici :

$$p(x) = a_0 + a_1(x-x_0) + a_2(x-x_0)(x-x_1) + a_3(x-x_0)(x-x_1)(x-x_2). \tag{1.21}$$

Exercice 1.27 Déterminer le polynôme de Newton de degré 3 qui passe par les points $(1, 0)$, $(1.5, 1)$, $(2, 2)$ et $(2.5, -1.5)$. (Pris sur internet.)

Réponse.

x_i	$f(x_i)$		$f[x_i, x_{i+1}]$		$f[x_i, x_{i+1}, x_{i+2}]$		$f[x_i, x_{i+1}, x_{i+2}, x_{i+3}]$
1	$0 = a_0$	$\to$	$\frac{1-0}{1.5-1} = 2 = a_1$	$\to$	$\frac{2-2}{2-} = 0 = a_2$	$\to$	$\frac{-9-0}{2.5-1} = -6 = a_3$
		$\nearrow$		$\nearrow$		$\nearrow$	
1.5	1	$\to$	$\frac{2-1}{2-1.5} = 2$	$\to$	$\frac{-7-2}{2.5-1.5} = -9$		
		$\nearrow$		$\nearrow$			
2	2	$\to$	$\frac{-1.5-2}{2.5-2} = -7$				
		$\nearrow$					
2.5	-1.5						

et donc $p(x) = 2(x-1) - 6(x-1)(x-1.5)(x-2)$.

Exercice 1.28 Déterminer le polynôme de degré 2 qui passe par les points $(-1, 9)$, $(2, 4)$, $(5, 1)$ par la méthode de Lagrange. Retrouver ce polynôme à l'aide de la méthode de Newton. Puis on ajoute le point $(0, 0)$: déterminer le polynôme de degré 3 qui passe par les 4 points donnés.

Remarque 1.29 Une fois les valeurs a_i calculées, il est intéressant, dans le cas où on doit calculer de nombreuses valeurs $p_n(x)$ en des points x, de se servir du schéma d'Horner ;

$$p_n(x) = a_0 + (x-x_0)\Big(a_1 + (x-x_1)\big(a_2 + (x-x_2)(a_3 + \ ...\ )\big)\Big),$$

qui réduit le nombre d'opérations et qui de plus est relativement insensible aux erreurs d'arrondis.

Remarque 1.30 On vient de voir deux décompositions polynômiales :

$$f(x_0) + f[x_0, x_1](x - x_0) + ... + f[x_0, ..., x_n](x - x_0)...(x - x_{n-1}) = \sum_{i=0}^{n} f(x_i) \prod_{j \neq i} \frac{(x - x_j)}{(x_i - x_j)}$$

les polynômes de gauche (décomposition de Newton) et de droite (décomposition de Lagrange) étant égaux car tous les deux de degré n, et prenant tous les deux les mêmes $n+1$ valeurs $y_i = f(x_i)$ aux points x_i.

1.6 * Polynômes d'Hermite

1.6.1 Introduction

Les polynômes de degré n de Lagrange permettent d'approximer une fonction f à l'aide de valeurs en $n+1$ points distincts $(x_i)_{i=0,...,n}$ connaissant les $y_i = f(x_i)$. On dit que chaque point est associé à un degré de liberté. Au total on a $n+1$ degrés de liberté (ddl ou dof = degree of freedom).

Les polynômes de développement limité de Taylor de degré n permettent d'approximer une fonction f au voisinage d'un seul point x_0 avec les $n+1$ valeurs $f^{(i)}(x_0)$ pour $i = 0, ..., n$, le point x_0 étant associé. On dit que le point x_0 est associé à $n+1$ degrés de liberté. Au total on a $n+1$ degrés de liberté.

L'idée de Hermite, pour obtenir un polynôme de degré n, est intermédiaire des deux précédentes : approcher une fonction f en $\gamma+1$ points $(x_i)_{i=0,...,\gamma}$ pour $\gamma < n$, où donc au moins un des x_i est associé à 2 ddl. On note n_i le nombre de ddl, et on suppose que $\sum_{i=0}^{\gamma} = n+1$.

Exemple 1.31 Lagrange : cas $n+1$ points x_i distincts et $n_i = 1$ pour tout i, chaque point étant associé à un seul ddl (degré de liberté), caractérisé par la valeur $y_i = f(x_i)$ pour la fonction f à interpoler. ▞

Exemple 1.32 Développement limité de Taylor à l'ordre n : cas un seul point x_0 et $n_0 = n+1$, l'unique point de x_0 étant associé aux $n+1$ ddl données par les valeurs $f(x_0), f'(x_0), ..., f^{(n)}(x_0)$. ▞

Exemple 1.33 Deux points x_0 et $x_1 \neq x_0$, deux ddl pour x_0 données par $f(x_0) =^{\text{noté}} y_0$ et $f'(x_0) =^{\text{noté}} y'_0$, et un ddl pour x_1 donné par $f(x_1) =^{\text{noté}} y_1$. ▞

Exercice 1.34 1- Donner les trois polynômes P_1, P_2, P_3 de degré 2 t.q. pour $x_0 \neq x_1$:

$$\left\{\begin{array}{l} P_1(x_0) = 1 \\ P_1'(x_0) = 0 \\ P_1(x_1) = 0 \end{array}\right. \qquad \left\{\begin{array}{l} P_2(x_0) = 0 \\ P_2'(x_0) = 1 \\ P_2(x_1) = 0 \end{array}\right. \qquad \left\{\begin{array}{l} P_3(x_0) = 0 \\ P_3'(x_0) = 0 \\ P_3(x_1) = 1 \end{array}\right.$$

et montrer qu'il forment une base de $\mathcal{P}_2$.

Réponse. x_1 est racine de P_1, donc $P_1(x) = (x - x_1)(\alpha x + \beta)$. Puis $P_1(x_0) = 1 = (x_0 - x_1)(\alpha x_0 + \beta)$. Comme $x_0 \neq x_1$ on a $\alpha x_0 + \beta = \frac{1}{x_0 - x_1}$. Puis $P_1'(x) = (\alpha x + \beta) + \alpha(x - x_1)$, d'où $P_1'(x_0) = 0 = (\alpha x_0 + \beta) + \alpha(x_0 - x_1) = \alpha(2x_0 - x_1) + \beta$. D'où $\beta = -\alpha(2x_0 - x_1)$, d'où $\alpha x_0 + \alpha(x_1 - 2x_0) = \frac{1}{x_0 - x_1}$, d'où $\alpha = -\frac{1}{(x_0 - x_1)^2}$, d'où $\beta = \frac{2x_0 - x_1}{(x_0 - x_1)^2}$.

x_0 et x_1 sont racines de P_2, d'où $P_2(x) = \alpha(x - x_0)(x - x_1)$, d'où $P_2'(x) = \alpha(2x - x_0 - x_1)$, d'où $P_2'(x_0) = 1 = \alpha(x_0 - x_1)$, d'où $\alpha = \frac{1}{x_0 - x_1}$.

x_0 est racine double de P_3, d'où $P_3(x) = \alpha(x - x_0)^2$, avec $P_3(x_1) = 1 = \alpha(x_1 - x_0)^2$, d'où $\alpha = \frac{1}{(x_1 - x_0)^2}$.

Finalement :

$$P_1(x) = \frac{(x - x_1)(-x + 2x_0 - x_1)}{(x_0 - x_1)^2}, \qquad P_2(x) = \frac{(x - x_0)(x - x_1)}{(x_0 - x_1)}, \qquad P_3(x) = \frac{(x - x_0)^2}{(x_1 - x_0)^2}. \tag{1.22}$$

▞

1.6.2 Exemple : splines C^1, vers la construction d'éléments finis C^1

Sur $[a, b] = \bigcup_{k=1}^{N_{\text{int}}} [a_{k-1}, a_k]$, cf. (1.1), on veut faire passer une fonction f qui est C^1 sur $[a, b]$ et t.q. $f_{|[a_{k-1}, a_k]}$ est un polynôme de degré 3 pour tout k (on peut voir qu'on ne peut pas prendre un polynôme de plus bas degré). Sur chaque intervalle $[a_{k-1}, a_k]$ on prend pour points de collocations les deux points $x_{k0} = a_{k-1}$ et $x_{k1} = a_k$ (extrémités de l'intervalle) tous les deux associés à 2 ddl, à savoir $f(x_{k0})$ et $f'(x_{k0})$ pour x_{k0} et $f(x_{k1})$ et $f'(x_{k1})$ pour x_{k1} (procédure d'Hermite). On a besoin du résultat élémentaire :

Proposition 1.35 *Soit deux réels distincts x_0 et x_1. Soit quatre réels y_0, z_0, y_1, z_1. Alors il existe un unique polynôme p de degré 3 tel que $p(x_0) = y_0$, $p'(x_0) = z_0$, $p(x_1) = y_1$, $p'(x_1) = z_1$.*

Preuve. Calcul direct. (Voir par exemple exercice 1.37 plus loin.) ▞

Application : spline C^1. On applique la proposition 1.35 pour chaque intervalle $[a_{k-1}, a_k]$ comme suit. On prend $x_{k0} = a_{k-1}$ et $x_{k1} = a_k$ (les extrémités de l'intervalle), on se donne des valeurs y_{k0}, z_{k0}, y_{k1}, z_{k1}, et on note p_k le polynôme de degré 3 obtenu sur cet intervalle, cf. proposition 1.35, i.e. le polynôme de degré 3 vérifiant $p_k(a_{k-1}) = y_{k0}$, $p_k'(a_{k-1}) = z_{k0}$, et $p_k(a_k) = y_{k1}$, $p_k'(a_k) = z_{k1}$.

Et on suppose que $y_{k1} = y_{(k+1)0}$ (continuité en x_{k0}) et $z_{k1} = z_{(k+1)0}$ (continuité de la dérivée en x_{k0}) pour tout $k = 1, ..., N_{\text{int}} - 1$. (Ou si on veut interpoler une fonction $f \in C^1([a, b])$, on prend $y_{k0} = f(a_{k-1})$, $y_{k1} = f(a_k)$ et $z_{k0} = f'(a_{k-1})$, $z_{k1} = f'(a_k)$, pour tout $k = 1, .., N_{\text{int}}$.)

Alors la fonction $g : [a, b] \to \mathbb{R}$ vérifiant $g_{[a_{k-1}, a_k]}$ est appelée une spline cubique C^1 : c'est une fonction C^1 sur tout $[a, b]$, qui est polynomiale de degré 3 sur chaque $[a_{k-1}, a_k]$ (mais elle n'est pas C^2, sauf si par miracle g est un polynôme de degré 3 sur tout $[a, b] = [a_0, a_{N_{\text{int}}}]$, ce qui est par exemple le cas si f est elle-même un polynôme de degré 3 sur $[a, b]$).

Remarque. Pour le calcul du p de la proposition 1.35, une démarche simple est de considérer les 4 polynômes $\varphi_1, \varphi_2, \varphi_3, \varphi_4$ de degré 3 qui vérifient (polynôme Hermite de base) :

$$\left\{\begin{array}{llll} \varphi_1(x_0) = 1, & \varphi_2(x_0) = 0, & \varphi_3(x_0) = 0, & \varphi_4(x_0) = 0, \\ \varphi_1'(x_0) = 0, & \varphi_2'(x_0) = 1, & \varphi_3'(x_0) = 0, & \varphi_4'(x_0) = 0, \\ \varphi_1(x_1) = 0, & \varphi_2(x_1) = 0, & \varphi_3(x_1) = 1, & \varphi_4(x_1) = 0, \\ \varphi_1'(x_1) = 0, & \varphi_2'(x_1) = 0, & \varphi_3'(x_1) = 0, & \varphi_4'(x_1) = 1. \end{array}\right. \tag{1.23}$$

Par exemple, φ_2 a pour racine double x_1, a pour racine simple x_0 et est donc de la forme $\varphi_2(x) =$

$\alpha(x-x_1)^2(x-x_0)$. Et α est déterminé par $\varphi_0^1(x_0)=1$. Les calculs donnent (exercice) :

$$\begin{cases} \varphi_1(x) = \dfrac{(x-x_1)^2}{(x_0-x_1)^2}\Big(2\dfrac{x-x_0}{x_1-x_0}+1\Big), \\ \varphi_2(x) = \dfrac{(x-x_1)^2}{(x_0-x_1)^2}\Big(x-x_0\Big), \end{cases} \tag{1.24}$$

et φ_3 et φ_4 en inversant les rôles de x_1 et de x_0.

Proposition 1.36 *Supposant $x_0 \neq x_1$, les polynômes $(\varphi_i)_{i=1,\dots,4}$ forment une base de $\mathcal{P}_3$ l'ensemble des polynômes de degré 3. Et tout polynôme de degré 3 dont on connaît les valeurs $p(x_0)=y_0$, $p'(x_0)=z_0$, $p(x_1)=y_1$ et $p'(x_1)=z_1$ se décompose sur cette base comme :*

$$p(x) = y_0\varphi_1(x) + z_0\varphi_2(x) + y_1\varphi_3(x) + z_1\varphi_4(x) \qquad (=\sum_{i=1}^{4} z_i\varphi_i(x)). \tag{1.25}$$

Et le polynôme p vérifie immédiatement : $p(x_0)=y_0$, $p(x_1)=y_1$, $p'(x_0)=z_0$ et $p'(x_1)=z_1$.

Preuve. Tout d'abord, les φ_i sont tous de degré 3, donc appartiennent à $\mathcal{P}_3$. Ensuite, ils forment une famille libre car si les α_i sont des réels tels que pour tout x :

$$(\alpha_1\varphi_1 + \alpha_2\varphi_2 + \alpha_3\varphi_3 + \alpha_4\varphi_4)(x) = 0,$$

alors au point x_0 on obtient $\alpha_1=0$, au point x_1 on obtient $\alpha_3=0$, puis en dérivant, au point x_0 on obtient $\alpha_2=0$, et au point x_1 on obtient $\alpha_4=0$. Puis c'est une famille génératrice de P_3 : 4 éléments indépendants dans un espace vectoriel de dimension 4. ∎

Exercice 1.37 Démontrer directement la proposition 1.35.

Réponse. Soit $p(x)=a_0+a_1x+a_2x^2+a_3x^3$; on a 4 inconnues les $(a_i)_{i=0,\dots,3}$ et 4 équations à satisfaire, sous forme matricielle :

$$\begin{pmatrix} 1 & x_0 & x_0^2 & x_0^3 \\ 0 & 1 & 2x_0 & 3x_0^2 \\ 1 & x_1 & x_1^2 & x_1^3 \\ 0 & 1 & 2x_1 & 3x_1^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} y_0 \\ z_0 \\ y_1 \\ z_1 \end{pmatrix}.$$

La matrice est inversible (pour $x_0 \neq x_1$), le déterminant de la matrice valant $(x_1-x_0)^4 \neq 0$. En effet, notant A la matrice, le déterminant étant une forme multilinéaire alternée :

$$\begin{aligned} \det(A) &= \begin{vmatrix} 1 & x_0 & x_0^2 & x_0^3 \\ 0 & 1 & 2x_0 & 3x_0^2 \\ 0 & x_1-x_0 & (x_1-x_0)(x_1+x_0) & (x_1-x_0)(x_1^2+x_0x_1+x_0^2) \\ 0 & 0 & 2(x_1-x_0) & 3(x_1-x_0)(x_1+x_0) \end{vmatrix} \\ &= (x_1-x_0)^2 \begin{vmatrix} 1 & 2x_0 & 3x_0^2 \\ 1 & (x_1+x_0) & (x_1^2+x_0x_1+x_0^2) \\ 0 & 2 & 3(x_1+x_0) \end{vmatrix} \\ &= (x_1-x_0)^2 \begin{vmatrix} 1 & 2x_0 & 3x_0^2 \\ 0 & (x_1-x_0) & (x_1^2+x_0x_1-2x_0^2) \\ 0 & 2 & 3(x_1+x_0) \end{vmatrix} \\ &= (x_1-x_0)^2[3(x_1-x_0)(x_1+x_0) - 2(x_1^2+x_0x_1-2x_0^2)] = (x_1-x_0)^4. \end{aligned}$$

D'où l'existence et l'unicité. ∎

1.6.3 Exemple : vers la formule d'intégration de Simpson

Soient 3 points x_0, x_1 et $x_{\frac{1}{2}} = \frac{x_0+x_1}{2}$ (avec $x_0 \neq x_1$) pour lesquels, soit 4 valeurs y_0, y_1, $y_{\frac{1}{2}}$ et $y_{\frac{1}{2}}^1$, et on veut toruver un polynôme p de degré 3 vérifiant :

$$y_0 = p(x_0), \quad y_1 = p(x_1), \quad y_{\frac{1}{2}} = p(x_{\frac{1}{2}}), \quad y_{\frac{1}{2}}^1 = p'(x_{\frac{1}{2}}).$$

Ici x_0 et x_1 ont un ddl (degré de liberté), alors que x_2 à associé deux ddl.

Proposition 1.38 *Un tel polynôme p de degré 3 existe et est unique.*

Preuve. Calcul direct : voir exercice suivant 1.40. ∎

Mais il est tout aussi simple dans la pratique de suivre l'idée de Lagrange : on cherche les 4 polynômes (dit de base) φ_i^j de degré 3 qui vérifient ;

$$\begin{cases} \varphi_0^0(x_0)=1, & \varphi_0^0(x_1)=0, & \varphi_0^0(x_{\frac12})=0, & \varphi_0^{0\prime}(x_{\frac12})=0,\\ \varphi_1^0(x_0)=0, & \varphi_1^0(x_1)=1, & \varphi_1^0(x_{\frac12})=0, & \varphi_1^{0\prime}(x_{\frac12})=0,\\ \varphi_{\frac12}^0(x_0)=0, & \varphi_{\frac12}^0(x_1)=0, & \varphi_{\frac12}^0(x_{\frac12})=1, & \varphi_{\frac12}^{0\,\prime}(x_{\frac12})=0,\\ \varphi_{\frac12}^1(x_0)=0, & \varphi_{\frac12}^1(x_1)=0, & \varphi_{\frac12}^1(x_{\frac12})=0, & \varphi_{\frac12}^{1\,\prime}(x_{\frac12})=1, \end{cases}$$

On a immédiatement (connaissant les racines) :

$$\begin{cases} \varphi_0^0(x)=\dfrac{(x-x_{\frac12})^2(x-x_1)}{(x_0-x_{\frac12})^2(x_0-x_1)},\\ \varphi_1^0(x)=\dfrac{(x-x_{\frac12})^2(x-x_0)}{(x_1-x_{\frac12})^2(x_1-x_0)},\\ \varphi_{\frac12}^0(x)=\dfrac{(x-x_0)(x-x_1)}{(x_{\frac12}-x_0)(x_{\frac12}-x_1)},\\ \varphi_{\frac12}^1(x)=\dfrac{(x-x_0)(x-x_1)(x-x_{\frac12})}{(x_{\frac12}-x_0)(x_{\frac12}-x_1)}, \end{cases}$$

Proposition 1.39 *Ces 4 polynômes forment une base de $\mathcal{P}_3$, et tout polynôme $p\in\mathcal{P}_3$ s'écrit :*

$$p(x)=y_0^0\varphi_0^0(x)+y_1^0\varphi_1^0(x)+y_{\frac12}^0\varphi_{\frac12}^0(x)+y_{\frac12}^1\varphi_{\frac12}^1(x), \tag{1.26}$$

où $y_0=p(x_0)$, $y_1=p(x_1)$, $y_{\frac12}=p(x_{\frac12})$ et $y_{\frac12}^1=p'(x_{\frac12})$.

Preuve. Ils forment une famille libre : si pour tout x :

$$y_0^0\varphi_0^0(x)+y_1^0\varphi_1^0(x)+y_{\frac12}^0\varphi_{\frac12}^0(x)+y_{\frac12}^1\varphi_{\frac12}^1(x)=0,$$

alors prenant (successivement) $x=x_0$, x_1 et $x_{\frac12}$ on trouve (successivement) $y_0^0=0=y_1^0=y_{\frac12}^0$. Et il reste $y_{\frac12}^1\varphi_{\frac12}^1(x)$ pour tout x, d'où $y_{\frac12}^1=0$.

Et ils sont au nombre de $4=\dim(\mathcal{P}_3)$. Donc ils forment une base. Et (1.26) est immédiat car p s'exprime sur cette base et les calculs des composantes sont triviaux. ∎

Exercice 1.40 Démontrer par un calcul direct la proposition 1.38.

Réponse. Soit $p(x)=a_0+a_1x+a_2x^2+a_3x^3$. On a 4 inconnues les $(a_i)_{i=0,\dots,4}$ et 4 équations à satisfaire :

$$\begin{pmatrix} 1 & x_0 & x_0^2 & x_0^3\\ 1 & x_1 & x_1^2 & x_1^3\\ 1 & x_{\frac12} & x_{\frac12}^2 & x_{\frac12}^3\\ 0 & 1 & 2x_{\frac12} & 3x_{\frac12}^2 \end{pmatrix}\begin{pmatrix} a_0\\ a_1\\ a_2\\ a_3 \end{pmatrix}=\begin{pmatrix} y_0\\ y_1\\ y_{\frac12}\\ y_{\frac12}^1 \end{pmatrix}$$

est inversible (pour $x_0\neq x_1$), le déterminant de la matrice notée A étant non nul. En effet :

$$\begin{aligned} \det(A)&=\begin{vmatrix} 1 & x_0 & x_0^2 & x_0^3\\ 0 & x_1-x_0 & x_1^2-x_0^2 & x_1^3-x_0^3\\ 0 & x_{\frac12}-x_0 & x_{\frac12}^2-x_0^2 & x_{\frac12}^3-x_0^3\\ 0 & 1 & 2x_{\frac12} & 3x_{\frac12}^2 \end{vmatrix}=\begin{vmatrix} x_1-x_0 & x_1^2-x_0^2 & x_1^3-x_0^3\\ x_{\frac12}-x_0 & x_{\frac12}^2-x_0^2 & x_{\frac12}^3-x_0^3\\ 1 & 2x_{\frac12} & 3x_{\frac12}^2 \end{vmatrix}\\ &=(x_1-x_0)(x_{\frac12}-x_0)\begin{vmatrix} 1 & x_1+x_0 & x_1^2+x_1x_0+x_0^2\\ 1 & x_{\frac12}+x_0 & x_{\frac12}^2+x_{\frac12}x_0+x_0^2\\ 1 & 2x_{\frac12} & 3x_{\frac12}^2 \end{vmatrix}\\ &=(x_1-x_0)(x_{\frac12}-x_0)\begin{vmatrix} 1 & x_1+x_0 & x_1^2+x_1x_0+x_0^2\\ 0 & x_{\frac12}-x_1 & x_{\frac12}^2-x_1^2+x_{\frac12}x_0-x_1x_0\\ 1 & 2x_{\frac12} & 3x_{\frac12}^2 \end{vmatrix}\\ &=(x_1-x_0)(x_{\frac12}-x_0)(x_{\frac12}-x_1)\begin{vmatrix} 1 & x_1+x_0 & x_1^2+x_1x_0+x_0^2\\ 0 & 1 & x_1+x_{\frac12}+x_0\\ 1 & 2x_{\frac12} & 3x_{\frac12}^2 \end{vmatrix}\\ &=(x_1-x_0)(x_{\frac12}-x_0)(x_{\frac12}-x_1)(x_{\frac12}^2+x_0x_1-x_1x_{\frac12}-x_{\frac12}x_0)\\ &=\frac14(x_0-x_1)^3(x_{\frac12}-x_0)(x_{\frac12}-x_1)=\frac{1}{16}(x_1-x_0)^5, \end{aligned}$$

sachant $x_{\frac{1}{2}} = \frac{x_0+x_1}{2}$. D'où $\det A \neq 0$. D'où A inversible, d'où l'existence de l'unicité. ∎

1.6.4 Cas général

Soit k+1 points $(x_i)_{i=0,...,k}$ et k+1 entiers $(\alpha_i)_{i=0,...,k}$. On se donne $n = k+\alpha_0+...+\alpha_k$ valeurs $y_i^j = f^{(j)}(x_i)$ pour $0 \leq j \leq \alpha_j$.

Proposition 1.41 *Il existe un unique polynôme p_n de degré n tel que :*

$$\forall (i,j) \in \{0 \leq i \leq k; 0 \leq j \leq \alpha_i\}, \quad p_n^{(j)}(x_i) = y_i^j, \tag{1.27}$$

et ce polynôme est appelé polynôme d'interpolation de Hermite.

Preuve. On cherche p_n sous la forme $p_n(x) = \sum_{m=0}^n a_m x^m$, et on pose $\vec{a} = (a_0, ..., a_n)$. Les équations (1.27) forment un système de n+1 équations à n+1 inconnues de la forme $A.\vec{a} = \vec{b}$ où le vecteur $\vec{b}$ est formé des y_i^j. On aura donc existence et unicité de la solution $\vec{a}$ dès que $A.\vec{a} = 0$ a pour unique solution $\vec{a} = 0$. Mais ce cas correspond à $y_i^j = 0$ pour tout i, j et au polynôme p_n tel que chaque x_i est racine de multiplicité α_i+1. I.e. p_n est un polynôme de degré n qui a n+1 racines. Donc $p_n = 0$. ∎

Remarque 1.42 D'un point de vue pratique, il est intéressant de construire des fonctions polynômes de degré n de base :

À i, $0 \leq i \leq k$, on se donne j_i, $0 \leq j_i \leq \alpha_i$. Au polynôme $\varphi_i^{j_i}$ de degré n, on demande de satisfaire les n conditions : $(\varphi_i^{j_i})^{(m_i)}(x_\ell) = \delta_{i\ell}\delta_{j_i m_i}$ pour $0 \leq \ell \leq k$ et $0 \leq m_i \leq \alpha_{j_i}$.

I.e., $(\varphi_i^{j_i})^{(m_i)}(x_\ell) = 0$ pour $0 \leq \ell \leq k$ et $0 \leq m \leq \alpha_j$ si $\ell \neq i$ et $m \neq j$, et $(\varphi_i^{j_i})^{(j_i)}(x_i) = 1$.

On montre immédiatement que ces polynômes $(\varphi_i^{j_i})$ forment une base de $\mathcal{P}_n$, puisqu'ils sont en nombre n+1 et qu'ils forment une famille libre : si :

$$0 = \sum_{i=0}^k \sum_{m=0}^{\alpha_i} z_i^m \, \varphi_i^m(x),$$

alors en particulier en x_ℓ, la dérivée d'ordre j donne $0 = z_\ell^j$. Donc, tout polynôme de degré n s'écrit :

$$p_n(x) = \sum_{i=0}^k \sum_{m=0}^{\alpha_i} y_i^j \, \varphi_i^m(x).$$

On renvoie à Crouzeix et Mignot [5] pour les expressions explicites des $\varphi_i^{j_i}$. ∎

1.7 Points équidistants : les oscillations

Il peut sembler naturel (simple) de reconstruire une fonction f connaissant ses valeurs en n+1 points (x_i, y_i) tels que les x_i soient équidistants : on se donne h>0 et :

$$x_i = x_0 + ih, \quad i = 0, ..., n.$$

Malheureusement, c'est en général une mauvaise idée lorsque n est assez grand.

Exemple : on veut faire passer un polynôme par les n points $(x_i, y_i = \frac{1}{1+x_i^2})$ où les $x_i \in [-5, +5]$: la fonction cherchée étant bien sûr la fraction rationnelle $f(x) = \frac{1}{1+x^2}$ (l'exemple est classique, pris par exemple dans Théodor [15]).

L'exemple précédent montre que, dans le cas des points équidistants, l'erreur :

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

peut être très grande. D'où les idées :

0- Prendre très peu de points (équidistants) : mais ce n'est guère satisfaisant, car on ne peut obtenir une bonne approximation.

1- Ne pas prendre les points équidistants : les choisir de telle sorte que $||f - p_n||_\infty$ soit minimale. Voir plus loin les polynômes de Chebyshev qui donneront une 'bonne approximation' du polynôme optimal cherché.

2- Ne pas prendre la norme $||.||_\infty$ comme instrument de mesure de l'erreur. On construit par exemple le polynôme de meilleure approximation au sens des moindres carrés : on trouvera ainsi de nouveaux points de collocation (x_i), voir par exemple la méthode de Gauss.

3- Partager l'intervalle $[a, b]$ en sous intervalles, cf. (1.1), utiliser un polynôme d'approximation avec peu de points sur chaque sous intervalle. La fonction approchée est alors une fonction polynomiale par morceaux. C'est l'optique des splines ou des méthodes type méthode des éléments finis.

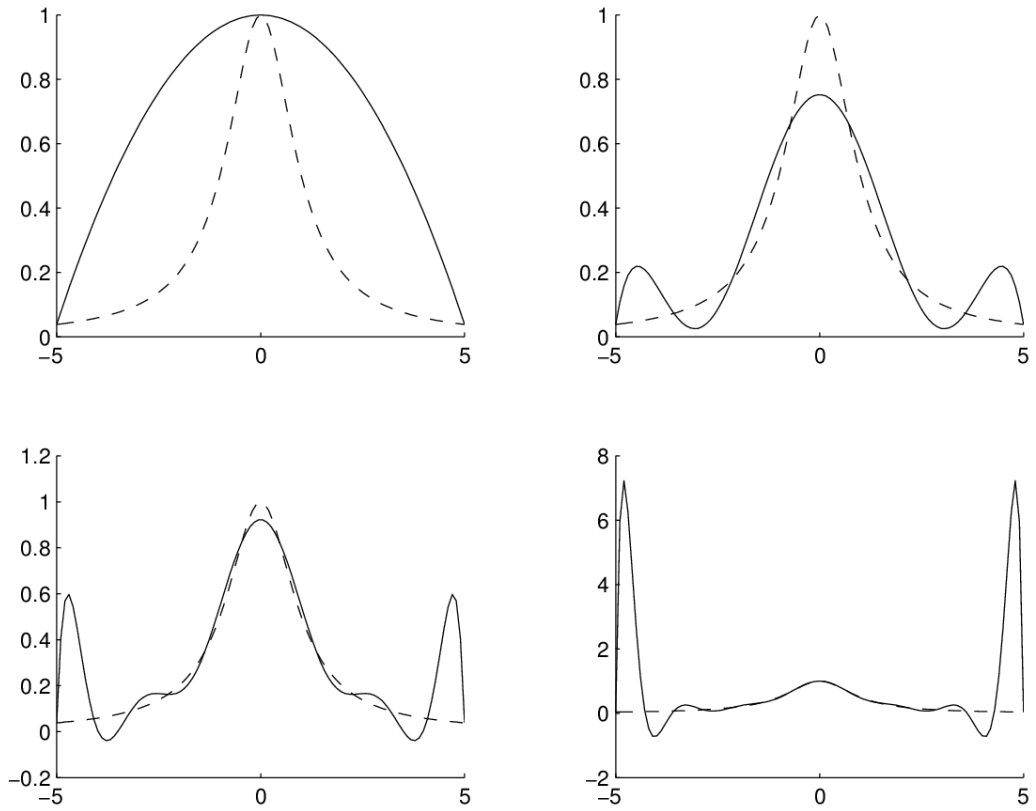

FIGURE 1.1 – Sur l'intervalle $[a,b] = [-5,+5]$. La fonction $y=\frac{1}{1+x^2}$ est en pointillés. Et les polynômes de Lagrange correspondant sont en trait plein, respectivement pour $n = 3$, 8, 12, puis 15 points équi-répartis : donc ici $x_i = -5+ih$, où $h = \frac{b-a}{n}$, et $y_i = \frac{1}{1+x_i^2}$, pour $i = 0, ..., n-1$, . Les polynômes de Lagrange ont tendance à osciller très fortement aux extrémités (phénomène de Runge) pour n grand au voisinage du bord.

Remarque 1.43 La recherche des points (éventuellement non équidistants) en lesquels il est intéressant d'évaluer f sera d'un intérêt pratique immédiat lorsqu'il s'agira par exemple d'évaluer $\int_a^b f(x)\,dx$ numériquement à l'aide d'une formule approchée du type $\sum_{i=0}^n \tilde{f}(x_i)w_i$, où $\tilde{f}$ est une évaluation de f. Voir le paragraphe sur l'intégration numérique. ∎

1.8 * Remarque "négative"

Les résultats présentés ici sont pris principalement dans Crouzeix et Mignot [5]. Pour des fonctions $f \in C^0([a,b])$ données, que l'on veut approximer par des polynômes de degré n, on se donne $n+1$ points distincts $(x_i^n)_{i=0,...,n}$ dans l'intervalle $[a,b]$. Et on note :

$$L_n : \begin{cases} C^0([a,b]) \to \mathcal{P}_n \\ \qquad f \mapsto L_n(f) = p_n \quad \text{polynôme d'interpolation de Lagrange de degré } n, \end{cases} \tag{1.28}$$

où donc p_n est défini par $p_n(x_i^n) = f(x_i^n)$ pour tout $i = 0, ..., n$. (Plus généralement, on peut choisir l'interpolation par les polynômes d'Hermite). On a alors le résultat défavorable :

> "Quels que soient les points d'interpolation de Lagrange choisis, il existe une fonction continue $f \in C^0([a,b])$ telle que $L_n(f)$ ne converge pas uniformément vers f quand $n \to \infty$."

(I.e. : pour chaque n on se donne un ensemble $B_n = (x_i)_{i=0,...,n-1}$, et on dispose ainsi d'une suite $(B_n)_{\mathbb{N}}$. Alors $\exists f \in C^0([a,b])$ telle que $||L_n f - f||_\infty \not\to_{n\to\infty} 0$.)

Remarque 1.44 Attention : ça ne veut pas dire que le théorème de Weierstrass 1.11 est faux. Mais qu'il faut faire attention quand on emploie l'approximation de Lagrange.

D'ailleurs, noter que le polynôme $\tilde{p}_n$ de "meilleure approximation", i.e. celui de degré n qui réalise $||f-\tilde{p}_n||_\infty \leq ||f-p||_\infty$ pour tout $p \in \mathcal{P}_n$ (celui qui réalise le minimum au sens de la norme $||.||_\infty$), n'est pas nécessairement un polynôme "de collocation", i.e. un polynôme qui aux points x_i prend les valeurs $y_i = f(x_i)$: le polynôme de meilleure approximation (au sens de la norme $||.||_\infty$) pourra prendre en les points x_i des valeurs $y_i \neq f(x_i)$. Voir figure 1.1. ∎

Lemme 1.45 *L'application L_n définie en (1.28) est linéaire.*

Preuve. Si f et g sont continues sur $[a,b]$, et si $\alpha \in \mathbb{R}$, alors le polynôme $L_n(f+\alpha g)$ prend les valeurs $L_n(f+\alpha g)(x_i) = (f+\alpha g)(x_i) = f(x_i) + \alpha g(x_i) = L_n(f)(x_i)+\alpha L_n(g)(x_i)$ aux points $n+1$ points de collocations x_i, et donc les polynômes $L_n(f+\alpha g)$ et $L_n(f)+\alpha L_n(g)$ sont égaux. ∎

L_n étant linéaire, on pose :

$$|||L_n||| = \sup_{g \in C^0([a,b])} \frac{||L_n g||_\infty}{||g||_\infty},$$

quantité qui ne dépend que des points de collocation au travers de L_n.

Proposition 1.46

$$|||L_n||| \underset{n\to\infty}{\longrightarrow} \infty.$$

Preuve. Voir Crouzeix et Mignot [5]. ∎

Et on pose :

$$E_n(t) = \inf_{q_n \in \mathcal{P}_n} ||f - q_n||_\infty,$$

quantité qui ne dépend que de n et n'a rien à voir avec les points de collocation (l'inf est réalisé par le polynôme dit de meilleur approximation).

Proposition 1.47 *Pour $f \in C^0([a,b])$ et $n \in \mathbb{N}$ fixé, on a (pour les polynômes de collocation) :*

$$||f - L_n(f)||_\infty \leq (1 + |||L_n|||)E_n(t), \tag{1.29}$$

Preuve. Voir Crouzeix et Mignot [5]. ∎

La plus faible erreur $||f - L_n(f)||_\infty$ que cette estimation d'erreur permet d'obtenir, est celle qui correspond aux choix des points de collocation qui minimisent $|||L_n|||$. On montre que les points $x_0^n, ..., x_n^n$ qui réalise le minimum de $|||L_n|||$ permettent d'obtenir :

$$|||L_n||| \simeq \frac{2}{\pi}\log(n) \quad \text{quand } n \to \infty.$$

Par contre, ces points sont difficiles à calculer.

On démontre alors également que pour les points de collocation x_i^n correspondant aux racines du polynôme de Chebyshev de degré n (paragraphe suivant), on a également :

$$|||L_n||| \simeq \frac{2}{\pi}\log(n) \quad \text{quand } n \to \infty.$$

Si on choisit comme points de collocation les points de Chebychev, qui sont très simples à calculer car égaux à $x_i^n = \frac{a+b}{2} + \frac{b-a}{2}\cos(\frac{2i+1}{2n+2}\pi)$, on a donc une approximation qui n'est pas loin d'être optimale. Cela explique pourquoi l'usage des points de Chebyshev est très vivement conseillé.

Par contre, si on choisit des points de collocation équidistants, on a :

$$|||L_n||| \simeq 2^{n+1}\frac{1}{e\, n \log n} \quad \text{quand } n \to \infty,$$

ce qui est plutôt mauvais ! D'où les résultats décevants comme les instabilités remarquées au voisinage des extrémités a et de b dans la figure 1.1.

1.9 * Erreur d'interpolation pour la norme $||.||_\infty$

Soit f une fonction $C^{n+1}([a,b];\mathbb{R})$, soient $(x_i)_{i=0,...,n}$ $n+1$ points 2 à 2 distincts dans l'intervalle $[a,b]$, et soit :

$$p_n = \sum_{i=0}^{n} f(x_i)L_i, \tag{1.30}$$

le polynôme d'interpolation de Lagrange de degré n (le polynôme de degré n qui vérifie $p_n(x_i) = f(x_i)$ pour tout i).

On suppose pour simplifier que $[a,b] = [\min_{i=0,...,n}(x_i), \max_{i=0,...,n}(x_i)]$. Et on note :

$$\pi_{n+1}(x) = \prod_{i=0}^{n}(x - x_i) \tag{1.31}$$

le polynôme de degré $n+1$ normalisé (le coefficient devant x^{n+1} est 1) dont les racines sont les x_i.

Alors, si $x \in [a,b]$, l'erreur ponctuelle $f(x) - p_n(x)$ (en chaque point x) est donnée par :

Théorème 1.48 *Si $f \in C^{n+1}([a,b];\mathbb{R})$, alors pour tout $x \in [a,b]$ il existe $\xi_x \in [a,b]$ tel que :*

$$f(x) - p_n(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!}\,\pi_{n+1}(x). \tag{1.32}$$

En particulier :

$$|f(x) - p_n(x)| \leq \frac{||f^{(n+1)}||_\infty}{(n+1)!}\,|\pi_{n+1}(x)|, \tag{1.33}$$

où "$||g||_\infty = \sup_{x\in[a,b]} |g(x)|$" (notation générique). Et l'erreur maximale est donc :

$$||f - p_n||_\infty \leq \frac{||f^{(n+1)}||_\infty}{(n+1)!}\,||\pi_{n+1}||_\infty. \tag{1.34}$$

(En particulier il sera intéressant de choisir les points x_i tels que $||\pi_{n+1}||_\infty$ soit le plus petit possible : voir les polynômes de Chebychev et le choix des x_i rendant $||\pi_{n+1}||_\infty$ petit.)

Preuve. Si f est un polynôme de degré $n+1$ alors la relation (1.32) est immédiate car $f^{(n+1)} =$ constante, $\pi_{n+1}^{(n+1)} = (n+1)!$ et $p_n^{(n+1)} = 0$.

Dans le cas général (f non polynomiale), le théorème généralise ce cas polynomial. Montrons-le.

S'il existe $i\in[0,n]$ tel que $x=x_i$ alors la relation est triviale : les deux membres sont nuls.

Supposons donc que pour tout i, $x\neq x_i$.

Établissons la formule (1.32) dans le cas $n=1$: cas où $f \in C^2([a,b];\mathbb{R})$, où $i = 0, 1$, et où $p_1(x) = f(x_0) + \frac{f(x_1)-f(x_0)}{x_1-x_0}(x-x_0)$, cf (1.12), (1.13) et (1.14). En particulier $p_1(x_0) = f(x_0)$ et $p_1(x_1) = f(x_1)$ (polynôme de Lagrange de f).

Fixons $x \in]a,b[$ avec $x \neq x_0$ et $x \neq x_1$ (ici x est fixé : ce n'est pas une variable). Définissons la valeur $\delta(x) \in \mathbb{R}$ telle que :

$$f(x) = p_1(x) + \delta(x)\pi_2(x) \qquad (= p_1(x) + \delta(x)(x-x_0)(x-x_1)),$$

i.e. posons, pour $x \neq x_0, x_1$:

$$\delta(x) = \frac{f(x) - p_1(x)}{(x-x_0)(x-x_1)}.$$

x étant fixé dans $]a,b[$, soit la fonction $R_x : [x_0,x_1] \to \mathbb{R}$ (reste) définie par :

$$R_x(y) = f(y) - p_1(y) - \delta(x)\pi_2(y).$$

Cette fonction vérifie $R_x(x_0) = 0 = R_x(x_1) = R_x(x)$, et R_x s'annulant en les 3 points x_0, x_1, x, le théorème de Rolle généralisé nous dit qu'il existe un point $\xi_x \in]\min(x_0,x_1,x), \max(x_0,x_1,x)[$ tel que $R_x''(\xi_x) = 0$, et donc, avec $p_1'' = 0$:

$$0 = R_x''(\xi_x) = f''(\xi_x) - 0 - 2\delta(x),$$

On a donc $\delta(x) = \frac{1}{2}f''(\xi_x)$ et donc :

$$R_x(y) = f(y) - p_1(y) - \frac{f''(\xi_x)}{2}(y-x_0)(y-x_1).$$

Comme $R_x(x) = 0$, on a (1.32).

Dans le cas général, à x fixé, on cherche $\delta(x) \in \mathbb{R}$ tel que

$$f(x) = p_n(x) + \delta(x)\pi_{n+1}(x).$$

La fonction (le reste) :

$$R_x(y) = f(y) - p_n(y) - \delta(x)\pi_{n+1}(y)$$

s'annule aux $n+2$ points $x_0, ..., x_n, x$, et le théorème de Rolle généralisé permet de conclure : il existe un point $\xi_x \in]\min(x_i,x), \max(x_i,x)[$ tel que $R_x^{n+1}(\xi_x) = 0$. Avec $\pi^{(n+1)}(x) = (n+1)!$ (constante), on obtient $f^{n+1}(\xi_x) = (n+1)!\,\delta(x)$, soit $\delta(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!}$, d'où (1.32). ∎

1.10 Polynômes de Chebyshev

Noter que Chebyshev (1821-1894) est également écrit Tchebycheff ou Tchebytcheff ou...

Le théorème 1.48 équation (1.34) indique que si on veut minimiser l'erreur (pour écrire a priori un "bon" code de calcul pour trouver un "bon" polynôme d'interpolation d'une fonction f quelconque), on a intérêt à ce que le polynôme π_{n+1} vérifie $||\pi_{n+1}||_\infty$ " petit" (i.e. $|\pi_{n+1}(x)|$ petit pour tout $x \in [a,b]$). Cela revient donc à choisir les points (x_i) d'interpolation en conséquence. Quitte à changer de repère, on prend $[a,b] = [-1,1]$, un bon choix pour x_i consistant à prendre les racines du n-ième polynôme de Chebyshev T_n.

Rappel : $\cos \left\{ \begin{array}{rl} [0,\pi] & \to [-1,1] \\ \theta & \to x = \cos(\theta) \end{array} \right\}$ est bijective d'inverse arccos : $\left\{ \begin{array}{rl} [-1,1] & \to [0,\pi] \\ x & \to \theta = \arccos(x) = \cos^{-1}(x) \end{array} \right\}$,
voir figure 1.2. Et pour $x \in]-1,1[$, la dérivée de arccos est donnée par, avec $\cos\theta = x$:

$$((\cos^{-1})'(x) =) \quad \arccos' x = -\frac{1}{\sqrt{1-x^2}} = -\frac{1}{\sin\theta} \qquad (= \frac{1}{\cos'\theta}), \tag{1.35}$$

puisque, par dérivation de la composée $(\cos^{-1} \circ \cos)(\theta) = \theta$ on obtient $(\cos^{-1})'(\cos\theta).\cos'\theta = 1$.

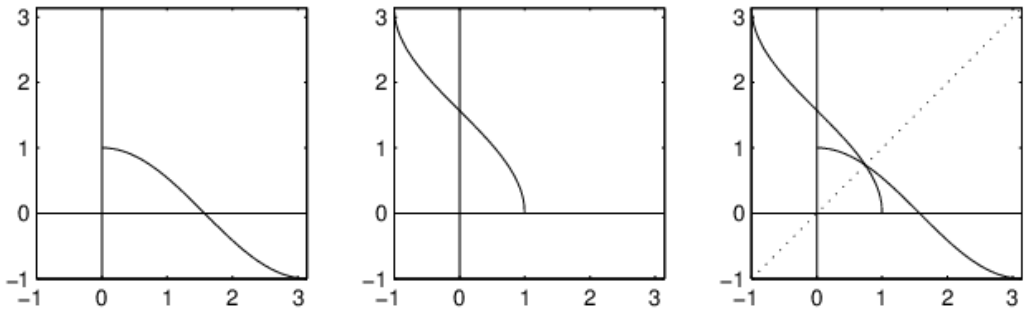

FIGURE 1.2 – fonctions $\cos : \theta \in [-\pi,\pi] \to x = \cos\theta \in [-1,1]$ (premier dessin), $\cos^{-1} = \arccos : x \in [-1,1] \to \arccos(x) = \theta \in [0,\pi]$ (second dessin), et les deux figures sur le même dessin (troisième dessin).

Définition 1.49 Le n-ième polynôme de Chebyshev est donné sur $[-1,1]$ par, pour $\theta \in [0,\pi]$:

$$T_n(\cos\theta) = \cos(n\theta), \tag{1.36}$$

i.e., pour $x \in [-1,1]$ et $\theta = \arccos(x) \in [0,\pi]$, on a $T_n(x) = \cos(n(\arccos x))$.

Proposition 1.50 *La fonction $T_n : [-1,1] \to \mathbb{R}$ est un polynôme de degré n : avec $E(\frac{n}{2})$ la partie entière de $\frac{n}{2}$,*

$$\begin{aligned} T_n(x) &= \sum_{m=0}^{E(\frac{n}{2})} (-1)^m C_n^{2m} x^{n-2m} (1-x^2)^m \\ &= x^n - C_n^2 x^{n-2}(1-x^2) + C_n^4 x^{n-4}(1-x^2)^2 - ... \end{aligned} \tag{1.37}$$

Preuve. On a $\cos(n\theta) = \Re(e^{in\theta})$ (partie réelle), avec $x = \cos\theta$ et (formule de Moivre) :

$$\cos(n\theta) + i\sin(n\theta) = e^{in\theta} = (e^{i\theta})^n = (\cos\theta + i\sin\theta)^n = (x + i\sqrt{1-x^2})^n,$$

d'où :

$$\cos(n\theta) = \Re(\sum_{k=0}^{n} C_n^k x^{n-k} i^k ((1-x^2)^{\frac{1}{2}})^k) = \sum_{\substack{k=0 \\ k \text{ pair}}}^{n} (-1)^{\frac{k}{2}} C_n^k x^{n-k} (1-x^2)^{\frac{k}{2}},$$

où $C_k^k = \binom{n}{k} = \frac{n!}{k!(n-k)!}$, puis on pose $m = 2k$. ∎

Exemple 1.51 $T_0(x) = 1$, $T_1(x) = x$, $T_2(x) = 2x^2 - 1$, $T_3(x) = 4x^3 - 3x$, $T_4(x) = 8x^4 - 8x^2 + 1$, $T_5(x) = 16x^5 - 20x^3 + 5x$ ... ∎

Exercice 1.52 Montrer que le coefficient devant le monôme x^n de T_n est 2^{n-1}.

Réponse. Il s'agit de montrer, cf. (1.37), que $1 + C_n^2 + C_n^4 + ... = 2^{n-1}$.

On a $(x+y)^n + (x-y)^n = 2 \sum_{\substack{k=0 \\ k \text{ pair}}}^{n} C_n^k x^{n-k} y^k$. On prend x=y=1. ∎

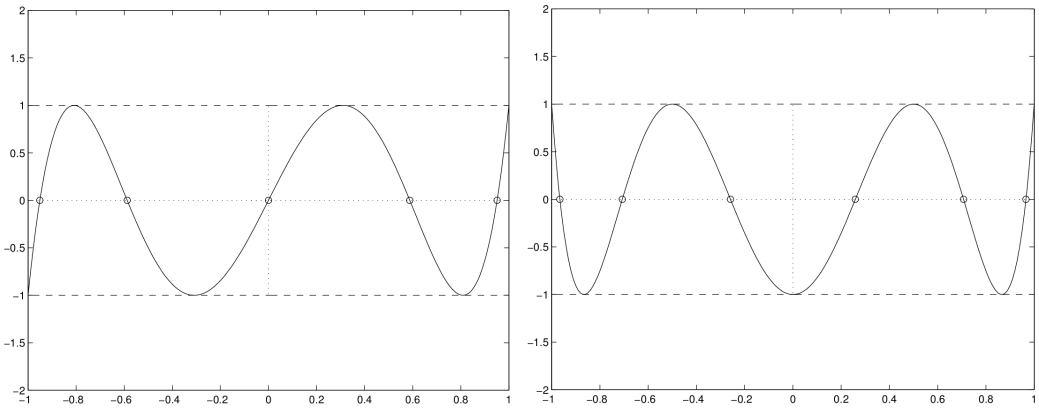

FIGURE 1.3 – Polynômes T_5 et T_6 représentés sur l'intervalle $[-1,1]$.

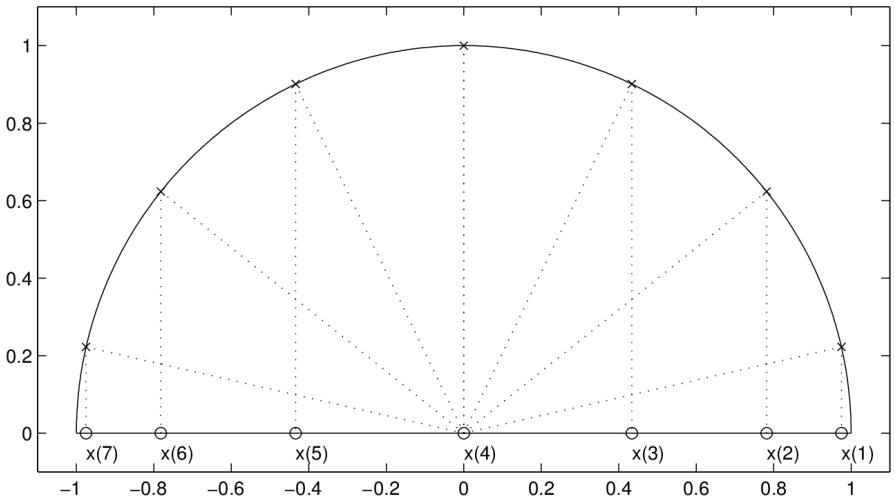

FIGURE 1.4 – Cas $n = 7$: racines de T_n dans l'intervalle $[-1,1]$: $x_k = \Re(e^{i\theta_k})$, cf. (1.41). En particulier les racines de T_n ne sont pas equiréparties : elles se "concentrent" vers les bords -1 et $+1$.

Proposition 1.53 *1- Pour tout $x \in [-1,1]$ on a $|T_n(x)| \leq 1$.*

2- Pour n pair T_n est pair, et pour n impair T_n est impair.

3- Pour tout n on a :

$$T_n(1) = 1, \quad T_n(-1) = (-1)^n, \quad et \quad T_{2n}(0) = (-1)^n, \quad T_{2n+1}(0) = 0. \tag{1.38}$$

4- Pour $x = \cos\theta \in]-1,1[$ et $n \neq 0$:

$$T_n'(x) = \frac{n \sin(n\theta)}{\sin\theta}, \tag{1.39}$$

et en particulier :

$$T_n'(1) = n^2, \quad T_n'(-1) = (-1)^n n^2, et \quad T_{2n}'(0) = 0, \quad T_{2n+1}'(0) = (-1)^n(2n+1). \tag{1.40}$$

5- T_n à n racines simples aux n points :

$$x_k = \cos(\theta_k), \quad \theta_k = \frac{2k-1}{2n}\pi, \quad k = 1, ..., n, \tag{1.41}$$

voir figure 1.4.

6- Dans l'intervalle $[-1,1]$, T_n a ses extrema aux $n+1$ points :

$$\tilde{x}_k = \cos(\frac{k}{n}\pi), \quad avec \quad T_n(\tilde{x}_k) = (-1)^k, \quad k=0, ..., n, \tag{1.42}$$

i.e. T_n vaut alternativement 1 et -1 aux extrema $\tilde{x}_k$ (donnés en ordre décroissant).

Preuve. 1- (1.36) donne directement $|T_n(x)| \leq 1$.

2- Avec $x = \cos\theta = -\cos(\theta+\pi)$, on a $T_n(-x) = T_n(\cos(\theta+\pi)) = \cos(n(\theta+\pi)) = \cos(n\theta+n\pi) = (-1)^n\cos(n\theta) = (-1)^n T_n(\cos\theta) = (-1)^n T_n(x)$.

3- $T_n(1) = \cos(n(\arccos 1)) = \cos n0 = \cos 0 = 1$, puis $T_n(-1)$ par parité ou imparité. Puis T_{2n+1} est impaire donc $T_{2n+1}(0) = 0$. Puis pour $x = 0$ on a $\theta = \frac{\pi}{2}$, d'où $T_{2n}(0) = \cos(2n\frac{\pi}{2}) = \cos(n\pi) = (-1)^n$.

4- Puis, posant $x(\theta) = \cos\theta$, on a $T_n(x(\theta)) = \cos(n\theta)$ qui donne par dérivation de fonctions composée $T_n'(x(\theta))x'(\theta) = -n\sin(n\theta)$, d'où (1.39) pour $x \in]-1,1[$.

Puis pour $x = 1$ on a $\theta = 0$ avec $\sin\alpha \sim \alpha$ au voisinage de $\alpha = 0$, donc $\frac{\sin(n\theta)}{\sin\theta} \sim \frac{n\theta}{\theta} = n$ au voisinage de $\theta = 0$. Et pour $x = -1$ on se sert de la parité ou imparité de T_n.

Comme T_{2n} est pair on a $T_{2n}'(0) = 0$, et pour $x = 0$ on a $\theta = \frac{\pi}{2}$, donc $T_{2n+1}'(0) = (2n+1)\sin((2n+1)\frac{\pi}{2}) = (2n+1)\sin(\frac{\pi}{2}+n\pi) = (-1)^n(2n+1)$.

5- On vérifie que $T_n(\cos(\frac{2k-1}{2n}\pi)) = \cos(\frac{2k-1}{2}\pi) = 0$, d'où les n racines simples. Comme T_n est de degré n et est non nul, ce sont les seules, et elles sont classées dans l'ordre décroissant (car cosinus est une fonction décroissante sur $[0,\pi]$).

6- Puis $T_n \in [-1,1]$ et $|T_n(\pm 1)| = 1$ donnent : en -1 et $+1$ T_n est extrémale.

Puis $T_n'(x) = 0$ pour $x \in]-1,1[$ ssi, cf. (1.39), $\sin(n\theta) = 0$ quand $x = \cos\theta$ et $\theta \in]0,\pi[$, i.e ssi $\theta = k\frac{\pi}{n}$ pour $k \in \mathbb{Z}$ et $\theta \in]0,\pi[$, i.e. ssi $\theta = k\frac{\pi}{n}$ pour $k = 1, ..., n-1$.

Puis $T_n(\tilde{x}_k) = \cos(n(\frac{k}{n}\pi)) = \cos(k\pi) = (-1)^k$. ▪

Exercice 1.54 Montrer la relation de récurrence, pour tout $n \in \mathbb{N}^*$ et $x \in [-1,1]$:

$$T_{n+1}(x) + T_{n-1}(x) = 2xT_n(x), \quad \text{noté} \quad T_{n+1} + T_{n-1} = 2xT_n \tag{1.43}$$

Réponse. C'est $\cos((n+1)\theta) + \cos((n-1)\theta) = 2\cos(n\theta)\cos(\theta)$, où on a posé $\theta = \arccos x$. ▪

Exercice 1.55 Soit $w :]-1,1[\to \mathbb{R}$ définie par $w(x) = \frac{1}{\sqrt{1-x^2}}$. Soit L^2_w l'ensemble des fonctions $f :]-1,1[\to \mathbb{R}$ telles que $\int_{-1}^1 f(x)^2 \frac{dx}{\sqrt{1-x^2}} < \infty$. On considère le produit scalaire sur L^2_w :

$$(f,g)_w = \int_{-1}^1 f(x)g(x)\,\frac{dx}{\sqrt{1-x^2}}. \tag{1.44}$$

1- Vérifier que $w : x \to w(x) = \frac{1}{\sqrt{1-x^2}}$ est intégrable dans $]-1,1[$.

2- Vérifier que $(\cdot,\cdot)_w$ est bien un produit scalaire dans L^2_w.

3- Montrer que les polynômes de Chebychev forment une base orthogonale dans l'ensemble des polynômes muni du produit scalaire $(\cdot,\cdot)_w$, et donner en une base orthonormale.

Réponse. 1- Comme $\sqrt{1-x^2} = \sqrt{1-x}\sqrt{1+x}$ et que $\frac{1}{\sqrt{x}} = x^{-\frac{1}{2}}$ est intégrable en 0+ (de primitive $2\sqrt{x} = 2x^{\frac{1}{2}}$), la fonction w est intégrable à gauche en 1 et à droite en -1. De plus la fonction w est continue dans $]-1,1[$: donc elle est intégrable sur $]-1,1[$.

2- Bilinéarité, symétrie et positivité immédiates. Définie positif : si $(f,f)_w = 0$, comme $\sqrt{1-x^2} > 0$ et $f^2(\vec{x}) \geq 0$ sur $]-1,1[$, nécessairement $f = 0$ sur $]-1,1[$ (au moins presque partout, voir cours d'intégration).

3- Le changement de variable $x = \cos\theta$ donnant $dx = -\sin\theta\,d\theta$ (avec $\sin\theta \geq 0$ sur $[0,\pi]$) on obtient :

$$(T_n,T_m)_w = \int_{x=-1}^1 T_n(x)T_m(x)\,\frac{dx}{\sqrt{1-x^2}} = \int_{\theta=0}^{\pi} \cos(n\theta)\cos(m\theta)\,d\theta.$$

Et $2\cos(n\theta)\cos(m\theta) = \cos((n+m)\theta)+\cos((n-m)\theta)$ donne $(T_n,T_m)_w = \begin{cases} 0 \text{ si } n \neq m, \\ \pi \text{ si } n = m = 0, \\ \frac{\pi}{2} \text{ si } n = m \neq 0. \end{cases}$ Donc la famille $(T_n)_{n=0,...,N}$ est une base orthogonale de $(\mathcal{P}_N([-1,1]),(\cdot,\cdot)_w)$, une base orthonormale étant donnée par $(\sqrt{\frac{1}{\pi}}\,T_0, (\sqrt{\frac{2}{\pi}}\,T_n)_{n=1,...,N})$. ▪

Exercice 1.56 Montrer :

$$T_n(x) = \frac{\sqrt{\pi}}{\Gamma(n+\frac{1}{2})}\frac{(-1)^n}{2^n}\sqrt{1+x^2}\,\frac{d^n}{dx^n}(\frac{(1-x^2)^n}{\sqrt{1+x^2}}),$$

où $\Gamma(1) = 1$ et $\Gamma(n+1) = n!$ et où :

$$\Gamma(\frac{1}{2}) = \sqrt{\pi} \quad \text{et} \quad \Gamma(n+\frac{1}{2}) = \frac{1.3.5...(2n-1)}{2^n}\Gamma(\frac{1}{2}).$$

Réponse. Vérifiez-le pour T_0 et T_1, et vérifier la relation de récurrence (1.43). ▪

Définition 1.57 On appelle polynôme de Chebyshev normalisé le polynôme :

$$\bar{T}_n = \frac{T_n}{2^{n-1}}.$$

(Le coefficient du terme x^n de plus haut degré vaut 1.)

En particulier, avec (1.42) on a $\sup_{x\in[-1,1]} |\bar{T}_n(x)| = \frac{1}{2^{n-1}}$.

Proposition 1.58 *Quel que soit le polynôme z_n normalisé de degré n (le coefficient devant x^n est 1), on a sur $[-1,1]$:*

$$||z_n||_\infty \geq ||\bar{T}_n||_\infty \qquad (= \frac{1}{2^{n-1}}),$$

i.e. $\sup_{x\in[-1,1]} |z_n(x)| \geq \sup_{x\in[-1,1]} |\bar{T}_n(x)|$ $(= \frac{1}{2^{n-1}})$.

Et $||z_n||_\infty = ||\bar{T}_n||_\infty$ sur $[-1,1]$ si et seulement si $z_n = \bar{T}_n$: le polynôme de Chebyshev normalisé $\bar{T}_n$ est le plus petit de tous les polynômes normalisés de degré n, au sens de $L^\infty([-1,1];\mathbb{R})$.

D'où, si les x_i sont les racines du polynômes de Chebyshev, l'erreur (1.34) donne sur $[-1,1]$:

$$||f - p_n||_\infty \leq \frac{1}{2^{n-1}} \frac{||f^{(n+1)}||_\infty}{(n+1)!},$$

où p_n est le polynôme d'interpolation de Lagrange de f aux point x_i de Chebyshev.

(Les racines des polynômes de Chebyshev sont des points de collocation qui permettent de réaliser une "bonne" approximation de Lagrange de f.)

Preuve. Supposons que $\sup_{x\in[-1,1]} |z_n(x)| < \frac{1}{2^{n-1}}$ $(= \sup_{x\in[-1,1]} |\bar{T}_n(x)|)$. Le polynôme z_n étant normalisé, $q = \bar{T}_n - z_n$ qui est de degré $n-1$ (les termes de degré n s'annulent). Et, les $\tilde{x}_k$ étant les abscisses où $T_n(\tilde{x}_k) = (-1)^k$, cf. (1.42), on a $q(\tilde{x}_k) = \bar{T}_n(\tilde{x}_k) - z_n(\tilde{x}_k) = (-1)^k \frac{1}{2^{n-1}} - z_n(\tilde{x}_k)$ change $n+1$ fois de signe avec k. Et a donc n racines au moins. Mais q étant de degré $\leq n-1$, on a alors $q = 0$. Donc $z_n = \bar{T}_n$. C'est absurde avec l'hypothèse $\sup_{x\in[-1,1]} |z_n(x)| < \frac{1}{2^{n-1}}$. Donc si $\sup_{x\in[-1,1]} |z_n(x)| = \frac{1}{2^{n-1}}$ alors $z_n = T_n$. ∎

Remarque 1.59 Noter que pour une fonction f particulière, ce n'est pas le choix des racines du polynôme de Chebychev qui donne $||f - p_n||_\infty$ réalise le minimum des $||f - q_n||_\infty$ pour q_n polynôme de degré n : le choix des x_i racines du polynôme de Chebyshev T_n nous assure seulement qu'on a une bonne approximation de f par son polynôme de Lagrange aux points x_i.

Pour avoir "le meilleur" polynôme q_n, celui qui réalise le minimum des $||f - q_n||_\infty$, on peut utiliser l'algorithme de Rémès ; et les racines (x_i) du polynôme de Chebychev permettent d'initialiser cet algorithme. Voir Crouzeix et Mignot [5]. ∎

Remarque 1.60 Si on souhaite se placer sur un intervalle $[a,b] \ni t$, on fait le changement de variable $x = \frac{2t-b-a}{b-a} \in [-1,1]$ pour se retrouver dans $[-1,1]$. ∎

1.11 * Moindres carrés

Pour les expressions explicites, ou les valeurs numériques, on renvoie à Abramowitz et Stegun [1].

On rappelle que si $f : [a,b] \to K$ avec $K = \mathbb{R}$ ou $\mathbb{C}$, alors f^2 est la fonction $[a,b] \to K$ définie par $f^2(x) = (f(x))^2$, $|f|$ est la fonction $[a,b] \to K$ définie par $|f|(x) = |f(x)|$, et, pour $K = \mathbb{C}$, $\bar{f}$ est la fonction $[a,b] \to \mathbb{C}$ définie par $\bar{f}(x) = \overline{f(x)}$ (et donc $|f|^2 = f\bar{f}$).

1.11.1 Polynômes de Legendre

Le polynômes de Chebychev permettent, pour une fonction f donnée, de trouver les $n+1$ points (x_i) pour lesquels on sait que si en ces points un polynôme p_n de degré n vérifie $p_n(x_i) = f(x_i)$ alors p_n est proche de f au sens de l'instrument de mesure $||.||_\infty$.

Il se trouve qu'un autre instrument de mesure est très utilisé : celui qui mesure l'énergie d'une fonction. Cette énergie est le carré de la norme définie sur un intervalle $[a,b]$ par :

$$||f||_2 = \Big(\int_a^b f^2(x)\,dx\Big)^{\frac{1}{2}}.$$

L'avantage de cette norme est qu'elle dérive du produit scalaire :

$$(f,g)_2 = \int_a^b f(x)g(x)\,dx$$

défini pour les fonctions de $L^2(]a,b[) = \{f : [a,b] \to \mathbb{R} \text{ t.q. } \int_a^b f^2(x)\,dx < \infty\}$.

Remarque 1.61 Pour les fonctions à valeurs complexe, on a $(f,g)_2 = \int_a^b f(x)\bar{g}(x)\,dx$, et donc $||f||_2 = (\int_a^b |f(x)|^2\,dx)^{\frac{1}{2}}$. ∎

Proposition 1.62 *Il existe un unique polynôme p_n de degré n (à savoir le projeté orthogonal de f sur $\mathcal{P}_n$ pour le produit scalaire $(\cdot,\cdot)_2$) qui réalise le minimum :*

$$||f-p_n||_2 = \inf_{q_n\in\mathcal{P}_n} ||f-q_n||_2, \tag{1.45}$$

i.e. $\exists! p_n \in \mathcal{P}_n$ t.q. $\forall q_n \in \mathcal{P}_n$ on a $||f-p_n||_2 \le ||f-q_n||_2$. Et p_n est caractérisé par :

$$\forall q_n \text{ de degré } n, \quad (f-p_n,q_n)_2 = 0. \tag{1.46}$$

Preuve. L'ensemble $\mathcal{P}_n$ des polynômes de degré n est un sous-espace de dimension finie donc fermé dans $(L^2,(\cdot,\cdot)_{L^2})$. Et on applique le théorème de projection sur des convexes fermés dans des espaces de Hilbert (= espaces vectoriels muni d'un produit scalaire qui sont complets pour la norme associée au produit scalaire). ∎

Exercice 1.63 Montrer que (1.46) implique (1.45).

Réponse. Sachant $p_n - q_n \in \mathcal{P}_n$:

$$(f-p_n,q_n-p_n)_2 = 0 \quad \text{d'où} \quad (f-q_n,q_n-p_n)_2 = (p_n-q_n,q_n-p_n)_2 = -||p_n-q_n||_2^2.$$

D'où :

$$\begin{aligned} ||f-p_n||_2^2 &= (f-q_n+q_n-p_n, f-q_n+q_n-p_n)_2 = ||f-q_n||_2^2 + 2(f-q_n,q_n-p_n)_2 + ||q_n-p_n||_2^2 \\ &\le ||f-q_n||_2^2 - ||p_n-q_n||_2^2 \le ||f-q_n||_2^2. \end{aligned}$$

∎

Le calcul de p_n peut se faire en cherchant ses composantes sur la base usuelle $(1,x,x^2,...,x^n)$ de $\mathcal{P}_n$, i.e. on cherche les coefficients $(a_i)_{i=0,...,n}$ de $p_n(x) = a_0 + ... + a_n x^n$. Le système est simple à former :

$$\begin{cases} (p_n,1)_2 = (f,1)_2 \\ (p_n,x)_2 = (f,x)_2 \\ \vdots \\ (p_n,x^n)_2 = (f,x^n)_2 \end{cases} \quad \text{soit} \quad \begin{cases} \sum_j (x^j,1)_2 a_j = (f,1)_2 \\ \sum_j (x^j,x)_2 a_j = (f,x)_2 \\ \vdots \\ \sum_j (x^j,x^n)_2 a_j = (f,x^n)_2 \end{cases} \quad \text{soit} \quad A.\vec{a} = \vec{f}$$

qui est un système de $n+1$ équations avec $n+1$ inconnues, où A est la matrice $[(x^j,x^i)_2]_{i,j=0,...,n}$, le vecteur $\vec{a} = \begin{pmatrix} a_0 \\ \vdots \\ a_n \end{pmatrix}$ est l'inconnue, et $\vec{f} = \begin{pmatrix} (f,1)_2 \\ \vdots \\ (f,x^n)_2 \end{pmatrix}$ est calculée à l'aide de la donnée f. Malheureusement, la matrice A est très mal conditionnée, et sa résolution pose problème (pour n grand).

Le problème vient du fait que la base $(x^i)_{i=0,...,n}$ de $\mathcal{P}_n$ n'est pas L^2-orthonormale, et la matrice A n'est en particulier pas unitaire (les matrices unitaires sont très bien conditionnées, toutes leurs valeurs propres étant de module égal à 1).

L'idée est donc de construire une base orthogonale de $\mathcal{P}_n$.

Proposition 1.64 *On se place dans le cas $[a,b] = [-1,1]$. Les polynômes M_n de degré $n \in \mathbb{N}$ donnés par :*

$$M_n(x) = \frac{d^n}{dx^n}((x^2-1)^n) \qquad (= \frac{(2n)!}{n!}\,x^n + ...) \tag{1.47}$$

forment une base orthogonale non normée de $(\mathcal{P},(\cdot,\cdot)_2)$ l'ensemble des polynômes muni du produit scalaire $(\cdot,\cdot)_{L^2}$ usuel.

Les polynômes M_n normalisés sont donnés par :

$$Q_n(x) = \frac{\sqrt{\frac{2n+1}{2}}}{2^n(n!)}\,M_n \qquad (= \sqrt{\frac{2n+1}{2}}\,\frac{1}{2^n(n!)}\,\frac{d^n}{dx^n}((x^2-1)^n)), \tag{1.48}$$

et forment une b.o.n. de $(\mathcal{P},(\cdot,\cdot)_2)$. (Les Q_n sont obtenus par le procédé d'orthonormalisation de Gram–Schmidt à partir de la base $(1,x,x^2,...)$.) En particulier $M_0 = 1$ polynôme constant, et pour tout $n \ge 1$:

$$\int_{-1}^{1} M_n(x)\,dx = 0. \tag{1.49}$$

Preuve. Vérifions que les Q_n forment une b.o.n. de $L^2([-1,1])$. Soit M_n donné par (1.47). Comme $(x^2-1)^n = x^{2n} + R_{2n-2}$ où R_{2n-2} est de degré $\leq 2n-2$, les M_n sont bien de degré n.

Montrons que $(M_n, M_m)_2 = 0$ pour tout $n \neq m$. On a par intégration par parties :

$$\begin{aligned}\int_{-1}^{1} \frac{d^n}{dx^n}((x^2-1)^n)\frac{d^m}{dx^m}((x^2-1)^m)\,dx = & -\int_{-1}^{1} \frac{d^{n-1}}{dx^{n-1}}((x^2-1)^n)\frac{d^{m+1}}{dx^{m+1}}((x^2-1)^m)\,dx \\ & + [\frac{d^{n-1}}{dx^{n-1}}((x^2-1)^n)\frac{d^m}{dx^m}((x^2-1)^m)]_{-1}^{1}.\end{aligned}$$

Comme $(x^2-1)^n = (x-1)^n(x+1)^n$, les racines $+1$ et -1 sont de multiplicité n, et donc $\frac{d^{n-1}}{dx^{n-1}}((x^2-1)^n)$ est nul en $x = \pm 1$, et donc le terme de bord est nul. D'où par récurrence, pour $p \leq n$:

$$\int_{-1}^{1} M_n(x)M_m(x)\,dx = (-1)^p \int_{-1}^{1} \frac{d^{n-p}}{dx^{n-p}}((x^2-1)^n)\frac{d^{m+p}}{dx^{m+p}}((x^2-1)^m)\,dx.$$

En particulier, pour $m < n$ et avec $p = m+1$, on a $\frac{d^{m+p}}{dx^{m+p}}((x^2-1)^m) = 0$ et donc $(M_n, M_m)_2 = 0$. Et $(M_n, M_m)_2 = (M_m, M_n)_2$, donc de même si $n < m$.

Et pour $m = n$, on prend $p = n$, et on a :

$$(M_n, M_n)_2 = (-1)^n \int_{-1}^{1} (x^2-1)^n (2n)!\,dx = (2n)!\,I_n, \qquad I_n = \int_{-1}^{1} (1-x^2)^n\,dx.$$

Par intégration par parties :

$$\begin{aligned} I_n = & \int_{-1}^{1} (1-x^2)^n\,dx = 2n\int_{-1}^{1} x^2(1-x^2)^{n-1} + [x(1-x^2)^n]_{-1}^{1} \\ & = 2n\int_{-1}^{1} (x^2-1)(1-x^2)^{n-1}\,dx + 2n\int_{-1}^{1} (1-x^2)^{n-1}\,dx + 0 \\ & = -2nI_n + 2nI_{n-1}, \end{aligned}$$

d'où $(2n+1)I_n = 2nI_{n-1}$. Et donc, comme $I_0 = 2$ et comme :

$$(2n-1)(2n-3)...3 = \frac{(2n)!}{2n(2n-2)(2n-4)...2} = \frac{(2n)!}{2^n\,n!},$$

on obtient :

$$I_n = \frac{2n}{2n+1}I_{n-1} = \frac{2n(2n-2)...2}{(2n+1)(2n-1)...3}\,2 = \frac{2^{n+1}\,n!}{(2n+1)\frac{(2n)!}{2^n\,n!}} = \frac{2^{2n+1}\,(n!)^2}{(2n+1)\,(2n)!}.$$

D'où :

$$(M_n, M_n)_2 = \frac{2^{2n+1}\,(n!)^2}{(2n+1)}, \qquad ||M_n|| = 2^n\,n!\,\sqrt{\frac{2}{(2n+1)}}.$$

Les Q_n forment donc une famille orthonormale de l'ensemble $\mathcal{P}$ des polynômes muni du produit scalaire L^2. Comme $\mathcal{P}$ est dense dans $L^2([-1,1])$, cette famille est une b.o.n.

Puis $M_0 = 1$, et pour $n \geq 1$ et $p \leq m < n$:

$$\begin{aligned}(M_n(x), x^m)_2 = & \int_{-1}^{1} \frac{d^n}{dx^n}((x^2-1)^n)x^m\,dx \\ & = -\int_{-1}^{1} \frac{d^{n-1}}{dx^{n-1}}((x^2-1)^n)mx^{m-1}\,dx + [\frac{d^{n-1}}{dx^{n-1}}((x^2-1)^n)x^m]_{-1}^{1} \\ & = (-1)^p \int_{-1}^{1} \frac{d^{n-p}}{dx^{n-p}}((x^2-1)^n)\frac{m!}{(m-p)!}x^{m-p}\,dx \\ & = (-1)^m \int_{-1}^{1} \frac{d^{n-m}}{dx^{n-m}}((x^2-1)^n)\,dx = (-1)^m[\frac{d^{n-m-1}}{dx^{n-m-1}}((x^2-1)^n)]_{-1}^{1} = 0,\end{aligned}$$

et donc M_n est orthogonal à toutes les fonctions x^m pour $m = 0, ..., n-1$: cette base suit le procédé de construction de Gram-Schmidt.

Enfin $M_n \perp M_0$ pour $n \geq 1$, i.e. $(M_0, M_n)_{L^2} = 0$ pour $n \geq 1$, i.e. (1.49). ∎

Définition 1.65 Les polynômes de Legendre de degré n sont les polynômes (formule de Rodrigues) :

$$P_n(x) = \frac{1}{n!\,2^n}\frac{d^n}{dx^n}((x^2-1)^n) \qquad (= \frac{1}{n!\,2^n}\,M_n(x)). \tag{1.50}$$

Exemple 1.66 On obtient $P_0(x) = 1$,
$P_1(x) = x$,
$P_2(x) = \frac{3}{2}x^2 - \frac{1}{2}$,
$P_3(x) = \frac{5}{2}x^3 - \frac{3}{2}x$,
$P_4(x) = \frac{35}{8}x^4 - \frac{15}{4}x^2 + \frac{3}{8}$,
$P_5(x) = \frac{63}{8}x^5 - \frac{35}{4}x^3 + \frac{15}{8}$,
$P_6(x) = \frac{231}{16}x^6 - \frac{315}{16}x^4 + \frac{105}{16}x^2 - \frac{5}{16}$, ... ▪

Proposition 1.67 *Les polynômes de Legendre vérifient :*

$$P_n(-1) = (-1)^n \qquad et \qquad P_n(1) = 1. \tag{1.51}$$

(Ce qui explique le choix de la constante $\frac{1}{n!\,2^n}$.)

Preuve. On a, formule de Leibniz :

$$\frac{d^n}{dx^n}((x^2-1)^n) = \frac{d^n}{dx^n}((x-1)^n(x+1)^n) = \sum_{k=0}^{n} C_n^k \frac{d^k}{dx^k}((x-1)^n)\frac{d^{n-k}}{dx^{n-k}}((x+1)^n).$$

Tous les termes de la somme pris en la valeur 1 sont nuls sauf pour k=n, car 1 est racine multiple d'ordre n du polynôme $(x-1)^n$. Et pour k=n on a $\frac{d^n}{dx^n}((x-1)^n) = n!$ et $((x+1)^n)_{|x=1} = 2^n$. Et donc $\frac{d^n}{dx^n}((x^2-1)^n)_{|x=1} = (n!)2^n$.

Tous les termes de la somme pris en la valeur -1 sont nuls sauf pour k=0, car -1 est racine multiple d'ordre n du polynôme $(x+1)^n$. Et pour k=0 on a $((x-1)^n)_{|x=-1} = (-2)^n$ et $\frac{d^n}{dx^n}((x+1)^n)_{|x=1} = n!$. Et donc $\frac{d^n}{dx^n}((x^2-1)^n)_{|x=-1} = (-1)^n(n!)2^n$. ▪

Proposition 1.68 *Les P_n sont de degré n et sont 2 à 2 orthogonaux dans $L^2(]-1,1[)$. Ils sont de la forme, pour $n \geq 2$:*

$$P_n(x) = a_n x^n + R_{n-2}(x), \qquad a_n = \frac{(2n)!}{2^n\,(n!)^2}, \tag{1.52}$$

où R_{n-2} est un polynôme de degré n−2. Et ils vérifient la relation de récurrence, pour $n \geq 1$:

$$\begin{cases} P_0(x) = 1, \qquad P_1(x) = x, \\ P_{n+1}(x) = \dfrac{2n+1}{n+1}\,x\,P_n(x) - \dfrac{n}{n+1}P_{n-1}(x). \end{cases} \tag{1.53}$$

Et les polynômes $\sqrt{\frac{2n+1}{2}}\,P_n$ forment une base orthonormale de $L^2(]-1,1[)$.

Preuve. On a $(x^2-1)^n = x^{2n} + C_n^{n-1}x^{2(n-1)} + ...$, d'où (1.52) sachant (1.47).

Puis $||M_n||_2 = \frac{2^n(n!)}{\sqrt{\frac{2n+1}{2}}}$ car $||Q_n||_2 = 1$, cf. (1.48), d'où $||P_n||_2 = \frac{1}{n!\,2^n}\frac{2^n(n!)}{\sqrt{\frac{2n+1}{2}}} = \sqrt{\frac{2}{2n+1}}$.

Montrons (1.53). Soit :

$$Z = P_{n+1}(x) - \frac{2n+1}{n+1}\,x\,P_n(x).$$

Son terme x^{n+1} a pour coefficient :

$$\frac{(2(n+1))!}{2^{n+1}\,((n+1)!)^2} - \frac{2n+1}{n+1}\frac{(2n)!}{2^n\,(n!)^2} = \frac{(2(n+1))! - 2(n+1)(2n+1)(2n)!}{2^{n+1}\,((n+1)!)^2} = 0.$$

Donc, avec (1.52), Z est un polynôme de degré $\leq n-1$, donc de la forme :

$$Z = \sum_{k=0}^{n-1} \alpha_k P_k.$$

Et les P_k étant 2 à 2 orthogonaux, pour tout $m \leq n-1$:

$$\alpha_m||P_m||_2^2 = (Z, P_m)_2 = (P_{n+1}, P_m)_2 - \frac{2n+1}{n+1}(xP_n, P_m)_2 = -\frac{2n+1}{n+1}(P_n, xP_m)_2.$$

Et pour $m \leq n-2$, le polynôme xP_m est de degré $\leq n-1$ et donc $(P_n, xP_m)_2 = 0$. Il reste :

$$Z = \alpha_{n-1}P_{n-1} \qquad \text{où} \qquad \alpha_{n-1}||P_{n-1}||_2^2 = -\frac{2n+1}{n+1}(P_n, xP_{n-1})_2.$$

Puis xP_{n-1} est de degré n, et avec (1.52) et identification des termes en x^n :

$$xP_{n-1} = \beta P_n + ... \qquad \text{où} \qquad \beta = \frac{a_{n-1}}{a_n} = \frac{\frac{(2(n-1))!}{2^{n-1}((n-1)!)^2}}{\frac{(2n)!}{2^n (n!)^2}} = \frac{2n^2}{2n(2n-1)} = \frac{n}{2n-1},$$

les ... étant une combinaison linéaire des P_k pour $k = 0, ..., n-1$. d'où :

$$\alpha_{n-1}||P_{n-1}||_2^2 = -\frac{2n+1}{n+1}\beta\,(P_n, P_n)_2,$$

avec $||P_n||_2^2 = \frac{2}{2n+1}$, d'où :

$$\alpha_{n-1} = -\frac{2n-1}{2}\frac{2n+1}{n+1}\,\frac{n}{2n-1}\,\frac{2}{2n+1} = -\frac{n}{n+1}.$$

■

Proposition 1.69 *Pour $n \geq 1$, les racines des polynômes de Legrendre sont toutes simples, toutes réelles, et sont toutes dans $]-1,1[$.*

Preuve. On note $x_1, ..., x_m$ les racines réelles distinctes de P_n de multiplicité impaire qui sont dans $]-1,1[$, i.e. P_n est de la forme $P_n(x) = \prod_{i=1}^m (x-x_i)^{2k_i-1} R(x)$ où $2k_i-1$ est la multiplicité de x_i et où R n'a pas de racine de multiplicité impaire dans $]-1,1[$. Il s'agit de montrer que $m = n$.

Posons $Q_0(x) = 1$ et $Q_m(x) = \prod_{i=1}^m (x-x_i)$ pour $m \geq 1$. Le polynôme non nul $P_n Q_m$ est donc de signe constant dans $]-1,1[$ puisque de la forme $P_n Q_m(x) = \prod_{i=1}^m (x-x_i)^{2k} R(x)$ où R n'a pas de racine de multiplicité impaire dans $]-1,1[$. Donc $(P_n, Q_m)_{L^2(]-1,1[)} \neq 0$. Si on suppose $m < n$ alors $(P_n, Q_m)_{L^2(]-1,1[)} = 0$ car P_n est orthogonal à tout polynôme de degré $< n$. C'est contradictoire, donc $m = n$. ■

Exercice 1.70 Montrer que :

$$P_n(x) = \frac{1}{2^n}\sum_{j=0}^n (C_n^j)^2 (x-1)^{n-j}(x+1)^j. \tag{1.54}$$

Preuve. On applique la formule de Leibniz $(fg)^{(n)} = \sum_{j=0}^n C_n^j f^{(n-j)} g^{(j)}$ avec $f(x) = (x-1)^n$ et $g(x) = (x+1)^n$. ■

■

Exercice 1.71 Montrer que $\frac{d}{dx}\Big((1-x^2)\frac{d}{dx}P_n\Big) + n(n+1)P_n = 0$ (équation du second ordre dont l'une des deux solutions est le polynôme de Legendre). ■

Exercice 1.72 Montrer que $P_{n+1}(x) = x\,P_n(x) - \frac{1}{n+1}(1-x^2)\frac{dP_n}{dx}(x)$. ■

Exercice 1.73 Montrer que $\frac{d}{dx}(P_{n+1} - P_{n-1}) = (2n+1)P_n$. ■

1.11.2 Généralisation

On peut également mesurer une énergie pondérée : sur un intervalle $[a,b]$, $a < b$, on se donne w est une fonction intégrable strictement positive (i.e. une densité, i.e. $w \in L^1(]a,b[)$ et $w > 0$), et on note :

$$L^2_w = \{f : \int_a^b f^2(x)\,w(x)dx < \infty\}.$$

C'est un espace vectoriel qu'on munit du produit scalaire :

$$(f,g)_w = \int_a^b f(x)g(x)\,w(x)dx.$$

Ou encore $(f,g)_w = \int_a^b f(x)g(x)\,d\mu$ où μ est la mesure de densité définie par $d\mu = w(x)dx$.

En particulier, si $[a,b] = [-1,1]$ et $w : x \to 1$ (la fonction constante 1), alors $d\mu = dx$ et on retrouve la b.o.n. des polynômes de Legendre.

Proposition 1.74 *A l'aide de Gram–Schmidt, on peut construire une b.o.n. $(P_n)_{n\in\mathbb{N}}$ de L^2_w (donc avec $(P_n, x^k)_w = 0$ pour tout $k < n$). Et les P_n vérifient la relation de récurrence :*

$$P_{n+1} = (A_n x + B_n)P_n - C_n P_{n-1}, \tag{1.55}$$

où :

$$A_n = \frac{a_{n+1}}{a_n}, \qquad B_n = \frac{b_{n+1}a_n - b_n a_{n+1}}{a_n}, \qquad C_n = \frac{a_{n-1}a_{n+1}}{a_n^2}, \tag{1.56}$$

où on a posé $P_n = a_n x^n + b_n x^{n-1} + ...$ pour tout n.

Preuve. Technique similaire au cas de Legendre. Voir par exemple Schatzman [13]. ∎

Proposition 1.75 *Pour $n \geq 1$, les racines des P_n sont toutes simples, toutes réelles, et sont toutes dans $]a, b[$.*

Preuve. Similaire au cas de Legendre. ∎

1.11.3 Retour sur les polynômes de Chebychev

Voir (1.44).

1.11.4 Polynômes de Laguerre

On prend $[a, b[= [0, \infty[$ et $w : x \in\to e^{-x}$, et donc

$$(f, g)_w = \int_0^\infty f(x)g(x)\, e^{-x} dx,$$

et on obtient les polynômes de Laguerre (base orthogonale dans L^2_w). Il sont définis par :

$$L_n(x) = \frac{1}{n!} e^x \frac{d^n}{dx^n}(e^{-x}x^n).$$

Et on vérifie que $L_n(0) = 1$.

1.11.5 Polynômes d'Hermite

On prend $]a, b[=]-\infty, \infty[$ et $w : x \in\to e^{-x^2}$, et donc

$$(f, g)_w = \int_{-\infty}^\infty f(x)g(x)\, e^{-x^2} dx,$$

et on obtient les polynômes d'Hermite (base orthogonale dans L^2_w). Il sont définis par :

$$H_n(x) = \frac{1}{n!} e^{x^2} \frac{d^n}{dx^n}(e^{-x^2}).$$

Et on vérifie que $H_{2n}(0) = (-1)^n \frac{(2n)!}{n!}$ (polynômes de degré pair) et $H_{2n+1}(0) = 0$.

1.12 Splines cubiques

1.12.1 Introduction

On traite le point 3 du § 1.7. Soit $[a, b] = \bigcup_{i=1}^n [x_{i-1}, x_i]$ (maillage) où $a{=}x_0 < x_1 < ... < b{=}x_n$ (où donc $n \geq 1$ est le nombre d'intervalles). Ici on utilisera uniquement les points a_{k-1} et a_k (pas de points intermédiaires), et $a_k \overset{\text{noté ici}}{=} x_k$. Donc ici les x_k ne sont pas des points d'interpolation de Lagrange mais les points définissant le maillage $[a, b] = \bigcup_{k=1}^{n} [x_{k-1}, x_k]$.

Une spline cubique nécessitera tout le maillage $[a, b] = \bigcup_{i=1}^n [x_{i-1}, x_i]$ pour être définie (définition globale : le point de départ de sa définition ne peut pas être un seul intervalle $[a_{k-1}, a_k] \overset{\text{noté ici}}{=} [x_{k-1}, x_k]$, contrairement aux polynômes de Lagrange ou d'Hermite).

1.12.2 Calculs

Soit $f \in C^0([a,b];\mathbb{R})$ et $y_i = f(x_i)$ pour $i = 0, ..., n$.

But : trouver une fonction $S \in C^2([a,b];\mathbb{R})$ t.q. :

$$\forall i = 0, ..., n, \quad S(x_i) = y_i, \quad \text{et} \quad \forall k = 1, ..., n, \quad S_{|]x_{k-1},x_k[} \overset{\text{noté}}{=} p_k \in \mathcal{P}_3 \tag{1.57}$$

(la restriction de S au k-ième intervalle est un polynôme $p_k \in \mathcal{P}_3$).

Un polynôme $p_k \in \mathcal{P}_3$ étant de la forme $a_{k0} + a_{k1}x + a_{k2}x^2 + a_{k3}x^3 = \sum_{j=0}^{3} a_{kj}x^j$, il y a $4n$ inconnues les $(a_{kj})_{\substack{k=1,...,n \\ j=0,...,3}}$. Exprimons les contraintes :

1. S vaut y_i aux x_i, donc $p_1(x_0) = y_0$, $p_1(x_1) = y_1$, $p_2(x_1) = y_1$, $p_2(x_2) = y_2$, ..., $p_n(x_{n-1}) = y_{n-1}$, $p_n(x_n) = y_n$, soit $2n$ équations (2 équations par intervalle).
2. $S \in C^1([a,b])$ donc $p_i'(x_i) = p_{i+1}'(x_i)$, pour i=1, ..., n−1, donc n−1 équations.
3. $S \in C^2([a,b])$ donc $p_i''(x_i) = p_{i+1}''(x_i)$, pour i=1, ..., n−1, donc n−1 équations.

Au total, on a $4n$−2 contraintes.

Définition 1.76 Une fonction S vérifiant (1.57) et les $4n$−2 contraintes ci-dessus est appelée une spline cubique (nom complet : spline cubique C^2).

Il nous manque deux contraintes pour avoir $4n$ équations (autant que d'inconnues les a_{kj}). Notons

$$S''(x_i) \overset{\text{noté}}{=} y_i'', \quad i = 0, ..., n \tag{1.58}$$

(les n+1 réels y_i'' sont inconnus pour le moment). En particulier $p_i''(x_i) = p_{i+1}''(x_i) = y_i''$ pour i=1, ..., n−1.

On impose souvent une courbure aux extrémités (donne des calculs simples) :

$$y_0'' \quad \text{et} \quad y_n'' \quad \text{donnés.} \tag{1.59}$$

Définition 1.77 Si on choisit d'introduire les deux contraintes supplémentaires $y_0'' = y_n'' = 0$, alors S est dite spline cubique naturelle.

Proposition 1.78 *Si on connaît les y_i'', $i = 0, ..., n$, alors les p_i sont donnés par :*

$$\begin{aligned} p_i(x) = {} & \frac{(x-x_i)^3}{(x_{i-1}-x_i)}\frac{y_{i-1}''}{6} + \frac{(x-x_{i-1})^3}{(x_i-x_{i-1})}\frac{y_i''}{6} \\ & + \left(\frac{y_{i-1}}{(x_{i-1}-x_i)} - \frac{y_{i-1}''}{6}(x_{i-1}-x_i)\right)(x-x_i) + \left(\frac{y_i}{(x_i-x_{i-1})} - \frac{y_i''}{6}(x_i-x_{i-1})\right)(x-x_{i-1}). \end{aligned} \tag{1.60}$$

Équation appelée équation de la spline dans l'intervalle $[x_{i-1}, x_i]$.

Preuve. Dans chaque $[x_{i-1}, x_i]$ les p_i sont de degré 3 donc leurs dérivées secondes p_i'' sont de degré 1. Et aux extrémités : $p_i''(x_{i-1}) = y_{i-1}''$ et $p_i''(x_i) = y_i''$. D'où dans $[x_{i-1}, x_i]$ (interpolation de Lagrange) :

$$p_i''(x) = \frac{(x-x_i)}{(x_{i-1}-x_i)}y_{i-1}'' + \frac{(x-x_{i-1})}{(x_i-x_{i-1})}y_i''. \tag{1.61}$$

(Immédiat avec les polynômes de Lagrange.) On intègre 2 fois, et on a 2 constantes d'intégrations α_i et β_i :

$$p_i(x) = \frac{(x-x_i)^3}{(x_{i-1}-x_i)}\frac{y_{i-1}''}{6} + \frac{(x-x_{i-1})^3}{(x_i-x_{i-1})}\frac{y_i''}{6} + \alpha_i(x-x_i) + \beta_i(x-x_{i-1}).$$

(Avec donc $\alpha_i(x-x_i) + \beta_i(x-x_{i-1}) = (\alpha_i+\beta_i)x - \alpha_i x_i - \beta_i x_{i-1} = \gamma_i x + \delta_i =$ notation plus usuelle mais qui ici ne serait pas pratique.)

On calcule ces constantes d'intégrations α_i et β_i à l'aide des valeurs connues y_{i-1}=$p_i(x_{i-1})$ et y_i=$p_i(x_i)$ qui donnent :

$$y_{i-1} = (x_{i-1}-x_i)^2\frac{y_{i-1}''}{6} + \alpha_i(x_{i-1}-x_i) \quad \text{et} \quad y_i = (x_i-x_{i-1})^2\frac{y_i''}{6} + \beta_i(x_i-x_{i-1}),$$

système de deux équations à deux inconnues α_i et β_i, de résolution immédiate qui donne (1.60). ∎

Proposition 1.79 *Supposons y_0'' et y_n'' connus, cf. (1.59). Alors les y_i'' pour $i = 1, ..., n-1$ existent et sont uniques : ils satisfont aux $n-1$ équations :*

$$\frac{(x_i-x_{i-1})}{(x_{i+1}-x_{i-1})}y''_{i-1} + 2y''_i + \frac{(x_{i+1}-x_i)}{(x_{i+1}-x_{i-1})}y''_{i+1} = 6f[x_{i-1}, x_i, x_{i+1}], \quad i = 1, ..., n-1, \tag{1.62}$$

où $f[x_{i-1}, x_i, x_{i+1}] = \frac{f[x_i,x_{i+1}]-f[x_{i-1},x_i]}{(x_{i+1}-x_{i-1})} = \frac{(y_{i+1}-y_i)(x_i-x_{i-1})-(y_i-y_{i-1})(x_{i+1}-x_i)}{(x_{i+1}-x_{i-1})(x_{i+1}-x_i)(x_i-x_{i-1})}$ (voir les polynômes de Newton). Soit :

$$A.\begin{pmatrix} y''_1 \\ y''_2 \\ \vdots \\ y''_{n-2} \\ y''_{n-1} \end{pmatrix} = 6\begin{pmatrix} f[x_0, x_1, x_2] \\ f[x_1, x_2, x_3] \\ \vdots \\ f[x_{n-3}, x_{n-2}, x_{n-1}] \\ f[x_{n-2}, x_{n-1}, x_n] \end{pmatrix} - \begin{pmatrix} \frac{(x_1-x_0)}{(x_2-x_0)}y''_0 \\ 0 \\ \vdots \\ 0 \\ \frac{(x_n-x_{n-1})}{(x_n-x_{n-2})}y''_n \end{pmatrix} \tag{1.63}$$

où :

$$A = \begin{pmatrix} 2 & \frac{(x_2-x_1)}{(x_2-x_0)} & 0 & \dots & & 0 \\ \frac{(x_2-x_1)}{(x_3-x_1)} & 2 & \frac{(x_3-x_2)}{(x_3-x_1)} & 0 & \dots & \\ 0 & & & & & \\ & & & \ddots & & \\ & & & & & 0 \\ & \dots & 0 & \frac{(x_{n-2}-x_{n-3})}{(x_{n-1}-x_{n-3})} & 2 & \frac{(x_{n-1}-x_{n-2})}{(x_{n-1}-x_{n-3})} \\ 0 & & \dots & 0 & \frac{(x_{n-1}-x_{n-2})}{(x_n-x_{n-2})} & 2 \end{pmatrix}, \tag{1.64}$$

système tri-diagonal inversible de résolution très rapide.

Preuve. On dérive (1.60), pour tout $i = 1, ..., n$:

$$\begin{aligned} p'_i(x) = {} & \frac{(x-x_i)^2}{(x_{i-1}-x_i)}\frac{y''_{i-1}}{2} + \frac{(x-x_{i-1})^2}{(x_i-x_{i-1})}\frac{y''_i}{2} \\ & + (\frac{y_{i-1}}{(x_{i-1}-x_i)} - \frac{y''_{i-1}}{6}(x_{i-1}-x_i)) + (\frac{y_i}{(x_i-x_{i-1})} - \frac{y''_i}{6}(x_i-x_{i-1})). \end{aligned}$$

Comme $p'_i(x_i) = p'_{i+1}(x_i)$, pour tout $i = 1, ..., n-1$, on déduit :

$$\begin{aligned} & (x_i-x_{i-1})\frac{y''_i}{2} + (\frac{y_{i-1}}{(x_{i-1}-x_i)} - \frac{y''_{i-1}}{6}(x_{i-1}-x_i)) + (\frac{y_i}{(x_i-x_{i-1})} - \frac{y''_i}{6}(x_i-x_{i-1})) \\ & = (x_i-x_{i+1})\frac{y''_i}{2} + (\frac{y_i}{(x_i-x_{i+1})} - \frac{y''_i}{6}(x_i-x_{i+1})) + (\frac{y_{i+1}}{(x_{i+1}-x_i)} - \frac{y''_{i+1}}{6}(x_{i+1}-x_i). \end{aligned}$$

D'où :

$$y''_{i-1}(x_i - x_{i-1}) + y''_i(2(x_i-x_{i-1}) - 2(x_i-x_{i+1})) + y''_{i+1}(x_{i+1}-x_i) = 6\frac{y_{i+1}-y_i}{(x_{i+1}-x_i)} - 6\frac{y_i-y_{i-1}}{(x_i-x_{i-1})}.$$

■

Corollaire 1.80 *La résolution de (1.63) donne les y_i'', d'où les p_i, cf. (1.60), d'où la spline cubique S.*

Remarque 1.81 Si le maillage est uniforme, i.e. $x_i - x_{i-1} = h = \frac{b-a}{n}$ (les n intervalles $]x_{i-1}, x_i[$ ont même longueur), alors :

$$A = \frac{1}{2}\begin{pmatrix} 4 & 1 & 0 & ... & & 0 \\ 1 & 4 & 1 & 0 & ... & \vdots \\ 0 & \ddots & \ddots & \ddots & & \\ \vdots & & & & & 0 \\ & & & & & \\ & & ... & 0 & 1 & 4 & 1 \\ 0 & & ... & & 0 & 1 & 4 \end{pmatrix}$$

■

Remarque 1.82 Une spline S est P_3 par morceaux, donc sa dérivée 3ème S''' est constante par morceau : donc en général S n'est pas C^3 sur tout $[a,b]$. ▪

Exercice 1.83 Montrer que si les y_i vérifient $y_i = g(x_i)$ avec $g(x) = c_0 + c_1 x + c_2 x^2 + c_3 x^3$ pour tout $x \in [a,b]$ (polynôme de degré 3 sur tout $[a,b]$), et si on suppose $y_0'' = g(x_0)$ et $y_n'' = g(x_n)$ (voir (1.59)), alors la spline S correspondante, cf. (1.57), vérifie $S = g$.

Réponse. Soit $p_i = g_{[x_{i-1},x_i]}$. Alors toutes les contraintes pour les p_i sont vérifiées. Et on a existence et unicité, cf. prop. 1.79. ▪

Définition 1.84 La k-ème spline cubique de base est la spline cubique naturelle φ_k vérifiant

$$\forall i = 0, ..., n, \quad \varphi_k(x_i) = \delta_{ik}. \tag{1.65}$$

Corollaire 1.85 *Toute spline cubique S s'exprime comme :*

$$S(x) = \sum_{k=0}^{n} y_k \varphi_k(x). \tag{1.66}$$

Et l'ensemble des splines cubiques est un espace vectoriel de dimension $n+1$ dont une base est $(\varphi_k)_{k=0,...,n}$.

Preuve. Les φ_k sont $C^2([a,b])$ et déterminées à l'aide de leurs restrictions $\varphi_k|[x_{i-1},x_i] = p_i$ à l'aide de (1.60).

Toute combinaison linéaire de fonction C^2 est C^2, et donc $x \in [a,b] \to \sum_{i=0}^n y_i \varphi_i(x)$ est une fonction C^2.

Toute combinaison linéaire de polynôme de degré 3 est un polynôme de degré 3, et donc, sur tout intervalle $[x_{i-1},x_i]$, $x \to \sum_{k=0}^n y_k \varphi_k(x)$ est un polynôme de degré 3.

Et $\sum_{k=0}^n y_k \varphi_k(x_i) = \sum_{k=0}^n y_k \delta_{ik} = y_i$, et donc la fonction $x \to \sum_{k=0}^n y_k \varphi_k(x)$ est la fonction spline S cherchée (existence et unicité) : celle qui prend la valeur y_i au point x_i. Donc $(\varphi_k)_{k=0,...,n}$ engendre l'espace des splines cubiques. Et si $\sum_{k=0}^n y_k \varphi_k = 0$, alors en particulier $\sum_{k=0}^n y_k \varphi_k(x_i) = 0$, et donc $y_i = 0$. Donc la famille $(\varphi_k)_{k=0,...,n}$ est libre. Libre et génératrice : c'est une base. ▪

Exercice 1.86 Déterminer le φ_0 du corollaire précédent.

Réponse. Ici on cherche φ_0 donné par (1.60), pour $i = 1, ..., n$:

$$\varphi_{0|[x_{i-1},x_i]}(x) = \frac{(x-x_i)^3}{(x_{i-1}-x_i)}\frac{y_{i-1}''}{6} + \frac{(x-x_{i-1})^3}{(x_i-x_{i-1})}\frac{y_i''}{6} + (-\frac{y_{i-1}''}{6}(x_{i-1}-x_i))(x-x_i) + (-\frac{y_i''}{6}(x_i-x_{i-1}))(x-x_{i-1}).$$

La matrice A est donnée en (1.64). Et (notation de Newton) $\varphi_0[x_0,x_1] = \frac{\varphi_0(x_1)-\varphi_0(x_0)}{x_1-x_0} = \frac{-1}{x_1-x_0}$ Et $\varphi_0[x_1,x_2] = 0 = \varphi_0[x_{i-1},x_i]$ pour tout $2, ..., n$. D'où $\varphi_0[x_0,x_1,x_2] = \frac{\varphi_0[x_1,x_2]-\varphi_0[x_0,x_1]}{x_2-x_0} = \frac{1}{(x_1-x_0)(x_2-x_0)}$ et $\varphi_0[x_{i-1},x_i,x_{i+1}] = 0$ pour tout $i = 2, ..., n-1$. Et donc, avec $y_0'' = 0 = y_n''$, les coefficients y_i'' pour déterminer φ_0 sont solutions de $A.\begin{pmatrix} y_1'' \\ \vdots \\ y_{n-1}'' \end{pmatrix} = 6\begin{pmatrix} \frac{1}{(x_1-x_0)(x_2-x_0)} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$. N.B. : pour trouver tous les y_i'', on décompose par exemple A en $L.L^T$ (Choleski : coût $\simeq 8n$ pour une matrice tridiagonale symétrique). ▪

2 Intégration numérique

2.1 Introduction

On souhaite trouver une formule d'intégration, pour $-\infty < a < b < \infty$, de type :

$$\int_a^b f(x)\,dx \simeq \text{somme finie} = \sum_{i=1}^{n} w_i f(\xi_i), \tag{2.1}$$

la 'somme finie' étant une 'bonne' approximation de l'intégrale. Les ξ_i seront des points où f est définie et les w_i (poids=weight en anglais) sont des réels.

Définition 2.1 La formule d'intégration numérique (2.1) où $w_i, \xi_i \in \mathbb{R}$ pour tout $i = 1, ..., n$ est dite d'ordre m si elle est exacte pour tout polynôme de degré $\leq m$.

Dans le cas où $[a, b]$ est un "grand" intervalle, une idée simple est de découper $[a, b]$ en n intervalles :

$$[a, b] = \bigcup_{i=1}^{n} [x_{i-1}, x_i] \quad \text{où} \quad a = x_0, \quad b = x_n, \quad \text{et} \quad x_{i-1} < x_i,\ \forall i = 1, ..., n.$$

Ayant :

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

on calcule alors chaque $\int_{x_{i-1}}^{x_i} f(x)\,dx$ à l'aide de (2.1).

Exemple. Lorsque f est continue, la formule approchée de Riemann est :

$$\int_{x_{i-1}}^{x_i} f(x)\,dx \simeq (x_i - x_{i-1}) f(\xi_i),$$

où ξ_i un point de $[x_{i-1}, x_i]$, ce qui donne :

$$\int_a^b f(x)\,dx \simeq \sum_{i=1}^{n} (x_i - x_{i-1}) f(\xi_i). \tag{2.2}$$

Dans le cas usuel où les intervalles sont de longueurs égales :

$$x_i - x_{i-1} = h = \frac{b-a}{n}, \tag{2.3}$$

on obtient :

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

Et dans ce cas, l'approximation de l'intégrale par la somme finie introduit une erreur :

$$E(f) = \Big(\int_a^b f(x)\,dx\Big) - \Big(h \sum_{i=1}^{n} f(\xi_i)\Big).$$

Le problème sera de connaître la valeur de cette erreur, i.e. la "précision" de l'approximation faite en remplaçant $\int_a^b f(x)\,dx$ par sa somme finie de Riemann. En particulier, si on divise les n intervalles par 2 pour avoir $2n$ intervalles (de longueur $\frac{1}{2}h$) et qu'on applique la somme de Riemann sur ces $2n$ intervalles, est-ce que l'erreur commise (a priori) est divisée par 2 (méthode en $O(h)$), par $4 = 2^2$ (méthode en $O(h^2)$), par 2^k (méthode en $O(h^k)$), ou par autre chose. Ce calcul d'erreur sera fait à l'aide du développement limité de f dans le cas des sommes de Riemann.

En particulier, on verra que si dans la somme de Riemann ci-dessus on prend pour ξ_i le point milieu $\xi_i = \frac{x_{i-1}+x_i}{2}$, alors l'erreur sera en h^2, et sinon l'erreur sera en h : la formule du point milieu est donc préférable, et sera du même ordre que la méthode des trapèzes.

Le problème suivant sera de voir s'il n'y a pas une méthode plus 'rapide' (en $O(h^k)$ avec $k \geq 3$) : on regardera ici en particulier les méthodes de Simpson et de Gauss.

La méthode générale pour faire le calcul d'erreur est : 1- on regarde le cas d'une fonction f polynôme de degré n dont on connaît l'intégrale et on calcule l'erreur $E(f)$ commise ; 2- on prend une fonction f quelconque et on se sert de son développement limité (ou d'un développement polynomial comme celui d'interpolation de Newton) pour se ramener au cas 1- avec un "petit reste" en plus.

2.2 Méthodes de Riemann à droite et à gauche

Dans la suite on prend les intervalles de même longueur h, cf. (2.3).

La somme de Riemann à gauche est l'approximation :

$$\int_{x_{i-1}}^{x_i} f(x)\,dx \simeq G_i \overset{\text{déf}}{=} (x_i{-}x_{i-1})f(x_{i-1}) = h\,y_{i-1}, \tag{2.4}$$

G_i étant l'aire du rectangle de largeur $x_i{-}x_{i-1} = h$ et de hauteur y_{i-1}. Ainsi $\int_a^b f(x)\,dx \simeq h\sum_{i=1}^n y_{i-1}$. Faire un dessin.

La somme de Riemann à droite est l'approximation :

$$\int_{x_{i-1}}^{x_i} f(x)\,dx \simeq D_i \overset{\text{déf}}{=} (x_i{-}x_{i-1})f(x_i) = h\,y_i, \tag{2.5}$$

D_i étant l'aire du rectangle de largeur $x_i{-}x_{i-1} = h$ et de hauteur y_i. Ainsi $\int_a^b f(x)\,dx \simeq h\sum_{i=1}^n y_i$. Faire un dessin.

Il est immédiat que dans le cas où f est constante sur les morceaux $[x_{i-1}, x_i]$ les formules (2.4) et (2.5) sont exactes : l'erreur commise est nulle. Cette méthode est donc au moins d'ordre 0.

2.2.1 Cas d'un polynôme de degré 1

On prend une fonction p_1 continue sur $[a, b]$, polynômiale de degré 1 par morceaux, et on note $p_{1|[x_{i-1},x_i]} = p_1^i$ (restriction à l'intervalle $[x_{i-1}, x_i]$). Si on note $y_j = p_1(x_j)$ pour tout j, on a :

$$p_1^i(x) = y_{i-1} + \frac{y_i - y_{i-1}}{x_i - x_{i-1}}\,(x{-}x_{i-1}) \qquad (= y_{i-1}\frac{x - x_i}{x_{i-1} - x_i} + y_i\frac{x - x_{i-1}}{x_i - x_{i-1}}).$$

(Polynôme de Newton et, entre parenthèses, polynôme de Lagrange.) Et l'aire sous le graphe de p_1 entre x_{i-1} et x_i est l'aire du trapèze :

$$\int_{x_{i-1}}^{x_i} p_1(x)\,dx = (x_i{-}x_{i-1})\frac{y_{i-1} + y_i}{2} = h\,\frac{y_{i-1} + y_i}{2}, \tag{2.6}$$

faire un dessin.

Il est immédiat que l'erreur locale commise (locale=sur l'intervalle $[x_{i-1}, x_i]$) est donnée par, pour la somme à gauche :

$$E_{G_i} = \int_{x_{i-1}}^{x_i} p_1(x)\,dx - G_i = h\,\frac{y_{i-1} + y_i}{2} - h\,y_{i-1} = +\frac{h}{2}\,(y_i{-}y_{i-1}),$$

et pour la somme à droite :

$$E_{D_i} = \int_{x_{i-1}}^{x_i} f(x)\,dx - D_i = h\,\frac{y_{i-1} + y_i}{2} - h\,y_i = -\frac{h}{2}\,(y_i{-}y_{i-1}).$$

Faire un dessin.

D'où sur l'intervalle $[a, b]$, pour la somme de Riemann à gauche :

$$E_G = \int_a^b p_1(x)\,dx - \sum_{i=1}^n G_i = \frac{h}{2}\,(y_n{-}y_0),$$

et pour la somme de Riemann à droite :

$$E_D = \int_a^b p_1(x)\,dx - \sum_{i=1}^n D_i = -\frac{h}{2}\,(y_n{-}y_0),$$

et l'erreur commise pour approximer une fonction affine par morceaux à l'aide des sommes de Riemann à gauche ou à droite est de l'ordre de h.

2.2.2 Cas d'une fonction

Maintenant si f est une fonction quelconque C^1, son développement limité au voisinage de x_{i-1} est donné par, pour $x \in]x_{i-1}, x_i[$:

$$f(x) = f(x_{i-1}) + (x - x_{i-1})f'(\xi_x) \tag{2.7}$$

pour un $\xi_x \in]x_{i-1}, x[$ (théorème des accroissements finis). Et donc l'erreur commise par la méthode de Riemann à gauche sur l'intervalle $]x_{i-1}, x_i[$ de longueur $h = x_i - x_{i-1}$ est :

$$\begin{aligned} E_{G_i} &= \Big(\int_{x_{i-1}}^{x_i} f(x)\,dx\Big) - \Big(hf(x_{i-1})\Big) \\ &= \Big(\int_{x_{i-1}}^{x_i} f(x_{i-1}) + (x - x_{i-1})f'(\xi_x)\,dx\Big) - \Big(\int_{x_{i-1}}^{x_i} f(x_{i-1})\,dx\Big) \\ &= \int_{x_{i-1}}^{x_i} (x - x_{i-1})f'(\xi_x)\,dx, \end{aligned}$$

D'où :

$$|E_{G_i}| \leq \max_{\xi \in [x_{i-1}, x_i]} |f'(\xi)| \, \Big|\int_{x_{i-1}}^{x_i} (x - x_{i-1})\,dx\Big| = \frac{h^2}{2} \max_{\xi \in [x_{i-1}, x_i]} |f'(\xi)|.$$

Et on trouve finalement, pour la somme de Riemann à gauche, sachant $n = \frac{b-a}{h}$:

$$|E_G| \leq \sum_{i=1}^{n} |E_{G_i}| \leq n\frac{h^2}{2} \max_{\xi \in [a,b]} |f'(\xi)| = (b-a)\frac{h}{2} \max_{\xi \in [a,b]} |f'(\xi)| = 0(h),$$

Même raisonnement et résultat pour la somme de Riemann à droite où on fait le développement limité en x_i.

Ces méthodes (somme de Riemann à gauche et somme de Riemann à droite) sont donc des méthodes d'approximation du premier ordre (erreur en $O(h)$) : si au lieu de prendre n intervalles on en prend $2n$, la taille d'un intervalle est alors $\frac{h}{2}$ et l'erreur est divisée par 2 (de manière asymptotique, i.e. pour h suffisamment petit, la notation $O(h)$ signifiant : de l'ordre de h lorsque h tend vers 0).

Exercice 2.2 Intégrer sin sur $[0, \pi]$ à l'aide de Riemann à droite et à gauche sur 2 sous-intervalles puis sur 4 sous-intervalles. ∎

2.3 Méthode du premier ordre des trapèzes

L'approximation par la méthode des trapèzes est l'approximation :

$$\int_{x_{i-1}}^{x_i} f(x)\,dx \simeq T_i \stackrel{\text{déf}}{=} (x_i - x_{i-1})\frac{f(x_{i-1}) + f(x_i)}{2} = h\,\frac{y_{i-1} + y_i}{2}, \tag{2.8}$$

Les calculs précédents ont montré que pour les polynômes de degré 1 on a $E_{D_i} = -E_{G_i}$: la méthode des trapèzes annule donc l'erreur commise pour les polynômes de degré 1, puisque l'erreur est $E_{T_i} = E_{G_i} + E_{D_i}$ dans ce cas. Autrement dit, la méthode des trapèzes compense les erreurs dans ce cas : c'est une méthode exacte pour les polynômes de degré 1.

Sur $[a, b]$ on obtient :

$$\int_a^b f(x)\,dx \simeq \sum_{i=1}^{n} h\,\frac{y_{i-1} + y_i}{2} = h(\frac{y_0}{2} + y_1 + \ldots + y_{n-1} + \frac{y_n}{2}).$$

2.3.1 Cas d'un polynôme de degré 2

Soit f t.q. $f_{|[x_{i-1}, x_i]} = p_2$ est un polynôme de degré 2 sur l'intervalle $[x_{i-1}, x_i]$: on commence par écrire que $p_2 = p_1 +$ 'reste' où p_1 est la fonction affine du trapèze :

$$p_1(x) = y_{i-1} + \frac{y_i - y_{i-1}}{h}(x - x_{i-1}), \quad \forall x \in [x_{i-1}, x_i].$$

Et tout polynôme p_2 de degré 2 qui passe par les points (x_{i-1}, y_{i-1}) et (x_i, y_i) est de la forme :

$$\forall x \in [x_{i-1}, x_i], \quad p_2(x) = p_1(x) + \delta(x - x_{i-1})(x - x_i),$$

cf. polynôme de Newton, puisque $(p_2 - p_1)$ est de degré 2 et s'annule en x_{i-1} et x_i qui sont donc racines de $(p_2 - p_1)$.

Et δ est donné par :

$$p_2''(x) = 0 + 2\delta, \quad \forall x \in [x_{i-1}, x_i].$$

(p_2'' est constante, p_2 étant quadratique.) On vient donc d'écrire que :

$$p_2(x) = p_1(x) + \frac{p_2''}{2}(x-x_{i-1})(x-x_i), \tag{2.9}$$

= (polynôme p_1 de degré 1 trapèze) + (monôme de degré 2 qui s'annule aux extrémités). Et :

$$\int_{x_{i-1}}^{x_i} p_2(x)\,dx = \int_{x_{i-1}}^{x_i} p_1(x)\,dx + \frac{p_2''}{2}\int_{x_{i-1}}^{x_i} (x-x_{i-1})(x-x_i)\,dx$$

d'où on déduit que (la formule du trapèze étant exacte pour les p_1) :

$$\begin{aligned}\int_{x_{i-1}}^{x_i} p_2(x)\,dx - T_i &= \int_{x_{i-1}}^{x_i} p_1(x)\,dx - T_i + \frac{p_2''}{2}\int_{x_{i-1}}^{x_i} (x-x_{i-1})(x-x_i)\,dx \\ &= \frac{p_2''}{2}\int_{x_{i-1}}^{x_i} (x-x_{i-1})(x-x_i)\,dx = -p_2''\frac{h^3}{12},\end{aligned} \tag{2.10}$$

puisque $\int_a^b (x-a)(x-b)\,dx = (b-a)^3\int_0^1 y(y-1)\,dy = -(b-a)^3\frac{1}{6}$ après avoir posé $y = \frac{x-a}{b-a}$.

D'où finalement, pour une fonction p_2 par morceaux, sachant que $n = \frac{b-a}{h}$:

$$\int_a^b p_2(x)\,dx - \sum_{i=1}^n T_i \le (b-a)\frac{h^2}{12}\max_{i=1,\dots,n}|p_2''(x_{i-\frac{1}{2}})| = O(h^2).$$

où $x_{i-\frac{1}{2}}$ est le point milieu $\frac{x_{i-1}+x_i}{2}$ (on rappelle que p_2'' est constant sur les intervalles $[x_{i-1}, x_i]$, et que prendre $p''(x_{i-\frac{1}{2}})$ est simplement pratique).

2.3.2 Cas d'une fonction

Dans le cas de la formule de Riemann à gauche, on avait approché f par une fonction de la forme :

$$f(x) \simeq p_0(x) + f'(\xi(x))(x-x_{i-1}) = y_{i-1} + f'(\xi(x))(x-x_{i-1}),$$

voir (2.7) (formule des accroissements finis). Ici, on va approcher f par :

$$f(x) \simeq p_1(x) + \frac{f''(\xi(x))}{2}(x-x_{i-1})(x-x_i), \quad \text{où } p_1(x) = y_{i-1} + \frac{y_i-y_{i-1}}{h}(x-x_{i-1}), \tag{2.11}$$

à comparer avec (2.9), pour un $\xi(x) = \xi_x$ bien choisi dans $[x_{i-1}, x_i]$: c'est une formule généralisée des accroissements finis, cf. (1.32), démontrée au théorème 1.48.

On en déduit que (la formule du trapèze étant exacte pour les p_1) :

$$\begin{aligned}\left|\int_{x_{i-1}}^{x_i} f(x)\,dx - T_i\right| &= \left|\int_{x_{i-1}}^{x_i} p_1(x)\,dx - T_i + \int_{x_{i-1}}^{x_i} \frac{f''(\xi_x)}{2}(x-x_{i-1})(x-x_i)\,dx\right| \\ &= \left|0 + \int_{x_{i-1}}^{x_i} \frac{f''(\xi_x)}{2}(x-x_{i-1})(x-x_i)\,dx\right| \le \frac{h^3}{12}\sup_{\xi\in[x_{i-1},x_i]}|f''(\xi)|\end{aligned}$$

D'où finalement, sachant que $n = \frac{b-a}{h}$:

$$\left|\int_a^b f(x)\,dx - \sum_{i=1}^n T_i\right| \le (b-a)\max_{\xi\in[a,b]}|f''(\xi)|\,\frac{h^2}{12} = O(h^2).$$

Et la formule des trapèzes est d'ordre 2.

Exercice 2.3 Intégrer sin sur $[0, \pi]$ à l'aide de la méthode des trapèzes sur 2 sous-intervalles puis sur 4 sous-intervalles. ∎

2.4 Méthode du premier ordre du point milieu

L'approximation par la méthode du point milieu est l'approximation :

$$\int_{x_{i-1}}^{x_i} f(x)\,dx \simeq M_i \stackrel{\text{déf}}{=} (x_i - x_{i-1}) f(x_{i-\frac{1}{2}}) = h\, y_{i-\frac{1}{2}}, \tag{2.12}$$

(méthode de Riemann pour le point milieu) où on a noté :

$$x_{i-\frac{1}{2}} = \frac{x_{i-1} + x_i}{2} \tag{2.13}$$

le point milieu et $y_{i-\frac{1}{2}} = f(x_{i-\frac{1}{2}})$. On vérifie immédiatement qu'on a égalité pour les polynômes de degré 1.

La formule du point milieu sur tout $[a, b]$ est donc :

$$\int_a^b f(x)\,dx \simeq h(y_{\frac{1}{2}} + \ldots + y_{n-\frac{1}{2}}).$$

Ici, au point $(x_{i-\frac{1}{2}}, f(x_{i-\frac{1}{2}}))$ on disposera implicitement de deux informations : la hauteur $y_{i-\frac{1}{2}}$ et la pente $f'(x_{i-\frac{1}{2}})$. À comparer avec la méthode du trapèze précédente où on avait 2 points, chacun ne contenant qu'une information de hauteur, soit également deux informations au total.

L'avantage d'avoir les informations de position et de pente en un seul point est qu'on peut utiliser la technique des développements limités. L'analyse va être simple.

2.4.1 Cas d'un polynôme de degré 2

Il est immédiat que cette formule est exacte pour les f affines, faire un dessin (ou faire le calcul).

Regardons le cas $f = p_2$ polynôme de degré 2 : on commence par considérer le polynôme de degré 1 qui passe par le point $(x_{i-\frac{1}{2}}, f(x_{i-\frac{1}{2}}))$ et de pente $f'(x_{i-\frac{1}{2}})$ en ce point :

$$p_1(x) = f(x_{i-\frac{1}{2}}) + (x - x_{i-\frac{1}{2}}) f'(x_{i-\frac{1}{2}}), \quad \forall x \in [x_{i-1}, x_i],$$

puis on développe $f = p_2$ au voisinage de $x_{i-\frac{1}{2}}$:

$$\begin{aligned} p_2(x) &= p_2(x_{i-\frac{1}{2}}) + (x - x_{i-\frac{1}{2}}) p_2'(x_{i-\frac{1}{2}}) + \frac{1}{2}(x - x_{i-\frac{1}{2}})^2 p_2''(x_{i-\frac{1}{2}}) \\ &= p_1(x) + \frac{1}{2}(x - x_{i-\frac{1}{2}})^2 p_2''(x_{i-\frac{1}{2}}). \end{aligned}$$

Et on en déduit donc que (la formule du point milieu étant exacte pour p_1) :

$$\int_{x_{i-1}}^{x_i} f(x)\,dx - M_i = 0 + \frac{1}{2} p_2''(x_{i-\frac{1}{2}}) \int_{x_{i-1}}^{x_i} (x - x_{i-\frac{1}{2}})^2\,dx = p_2''(x_{i-\frac{1}{2}}) \frac{h^3}{24}.$$

D'où finalement, sachant que $n = \frac{b-a}{h}$:

$$\left| \int_a^b p_2(x)\,dx - \sum_{i=1}^n M_i \right| \leq (b-a) \max_{\xi \in [a,b]} (p_2''(\xi))\ \frac{h^2}{24} = O(h^2).$$

2.4.2 Cas d'une fonction

Maintenant, pour f quelconque C^2, il suffit de considérer son développement limité autour de $x_{i-\frac{1}{2}}$:

$$f(x) \simeq f(x_{i-\frac{1}{2}}) + (x - x_{i-\frac{1}{2}}) f'(x_{i-\frac{1}{2}}) + \frac{(x - x_{i-\frac{1}{2}})^2}{2} f''(\xi_x) \tag{2.14}$$

pour un $\xi_x \in [x_{i-\frac{1}{2}}, x]$, voir exercice suivant. On procède alors comme pour les sommes de Riemann et on obtient : la formule du point milieu est d'ordre 2.

Exercice 2.4 Démontrer (2.14).

Réponse. On dispose du développement de Taylor avec reste intégral :

$$f(b) = f(a) + (b-a) f'(a) + \int_a^b f''(t)(b-t)\,dt \qquad (= f(a) + \int_a^b f'(t)\,dt),$$

puisque $\int_a^b 1 f'(t)\,dt = -\int_a^b (t-b) f''(t)\,dt + [(t-b) f'(t)]_a^b$ (intégration par parties après avoir posé $u' = 1$ et $v = f'$ et avoir choisi $u = t-b$). Puis, avec $b > a$ (même démarche avec $b < a$), pour $t \in [a, b]$ on a $b - t > 0$, et donc :

$$(\min_{t \in [a,b]} f''(t)) \int_a^b (b-t)\,dt \leq \int_a^b f''(t)(b-t)\,dt \leq (\max_{t \in [a,b]} f''(t)) \int_a^b (b-t)\,dt.$$

Puis, f'' étant supposé continu, on dispose du théorème des valeurs intermédiaires, et donc il existe $\xi \in [a, b]$ tel que $f''(\xi)) \int_a^b (b-t)\,dt = \int_a^b f''(t)(b-t)\,dt$. Et avec $\int_a^b (b-t)\,dt = [\frac{-(b-t)^2}{2}]_a^b = \frac{(b-a)^2}{2}$, on obtient (2.14). ∎

Exercice 2.5 Intégrer sin sur $[0,\pi]$ à l'aide de la méthode du point milieu sur 2 sous-intervalles puis sur 4 sous-intervalles.

2.5 Méthode du troisième ordre : Simpson

La formule du trapèze donnait une erreur locale de l'ordre de $-\frac{h^3}{12}$ alors que la formule du point milieu donnait une erreur locale de l'ordre de $+\frac{h^3}{24}$. L'idée naturelle est donc de proposer l'annulation de ces erreurs à l'aide de la formule d'intégration :

$$\int_a^b f(x)\,dx \simeq \sum_{i=1}^{n} S_i \quad \text{où} \quad S_i = \frac{T_i + 2M_i}{3} = h\frac{f(x_i) + 4f(x_{i-\frac{1}{2}}) + f(x_{i-1})}{6},$$

barycentre de T_i et M_i avec coefficients barycentriques $\frac{1}{3}$ et $\frac{2}{3}$. C'est l'idée de Simpson. Et donc :

$$\int_a^b f(x)\,dx \simeq \frac{h}{6}(y_0 + 4y_{\frac{1}{2}} + 2y_1 + 4y_{\frac{3}{2}} + 2y_2 + \ldots + 4y_{n-\frac{1}{2}} + y_n).$$

C'est la formule de Simpson. On vérifie immédiatement, à l'aide des deux paragraphes précédents, que cette formule est exacte pour toute fonction $f = p_2$ polynôme de degré 2.

2.5.1 Cas d'un polynôme de degré 3

Regardons ce qui se passe pour les f fonctions polynômes de degré 3 : les formules du trapèze et du point milieu se servent de 4 valeurs, à savoir $f(x_{i-1})$, $f(x_i)$, $f(x_{i-\frac{1}{2}})$, $f'(x_{i-\frac{1}{2}})$. Un polynôme est donné par :

$$p_3(x) = p_2(x) + m_3(x) \quad \text{où} \quad m_3(x) = \delta_3(x - x_{i-1})(x - x_{i-\frac{1}{2}})(x - x_i)$$

où p_2 est l'unique polynôme de degré 2 qui passe par les points $(x_{i-1}, p_3(x_{i-1}))$, $(x_{i-\frac{1}{2}}, p_3(x_{i-\frac{1}{2}}))$ et $(x_i, p_3(x_i))$, et où on a forcément $\delta_3 = \frac{p_3'''}{6}$. Et ayant d'une part :

$$\int_{x_{i-1}}^{x_i} m_3(x)\,dx = \int_{x_{i-1}}^{x_i} (x - x_{i-1})(x - x_{i-\frac{1}{2}})(x - x_i)\,dx = \int_{-\frac{h}{2}}^{\frac{h}{2}} (z + \frac{h}{2})z(z - \frac{h}{2})\,dz = 0$$

(l'intégrant en z est impaire), et d'autre part avec la formule de Simpson :

$$h\frac{m_3(x_i) + 4m_3(x_{i-\frac{1}{2}}) + m_3(x_{i-1})}{6} = 0,$$

(les x_i, x_{i-1} et $x_{i-\frac{1}{2}}$ sont racines de m_3) on en déduit que la formule de Simpson est également exacte pour tout polynôme p_3 (de degré 3).

2.5.2 Cas d'un polynôme de degré 4

Regardons ce qui se passe pour les polynômes p_4 de degré 4 : on l'écrit comme :

$$p_4(x) = p_3(x) + \delta(x - x_{i-1})(x - x_{i-\frac{1}{2}})^2(x - x_i)$$

où p_3 est le polynôme d'interpolation (de Hermite) de degré 3 défini par :

$$p_3(x_{i-1}) = f(x_{i-1}), \quad p_3(x_i) = f(x_i), \quad p_3(x_{i-\frac{1}{2}}) = f(x_{i-\frac{1}{2}}), \quad p_3'(x_{i-\frac{1}{2}}) = f'(x_{i-\frac{1}{2}}),$$

voir paragraphe 1.6.3, et où donc $\delta = \frac{p_4''''}{24}$.

D'où, la formule de Simpson (exacte pour le polynômes de degré 3) :

$$\begin{aligned} E_i &= \int_{x_{i-1}}^{x_i} p_4(x)\,dx - h\frac{p_4(x_i) + 4p_4(x_{i-\frac{1}{2}}) + p_4(x_{i-1})}{6} \\ &= \frac{1}{24}p'''' \int_{x_{i-1}}^{x_i} (x - x_{i-1})(x - x_{i-\frac{1}{2}})^2(x - x_i)\,dx = -\frac{h^5}{2880}p'''' \end{aligned}$$

Et donc :

$$E = -(b-a)\frac{h^4}{2880}p'''' = O(h^4).$$

2.5.3 Cas d'une fonction

On écrit, à x fixé :

$$f(x) = p_3(x) + \frac{f''''(\xi_x)}{24}(x - x_{i-1})(x - x_{i-\frac{1}{2}})^2(x - x_i) \tag{2.15}$$

où ξ_x dépend de x.

Et on peut montrer (calcul un peu long) que le résultat est conservé pour les fonctions C^4 : la formule de Simpson est d'ordre 4.

C'est une des formules les plus utilisées pour calculer des intégrales.

Exercice 2.6 Intégrer sin sur $[0, \pi]$ à l'aide de la méthode de Simpson sur 2 sous-intervalles puis sur 4 sous-intervalles. ▪

Exercice 2.7 Intégrer e^x sur $[0, 1]$ à l'aide de la méthode de Simpson sur 2 sous-intervalles puis sur 4 sous-intervalles. ▪

2.6 * Méthode d'intégration de Gauss

Cette méthode est également appelée méthode d'intégration de Gauss–Legendre.

Le but est le calcul de $\int_{x_{i-1}}^{x_i} f(x)\,dx$ où f est une fonction intégrable. Quitte à faire un changement de variables, on se ramène à $\int_{-1}^{1} f(x)\,dx$. Gauss propose alors la formule de la forme :

$$\int_{-1}^{1} f(x)\,dx \simeq \sum_{i=1}^{m} f(\xi_i) w_i, \tag{2.16}$$

où les $\xi_i \in [-1, 1]$ sont m points bien choisis, et les $w_i > 0$ sont les poids="weights" correspondants, de telle sorte que la formule soit exacte pour les polynômes de degré $\leq 2m - 1$. (Exemple : avec 4 points on intègre de manière exacte tout polynôme de degré ≤ 7.)

Exemple 2.8 Pour la méthode des trapèzes, $\xi_1 = -1$, $\xi_2 = 1$, $w_1 = w_2 = 1$. Mais ce choix conduit à l'exactitude de la formule (2.16) uniquement pour les polynômes de degré 1. Alors que Gauss montre qu'on peut intégrer de manière exacte les polynômes de degré 3 à l'aide de (2.16) quand on choisit de manière judicieuse ξ_1, ξ_2, w_1 et w_2. ▪

Proposition 2.9 *Soit un entier $m \geq 1$. Soit P_m le polynôme de Legendre de degré m, cf. (1.50). On note $(\xi_i)_{i=1,\ldots,m}$ ses racines et c la constante de normalisation :*

$$P_m(x) = c \prod_{i=1,\ldots,m} (x - \xi_i). \tag{2.17}$$

On note $(L_k)_{k=1,\ldots,m}$ les m polynômes de Lagrange (de degré $m-1$) associés aux m racines $(\xi_i)_{i=1,\ldots,m}$:

$$L_k(x) = \prod_{\substack{i=1,\ldots,m \\ i \neq k}} \frac{(x - \xi_i)}{(\xi_k - \xi_i)} \tag{2.18}$$

On pose, pour $i = 1, ..., m$:

$$w_i = \int_{-1}^{1} L_i(x)\,dx. \tag{2.19}$$

Alors la formule (2.16) est exacte pour tout $f = p_n$ polynôme de degré $n \leq 2m-1$.

Preuve. 1- Analyse.

11- Les points ξ_i. Si la formule (2.16) est exacte pour tout polynôme de degré $2m-1$, elle l'est en particulier pour tout polynôme de la forme $f(x) = q_{m-1}(x) \prod_{j=1}^{m} (x - \xi_j)$ où q_{m-1} est un polynôme quelconque de degré $m-1$. Et donc, on a l'égalité dans (2.16) :

$$\int_{-1}^{1} q_{m-1}(x) \prod_{j=1}^{m} (x - \xi_j)\,dx = \sum_{i=1}^{m} w_i\, q_{m-1}(\xi_i) \underbrace{\prod_{j=1}^{m} (\xi_i - \xi_j)}_{=0} = 0.$$

Donc :

$$(\prod_{j=1}^{m} (x - \xi_j), q_{m-1})_{L^2} = 0$$

pour tout polynôme de degré $m-1$, i.e. $\prod_{j=1}^{m} (x - \xi_j)$ est orthogonal à tout polynôme de degré $m-1$ pour le

produit scalaire L^2, et donc $\prod_{j=1}^m (x-\xi_j)$ est proportionnel au polynôme de Legendre P_m de degré m, cf. (2.17) : les ξ_j sont nécessairement les racines de P_m.

12- Les poids w_i. Si la formule (2.16) est exacte pour tout polynôme de degré $2m-1$, elle l'est pour les polynômes L_k donnés en (2.18) (de degré $m-1$). Comme $L_k(\xi_k)=1$ et $L_k(\xi_j)=0$ pour tout $j\neq k$, on a :

$$\int_{-1}^{1} L_k(x)\,dx = \sum_{j=1}^{m} L_k(\xi_j)w_j = \sum_{j=1}^{m} \delta_{kj} w_j = w_k,$$

i.e. les poids w_k pour $1\leq k\leq m$ sont donnés par (2.19).

2- Synthèse. Notons ξ_i et w_i les points et poids trouvés ci-dessus, et vérifions que (2.16) est exacte pour tout polynôme de degré $2m-1$.

21- Montrons que la formule est exacte pour les polynômes de degré $\leq m-1$. Comme $\mathcal{P}_{m-1} = \mathrm{Vect}\{1,x,...,x^{m-1}\} = \mathrm{Vect}\{L_1,...,L_m\}$, si q_{m-1} est un polynôme de degré $\leq m-1$, alors q_{m-1} est combinaison linéaire de L_k, et comme $L_k(\xi_i)=\delta_{ki}$ on a $q_{m-1}(x)=\sum_{i=1}^m q_m(\xi_i)L_i(x)$ (voir polynômes de Lagrange). Et donc :

$$\int_{-1}^{1} q_{m-1}(x)\,dx = \sum_{i=1}^{m} q_{m-1}(\xi_i)\int_{-1}^{1} L_i(x)\,dx = \sum_{i=1}^{m} q_{m-1}(\xi_i)w_i,$$

et donc la formule (2.16) est exacte pour tout polynôme de degré $\leq m-1$.

22- Montrons que la formule est exacte pour les polynômes de degré $\leq 2m-1$.

Soit q_{2m-1} un polynôme de degré $\leq 2m-1$. La division euclidienne (cf. (1.4)) de q_{2m-1} par $\prod_{i=1}^m (x-\xi_i)$ (polynôme de degré m) donne :

$$q_{2m-1}(x) = q_{m-1}(x)\prod_{i=1}^{m}(x-\xi_i) + r_{m-1}(x), \tag{2.20}$$

où q_{m-1} et r_{m-1} tous deux de degré $\leq m-1$. D'où :

$$\int_{-1}^{1} q_{2m-1}(x)\,dx = (q_{m-1}, \prod_{j=1}^{m}(x-\xi_j))_{L^2} + \int_{-1}^{1} r_{m-1}(x)\,dx = 0 + \sum_{j=1}^{m} r_{m-1}(\xi_j)w_j.$$

Et (2.20) donne $q_{2m-1}(\xi_i) = r_{m-1}(\xi_i)$. Et (2.16) est exacte pour q_{2m-1}. ▪

Exemple 2.10 On vérifiera (voir [1] pour les valeurs suivantes avec 15 chiffres après la virgule) :

	$\pm\xi_i$	w_i
m=1	0.0	2.0
m=2	0.57735 02691 89626	1.00000 00000 00000
m=3	0.0	0.88888 88888 88889
	0.77459 66692 41483	0.55555 55555 55556
m=4	0.33998 10435 84856	0.65214 51548 62546
	0.86113 63115 94053	0.34785 48451 37454

la ligne $m=4$ permettant d'intégrer de manière exacte tout polynôme de degré $2m-1=7$. ▪

Remarque 2.11 Pour calculer $\int_a^b g(x)\,dx$ par la méthode de Gauss, on commence par faire le changement de variable $y=\frac{2x-a-b}{b-a}$, d'où avec $x=\frac{1}{2}((b-a)y+a+b)$:

$$\int_{x=a}^{b} g(x)\,dx = \int_{y=-1}^{1} g(\frac{(b-a)y+a+b}{2})\,\frac{b-a}{2}\,dy = \int_{y=-1}^{1} f(y)\,dy,$$

et on applique la formule (2.16) à la fonction $f: y\to f(y)=\frac{b-a}{2}g(\frac{(b-a)y+a+b}{2})$. ▪

Remarque 2.12 Si l'intervalle $[a,b]$ est "grand", on partitionne l'intervalle, et on calcule $\int_a^b f(x)\,dx$ comme $=\sum_{i=1}^N(\int_{x_{i-1}}^{x_i} f(x)\,dx)$. Par changement de variable, on calcule chaque $\int_{x_{i-1}}^{x_i} f(x)\,dx$ en se ramenant au calcul à $\int_{y=-1}^{1}$. Et donc, avec $h=x_i-x_{i-1}$ supposé constant :

$$\int_a^b f(x)\,dx = \sum_{i=1}^{N}\int_{-1}^{1} f(\frac{h}{2}y+x_{i-\frac{1}{2}})\,\frac{h}{2}\,dy \simeq \frac{h}{2}\sum_{i=1}^{N}\sum_{j=1}^{m} w_j\, f(\xi_j\frac{h}{2}+x_{i-\frac{1}{2}}),$$

où on a noté $x_{i-\frac{1}{2}}=\frac{x_{i-1}+x_i}{2}$ le milieu de x_{i-1} et de x_i. ▪

Remarque 2.13 Rappel. Toute fonction f s'écrit comme la somme d'une fonction paire f_p et d'une fonction impaire f_i :

$$f(x) = f_p(x) + f_i(x) \quad \text{où} \quad f_p(x) = \frac{f(x)+f(-x)}{2}, \quad f_i(x) = \frac{f(x)-f(-x)}{2}.$$

Et donc :

$$\int_{-1}^{1} f(x)\,dx = \int_{-1}^{1} f_p(x)\,dx + \int_{-1}^{1} f_i(x)\,dx = 2\int_{0}^{1} f_p(x)\,dx + 0.$$

▪

Exercice 2.14 Proposer des calculs simples pour déterminer les points de poids de Gauss permettant d'intégrer de manière exacte les polynômes de degré ≤ 5.

Réponse. Commençons par la remarque suivante : si f est impaire (i.e. $f(x) = -f(-x)$ pour tout x), alors $\int_{-1}^{1} f(x)\,dx = 0$. Et donc, si la formule (2.16) est vraie pour tout polynôme de degré $2k$ est l'est pour tout polynôme de degré $2k+1$ (ajout d'un monôme impaire αx^{2k+1}).

Et pour une fonction paire on a $\int_{-1}^{1} f_p(x)\,dx = 2\int_{0}^{1} f_p(x)\,dx$. Il est alors naturel d'essayer de prendre les points ξ_i symétriques par rapport à 0 avec des poids identiques aux points symétriques :

$$\text{pour } i = 1, ..., m \ : \qquad \xi_{m+1-i} = -\xi_i, \qquad w_{m+1-i} = w_i.$$

On aura alors :

$$\int_{-1}^{1} p_n(x)\,dx = \begin{cases} \displaystyle\sum_{i=1}^{\frac{m}{2}} (p_n(\xi_i) + p_n(-\xi_i))w_i, & \text{si } m \text{ est pair,} \\ \\ \displaystyle p_n(0)w_0 + \sum_{i=1}^{\frac{m-1}{2}} (p_n(\xi_i) + p_n(-\xi_i))w_i, & \text{si } m \text{ est impair,} \end{cases}$$

Pour $n = 0$ (et $= 1$), on veut intégrer de manière exacte les polynômes de degré 0, i.e. de la forme $p_0 : x \to p_0(x) = a_0$. On prend alors un seul point ξ_1, ce qui impose par symétrie $\xi_1 = 0$, et un seul poids w_1, et on veut que $\int_{-1}^{1} p_0(x)\,dx = a_0 w_1$ avec $\int_{-1}^{1} p_0(x)\,dx = \int_{-1}^{1} a_0\,dx = 2a_0$. D'où $w_1 = 2$. Sachant que $\int_{-1}^{1} x\,dx = 0$, avec le point $x_1 = 0$ et le poids $w_1 = 2$, la formule (2.16) est exacte pour tout polynôme de degré $n \leq 1$.

Pour $n = 2$ (et $= 3$), on veut intégrer de manière exacte les polynômes de degré 2 de la forme $p_2 : x \to p_2(x) = a_0 + a_2 x^2$. On prend alors deux points ξ_1 et ξ_2, et par symétrie on impose $\xi_1 = -\xi_2$ et $w_1 = w_2$, soit 2 équations. Et on veut que (2.16) soit exacte pour tout polynôme de degré ≤ 2 donc pour les polynômes $x \to 1$ (ce qui donne une 3-ième équation) et $x \to x^2$ (ce qui donne une 4-ième équation) : avec $\int_{-1}^{1} 1\,dx = 2$ et avec $\int_{-1}^{1} x^2\,dx = \frac{2}{3}$ on obtient $w_1 + w_2 = 2$ et $\xi_1^2 w_1 + \xi_2^2 w_2 = \frac{2}{3}$, d'où $w_1 = w_2 = 1$ et $\xi_2 = -\xi_1 = \sqrt{\frac{1}{3}}$. Et $x \to x^3$ étant impaire, la formule est exacte pour tout polynôme de degré $n \leq 3$.

Pour $n = 4$ (et $= 5$), on obtient $\xi_1 = -\xi_3$, $\xi_2 = 0$, $w_1 = w_3$, $w_1 + w_2 + w_3 = 2$, $w_1\xi_1^2 + w_2\xi_2^2 + w_3\xi_3^2 = \frac{2}{3}$ et $w_1\xi_1^4 + w_2\xi_2^4 + w_3\xi_3^4 = \frac{2}{5}$ $(= \int_{-1}^{1} x^4\,dx)$. On a 6 équations et 6 inconnues, système (non linéaire) qu'on peut résoudre, et qui donne l'existence et l'unicité.

Pour n quelconque, la résolution du système non linéaire obtenu n'est pas simple. ▪

2.7 * Noyau de Péano

2.7.1 Rappel : développement de Taylor avec reste intégral

Pour $g : \mathbb{R} \to \mathbb{R}$, on note $g_+ = \sup(0, g)$ la partie positive de g :

$$g_+(y) = \sup(0, g(y)) = \begin{cases} g(y) & \text{si } g(y) \geq 0, \\ 0 & \text{si } g(y) \leq 0. \end{cases} \tag{2.21}$$

En particulier, à c fixé :

$$y \to (c-y)_+ = \left\{ \begin{array}{rl} c-y & \text{si } y \leq c, \\ 0 & \text{si } y \geq c, \end{array} \right\} = (c-y)1_{]-\infty,c]}(y), \tag{2.22}$$

faire un dessin. De même :

$$z \to (z-c)_+ = \left\{ \begin{array}{rl} z-c & \text{si } z \geq c, \\ 0 & \text{si } z \leq c, \end{array} \right\} = (z-c)1_{[c,\infty[}(z), \tag{2.23}$$

Et on notera, pour $n \in \mathbb{N}$:

$$(c-y)_+^n \stackrel{\text{déf}}{=} ((c-y)_+)^n \qquad \text{et} \qquad (z-c)_+^n \stackrel{\text{déf}}{=} ((z-c)_+)^n. \tag{2.24}$$

Proposition 2.15 *(Formule de Taylor avec reste intégrale). Soit $\alpha \in \mathbb{R}$ et $f \in C^{n+1}(\mathbb{R})$. Le développement limité de f au voisinage de α est donné par :*

$$f(x) = \sum_{k=0}^{n} \frac{(x-\alpha)^k}{k!} f^k(\alpha) + \frac{1}{n!} \int_{t=\alpha}^{x} (x-t)^n f^{n+1}(t)\, dt, \tag{2.25}$$

en particulier, pour $\beta \geq x \geq \alpha$:

$$f(x) = \sum_{k=0}^{n} \frac{(x-\alpha)^k}{k!} f^k(\alpha) + \frac{1}{n!} \int_{t=\alpha}^{\beta} (x-t)_+^n f^{n+1}(t)\, dt. \tag{2.26}$$

Preuve. Connu (démonstration directe par récurrence) :

$f(x) = f(\alpha) + \int_\alpha^x f'(t)\, dt$, puis, par intégration par parties,

$\int_\alpha^x 1.f'(t)\, dt = \int_\alpha^x (-1).(-f'(t))\, dt = \int_\alpha^x u'(t).v(t)\, dt = [u(t)v(t)]_\alpha^x - \int_\alpha^x u(t).v'(t)\, dt$, où on a posé $u'(t) = -1$ et $v(t) = -f'(t)$, d'où $v'(t) = -f''(t)$ et on prend $u(t) = x-t$ (ici x est fixé).

D'où $\int_\alpha^x 1.f'(t)\, dt = [(x-t)(-f'(t))]_\alpha^x + \int_\alpha^x (x-t).f''(t)\, dt = (x-\alpha)f'(\alpha) + \int_\alpha^x (x-t).f''(t)\, dt$.

Puis de même $\int_\alpha^x (x-t).f''(t)\, dt = \frac{(x-\alpha)^2}{2} f''(\alpha) + \int_\alpha^x \frac{(x-t)^2}{2}.f'''(t)\, dt$. Puis récurrence.

Puis $\int_\alpha^x g(t)\, dt = \int_\alpha^\beta g(t) 1_{]-\infty,x]}(t)\, dt$ quand $\beta \geq x \geq \alpha$, et $g(t) = (x-t)^n f^{n+1}(t)$ donne $(x-t)^n f^{n+1}(t) 1_{]-\infty,x]}(t) = (x-t)_+^n f^{n+1}(t)$. ∎

2.7.2 Calcul de l'erreur avec le noyau de Péano

On pose, pour $-\infty < a < b < \infty$, quand f est intégrable sur $[a,b]$:

$$I(f) = \int_a^b f(x)\, dx, \tag{2.27}$$

et on note, pour des $x_i \in \mathbb{R}$ et des $\lambda_i \in \mathbb{R}$, lorsque la fonction f est définie en tous les x_i supposés deux à deux distincts :

$$I_n(f) = \sum_{k=0}^{n} \lambda_k f(x_k), \tag{2.28}$$

la valeur approchée par une formule d'intégration numérique (donc $I(f) \simeq I_n(f)$).

On note $E(f)$ l'erreur commise :

$$E(f) = I(f) - I_n(f) \tag{2.29}$$

La fonction $E : C^{n+1} \to \mathbb{R}$ est trivialement linéaire puisque I et I_n le sont. Et si $E(q) = 0$ pour tout polynôme q de degré $\leq n$, posant $g(x) = \int_{t=a}^b (x-t)_+^n f^{n+1}(t)\, dt$, avec (2.26) où $\alpha = a$, $\beta = b$, $q_n(x) = \sum_{k=0}^n \frac{(x-a)^k}{k!} f^k(a)$, on a :

$$E(f) = E(q_n) + \frac{1}{n!} E(g) = 0 + \frac{1}{n!}(I(g) - I_n(g)). \tag{2.30}$$

Définition 2.16 Le noyau de Péano est la fonction :

$$K_n(t) = E(x \mapsto (x-t)_+^n), \tag{2.31}$$

i.e. c'est la fonction erreur commise $E(f)$ pour $f(x) = (x-t)_+^n$, i.e. :

$$K_n(t) = \int_{x=a}^b (x-t)_+^n\, dx - \sum_{k=0}^n \lambda_k (x_k - t)_+^n. \tag{2.32}$$

(Attention aux notations des variables.) Soit encore :

$$K_n(t) = \int_{x=t}^b (x-t)^n\, dx - \sum_{k=0}^n \lambda_k (x_k - t)_+^n. \tag{2.33}$$

Proposition 2.17 *Supposant $E(q) = 0$ pour tout polynôme q de degré $\leq n$ (la formule d'intégration numérique (2.28) est supposée exacte pour les polynômes de degré $\leq n$), pour $f \in C^{n+1}([a,b])$ on a :*

$$E(f) = \frac{1}{n!} \int_{t=a}^b K_n(t) f^{n+1}(t)\, dt \tag{2.34}$$

Preuve. On intègre (2.26) : avec (2.30) :

$$\begin{aligned} n!\,E(f) &= \int_{x=a}^{b} \Big(\int_{t=a}^{b} (x-t)_+^n f^{n+1}(t)\,dt\Big)\,dx - \sum_{k=0}^{n} \lambda_k \int_{t=a}^{b} (x_k-t)_+^n f^{n+1}(t)\,dt \\ &= \int_{t=a}^{b} f^{n+1}(t)\Big(\int_{x=a}^{b} (x-t)_+^n\,dx - \sum_{k=0}^{n} \lambda_k (x_k-t)_+^n\Big)\,dt = \int_{t=a}^{b} K_n(t) f^{n+1}(t)\,dt. \end{aligned}$$

On a pu appliquer le théorème de Fubini car les fonctions sont toutes continues sur $[a,b]$. ∎

On en déduit immédiatement l'ordre des méthodes d'intégrations numériques de type (2.28) :

Corollaire 2.18

$$|E(f)| \le \frac{\int_a^b |K_n(t)|\,dt}{n!} \max_{t\in[a,b]} |f^{(n+1)}(t)|. \tag{2.35}$$

Exemple 2.19 Sur $[a,b] = [0,1]$, et la formule pour Riemann à gauche, calculer le noyau de Péano et $\int_a^b |K_n(t)|\,dt$.

Réponse. Ici dans (2.28) on a $n=0$, $x_0=0$, $\lambda_0=1$, et $I_0(f)=f(0)$. Avec (2.33) on a $K_0(t) = \int_t^1 dx - 0$ pour $t\in[0,1]$, soit $K_0(t) = 1-t$. D'où $\int_0^1 |K_0(t)|\,dt = \frac{1}{2}$. ∎

Exemple 2.20 Sur $[a,b] = [0,1]$, et la formule pour Riemann à droite, calculer le noyau de Péano et $\int_a^b |K_n(t)|\,dt$.

Réponse. Ici dans (2.28) on a $n=0$, $x_0=1$, $\lambda_0=1$, et $I_0(f)=f(1)$. Avec (2.33) on a $K_0(t) = \int_t^1 dx - \lambda_0$ pour $t\in[0,1]$, soit $K_0(t) = 1-t-1 = -t$. D'où $\int_0^1 |K_0(t)|\,dt = \frac{1}{2}$. ∎

Exemple 2.21 Sur $[a,b]=[0,1]$, et la formule du trapèze, calculer le noyau de Péano et $\int_a^b |K_n(t)|\,dt$.

Réponse. Ici $I_1(f) = \frac{f(0)+f(1)}{2}$: $n=1$, $x_0=0$, $x_1=1$, $\lambda_0=\lambda_1=\frac{1}{2}$. Avec (2.33) on a $K_1(t) = \int_t^1 (x-t)\,dx - \frac{1}{2}(x_0-t)_+ - \frac{1}{2}(x_1-t)_+ = \frac{(1-t)^2}{2} - \frac{1}{2}(1-t) = \frac{(1-t)}{2}(-t)$ pour $t\in[0,1]$. D'où $\int_0^1 |K_1(t)|\,dt = \frac{1}{2}\int_0^1 t(1-t)\,dt = \frac{1}{2}(\frac{1}{2}-\frac{1}{3}) = \frac{1}{12}$. ∎

Exemple 2.22 Sur $[a,b]=[0,1]$, et la formule du point milieu, calculer le noyau de Péano et $\int_a^b |K_n(t)|\,dt$.

Réponse. Ici dans (2.28) on a $n=0$, $x_0=\frac{1}{2}$, $\lambda_0=1$, et $I_0(f)=f(\frac{1}{2})$. La méthode du point milieu est exacte pour les polynômes d'ordre 1. Avec (2.33) on a $K_1(t) = \int_t^1 (x-t)\,dx - \lambda_0(\frac{1}{2}-t)1_{[-\infty,\frac{1}{2}]}(t)$, soit $K_1(t) = \frac{(1-t)^2}{2} - (\frac{1}{2}-t)1_{[0,\frac{1}{2}]}(t)$ pour $t\in[0,1]$. Soit $K_1(t) = \frac{(1-t)^2}{2} - (\frac{1}{2}-t) = \frac{t^2}{2}$ pour $t\in[0,\frac{1}{2}]$ et $K_1(t) = \frac{(1-t)^2}{2}$ pour $t\in[\frac{1}{2},1]$. D'où $\int_0^1 |K_1(t)|\,dt = \frac{1}{6}(\frac{1}{2})^3 - [\frac{(1-t)^3}{6}]_{\frac{1}{2}}^1 = 2\frac{1}{6}(\frac{1}{2})^3 = \frac{1}{24}$. ∎

2.8 Exercices

Exercice 2.23 Soit $f\in C^1([0,1])$. On veut estimer $I(f) = \int_0^1 f(x)\,dx$. On pose :

$$J(f) = w_0 f(0) + w_1 f'(0) + w_2 f'(\xi), \tag{2.36}$$

où $\xi\in]0,1]$ et les $w_i\in\mathbb{R}$.

1- Déterminer ξ et les w_i pour que $J(f)=I(f)$ pour tout polynôme de degré ≤ 3.

2- Soit $E(f) = I(f) - J(f)$. Calculer $E(x\to x^4)$ et en déduire l'ordre de la méthode.

3- Déterminer le noyau de Péano. ∎

Exercice 2.24 On reprend les polynômes $(\varphi_i)_{i=1,\dots,4}$ donnés en (1.24). On veut estimer $I(f) = \int_0^1 f(x)\,dx$. On pose pour $f\in C^1([0,1])$:

$$J(f) = w_1 f(0) + w_2 f'(0) + w_3 f(1) + w_4 f'(1). \tag{2.37}$$

1- Déterminer les w_i pour que J soit exacte pour tout polynôme de degré ≤ 3 (on montrera que $w_i = \int_0^1 \varphi_i(x)\,dx$).

2- Pour $f\in C^1$, montrer que $J(f) = J(p_f)$ où p_f est le polynôme (d'interpolation) de degré ≤ 3 vérifiant $p_f(0)=f(0)$, $p_f'(0)=f'(0)$, $p_f(1)=f(1)$, $p_f'(1)=f'(1)$. ∎

Exercice 2.25 Soit $f\in C^3([-1,1];\mathbb{R})$. Soit :

$$I(f) = \int_{-1}^{1} f(x)\,dx, \qquad J(f) = \frac{1}{3}(f(-1)+4f(0)+f(1)). \tag{2.38}$$

1- Montrer que $J(p)=I(p)$ pour tout polynôme p de degré ≤ 2.

2- Soit P le polynôme d'interpolation de Lagrange de f aux point $-1, 0, 1$. Donner les polynômes de Lagrange de base et l'expression de P dans cette base.

3- Soit $x \in]-1,1[$ et soit $g_x(t) = f(t) - P(t) - \dfrac{f(x)-P(x)}{x(x^2-1)}\, t\,(t^2-1)$. Montrer qu'il existe $\xi \in]-1,1[$ t.q. $g'''(\xi) = 0$. (On montrera que g s'annule en 4 points.)

4- En déduire une majoration de $||f-P||_\infty$ en fonction de $|||f'''|||_\infty$, puis une majoration de $|I(f)-J(f)|$. ▪

Exercice 2.26 1- Donner les trois polynômes P_1, P_2, P_3 de degré 2 t.q. :

$$\begin{cases} P_1(a) = 1 \\ P_1'(a) = 0 \\ P_1(b) = 0 \end{cases} \qquad \begin{cases} P_2(a) = 0 \\ P_2'(a) = 1 \\ P_2(b) = 0 \end{cases} \qquad \begin{cases} P_3(a) = 0 \\ P_3'(a) = 0 \\ P_3(b) = 1 \end{cases}$$

et montrer qu'il forment une base de $\mathcal{P}_2$. (Voir (1.22).)

2- Soit $f \in C^3([a,b];\mathbb{R})$. Montrer qu'il existe un unique polynôme $P \in \mathcal{P}_2$ t.q. :

$$P(a) = f(a), \qquad P'(a) = f'(a), \qquad P(b) = f(b),$$

et l'exprimer sur la base (P_1, P_2, P_3).

3- Calculer $J = \int_a^b P(x)\,dx$.

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

5- Soit $I = \int_a^b f(x)\,dx$. Donner une estimation de l'erreur $|I-J|$. ▪

3 Résolution numérique des équations différentielles

3.1 L'équation différentielle considérée

Soit E un espace de Banach (dans la suite on prendra $E = \mathbb{R}^n$, et souvent $E = \mathbb{R}$).

Soit $f : \left\{ \begin{array}{ll} \mathbb{R} \times E & \to E \\ (t,x) & \to f(t,x) \end{array} \right\}$ une fonction qu'on suppose lipschitzienne uniformément en t :

$$\exists L > 0, \quad \forall (t, x_1, x_2) \in [t_0, T] \times E^2, \quad ||f(t, x_1) - f(t, x_2)||_E \leq L||x_1 - x_2||_E.$$

(Les pentes moyennes $\frac{||f(t,x_1)-f(t,x_2)||_E}{||x_1-x_2||_E}$ sont uniformément bornées une constante L.)

Et on cherche une fonction $u : [t_0, T] \to E$ qui vérifie l'équation implicite

$$\left\{ \begin{array}{l} \dfrac{du}{dt}(t) = f(t, u(t)), \quad \forall t \in [t_0, T], \\ u(t_0) = u_0 \quad \text{(condition initiale)}, \end{array} \right\} \quad = \text{E.D.O.} = \text{Equation Différentielle Ordinaire.} \tag{3.1}$$

Le théorème de Cauchy–Lipschitz donne l'existence et l'unicité d'une solution $u \in C^1([t_0, T]; E)$.

Remarque 3.1 L'équation différentielle (3.1) est souvent notée :

$$\frac{dx}{dt} = f(t, x), \quad x(t_0) = x_0.$$

Attention à cette notation : ici x est une fonction $t \to x(t)$, et c'est bien $\frac{dx}{dt}(t) = f(t, x(t))$ qu'il faut lire. Pour éviter toute confusion, on appellera u (et non x) la fonction inconnue cherchée. ▪

On ne peut pas espérer, en général, trouver une expression simple de la solution u de (3.1). On va donc en chercher une solution approchée.

3.2 Méthodes à un pas

3.2.1 Notations

On va chercher une solution approchée $\tilde{u}$ de u qui vérifie une équation approchée :

$$\frac{d\tilde{u}}{dt}(t) \simeq \tilde{f}(t, \tilde{u}(t)), \quad \tilde{u}(t_0) \simeq \tilde{u}_0,$$

où $\tilde{f}$, $\tilde{u}_0$ approchent f et u_0 (on n'oubliera pas par exemple les erreurs d'arrondi).

Soit $n \in \mathbb{N}^*$. On partitionne $[t_0, T]$ en $\bigcup_{k=1,\dots,n}[t_{k-1}, t_k]$, avec $t_{k-1} < t_k$ et $t_n = T$, et on note $h_k = t_k - t_{k-1}$ le k-ème pas de temps. Et on notera $\tilde{u}_k := \tilde{u}(t_k)$. Et pour des temps intermédiaires moitiés par exemple, on notera :

$$t_{k+\frac{1}{2}} = \frac{t_k + t_{k+1}}{2} \ (= t_k + \frac{h_k}{2}), \quad \text{et} \quad \tilde{u}_{k+\frac{1}{2}} = \tilde{u}(t_{k+\frac{1}{2}}).$$

Remarque 3.2 Exemple d'un maillage uniforme : intervalles de temps de même longueur, i.e. $h_k = h = \frac{T-t_0}{n}$ et $t_k = t_0 + kh$.

Dans les codes industriels, pour lesquels les temps de calcul doivent être le plus faible possible, on ne prend pas h constant. Prendre h constant permet d'alléger les notations (et de faire les calculs d'erreurs). On renvoie à Crouzeix et Mignot [5] (par exemple) pour le choix optimal des pas de temps. ▪

3.2.2 Méthode d'Euler (explicite)

L'idée d'Euler est la suivante : à l'instant initial $t = t_0$ on connaît la position $u_0 = u(t_0)$, donc on connaît $f(t_0, u_0)$, donc on connaît $u'(t_0) = f(t_0, u_0)$ la pente en t_0. Et connaissant la position et la pente au temps t_0, on évalue les valeurs suivantes $u(t)$ (pour $t > t_0$) à l'aide du développement limité au premier ordre

$$u(t_0+h) = \underbrace{u(t_0) + hu'(t_0)}_{\text{partie affine}} + o(h) = \underbrace{u(t_0) + hf(t_0, x_0)}_{\text{partie affine}} + o(h) \tag{3.2}$$

en gardant la partie affine de u ; i.e., on pose

$$t_1 = t_0 + h_1 \quad \text{et} \quad \tilde{u}(t_1) = \tilde{u}(t_0) + h_1 f(t_0, \tilde{u}_0), \quad \text{noté} \quad \tilde{u}_1 = \tilde{u}_0 + h_1 f(t_0, \tilde{u}_0). \tag{3.3}$$

C'est la première étape de la méthode d'Euler explicite (ou méthode d'Euler progressive). (Dans la pratique $\tilde{u}(t_0)$ vaut $u(t_0)$ aux erreurs d'arrondi ou de mesure près.)

Puis à partir de la position estimée $\tilde{u}(t_1)$, on calcule la pente estimée $f(t_1, \tilde{u}(t_1))$, on pose

$$t_2 = t_1 + h_2 \quad \text{et} \quad \tilde{u}(t_2) = \tilde{u}(t_1) + h_2\, f(t_1, \tilde{u}(t_1)), \quad \text{noté} \quad \tilde{u}_2 = \tilde{u}_1 + h_2 f(t_1, \tilde{u}_1). \tag{3.4}$$

Et on poursuit le processus, ce qui permet de calculer $\tilde{u}(t_{k+1})$ en fonction de $\tilde{u}(t_k)$:

$$t_{k+1} = t_k + h_{k+1} \quad \text{et} \quad \tilde{u}_{k+1} = \tilde{u}_k + h_{k+1}\, f(t_k, \tilde{u}_k). \tag{3.5}$$

D'où l'algorithme élémentaire de calcul de la solution $\tilde{u}$ approchée, donné ici pour un pas de temps $h = h_k$ constant :

1. Initialisation : on se donne la fonction $f : \mathbb{R}^2 \to \mathbb{R}$, le temps initial t_0, le temps final T, la condition initiale (approchée) $\tilde{u}_0$.
2. On se donne $n \in \mathbb{N}^*$ et le pas temps $h = \frac{T-t_0}{n}$ (ici maillage temporel uniforme).
3. Boucle pour $k = 0$ à $n-1$:

$$\begin{cases} p_g = f(t_k, \tilde{u}_k) \quad \text{(pente à gauche approchée dans } [t_k, t_{k+1}]), \\ \tilde{u}_{k+1} = \tilde{u}_k + h\, p_g \quad \text{(valeur approchée à } t_{k+1}), \\ t_{k+1} = t_k + h \quad \text{(temps suivant).} \end{cases} \tag{3.6}$$

On a ainsi calculé des valeurs ponctuelles $\tilde{u}_k$ de $\tilde{u}$ aux points t_k, et la solution approchée $\tilde{u}$ a pour graphe la fonction affine par morceaux qui passe par les points $(t_k, \tilde{u}_k)$ pour $0 \le k \le n$.

Méthode très facile à mettre en œuvre et qu'on utilise en première approche, au moins pour déboguer.

Remarque 3.3 Malheureusement, cet algorithme très simple n'est pas toujours "performant" : 1- pour avoir une bonne approximation de u, on doit parfois choisir h très petit, ou si on préfère n très grand, et le calcul peut être coûteux, et 2- cette méthode n'est pas inconditionnellement A-stable, voir paragraphe 3.5.1 : si h n'est pas suffisamment petit, on peut obtenir une solution absurde.

Ces problèmes de précision et de stabilité (ou d'instabilité) sont traités dans la suite. ▞

Autre présentation. On approxime la pente $u'(t_0) = \frac{du}{dt}(t_0)$ dans l'intervalle $[t_0, t_0{+}h]$ pour $h > 0$:

$$\frac{du}{dt}(t_0) = f(t_0, u(t_0)) = \lim_{h \to 0,\, h>0} \frac{u(t_0{+}h) - u(t_0)}{h} \text{ (pente à gauche dans } [t_0, t_0{+}h]), \tag{3.7}$$

donc $f(t_0, u(t_0)) \simeq \frac{u(t_0+h)-u(t_0)}{h}$, et, connaissant $\tilde{u}(t_0)$ (une approximation de $u(t_0)$ imposée comme condition initiale) on définit $\tilde{u}(t_0{+}h)$ par

$$f(t_0, \tilde{u}(t_0)) = \frac{\tilde{u}(t_0{+}h) - \tilde{u}(t_0)}{h}. \tag{3.8}$$

On a retrouvé (3.8).

Remarque 3.4 La méthode d'Euler explicite découle également immédiatement de la méthode d'intégration de Riemann à gauche :

$$u(t_1) - u(t_0) = \int_{t_0}^{t_1} u'(t)\, dt \simeq \int_{t_0}^{t_1} u'(t_0)\, dt = u'(t_0)(t_1 - t_0) = f(t_0, u(t_0))(t_1 - t_0) \overset{\text{noté}}{=} \tilde{u}(t_1) - \tilde{u}(t_0).$$

Et on retrouve (3.8) et (3.3). (On dit que "résoudre une E.D.O." c'est "intégrer l'E.D.O.".) ▞

Exercice 3.5 Soit $f : \mathbb{R}^2 \to \mathbb{R}$ donnée par $f(x, y) = x$. Donner la solution exacte de (3.1) (ici on sait la calculer). Résoudre de manière approchée (3.1) à l'aide du schéma d'Euler explicite. Comparer les solutions à T.

Notations imposées : $n \in \mathbb{N}^*$, $h = \frac{T-t_0}{n}$ (le pas de temps), $t_k = t_0 + kh$ et $\tilde{u}_k$ les valeurs approchées de $u(t_k)$.

Réponse. Ici $u : [t_0, T] \to \mathbb{R}$ est solution de $u'(t) = t$ avec $u(t_0) = u_0$, donc $u(t) = u(t_0) + \int_{t_0}^{t} \tau\, d\tau = u_0 + \frac{1}{2}(t^2 - t_0^2)$.

Schéma d'Euler explicite : $\tilde{u}_{k+1} = \tilde{u}_k + hf(t_k, \tilde{u}_k) = \tilde{u}_k + ht_k$, pour $k = 0, ..., n{-}1$.

Donc $\tilde{u}_n = \tilde{u}_{n-1} + ht_{n-1} = \tilde{u}_{n-2} + ht_{n-2} + ht_{n-1} = ... = \tilde{u}_0 + h(t_0 + t_1 + ... + t_{n-1})$. Avec $(t_k)_{k=0,...,n-1}$ suite arithmétique de raison h donc $t_0 + t_1 + ... + t_{n-1} = \frac{n}{2}(t_0 + t_{n-1}) = \frac{n}{2}(t_0 + T - h)$. Donc $\tilde{u}_n = \tilde{u}_0 + h\frac{n}{2}(t_0 + T - h)$ avec $nh = T - t_0$, donc Donc $\tilde{u}_n = \tilde{u}_0 + \frac{1}{2}(T - t_0)(T + t_0 - h)$,

A comparer avec la solution exacte $u(T) = u_0 + \frac{1}{2}(T - t_0)(T + t_0)$: erreur $u(T) - \tilde{u}_n$ d'ordre h. ▞

Exercice 3.6 Soit $f : \mathbb{R}^2 \to \mathbb{R}$ donnée par $f(x, y) = y$. Mêmes questions.

Réponse. Ici $u : [t_0, T] \to \mathbb{R}$ est solution de $u'(t) = u(t)$ avec $u(t_0) = u_0$, donc $u(t) = u_0 e^{t-t_0}$.

Schéma d'Euler explicite, pour $k = 0, ..., n{-}1$: on a $\tilde{u}_{k+1} = \tilde{u}_k + h\tilde{u}_k = (1 + h)\tilde{u}_k$.

Donc $\tilde{u}_n = (1 + h)^n \tilde{u}_0 = (1 + \frac{T-t_0}{n})^n \tilde{u}_0$.

A comparer avec la solution exacte : $(1 + \frac{T-t_0}{n})^n \longrightarrow_{n\to\infty} e^{T-t_0}$ (convergence quand $n \to \infty$, i.e. $h \to 0$). ▞

Exercice 3.7 Soit $f : \mathbb{R}^2 \to \mathbb{R}$ donnée par $f(x, y) = y^2$. Mêmes questions.

Réponse. Fonction inconnue : $u : [t_0, T] \to \mathbb{R}$ solution de $u'(t) = u(t)^2$ et $u(t_0) = u_0$. Solution formelle par séparation des variables dans le cas $u_0 \neq 0$ (si $u_0 = 0$ alors la fonction nulle est une solution mais attention aux erreurs d'arrondi). On a $\frac{du}{dt} = u^2$, d'où $\frac{du}{u^2} = dt$, d'où $\frac{-1}{u} = t + c$, soit $u = -\frac{1}{t+c}$, où c est la constante d'intégration donnée à l'aide de la condition initiale : $c = \frac{-1}{u_0} - t_0$. Vérifions que ce calcul formel donne la solution : on pose $u(t) = -\frac{1}{t+c}$ avec $c = \frac{-1}{u_0} - t_0$, quand t est dans l'un des intervalles $]-\infty, -c[$ ou $]-c, \infty[$. Dans ces intervalles, $u'(t) = \frac{1}{(t+c)^2}$. Et on a bien $u'(t) = u(t)^2$.

Cette solution n'a de sens que si : 1- soit $]-c, \infty[\ni t_0$, i.e. $-c = \frac{1}{u_0} + t_0 < t_0$, i.e. $u_0 < 0$; 2- soit $]-\infty, -c[\ni t_0$, i.e. $-c = \frac{1}{u_0} + t_0 > t_0$, i.e. $u_0 > 0$, et alors le calcul s'arrête à la limite à $T = -c$ (valeur infinie pour $u(T)$).

Schéma d'Euler explicite, pour $k = 0, ..., n-1$: on a $\tilde{u}_{k+1} = \tilde{u}_k + h\tilde{u}_k^2 = \tilde{u}_k(1 + h\tilde{u}_k)$.

Comparaison graphique. ▪

Exercice 3.8 Résoudre de manière approchée à l'aide du schéma d'Euler explicite l'équation différentielle du second ordre $x'' + a_1 x' + a_2 x = g$ dans $[0, T]$, avec $x(0) = x_0$, $x'(0) = v_0$ et g une fonction continue donnés.

Réponse. La fonction inconnue est la fonction $x : [0, T] \to \mathbb{R}$ solution de $x''(t) + a_1 x'(t) + a_2 x(t) = g(t)$ avec les conditions initiales $x(0) = x_0$ et $x'(0) = v_0$. On transforme l'ED de 2nd ordre sous forme Système Différentiel (SD) du premier ordre. On pose $\vec{X}_0 = \begin{pmatrix} x_0 \\ v_0 \end{pmatrix} \in \mathbb{R}^2$ (la condition initiale vectorielle), $\vec{X} = \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \\ x' \end{pmatrix} \in \mathbb{R}^2$, et donc $\left\{ \begin{array}{l} x' = x', \\ x'' = -a_2 x - a_1 x' = g, \end{array} \right\}$ soit $\vec{X}\,'(t) = A.\vec{X}(t) + \vec{b}(t)$ où $A = \begin{pmatrix} 0 & 1 \\ -a_2 & -a_1 \end{pmatrix}$ et $\vec{b} = \begin{pmatrix} 0 \\ g \end{pmatrix}$. Donc $\vec{X}$ est solution de l'ED $\vec{X}\,' = \vec{f}(t, \vec{X})$ où $\vec{f}(t, \vec{X}) = A.\vec{X} + \vec{b}(t)$ avec CI $\vec{X}(0) = \vec{X}_0$. D'où le schéma d'Euler explicite $\vec{X}_{k+1} = \vec{X}_k + hA.\vec{X}_k + \vec{b}_k$. Soit sous forme système

$$\left\{ \begin{array}{l} x_{k+1} = x_k + hy_k, \\ y_{k+1} = -a_2 x_k - a_1 y_k + g(t_k), \end{array} \right.$$

système initialisé à l'aide de x_0 et $y_0 = v_0$. Et on retient toutes les valeurs x_k (les y_k sont des intermédiaires de calcul). ▪

3.2.3 Méthode d'Euler implicite

Pour pallier au problème de la stabilité asymptotique dans la méthode d'Euler explicite, on change l'expression de l'approximation de la dérivée : avec $h > 0$,

$$u'(t) = f(t, u(t)) \simeq \frac{u(t) - u(t-h)}{h} \quad \text{(pente à droite dans } [t-h, t]) \tag{3.9}$$

Cela revient à utiliser le développement limité (où donc $-h < 0$) :

$$u(t-h) = u(t) - h\,u'(t) + o(h),$$

soit encore :

$$u(t) = u(t-h) + h\,f(t, u(t)) + o(h).$$

Puis on néglige le "petit terme" $o(h)$ pour définir la solution approchée à $t_1 = t_0 + h_1$:

$$\tilde{u}_1 = u_0 + h_1 f(t_1, \tilde{u}_1), \quad \text{où donc} \quad f(t_1, \tilde{u}_1) \simeq u'(t_1) \quad \text{(pente à droite dans } [t_0, t_1]).$$

Puis on poursuit :

$$\tilde{u}_{k+1} = \tilde{u}_k + h_{k+1} f(t_k, \tilde{u}_k), \quad \text{où donc} \quad f(t_k, \tilde{u}_k) \simeq u'(t_k) \quad \text{(pente à droite dans } [t_k, t_{k+1}]). \tag{3.10}$$

Le problème de cette méthode est qu'il faut résoudre l'équation $\tilde{u}_1 = u_0 + h_1 f(t_1, \tilde{u}_1)$ en l'inconnue $\tilde{u}_1$: c'est une équation est implicite. On utilisera, par exemple la méthode de Newton, ou tout programme de recherche de 0 de la fonction

$$g(z) = z - hf(t_{k+1}, z) - \tilde{u}_k. \tag{3.11}$$

Le z solution de "$g(z) = 0$" donnera $\tilde{u}_{k+1} = z$.

D'où l'algorithme élémentaire de calcul de la solution $\tilde{u}$ approchée :

1. Initialisation : on se donne la fonction $f : \mathbb{R}^2 \to \mathbb{R}$, le temps initial t_0, le temps final T, la condition initiale (approchée) $\tilde{u}_0$.
2. On se donne $n \in \mathbb{N}^*$ et le pas temps $h = \frac{T - t_0}{n}$ (ici maillage temporel uniforme).
3. Boucle pour $k = 0$ à $n-1$:

$$\left\{ \begin{array}{l} t_{k+1} = t_k + h, \\ \text{résolution de } g(z) = 0 \text{ dans (3.11)}. \end{array} \right. \tag{3.12}$$

Remarque 3.9 Cette méthode est plus coûteuse que la méthode d'Euler (explicite) mais est utile dans le cas où il y a des problèmes de stabilité. C'est également une méthode d'ordre 1. ▞

Remarque 3.10 La méthode d'Euler implicite découle également immédiatement de la méthode d'intégration de Riemann à droite :

$$u(t_1) - u(t_0) = \int_{t_0}^{t_1} u'(t)\,dt \simeq (t_1 - t_0)u'(t_1) = (t_1 - t_0)f(t_1, u(t_1)). \tag{3.13}$$

▞

Remarque 3.11 De manière générale, l'équation $g(z) = 0$ en l'inconnue $z = \tilde{u}_{k+1}$ est de la forme :

$$\tilde{u}_{k+1} = \tilde{g}(h, t_{k+1}, \tilde{u}_k, \tilde{u}_{k+1}).$$

Pour h, $s = t_{k+1}$, $u = \tilde{u}_k$ donnés, on cherche donc $v = \tilde{u}_{k+1}$ tel que :

$$v = g_{s,u,h}(v) \quad \text{où} \quad g_{s,u,h}(v) = u + hf(s, v).$$

C'est un problème de point fixe en v, qui admet une solution dès que $g_{s,u,h}$ est une contraction (lipschitzienne de rapport <1 en la variable v). Mais par hypothèse, f est une fonction lipschitzienne de rapport $L > 0$, et donc :

$$|g_{s,u,h}(v) - g_{s,u,h}(w)| \leq hL|v - w|,$$

et donc $g_{s,u,h}$ est une contraction dès que $h < \frac{1}{L}$. On fera cette hypothèse dans la suite (en prenant par exemple $h_{\max} = \frac{95}{100}\frac{1}{L}$, et le schéma implicite donnera une solution unique. ▞

Exercice 3.12 Soit $f : \mathbb{R}^2 \to \mathbb{R}$ donnée par $f(x, y) = x$. Résoudre de manière approchée (3.1) à l'aide du schéma d'Euler implicite.

Réponse. Fonction inconnue : $u : [t_0, T] \to \mathbb{R}$ solution de $u'(t) = t$ et $u(t_0) = u_0$.

Schéma d'Euler implicite : $\tilde{u}_{k+1} = \tilde{u}_k + hf(t_{k+1}, \tilde{u}_{k+1}) = \tilde{u}_k + ht_{k+1}$, pour $k = 0, ..., n-1$.

N.B. : la solution exacte est $u(t) = u(t_0) + \int_{t_0}^{t} \tau\,d\tau = u_0 + \frac{t^2 - t_0^2}{2}$. D'où $u(T) = u_0 + \frac{1}{2}(T - t_0)(T + t_0)$.

N.B. : ici on a donc $\tilde{u}_n = \tilde{u}_{n-1} + ht_n = \tilde{u}_{n-2} + ht_{n-1} + ht_n = ... = \tilde{u}_0 + h(t_1 + t_1 + ... + t_n)$. Avec (t_k) suite arithmétique de raison h donc $t_1 + t_2 + ... + t_n = \frac{n}{2}(t_1 + t_n)$, et donc $\tilde{u}_n = \tilde{u}_0 + \frac{1}{2}nh(t_0 + T + h) = \tilde{u}_0 + \frac{1}{2}(T - t_0)(t_0 + T + h)$, à comparer avec la solution exacte : l'erreur commise est d'ordre h. ▞

Exercice 3.13 Soit $f : \mathbb{R}^2 \to \mathbb{R}$ donnée par $f(x, y) = y$. Résoudre de manière approchée (3.1) à l'aide du schéma d'Euler implicite.

Réponse. Fonction inconnue : $u : [t_0, T] \to \mathbb{R}$ solution de $u'(t) = u(t)$ et $u(t_0) = u_0$.

Schéma d'Euler implicite : $\tilde{u}_{k+1} = \tilde{u}_k + h\tilde{u}_{k+1}$, pour $k = 0, ..., n-1$ donc $\tilde{u}_{k+1} = \frac{\tilde{u}_k}{1-h}$ (à comparer au schéma explicite).

N.B. : ici la solution exacte est $u(t) = u_0 e^{t - t_0}$. En particulier $u(T) = u_0 e^{T - t_0}$.

N.B. : ici on a donc $\tilde{u}_n = \frac{\tilde{u}_0}{(1-h)^n}$. Avec $\frac{1}{(1-h)^n} = \frac{1}{(1 - \frac{T - t_0}{n})^n}$ approximation de $\frac{1}{e^{-(T - t_0)}} = e^{(T - t_0)}$, à comparer avec la solution exacte. ▞

Exercice 3.14 Soit $f : \mathbb{R}^2 \to \mathbb{R}$ donnée par $f(x, y) = y^2$. Résoudre de manière approchée (3.1) à l'aide du schéma d'Euler implicite.

Réponse. Fonction inconnue : $u : [t_0, T] \to \mathbb{R}$ solution de $u'(t) = u(t)^2$ et $u(t_0) = u_0$.

Schéma d'Euler implicite : $\tilde{u}_{k+1} = \tilde{u}_k + h\tilde{u}_{k+1}^2$, pour $k = 0, ..., n-1$. Donc u_{k+1} est solution de $h\tilde{u}_{k+1}^2 - \tilde{u}_{k+1} + \tilde{u}_k = 0$, de discriminant $\Delta = 1 - 4h\tilde{u}_k$. D'où deux solutions possibles, $\tilde{u}_{k+1} = \frac{1 \pm \sqrt{1 - 4h\tilde{u}_k}}{2h}$. la solution avec + n'est pas admissible car quand $h \to 0$ on a $u_{k+1} \to \infty$. Donc la solution à retenir est $\tilde{u}_{k+1} = \frac{1 - \sqrt{1 - 4h\tilde{u}_k}}{2h}$. ▞

Exercice 3.15 Reprendre l'exercice précédent en remplaçant f par sa linéarisée en y.

Puis comparer les solutions.

Réponse. La fonction $f(x, y) = y^2$ s'écrit $f_x(y) = y^2$. Linéarisons f_x au voisinage d'un point y_0 : $f_x(y) = f_x(y_0) + (y - y_0)f_x'(y_0) + o(y - y_0) = y_0^2 + 2(y - y_0)y_0 + o(y - y_0) = 2y_0 y - y_0^2 + o(y - y_0)$. Approchons f_x par sa linéarisée g_x donnée par $g_x(y) = 2y_0 y - y_0^2$ (Le graphe de g est la droite de pente $2y_0$ qui passe par le point $(y_0, y_0^2 = g_x(y_0))$, droite tangente à la parabole en ce point).

L'équation différentielle linéarisée au voisinage de u_0 est $v'(t) = g_t(v(t)) = 2v_0 v(t) - v_0^2$ où on a posé $v_0 = u_0 = v(t_0)$ qu'on notera également $\tilde{v}_0$.

Schéma d'Euler implicite, pour le premier pas de temps : $\tilde{v}_1 = \tilde{v}_0 + h(2v_0\tilde{v}_1 - \tilde{v}_0^2)$. D'où $\tilde{v}_1 = \tilde{v}_0 \frac{1 - h\tilde{v}_0}{1 - 2h\tilde{v}_0}$.

D'où le schéma générique $\tilde{v}_{k+1} = \tilde{v}_k \frac{1 - h\tilde{v}_k}{1 - 2h\tilde{v}_k}$ (après linéarisation au voisinage du point $\tilde{v}_k$ à chaque étape).

Comparons cette solution avec celle de l'exercice précédent.

Ici on a $\frac{1}{1-y} = 1 + y + o(y)$. D'où $\tilde{v}_{k+1} = \tilde{v}_k(1 - h\tilde{v}_k)(1 + 2h\tilde{v}_k + o(h)) = \tilde{v}_k + h\tilde{v}_k^2 + o(h)$.

Pour l'exercice précédent, soit $\varphi(y) = (1+y)^{\frac{1}{2}}$. On a $\varphi'(y) = \frac{1}{2}(1+y)^{-\frac{1}{2}}$ et $\varphi''(y) = -\frac{1}{4}(1+y)^{-\frac{3}{2}}$. D'où le développement limité de φ au voisinage de $y=0$: $\sqrt{1+y} = 1 + \frac{y}{2} - \frac{y^2}{8} + o(y^2)$.

D'où $\tilde{u}_{k+1} = \frac{-\frac{-4h\tilde{u}_k}{2} + \frac{(4h\tilde{u}_k)^2}{8} + o(h^2)}{2h} = \tilde{u}_k + h\tilde{u}_k^2 + o(h)$ au voisinage de $h=0$.

Donc au premier ordre, les méthodes sont comparables (il n'est pas nécessaire d'aller plus loin que le 1er ordre : on verra que la méthode d'Euler implicite est d'ordre 1). ∎

3.2.4 Schéma de Crank–Nicholson et θ-schémas

Sur l'intervalle $[t_0, t_1]$, la méthode d'Euler explicite se sert de la pente à gauche $p_g = u'(t_0) = f(t_0, u(t_0))$ et celle d'Euler implicite se sert de la pente à droite $p_d = u'(t_1) = f(t_1, u(t_1))$. L'idée suivante est de se servir d'une pente intermédiaire $p_\theta = (1-\theta)p_g + \theta p_d$ pour $0 \leq \theta \leq 1$.

Le schéma qui en résulte s'écrit :

$$\tilde{u}_1 = \tilde{u}_0 + h\Big((1-\theta)p_g + \theta p_d\Big),$$

soit :

$$\tilde{u}_1 = \tilde{u}_0 + h\Big((1-\theta)f(t_0, \tilde{u}_0) + \theta f(t_1, \tilde{u}_1)\Big),$$

et au pas de temps t_{k+1} :

$$\tilde{u}_{k+1} = \tilde{u}_k + h\Big((1-\theta)f(t_k, \tilde{u}_k) + \theta f(t_{k+1}, \tilde{u}_{k+1})\Big). \tag{3.14}$$

C'est le θ-schéma. Il est très utilisé pour $\frac{1}{2} \leq \theta \leq 1$ car il est A-stable. ($\theta = 0$ donne Euler explicite et $\theta = 1$ donne Euler implicite). (Le calcul de $\tilde{u}_{k+1}$ est implicite pour $\theta \neq 0$.)

Quand $\theta = \frac{1}{2}$:

$$\tilde{u}_{k+1} = \tilde{u}_k + \frac{h}{2}\Big(f(t_k, \tilde{u}_k) + f(t_{k+1}, \tilde{u}_{k+1})\Big), \tag{3.15}$$

on obtient le schéma de Crank–Nicholson.

Remarque 3.16 On peut se rendre compte simplement que le schéma de Crank–Nicholson est meilleur en termes de précision : on a, pour $u \in C^2$: d'une part :

$$\begin{aligned} u(t_{k+1}) &= u(t_k) + hu'(t_k) + \frac{h^2}{2}u''(t_k) + o(h^2) \\ &= u(t_k) + hf(t_k, u(t_k)) + \frac{h^2}{2}u''(t_k) + o(h^2), \end{aligned} \tag{3.16}$$

et d'autre part :

$$\begin{aligned} u(t_k) &= u(t_{k+1}) - hu'(t_{k+1}) + \frac{h^2}{2}u''(t_{k+1}) + o(h^2) \\ &= u(t_{k+1}) - hf(t_{k+1}, u(t_{k+1})) + \frac{h^2}{2}u''(t_k) + o(h^2), \end{aligned} \tag{3.17}$$

car $u''(t_{k+1}) = u''(t_k) + o(1)$. D'où par différence :

$$u(t_{k+1}) = u(t_k) + \frac{h}{2}\Big(f(t_k, u(t_k)) + f(t_{k+1}, u(t_{k+1}))\Big) + o(h^2),$$

à comparer avec (3.15). Et donc l'approximation par $\tilde{u}$ donne une erreur en $o(h^2)$ (à comparer avec l'erreur en $o(h)$ pour Euler explicite, voir (3.2), ou encore pour le θ-schéma pour $\theta \neq \frac{1}{2}$). ∎

Remarque 3.17 Le schéma de Crank–Nicholson (cas $\theta = \frac{1}{2}$) est inconditionnellement A-stable, mais pour $\theta < \frac{1}{2}$, le θ-schéma n'est pas inconditionnellement A-stable. Des erreurs d'arrondis peuvent donc rendre le schéma de Crank–Nicholson instable. Si cela arrive, on pourra privilégier un θ-schéma avec par exemple $\theta = \frac{3}{5}$. ∎

Remarque 3.18 Le schéma de Crank–Nicholson découle également de la formule d'intégration des trapèzes (d'ordre 2) :

$$u(t_1) - u(t_0) = \int_{t_0}^{t_1} u'(t)\,dt \simeq \frac{u'(t_0) + u'(t_1)}{2}(t_1 - t_0),$$

soit avec $h = t_1 - t_0$:

$$u(t_1) - u(t_0) \simeq h\,\frac{f(t_0, u(t_0)) + f(t_1, u(t_1))}{2}.$$

∎

3.2.5 Méthode d'Euler améliorée

La méthode de Crank–Nicholson est implicite à cause de $\tilde{u}_{k+1}$ dans $f(t_{k+1}, \tilde{u}_{k+1})$ dans le membre de droite de (3.15). L'idée est alors d'estimer $\tilde{u}_{k+1}$ (prédiction) à l'aide de la méthode d'Euler explicite :

$$\tilde{u}_{k+1}^{\text{prédiction}} = \tilde{u}_k + hf(t_k, \tilde{u}_k).$$

Et on améliore cette prédiction (étape de correction) en s'inspirant du schéma de Crank-Nicholson :

$$\tilde{u}_{k+1} = \tilde{u}_k + \frac{h}{2}\Big(f(t_k, \tilde{u}_k) + f(t_{k+1}, \tilde{u}_{k+1}^{\text{prédiction}})\Big) \qquad (= \tilde{u}_{k+1}^{\text{correction}}). \tag{3.18}$$

(Prédiction–correction = méthode d'Heun.) Autrement dit :

$$\tilde{u}_{k+1} = \tilde{u}_k + \frac{h}{2}\Big(f(t_k, \tilde{u}_k) + f(t_{k+1}, \tilde{u}_k + hf(t_k, \tilde{u}_k))\Big).$$

Ce schéma est explicite et on montrera qu'il est d'ordre 2.

En revanche, il n'est pas inconditionnellement A-stable.

3.2.6 Méthode du point milieu

Sur l'intervalle $[t_k, t_{k+1}]$, plutôt que de se servir de la pente p_g à gauche et de la pente p_d à droite, et on peut essayer de se servir de la pente au milieu et proposé le schéma :

$$\tilde{u}_{k+1} = \tilde{u}_k + hf(t_{k+\frac{1}{2}}, u_{k+\frac{1}{2}}), \tag{3.19}$$

où $t_{k+\frac{1}{2}} = \frac{t_k + t_{k+1}}{2}$ et $u_{k+\frac{1}{2}} = u(t_{k+\frac{1}{2}})$.

Remarque 3.19 Ce schéma découle également de la formule d'intégration du point milieu (d'ordre 2) :

$$u(t_1) - u(t_0) = \int_{t_0}^{t_1} u'(t)\,dt \simeq u'(t_{\frac{1}{2}})(t_1 - t_0),$$

i.e., avec $h = t_1 - t_0$:

$$u(t_1) - u(t_0) \simeq h\, f(t_{\frac{1}{2}}, \tilde{u}_{\frac{1}{2}}).$$

∎

Remarque 3.20 On peut se rendre compte simplement que le schéma du point milieu est, comme le schéma de Crank-Nicholson, d'ordre 2 : on a, pour $u \in C^2$: d'une part :

$$\begin{aligned} u(t_{k+1}) &= u(t_{k+\frac{1}{2}}) + \frac{h}{2}u'(t_{k+\frac{1}{2}}) + \frac{(\frac{h}{2})^2}{2}u''(t_{k+\frac{1}{2}}) + o(h^2), \\ &= u(t_{k+\frac{1}{2}}) + \frac{h}{2}f(t_{k+\frac{1}{2}}, u(t_{k+\frac{1}{2}})) + \frac{h^2}{8}u''(t_{k+\frac{1}{2}}) + o(h^2), \end{aligned} \tag{3.20}$$

et d'autre part :

$$\begin{aligned} u(t_k) &= u(t_{k+\frac{1}{2}}) - \frac{h}{2}u'(t_{k+\frac{1}{2}}) + \frac{(\frac{h}{2})^2}{2}u''(t_{k+\frac{1}{2}}) + o(h^2), \\ &= u(t_{k+1}) - \frac{h}{2}f(t_{k+\frac{1}{2}}, u(t_{k+\frac{1}{2}})) + \frac{h^2}{8}u''(t_{k+\frac{1}{2}}) + o(h^2), \end{aligned} \tag{3.21}$$

D'où par différence :

$$u(t_{k+1}) = u(t_k) + hf(t_{k+\frac{1}{2}}, u(t_{k+\frac{1}{2}})) + o(h^2),$$

à comparer avec (3.19). ∎

3.2.7 * Méthode de Runge-Kutta d'ordre 2 (méthode d'Heun)

Le problème est qu'on ne connaît pas $u(t_{k+\frac{1}{2}})$. On va l'estimer (prédiction). On se sert alors de la méthode d'Euler (explicite) :

$$\tilde{u}_{k+\frac{1}{2}}^{\text{prédiction}} = \tilde{u}_k + \frac{h}{2}f(t_k, \tilde{u}_k),$$

d'où la pente estimée $= f(t_{k+\frac{1}{2}}, \tilde{u}_{k+\frac{1}{2}}^{\text{prédiction}})$. On pose donc :

$$\tilde{u}_{k+1} = \tilde{u}_k + hf(t_{k+\frac{1}{2}}, \tilde{u}_{k+\frac{1}{2}}^{\text{prédiction}}) \qquad (= \tilde{u}_{k+1}^{\text{correction}}). \tag{3.22}$$

(Prédiction–correction = méthode d'Heun.)

Le calcul de $\tilde{u}_{k+1}$ est explicite, effectué en 2 étapes simples. On montrera que le schéma obtenu (de Heun) est d'ordre 2.

3.2.8 * Méthode de Runge-Kutta (classique d'ordre 4)

On souhaite toujours une méthode explicite, mais convergeant plus rapidement. On peut s'inspirer de la méthode d'intégration de Simpson d'ordre 4 :

$$\int_{t_k}^{t_{k+1}} g(x)\,dx \simeq h\frac{g(t_k)+4g(t_{k+\frac{1}{2}})+g(t_{k+1})}{6}. \tag{3.23}$$

Malheureusement, appliquée à l'équation différentielle, la méthode qui en résulte est implicite.

Et pour la modifier et la rendre explicite, les estimations précédentes de $\tilde{u}_{k+\frac{1}{2}}$ (méthode d'Heun) et de $\tilde{u}_{k+1}$ (par la méthode d'Euler améliorée) ne sont cependant pas suffisantes pour avoir une méthode d'ordre 4 explicite. On définit, relativement à l'intervalle $[t_k, t_{k+1}]$:

1- une estimation de la pente p_1 amont (en t_k) :

$$p_1 = f(t_k, \tilde{u}_k).$$

2- Cette valeur donne une valeur estimée au point milieu :

$$\tilde{u}_{k+\frac{1}{2},1} = \tilde{u}_k + \frac{h}{2}\,p_1,$$

(méthode d'Euler explicite sur $[t_k, t_{k+\frac{1}{2}}]$) et donc une estimation de la pente p_2 au point milieu (en $t_{k+\frac{1}{2}}$) :

$$p_2 = f(t_{k+\frac{1}{2}}, \tilde{u}_{k+\frac{1}{2},1}) \qquad (= f(t_k + \frac{h}{2}, \tilde{u}_k + \frac{h}{2}\,p_1)).$$

3- On a alors une autre estimation au point milieu :

$$\tilde{u}_{k+\frac{1}{2},2} = \tilde{u}_k + \frac{h}{2}\,p_2,$$

(donné par la méthode d'Euler implicite sur $[t_k, t_{k+\frac{1}{2}}]$ modifié en utilisant p_2) d'où une autre estimation de la pente milieu p_3 (en $t_{k+\frac{1}{2}}$) :

$$p_3 = f(t_{k+\frac{1}{2}}, \tilde{u}_{k+\frac{1}{2},2}) \qquad (= f(t_k + \frac{h}{2}, \tilde{u}_k + \frac{h}{2}\,p_2)).$$

4- Et la pente estimée en aval (en t_{k+1}) est prise valant :

$$p_4 = f(t_k + h, \tilde{u}_k + hp_3),$$

Enfin, la valeur retenue pour la pente permettant de passer de $u(t_k)$ à $u(t_{k+1})$ est le barycentre $p = \frac{p_1+2p_2+2p_3+p_4}{6}$ (les estimations au milieu ont plus de poids que celles des extrémités, comme dans la méthode de Simpson (3.23)). D'où la valeur calculée $\tilde{u}_{k+1}$:

$$\tilde{u}_{k+1} = \tilde{u}_k + h\frac{p_1+2p_2+2p_3+p_4}{6}. \tag{3.24}$$

Cette méthode est très utilisée, et sa programmation est simple :

1. on se donne $h > 0$ (le pas de temps),
2. on calcule $n =$ partie entière de $\frac{T-t_0}{h}$ (le nombre de pas de temps pour atteindre T),
3. on pose $t_k = t_0 + kh$ pour $k = 1$ à n,
4. pour $k = 1$ à n,

$$\begin{aligned} p_1 &= f(t_k, \tilde{u}_k),\\ p_2 &= f(t_{k+\frac{1}{2}}, \tilde{u}_k + \frac{h}{2}p_1),\\ p_3 &= f(t_{k+\frac{1}{2}}, \tilde{u}_k + \frac{h}{2}p_2),\\ p_4 &= f(t_{k+1}, \tilde{u}_k + hp_3),\\ \tilde{u}_{k+1} &= \tilde{u}_k + \frac{h}{6}(p_1+2p_2+2p_3+p_4). \end{aligned} \tag{3.25}$$

On peut montrer que cette méthode est d'ordre 4, mais qu'elle n'est pas inconditionnellement A-stable.

Par contre, cette méthode demande le calcul des $f((t_{k+\frac{1}{2}}, \tilde{u}_k + \frac{h}{2}p_i)$ (pour $i = 1, 2$) et de $f(t_{k+1}, \tilde{u}_k + hp_3)$, ce qui peut être coûteux, par exemple dans le cas des systèmes différentiels.

Par ailleurs, dans le cas des problèmes "raides", cette méthode ne donne en général pas de bon résultat (de même que les méthodes explicites en général).

3.2.9 * Méthodes de Runge-Kutta généralisées

Revenons sur l'idée de base de la méthode de Runge–Kutta classique : l'obtention de valeurs intermédiaires non directement utilisées à l'aide de :

1. des temps intermédiaires
$$t_{k,i} = t_k + c_i h,$$
avec $0 \leq c_i \leq 1$ pour tout i, et un temps final $t_{k+1} = t_k + h$.
2. Une formule de quadrature pour les valeurs intermédiaires (non directement utilisées) :
$$\tilde{u}(t_{k,i}) - \tilde{u}(t_k) = \int_{t_k}^{t_{k,i}} f(t, u(t))\, dt \simeq h \sum_{j=1}^{q} a_{ij} f(t_{k,j}, u(t_{k,j})).$$
3. Une formule de quadrature pour la valeur finale retenue :
$$\tilde{u}(t_{k+1}) - \tilde{u}(t_k) = \int_{t_k}^{t_{k+1}} f(t, u(t))\, dt \simeq h \sum_{j=1}^{q} b_j f(t_{k,j}, u(t_{k,j})).$$

On présente les quantités introduites sous la forme du tableau :

$$\begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \dots & a_{1q} \\ c_2 & a_{21} & a_{22} & \dots & a_{2q} \\ \vdots & \vdots & \vdots & & \vdots \\ c_q & a_{q1} & a_{q2} & \dots & a_{qq} \\ \hline & b_1 & b_2 & \dots & b_q \end{array}$$

Notez que pour toutes les méthodes explicites, la matrice carrée $[a_{ij}]_{1 \leq i,j \leq q}$ est triangulaire inférieure stricte.

Exemple 3.21 Méthode d'Euler explicite (3.6) : la valeur finale est $\tilde{u}_{k+1} = \tilde{u}_k + hf(t_k, \tilde{u}_k)$. Donc $q = 1$, $b_1 = 1$ et $t_{k,1} = t_k$, donc $c_1 = 0$ et $a_{1,1} = 0$. Donc une étape, qui est en fait l'étape finale. Le tableau associé est $\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$. ▚

Exemple 3.22 Méthode d'Euler implicite (3.10) : la valeur finale est $\tilde{u}_{k+1} = \tilde{u}_k + hf(t_{k+1}, \tilde{u}_{k+1})$. Donc $q = 1$, $b_1 = 1$, $t_{k,1} = t_{k+1}$, $c_1 = 1$ et $a_{1,1} = 1$. D'où le tableau $\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$. ▚

Exemple 3.23 θ-schéma (3.14) : pour $\theta \neq 0$, la valeur finale retenue donne $q = 2$, $b_1 = (1-\theta)$, $b_2 = \theta$, $c_1 = 0$, $c_2 = 1$, $a_{11} = 0 = a_{12} = a_{21}$ et $a_{22} = 1$. D'où le tableau $\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 0 & 1 \\ \hline & (1-\theta) & \theta \end{array}$. La méthode est implicite pour $\theta \neq 0$. Noter que si $\theta = 0$, le tableau ci-dessus n'a plus d'intérêt : on a dans ce cas $a_{22} = 0$ et on retrouve la méthode d'Euler explicite. ▚

Exemple 3.24 Méthode d'Euler améliorée (3.18) : la valeur retenue donne $q = 2$, $b_1 = \frac{1}{2}$, $b_2 = \frac{1}{2}$, $c_1 = 0$, $c_2 = 1$ d'où $a_{11} = a_{12} = a_{22} = 0$ et $a_{21} = \frac{1}{2}$. D'où le tableau $\begin{array}{c|cc} 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$. ▚

Exemple 3.25 Méthode du point milieu (3.19) : la valeur retenue est calculée avec un temps intermédiaire donne $q = 1$, $b_1 = 1$, $c_1 = \frac{1}{2}$, $a_{11} = \frac{1}{2}$. D'où le tableau $\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}$. ▚

Exemple 3.26 Méthode du point milieu rendue explicite (3.22) (méthode d'Heun) : la valeur retenue est calculée avec un temps intermédiaire. Cela donne $q = 2$, $b_1 = 1$, $b_2 = 0$, $c_1 = \frac{1}{2}$, $c_2 = 0$, $a_{11} = a_{12} = a_{22} = 0$ et $a_{21} = \frac{1}{2}$. D'où le tableau $\begin{array}{c|cc} 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \hline & 0 & 1 \end{array}$. ▚

Exemple 3.27 Runge–Kutta classique (3.24) : avec 3 pas de temps intermédiaires. Le tableau associé est donné par : $\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & \frac{1}{6} & \frac{2}{6} & \frac{2}{6} & \frac{1}{6} \end{array}$. ▚

Exemple 3.28 On peut montrer que la méthode implicite définie par le tableau $\begin{array}{c|cc} \alpha & \alpha & 0 \\ 1-\alpha & 1-2\alpha & \alpha \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$ avec $\alpha = \frac{1}{2} + \frac{1}{2\sqrt{3}}$ est d'ordre 3. Voir Crouzeix et Mignot [5]. ▪

3.3 * Formulation générique des méthodes à un pas

3.3.1 Formulation

Les méthodes présentées ci-dessus sont des méthodes à un pas, i.e. des méthodes qui ne se servent que des deux valeurs $\tilde{u}_k$ et $\tilde{u}_{k+1}$ correspondant à un unique pas de temps $h_k = t_{k+1} - t_k$. Elles se présentent toutes sous la forme séquentielle, pour $\tilde{u}_0$ donné :

$$\frac{\tilde{u}_{k+1} - \tilde{u}_k}{h_k} = \Phi(t_k, \tilde{u}_k, h_k), \tag{3.26}$$

soit :

$$\tilde{u}_{k+1} = \tilde{u}_k + h_k \Phi(t_k, \tilde{u}_k, h_k). \tag{3.27}$$

(Ou encore sous la forme $\tilde{u}_{k+1} = \tilde{u}_k + h_k \bar{\Phi}(t_k, \tilde{u}_k, t_{k+1})$, cette dernière forme avec $\bar{\Phi}$ n'étant pas pratique pour les calculs d'erreurs et de convergence où $h_k = t_{k+1} - t_k$ est une variable primordiale.)

Exemple 3.29 Les méthodes explicites :

Méthode d'Euler : $\Phi(t_k, \tilde{u}_k, h_k) = f(t_k, \tilde{u}_k)$.

Méthode d'Euler améliorée : $\Phi(t_k, \tilde{u}_k, h_k) = \frac{1}{2}\big(f(t_k, \tilde{u}_k) + f(t_{k+1}, \tilde{u}_k + h_k f(t_k, \tilde{u}_k))\big)$.

Méthode du point milieu : $\Phi(t_k, \tilde{u}_k, h_k) = f(t_k + \frac{h_k}{2}, \tilde{u}_k + \frac{h_k}{2} f(t_k, \tilde{u}_k))$.

Méthode de Runge-Kutta : $\Phi(t_k, \tilde{u}_k, h_k) = \frac{1}{6}(p_1 + 2p_2 + 2p_3 + p_4)$, où les p_i sont donnés en (3.25) et ne dépendent que de t_k, $\tilde{u}_k$ et h_k (calcul immédiat). ▪

Exemple 3.30 Les méthodes implicites :

Méthode d'Euler implicite : on a $\tilde{u}_{k+1} = \tilde{u}_k + h_k f(t_{k+1}, \tilde{u}_{k+1})$ d'inconnue $\tilde{u}_{k+1}$, dont la résolution donne $\tilde{u}_{k+1}$: on note $\tilde{u}_{k+1} = G(t_k, \tilde{u}_k, h_k)$ la solution trouvée.

D'où $\Phi(t_k, \tilde{u}_k, h_k) = f(t_k + h_k, G(t_k, \tilde{u}_k, h_k))$.

Même démarche pour la méthode de Crank–Nicholson. ▪

3.3.2 Une méthode générale de construction des schémas

Elle est basée sur le développement de Taylor de la solution cherchée :

$$\frac{u(t_0 + h) - u(t_0)}{h} = u'(t_0) + \frac{h}{2} u''(t_0) + \frac{h^2}{3!} u'''(t_0) + ...$$

De $u'(t) = f(t, u(t))$, équation satisfaite par la solution cherchée, on déduit, supposant f C^2 :

$$\begin{aligned} u''(t) &= f_{,t}(t, u(t)) + f_{,x}(t, u(t)).u'(t) \\ &= (f_{,t} + f_{,x}.f)(t, u(t)), \end{aligned}$$

et

$$\begin{aligned} u'''(t) &= f_{,tt}(t, u(t)) + 2f_{,tx}(t, u(t)).u'(t) + f_{,xx}(t, u(t)).u'(t)^2 + f_{,x}(t, u(t)).u''(t) \\ &= (f_{,tt} + 2f_{,tx}.f + f_{,xx}.f^2 + f_{,x}(f_{,t} + f_{,x}.f))(t, u(t)), \end{aligned}$$

où on a noté $f_{,t} = \frac{\partial f}{\partial t}$, $f_{,x} = \frac{\partial f}{\partial x}$, $f_{,tt} = \frac{\partial^2 f}{\partial t^2}$...

Et ainsi le développement de Taylor de u donne :

$$\begin{aligned} \frac{u(t_0 + h) - u(t_0)}{h} = f(t_0, u(t_0)) &+ \frac{h}{2}(f_{,t} + f_{,x}.f)(t_0, u(t_0)) \\ &+ \frac{h^2}{3!}(f_{,tt} + 2f_{,tx}.f + f_{,xx}.f^2 + f_{,x}f_{,t} + f_{,x}^2.f))(t_0, u(t_0)) + ... \end{aligned} \tag{3.28}$$

3.3.3 Application aux schémas explicites à un pas

Le gros problème dans cette expression est d'évaluer les dérivées partielles de f, et c'est, en général, coûteux (et parfois laborieux). L'idée est de remplacer ces dérivées partielles par des quantités de type $f(t_0+\alpha, x_0+\beta)$ avec α et β réels bien choisis : on veut écrire le membre de droite de (3.28) sous une forme donnant le schéma de Runge–Kutta :

$$\frac{u(t_0+h)-u(t_0)}{h} = ap_1 + bp_2 + cp_3 + dp_4... \tag{3.29}$$

où les p_i sont des estimations des pentes :

$$\begin{aligned}
p_1 &= f(t_0, x_0),\\
p_2 &= f(t_0+\alpha_1 h, x_0+\alpha_1 p_1) \quad \text{(dépend de } p_1),\\
p_3 &= f(t_0+\alpha_2 h, x_0+\alpha_2 p_2) \quad \text{(dépend de } p_2),\\
p_4 &= f(t_0+\alpha_3 h, x_0+\alpha_3 p_3) \quad \text{(dépend de } p_3).
\end{aligned}$$

avec les $\alpha_i \in \mathbb{R}$. On se sert pour cela du développement de la fonction f :

$$\begin{aligned}
f(t_0+\gamma, x_0+\delta) = f(t_0,x_0)&\\
&+(\gamma f_{,t} + \delta f_{,x})\\
&+\frac{1}{2}(\gamma^2 f_{,tt} + 2\gamma\delta f_{,tx} + \delta^2 f_{,xx})(t_0,x_0)\\
&+\frac{1}{3!}(\gamma^3 f_{,ttt} + 3\gamma^2\delta f_{,ttx} + 3\gamma\delta^2 f_{,txx} + \delta^3 f_{,xxx})(t_0,x_0) + ...
\end{aligned}$$

Les identifications des termes ayant même exposant en h entre (3.28) et (3.29) donne :

$$\begin{aligned}
&a+b+c+d = 1,\\
&b\alpha_1 + c\alpha_2 + d\alpha_3 = \frac{1}{2},\\
&b\alpha_1^2 + c\alpha_2^2 + d\alpha_3^2 = \frac{1}{3},\\
&c\alpha_1\alpha_2 + c\alpha_2\alpha_3 = \frac{1}{6}.
\end{aligned}$$

On obtient ainsi des méthodes exactes au troisième ordre.

Exemple 3.31 La méthode de Runge–Kutta est donné par $\alpha_1 = \alpha_2 = \frac{1}{2}$ et $\alpha_3 = 1$, ce qui donne $a = d = \frac{1}{6}$ et $b = c = \frac{1}{3}$. On peut montrer que cette méthode est en fait d'ordre 4 en poursuivant les calculs. ▪

Exemple 3.32 La méthode de Kutta–Simpson $\frac{3}{8}$ est donnée par $\alpha_1 = \frac{1}{3}$, $\alpha_2 = \frac{2}{3}$ et $\alpha_3 = 1$, ce qui donne $a = d = \frac{1}{8}$ et $b = c = \frac{3}{8}$. Mais on lui préfère en général la méthode de Runge–Kutta. ▪

3.4 * Étude des Méthodes à un pas, définitions générales

On souhaite résoudre (3.1) de manière approchée à l'aide d'un schéma de type (3.26). Il s'agit de savoir si les méthodes proposées sont 'bonnes'.

On suppose que l'équation différentielle (3.1) admet une solution $u : [t_0, T] \to \mathbb{R}$. Les schémas calculent des valeurs (discrètes) $\tilde{u}(t_k) = \tilde{u}_k$ pour $t_k = t_0 + kh$ où $h = \frac{T-t_0}{n}$ et $1 \le k \le n$, où n est le nombre de pas de temps.

Il s'agit de comparer $|u(t_k) - \tilde{u}_k|$ (différence entre la valeur exacte cherchée et la valeur calculée).

3.4.1 Convergence

Définition 3.33 L'approximation $(\tilde{u}_k)_{k=1,...,n}$ est dite convergente si, lorsque $\tilde{u}_0 \underset{h\to 0}{\longrightarrow} u(t_0)$ (l'erreur sur l'approximation de la condition initiale s'annule avec h), on a, avec $n = \frac{T-t_0}{h}$:

$$\max_{1\le k\le n} |u(t_k) - \tilde{u}_k| \underset{h\to 0}{\longrightarrow} 0.$$

Autrement dit, on a convergence de la solution approchée $\tilde{u}$ vers la solution cherchée u ssi :

$$\lim_{h\to 0}\Big(\max_{0\le k\le n} |u(t_k) - \tilde{u}_k|\Big) = 0. \tag{3.30}$$

Il s'agit maintenant de savoir si les schémas proposés donnent une approximation convergente. On décompose la recherche de cette convergence en 2 étapes : la stabilité et la consistance.

3.4.2 Stabilité (faible sensibilité aux erreurs)

Définition 3.34 Le schéma (3.26) est dit stable s'il n'est pas trop sensible aux erreurs (comme les erreurs d'arrondi), i.e. s'il existe $M > 0$ tel que si deux suites quelconques $(\tilde{u}_k)_{0\leq k\leq n}$ et $(\tilde{v}_k)_{0\leq k\leq n}$ vérifient :

$$\begin{cases} \tilde{u}_{k+1} = \tilde{u}_k + h\Phi(t_k, \tilde{u}_k, h), \\ \tilde{v}_{k+1} = \tilde{v}_k + h\Phi(t_k, \tilde{v}_k, h) + \varepsilon_k, \end{cases} \tag{3.31}$$

où $(\varepsilon_k)_{0\leq k\leq n}$ est une suite donnée quelconque, alors :

$$\max_{0\leq k\leq n}\Big(|\tilde{u}_k - \tilde{v}_k|\Big) \leq M\Big(|\tilde{u}_0 - \tilde{v}_0| + \sum_{j=0}^{n-1}|\varepsilon_j|\Big). \tag{3.32}$$

(Le max de la différence est majoré, à une constante multiplicative près, par l'erreur initiale et la somme des erreurs accumulées au cours du temps.)

Exemple 3.35 Soit $\tilde{v}_k = u(t_k)$: i.e. $(\tilde{v}_k)$ est la suite des valeurs exactes aux temps t_k ; ici ε_k est l'erreur commise en utilisant le schéma, à savoir :

$$\varepsilon_k = u(t_{k+1}) - \tilde{u}_{k+1} + h\left(\Phi(t_k, u(t_k), h) - \Phi(t_k, \tilde{u}_k, h)\right). \tag{3.33}$$

Et (3.32) compare la solution approchée et la solution exacte. ▟

Exemple 3.36 On vérifie que les schémas proposés, Euler explicite et implicite, θ-schémas, Crank-Nicholson, Runge–Kutta... sont tous stables au sens de cette définition.

Par contre, ces schémas ne sont pas tous inconditionnellement stables, voir paragraphe sur la stabilité numérique (dite A-stabilité ou stabilité asymptotique). En particulier pour les schémas explicites, la stabilité numérique n'a lieu que si le pas de temps $h = \Delta t$ est suffisamment petit (stabilité numérique sous condition). ▟

3.4.3 Consistance (annulation de la somme des erreurs avec h)

Ici on s'intéresse au cas de l'exemple 3.35. En particulier ε_k est donné par, cf. $(3.31)_2$:

$$\varepsilon_k = u(t_{k+1}) - u(t_k) - h\Phi(t_k, u(t_k), h), \tag{3.34}$$

appelée erreur locale à l'instant t_{k+1}, ou si on préfère, avec $(3.31)_1$ et (3.34) :

$$\varepsilon_k = \big(u(t_{k+1}) - \tilde{u}_{k+1}\big) - \big(u(t_k) - \tilde{u}_k\big). \tag{3.35}$$

Définition 3.37 Le schéma (3.26) est dit consistant ssi la somme des erreurs locales, en valeur absolue, s'annule avec h :

$$\lim_{h\to 0}\Big(\sum_{k=0}^{n-1}|\varepsilon_k|\Big) = 0, \tag{3.36}$$

i.e. :

$$\lim_{h\to 0}\Big(\sum_{k=0}^{n-1}|u(t_{k+1}) - u(t_k) - h\Phi(t_k, u(t_k), h)|\Big) = 0. \tag{3.37}$$

Remarque 3.38 On a également $\frac{\varepsilon_k}{h} = |\frac{u(t_{k+1})-u(t_k)}{h} - \frac{\tilde{u}_{k+1}-\tilde{u}_k}{h}|$ = valeur absolue de la "différence des pentes moyennes". ▟

3.4.4 Stabilité + consistance ⇒ convergence

On a le théorème fondamental : stabilité + consistance ⇒ convergence :

Théorème 3.39 *Si le schéma (3.26) à un pas est stable et consistant, et si* $|\tilde{u}_0 - u(0)| \underset{h\to 0}{\longrightarrow} 0$ *(l'erreur sur la condition initiale s'annule avec h), alors il est convergent.*

Preuve. Par hypothèses de consistance, ayant $|\varepsilon_k| = |u(t_{k+1}) - u(t_k) - h\Phi(t_k, u(t_k), h)|$, on a :

$$\lim_{h\to 0}\Big(\sum_{k=0}^{n-1}|\varepsilon_k|\Big) = 0.$$

Ayant également $|\varepsilon_k| = |(u(t_{k+1}) - \tilde{u}_{k+1}) - (u(t_k) - \tilde{u}_k)|$, posant $\tilde{v}_k = u(t_k)$, alors $(\tilde{u}_k)_{0\leq k\leq n}$ et $(\tilde{v}_k)_{0\leq k\leq n}$ vérifient (3.32), on a $\max_{0\leq k\leq n}\Big(|\tilde{u}_k - \tilde{v}_k|\Big) \leq M\Big(|\tilde{u}_0 - \tilde{v}_0| + \sum_{j=0}^{k-1}|\varepsilon_j|\Big)$, d'où $\max_{0\leq k\leq n}\Big(|\tilde{u}_k - \tilde{v}_k|\Big) \longrightarrow_{h\to 0} 0$, i.e. (3.30). ▟

3.4.5 Critère de stabilité

On a la condition suffisante :

Théorème 3.40 *Si Φ est lipschitzienne en u, uniformément en t et h, i.e. si :*

$$\exists L_\Phi > 0, \quad \forall (t,u,v,h) \in [t_0,T]\times\mathbb{R}\times\mathbb{R}\times[0,h_{\max}], \quad |\Phi(t,u,h)-\Phi(t,v,h)| \le L_\Phi |u-v|, \tag{3.38}$$

alors le schéma (3.26) est stable.

Preuve. C'est une application du lemme de Gronwall discret, voir cours d'équations différentielles : on se donne deux suites $(\tilde{u}_k)$ et $(\tilde{v}_k)$ vérifiant (3.31). L'hypothèse (3.38) donne avec (3.26) :

$$|\tilde{u}_{k+1}-\tilde{v}_{k+1}| \le (1+hL_\Phi)|\tilde{u}_k-\tilde{v}_k| + \varepsilon_k.$$

Et le lemme de Gronwall discret donne :

$$|\tilde{u}_k-\tilde{v}_k| \le e^{L_\Phi(t_k-t_0)}|\tilde{u}_0-\tilde{v}_0| + \sum_{i=0}^{k-1} e^{L_\Phi(t_k-t_i)}|\varepsilon_i|.$$

D'où le résultat (3.32) avec $M = e^{L_\Phi(T-t_0)}$. ∎

3.4.6 Critère de consistance

Théorème 3.41 *Une condition nécessaire et suffisante pour que le schéma (3.26) soit consistant est :*

$$\forall (t,\tilde{u}) \in [0,T]\times\mathbb{R}, \quad \Phi(t,\tilde{u},0) = f(t,\tilde{u}). \tag{3.39}$$

Preuve. On pose $\varepsilon_k = u(t_{k+1}) - u(t_k) - h\Phi(t_k,u(t_k),h)$, quantité dans laquelle on fait intervenir f et $\Phi(t_k,u(t_k),0)$:

$$\begin{aligned}\varepsilon_k = &\int_{t_k}^{t_{k+1}} [f(s,u(s)) - f(t_k,u(t_k))]\,ds \\ &+ h[f(t_k,u(t_k)) - \Phi(t_k,u(t_k),0)] + h[\Phi(t_k,u(t_k),0) - \Phi(t_k,u(t_k),h)]\end{aligned}$$

Ayant supposé f et Φ C^1 et posant $g(s) = f(s,u(s))$ et $G(h) = \Phi(t_k,u(t_k),0) - \Phi(t_k,u(t_k),h)$ on en déduit que (théorème des accroissements finis) :

$$\begin{cases} |\displaystyle\int_{t_k}^{t_{k+1}} [f(s,u(s)) - f(t_k,u(t_k))]\,ds| \le h^2 \max_{s\in[t_k,t_{k+1}]} |g'(s)|, \\ h[\Phi(t_k,u(t_k),0) - \Phi(t_k,u(t_k),h)] \le h^2 \max\limits_{l\in]0,h_{\max}]} |G'(l)|, \end{cases}$$

d'où :

$$h|f(t_k,u(t_k)) - \Phi(t_k,u(t_k),0)| \le |\varepsilon_k| + h^2 \max_{s\in[t_k,t_{k+1}]} |g'(s)| + h^2 \max_{l\in]0,h_{\max}]} |G'(l)|.$$

D'où :

1- on suppose le schéma consistant : alors :

$$\sum_{k=0}^{n-1} h|f(t_k,u(t_k)) - \Phi(t_k,u(t_k),0)| \underset{h\to 0}{\longrightarrow} 0.$$

Mais la somme est une somme de Riemann qui tend vers $\int_{t_0}^T |f(s,u(s)) - \Phi(s,u(s),0)|\,ds$, et cette intégrale est donc nulle. On en déduit que pour tout $t\in[t_0,T]$ on a $|f(t,u(t)) - \Phi(t,u(t),0)| = 0$, et donc (3.39) (en prenant pour (t,u) un condition initiale quelconque (t_0,u_0)).

2- Réciproquement, on suppose (3.39). Alors :

$$|\varepsilon_k| \le h^2 \max_{s\in[t_k,t_{k+1}]} |g'(s)| + h^2 \max_{l\in]0,h_{\max}]} |G'(l)|.$$

D'où, avec $\sum_{k=1}^n h^2 = nh^2 = h(T-t_0)$:

$$\sum_k |\varepsilon_k| \le h(T-t_0)(\max_{s\in[t_0,T]} |g'(s)| + \max_{l\in]0,h_{\max}]} |G'(l)|),$$

d'où la consistance. ∎

3.4.7 Ordre d'un schéma

Les critères de stabilité et consistance ci-dessus permettent d'établir la convergence, mais pas la 'rapidité' de la convergence. La définition suivante précise le critère de consistance :

Définition 3.42 Le schéma (3.26) est d'ordre p, avec $p \in \mathbb{N}$, si étant donnée une solution u de (3.1) :

$$\exists C > 0, \qquad \sum_{k=0}^{n-1} |\varepsilon_k| \leq Ch^p, \tag{3.40}$$

i.e. :

$$\exists C > 0, \qquad \sum_{k=0}^{n-1} |u(t_{k+1}) - u(t_k) - h\Phi(t_k, u(t_k), h)| \leq Ch^p, \tag{3.41}$$

ou encore si cette somme est $O(h^p)$.

On en déduit immédiatement :

Théorème 3.43 *Si le schéma (3.26) est stable et d'ordre p, alors :*

$$\max_{k=0,n} (|u(t_k) - \tilde{u}_k|) = O(h^p).$$

Preuve. Similaire à la démonstration du théorème 3.39. ∎

Ainsi, pour un schéma d'ordre p, si on refait un calcul avec un pas de temps égal à $\frac{h}{2}$ (après avoir fait un calcul avec un pas de temps égal à h), alors on sait qu'a priori l'erreur va être divisée par 2^p. Ainsi pour la méthode d'Euler explicite, l'erreur est divisée par 2 (schéma d'ordre 1) alors que pour Runge-Kutta, l'erreur est divisée par 16, ce qui explique la différence de performance entre ces 2 méthodes.

3.5 Schéma A-stable

3.5.1 A-stabilité, rayon de stabilité

Pour cette notion, on ne considère que l'équation différentielle avec $f(x, y) = -\lambda y$ dans (3.1), i.e.

$$\begin{cases} u'(t) = -\lambda u(t), \quad t \in [0, \infty[, \\ u(0) = u_0, \end{cases} \tag{3.42}$$

La solution analytique est $u(t) = u_0 e^{-\lambda t}$. En particulier $u(t) \underset{t\to\infty}{\longrightarrow} 0$.

On veut que les schémas numériques utilisés aient la même propriété de convergence vers 0 que $e^{-\lambda t}$.

Soit un pas de temps $h{=}\Delta t$ fixé, et soit $(\tilde{u}_k)_{k\in\mathbb{N}}$ la suite des valeurs numériques obtenus par un schéma (3.27) donné (où on espère que $\tilde{u}_k \simeq u(t_k)$).

Définition 3.44 Un schéma numérique est dit A-stable, ou asymptotiquement stable, si les approximations $\tilde{u}_k$ de l'équation différentielle (3.42) vérifient :

$$\forall h > 0, \quad \tilde{u}_k \underset{k\to\infty}{\longrightarrow} 0.$$

Définition 3.45 Un schéma numérique est dit conditionnellement A-stable ssi les approximations $\tilde{u}_k$ de l'équation différentielle (3.42) vérifient :

$$\exists h^* > 0, \quad \forall h < h^*, \quad \tilde{u}_k \underset{k\to\infty}{\longrightarrow} 0, \quad \text{et} \quad h^* = \infty \text{ ne convient pas.}$$

Et le plus grand des h^* est appelé le rayon de stabilité et noté R.

3.5.2 Exemples

Exemple 3.46 Pour Euler explicite :

$$u_{n+1} = u_n - h\,\lambda u_n \quad \text{donc} \quad = (1 - h\,\lambda)u_n,$$

soit, pour tout $n \in \mathbb{N}$:

$$u_n = (1 - h\,\lambda)^n\, u_0.$$

Le schéma est conditionnellement A-stable : on a besoin de $|1 - h\lambda| < 1$ (avec $h > 0$), i.e. $h < \frac{2}{\lambda}$. Ici, le rayon de convergence est $R = \frac{2}{\lambda}$. Voir `http://yallouz.arie.free.fr/mathml/euler_instabilite.php`.

On peut également souhaiter que le schéma soit monotone, ici décroissant (comme la fonction $e^{-\lambda x}$ qu'on approche), auquel cas on prend $h < \frac{1}{\lambda}$. ∎

Exemple 3.47 Pour Euler implicite :

$$u_{n+1} = u_n - h\,\lambda u_{n+1},$$

et donc :

$$u_{n+1} = \frac{1}{1+h\lambda}\,u_n = (\frac{1}{1+h\lambda})^{n+1}\,u_0.$$

Comme pour tout $h > 0$ on a $|\frac{1}{1+h\lambda}| < 1$, le rayon de convergence est $R = \infty$: le schéma est A-stable. ▟

Exercice 3.48 Pour Crank-Nicholson, on montre que $R = \infty$: schéma inconditionnellement A-stable. ▟

Exercice 3.49 Pour Euler amélioré, on montre que $R = \frac{2}{\lambda}$: schéma conditionnellement A-stable. ▟

Exercice 3.50 Pour Runge Kutta, on montre que $R \simeq \frac{2{,}78}{\lambda}$. ▟

3.6 * Introduction : méthodes à plusieurs pas

Ces méthodes sont surtout utilisées (et efficaces) lorsqu'on sait a priori que la solution u n'a pas de variations u' importantes (problème non raide). Dans ce cas, on essaie de prévoir la valeur $\tilde{u}_{k+1}$ à l'aide de plusieurs valeurs précédentes $\tilde{u}_k$,$\tilde{u}_{k-1}$,...,$\tilde{u}_{k-n}$ (méthode à $n+1$ pas).

3.6.1 Idée de base

Les méthodes à un pas sont basées sur :

$$\tilde{u}(t_{k+1}) - \tilde{u}(t_k) = \int_{t_k}^{t_{k+1}} g(t)\,dt \quad \text{où} \quad g(t) = f(t, \tilde{u}(t)),$$

où $g(t) = f(t, \tilde{u}(t))$ est une valeur approchée de la vraie pente $f(t, u(t))$.

Et on applique une formule d'intégration numérique, pour $n_p \in \mathbb{N}$ (nombre de points d'intégration) :

$$\int_{t_k}^{t_{k+1}} g(t)\,dt \simeq \sum_{i=0}^{n_p} g(t_{ki}) w_i, \quad \forall i = 0, ..., n_p,\ t_k \le t_{ki} \le t_{k+1},\ w_i \in \mathbb{R}$$

En d'autres termes, on a remplacé la fonction $t \to g(t)$ par son approximation polynomiale $t \to p_N(t)$ polynôme (de degré $N \ge n_p$), qui vérifie $p_N(t_{ki}) = g(t_{ki})$ pour $i = 0, ..., n_p$, et on a écrit :

$$\int_{t_k}^{t_{k+1}} g(t)\,dt \simeq \int_{t_k}^{t_{k+1}} p_N(t)\,dt = \sum_{i=0}^{N} g(t_{ki}) w_i, \quad \forall i = 0, ..., N,\ t_k \le t_{ki} \le t_{k+1},\ w_i \in \mathbb{R}$$

la formule d'intégration étant supposée exacte pour les polynôme de degré N.

Dans ces formules, on ne s'est servi que des valeurs de $g(t) = f(t, \tilde{u}(t))$ entre les temps t_k et t_{k+1}. Or on connaît les valeurs de $g(t)$ au temps précédents (qu'on vient de calculer). Il pourrait donc être intéressant de profiter de cette connaissance : i.e. de considérer le polynôme d'interpolation de g à partir des points t_k, t_{k+1} et des temps précédents t_{k-1}, t_{k-2}..., et d'intégrer le polynôme exactement sur l'intervalle $[t_k, t_{k+1}]$.

On continuera à noter $g(t) = f(t, \tilde{u}(t))$.

3.6.2 Formules d'Adams–Bashforth

On considère le polynôme de degré 0 qui passe par le point $(t_k, g(t_k) = f(t_k, \tilde{u}_k))$: c'est $p_0(t) = g(t_k)$. On obtient :

$$\tilde{u}_{k+1} - \tilde{u}_k = \int_{t_k}^{t_{k+1}} p_0(t)\,dt = g(t_k)(t_{k+1} - t_k),$$

ce qui n'est autre que la formule d'Euler explicite (3.6).

On considère le polynôme de degré 1 qui passe par les points $(t_k, g(t_k))$ et $(t_{k-1}, g(t_{k-1}))$:

$$p_1(t) = g(t_k) + g[t_k, t_{k-1}](t - t_k) = g(t_k) + \frac{g(t_{k-1}) - g(t_k)}{t_{k-1} - t_k}(t - t_k),$$

où on a posé $g[t_k, t_{k-1}] = \frac{g(t_{k-1}) - g(t_k)}{t_{k-1} - t_k}$ (expression du polynôme de Newton). On obtient :

$$\tilde{u}_{k+1} - \tilde{u}_k = \int_{t_k}^{t_{k+1}} p_1(t)\,dt = g(t_k)(t_{k+1} - t_k) + \frac{g(t_{k-1}) - g(t_k)}{t_{k-1} - t_k}\,\frac{(t_{k+1} - t_k)^2}{2}.$$

On suppose pour simplifier que $h = t_{k+1} - t_k = t_{k-1} - t_k$ (schéma à pas constant). Alors :

$$\tilde{u}_{k+1} = \tilde{u}_k + \frac{h}{2}(3g(t_k) - g(t_{k-1})), \quad k \ge 1.$$

C'est une formule explicite qui donne un schéma d'ordre 1 : c'est la formule d'Adams–Bashforth à 2 pas.

Remarque 3.51 Ce schéma ne peut pas s'appliquer pour $k = 0$ car on ne connaît pas $\tilde{u}_{-1}$. Il faut donc une étape d'initialisation. On se sert par exemple du schéma d'Euler explicite, également d'ordre 1, pour calculer $\tilde{u}_1$, puis on se sert de la formule d'Adams–Bashforth pour les temps suivants. ▪

On considère le polynôme de degré 2 qui passe par les points $(t_k, g(t_k))$, $(t_{k-1}, g(t_{k-1}))$ et $(t_{k-2}, g(t_{k-2}))$, sous sa forme polynôme de Newton :

$$p_2(t) = g(t_k) + g[t_k, t_{k-1}](t - t_k) + g[t_k, t_{k-1}, t_{k-2}](t - t_k)(t - t_{k-1}),$$

où $g[t_k, t_{k-1}, t_{k-2}] = \frac{g[t_k, t_{k-1}] - g[t_{k-1}, t_{k-2}]}{(t_{k-2} - t_k)}$ (expression du polynôme de Newton). On suppose pour simplifier que $h = t_{k+1} - t_k = t_{k-1} - t_k$ (schéma à pas constant). On obtient :

$$\tilde{u}_{k+1} = \tilde{u}_k + \int_{t_k}^{t_{k+1}} p_2(t)\, dt,$$

soit :

$$\tilde{u}_{k+1} = \tilde{u}_k + \frac{h}{12}(23g(t_k) - 16g(t_{k-1}) + 5g(t_{k-2}), \quad k \geq 2.$$

C'est une formule explicite qui donne un schéma d'ordre 2 : c'est la formule d'Adams–Bashforth à 3 pas.

Remarque 3.52 Ce schéma ne peut pas s'appliquer pour $k = 0, 1$ car on ne connaît pas $\tilde{u}_{-1}$ et $\tilde{u}_{-2}$. Il faut donc 2 étapes d'initialisation. On se sert par exemple du schéma d'Euler amélioré, également d'ordre 2, pour calculer $\tilde{u}_1$ et $\tilde{u}_2$, puis on se sert de la formule d'Adams–Bashforth pour les temps suivants. ▪

Puis pour avoir des méthodes d'ordres supérieurs, on se sert des polynômes de degré n qui passe par les points $(t_k, g(t_k))$, $(t_{k-1}, g(t_{k-1}))$,...,$(t_{k-n}, g(t_{k-n}))$, et on obtient les formules d'Adams–Bashforth à n pas.

On renvoie à Crouzeix et Mignot [5] pour les valeurs obtenues.

Remarque 3.53 On montre que les formules d'Adams–Bashforth à $n+1$ pas donnent des méthodes d'ordre n, et qu'il faut n étapes d'initialisation pour déterminer les n premières valeurs $\tilde{u}_1$,...,$\tilde{u}_n$. Ces premières valeurs sont obtenues en utilisant également une méthode d'ordre n : par exemple pour $n = 4$, on se sert de la méthode de Runge–Kutta. ▪

3.6.3 Formules d'Adams–Moulton

Les méthodes d'Adams-Bashforth sont parfois soumises à des problèmes de stabilité. On peut alors préférer des méthodes implicites.

La démarche est la même que précédemment, pour obtenir des méthodes implicites : par exemple pour $n = 0$:

$$\tilde{u}_{k+1} - \tilde{u}_k = \int_{t_k}^{t_{k+1}} g(t)\, dt \simeq g(t_{k+1})(t_{k+1} - t_k),$$

ce qui n'est autre que la formule d'Euler implicite (3.10).

On considère le polynôme de degré 1 qui passe par les points $(t_{k+1}, g(t_{k+1}))$ et $(t_k, g(t_k))$:

$$p_1(t) = g(t_{k+1}) + g[t_{k+1}, t_k](t - t_{k+1}) = g(t_{k+1}) + \frac{g(t_k) - g(t_{k+1})}{t_k - t_{k+1}}(t - t_{k+1}),$$

qui est une interpolation linéaire de la pente $g(t) = f(t, u(t))$ sur l'intervalle $[t_k, t_{k+1}]$ (en particulier $p_1(t_{k+1}) = g(t_{k+1})$ et $p_1(t_k) = g(t_k)$). On obtient l'approximation :

$$\tilde{u}_{k+1} - \tilde{u}_k = \int_{t_k}^{t_{k+1}} g(t)\, dt \simeq \int_{t_k}^{t_{k+1}} p_1(t)\, dt.$$

On suppose pour simplifier que $h = t_{k+1} - t_k = t_k - t_{k-1}$ (schéma à pas constant). Alors :

$$\tilde{u}_{k+1} = \tilde{u}_k + \frac{h}{2}(g(t_{k+1}) + g(t_k)), \quad k \geq 1.$$

C'est une formule implicite qui donne un schéma d'ordre 1 : c'est la formule d'Adams–Moulton à 2 pas. C'est une méthode implicite car $g(t_{k+1}) = f(t_{k+1}, \tilde{u}(t_{k+1}))$ et pour trouver $\tilde{u}_{k+1}$, il faut résoudre l'équation "$x = \tilde{u}_k + \frac{h}{2}(f(t_{k+1}, x) + f(t_k, \tilde{u}(t_k)))$".

Remarque 3.54 Ce schéma ne peut pas s'appliquer pour $k = 0$ car on ne connaît pas $\tilde{u}_{-1}$. Il faut donc une étape d'initialisation. On se sert par exemple du schéma d'Euler explicite, également d'ordre 1, pour calculer $\tilde{u}_1$. ▪

Même démarche pour le polynôme d'ordre 2 : on obtient :

$$\tilde{u}_{k+1} = \tilde{u}_k + \frac{h}{12}(5g(t_{k+1}) + 8g(t_k) - g(t_{k-1})), \quad k \geq 2.$$

On renvoie à Crouzeix et Mignot [5] pour les valeurs obtenues pour les polynômes d'ordre n.

3.6.4 Schémas de prédiction–évaluation–correction (PEC)

Le prototype est donné par la méthode d'Heun. Ici pour éviter les calculs parfois difficiles et long des méthodes implicites d'Adams–Moulton, on peut estimer la valeur $\tilde{u}_{k+1}$ à l'aide d'une méthode explicite. On obtient les méthodes PEC de Prédiction, Évaluation, Correction.

Cas des méthodes d'ordre 1 d'Adams (pour $k \geq 1$, i.e. après l'étape d'initialisation) :
(P)- Prédiction d'Adams–Bashforth : $\tilde{u}_{k+1}^{\text{pred}} = \tilde{u}_k + \frac{h}{2}(3g(t_k) - g(t_{k-1}))$,
(E)- Évaluation pente à droite $g_{k+1} = f(t_{k+1}, \tilde{u}_{k+1}^{\text{pred}})$,
(C)- Correction d'Adams–Moulton : $\tilde{u}_{k+1} = \tilde{u}_k + \frac{h}{2}(g_{k+1} + g(t_k))$.

Cas des méthodes d'ordre 2 d'Adams (pour $k \geq 2$, i.e. après les étapes d'initialisation) :
(P)- Prédiction d'Adams–Bashforth : $\tilde{u}_{k+1}^{\text{pred}} = \tilde{u}_k + \frac{h}{12}(23g(t_k) - 16g(t_{k-1}) + 5g(t_{k-2})$,
(E)- Évaluation pente à droite $g_{k+1} = f(t_{k+1}, \tilde{u}_{k+1}^{\text{pred}})$,
(C)- Correction d'Adams–Moulton : $\tilde{u}_{k+1} = \tilde{u}_k + \frac{h}{12}(5g_{k+1} + 8g(t_k) - g(t_{k-1}))$.

Ces schémas sont en général améliorés avec la méthode P(EC)^t, de Prédiction et t étapes de "Évaluation+Correction". Par exemple, dans le cas des méthodes d'ordre 1 d'Adams (pour $k \geq 1$, i.e. après l'étape d'initialisation) :
(P)- Prédiction d'Adams–Bashforth : $\tilde{u}_{k+1}^0 = \tilde{u}_k + \frac{h}{2}(3g(t_k) - g(t_{k-1}))$,
(E)- Évaluation pente $g_{k+1}^0 = f(t_{k+1}, \tilde{u}_{k+1}^0)$,
(C)- Correction d'Adams–Moulton $\tilde{u}_{k+1}^1 = \tilde{u}_k + \frac{h}{2}(g_{k+1}^0 + g(t_k))$,
(E)- Évaluation pente $g_{k+1}^1 = f(t_{k+1}, \tilde{u}_{k+1}^1)$,
(C)- Correction d'Adams–Moulton $\tilde{u}_{k+1}^2 = \tilde{u}_k + \frac{h}{2}(g_{k+1}^1 + g(t_k))$,...
... t étapes de (EC) = Évaluation–Correction donnant g_{k+1}^{t-1} et $\tilde{u}_{k+1}^t$ qu'on choisit comme valeur d'approximation $= \tilde{u}_{k+1}$.

Une méthode usuelle, notée P(EC)^t consiste à faire t étapes (EC) avec $t = 1, 2$ ou 3.
On renvoie à Crouzeix et Mignot [5] pour les calculs d'erreur qui justifie cette approche.
Noter également qu'on appelle méthode PECE la méthode PEC où on ajoute également l'étape final d'évaluation g_{k+1}^n.

3.7 * Equations du second ordre : méthode de Newmark

3.7.1 Introduction

On s'intéresse à la résolution numérique d'une équation différentielle du second ordre de type : trouver une fonction $u \in \mathcal{F}(\mathbb{R}; \mathbb{R})$ telle que :

$$u''(t) = f(t, u(t), u'(t)), \quad u(0) = u_0, \quad u'(0) = u_1, \tag{3.43}$$

où $f : \left\{ \begin{array}{rl} \mathbb{R}^3 & \to \mathbb{R} \\ (t,x,y) & \mapsto f(t,x,y) \end{array} \right\}$ est une fonction régulière donnée, et où les conditions initiales u_0 et u_1 sont des réels donnés.

On réécrit l'équation (3.43) sous la forme équation linéaire du premier ordre :

$$\left\{ \begin{array}{l} u'(t) = u'(t), \\ u''(t) = f(t, u(t), u'(t)), \end{array} \right. \tag{3.44}$$

i.e. :

$$U' = F(t, U), \tag{3.45}$$

où on a posé :

$$U = \begin{pmatrix} u \\ u' \end{pmatrix}, \quad F(t, U) = F(t, u, u') = \begin{pmatrix} u' \\ f(t, u, u') \end{pmatrix}. \tag{3.46}$$

La résolution de (3.45) donnera une solution $U : \left\{ \begin{array}{rl} \mathbb{R}^3 & \to \mathbb{R}^2 \\ (t,x,y) & \mapsto U(t,x,y) = \begin{pmatrix} u_1(t,x,y) \\ u_2(t,x,y) \end{pmatrix} \end{array} \right\}$ telle que sa première composante est la solution $u_1 = u$ cherchée, et la seconde composantes u_2 n'est autre que la dérivée u' de u.

On peut donc appliquer les méthodes de résolution exposées précédemment (à un pas ou à pas multiples). On regardera ici une méthode particulièrement adaptée aux systèmes du second ordre, la méthode de Newmark.

Exemple 3.55 L'équation des ondes "$u''(t) - \nu u'(t) - ku(t) = g(t)$" est de cette forme, où $f(t, x, y) = kx + \nu y + g(t)$ est ici une fonction affine en x et y, et où $F(t, U) = A.U + G$ où on a posé $A = \begin{pmatrix} 0 & 1 \\ k & \nu \end{pmatrix}$ et $G = \begin{pmatrix} 0 \\ g \end{pmatrix}$. Cette équation est linéaire. ∎

3.7.2 Méthode de Newmark

On s'intéresse à la résolution de (3.45) lorsqu'elle est de la forme (3.44). L'inconnue du problème est la fonction $U : t \to U(t) = \begin{pmatrix} u_1(t) \overset{\text{noté}}{=} u(t) \\ u_2(t) \overset{\text{noté}}{=} v(t) \end{pmatrix}$ (à valeurs vectorielles), i.e. les inconnues du problème sont les composantes u et v de U. La fonction F de (3.45) donnera bien sûr (après calcul) $v = u'$.

On commence par la partie usuelle : l'estimation de $v(= u')$ en fonction de $v'(= u'' = f)$: Newmark propose le θ-schéma, i.e. si $\tilde{v}(t_n) = \tilde{v}_n$ est une estimation de v à l'instant t_n :

$$\tilde{v}(t_{n+1}) = \tilde{v}(t_n) + h\big((1-\theta)\tilde{v}'(t_n) + \theta\tilde{v}'(t_{n+1})\big), \quad 0{\leq}\theta{\leq}1. \tag{3.47}$$

Et on verra que $\theta = \frac{1}{2}$ sera un bon choix (pour la stabilité et la précision, choix de la méthode de Crank–Nicholson). Notant $\tilde{f}_n = f(t_n, \tilde{u}_n, \tilde{v}_n)$, où $\tilde{u}_n$ est une estimation de $u(t_n)$, on a dans ce cas :

$$\tilde{v}_{n+1} = \tilde{v}_n + \frac{h}{2}(\tilde{f}_n + \tilde{f}_{n+1}). \tag{3.48}$$

Puis on s'intéresse à u en fonction de u' et u'' (on rappelle que u'' est "connu" : c'est la donnée f du problème). On se sert des deux développements limités au second ordre (l'un à droite en t_n et l'autre à gauche en t_{n+1}), notant $h = \Delta t$:

$$\begin{cases} u(t_{n+1}) = u(t_n) + hu'(t_n) + \dfrac{h^2}{2}u''(t_n) + O(h^3), \\ u(t_n) = u(t_{n+1}) - hu'(t_{n+1}) + \dfrac{h^2}{2}u''(t_{n+1}) + O(h^3). \end{cases} \tag{3.49}$$

Puis la deuxième équation $(3.49)_2$ est transformée avec $u'(t_{n+1}) = u'(t_n) + hu''(t_n) + O(h^2)$ en :

$$u(t_n) = u(t_{n+1}) - hu'(t_n) - h^2u''(t_n) + \frac{h^2}{2}u''(t_{n+1}) + O(h^3).$$

On multiplie la première équation de (3.49) par $1{-}\alpha$ et la seconde par $-\alpha$ et on somme :

$$u(t_{n+1}) = u(t_n) + hu'(t_n) + \frac{h^2}{2}((1{+}\alpha)u''(t_n) - \alpha u''(t_{n+1})) + O(h^3).$$

On pose $\beta = -\dfrac{\alpha}{2}$, et on obtient :

$$u(t_{n+1}) = u(t_n) + hu'(t_n) + h^2((\frac{1}{2}{-}\beta)u''(t_n) + \beta u''(t_{n+1})) + O(h^3).$$

Puis, comme $u''(t) = f(t, u(t), u'(t))$, Newmark propose de calculer la solution approchée $\tilde{u}(t_n) = \tilde{u}_n$, à l'aide du schéma implicite, posant $\tilde{f}_n = f(t_n, \tilde{u}_n, \tilde{v}_n)$:

$$\tilde{u}_{n+1} = \tilde{u}_n + h\tilde{v}_n + h^2((\frac{1}{2}{-}\beta)\tilde{f}_n + \beta\tilde{f}_{n+1})), \tag{3.50}$$

sachant que $\tilde{v}_n$ est donné par (3.47). Une "bonne valeur" de β sera $\beta = \frac{1}{4}$ (pour la stabilité et la précision), ce qui avec (3.48) donne dans ce cas le système à résoudre :

$$\begin{cases} \tilde{u}_{n+1} = \tilde{u}_n + h\tilde{v}_n + \dfrac{h^2}{4}(\tilde{f}_n + \tilde{f}_{n+1})), \\ \tilde{v}_{n+1} = \tilde{v}_n + \dfrac{h}{2}(\tilde{f}_n + \tilde{f}_{n+1})). \end{cases} \tag{3.51}$$

C'est un système implicite (si on avait pris $\theta = 0 = \beta$, on aurait obtenu un schéma explicite mais peu précis et conditionnellement stable.)

Remarque 3.56 Cas particulier où $f(t, u, u') = f(t, u)$, i.e. où $f(t, x, y) = f(t, x)$ est indépendant de y. Voir par exemple l'équation des ondes de l'exemple 3.55 lorsque $\nu = 0$. On élimine alors v_n, sachant que $\tilde{f}_n$ ne dépend pas de $\tilde{v}_n$, à l'aide de (3.50) :

$$(\tilde{u}_{n+2} - \tilde{u}_{n+1}) - (\tilde{u}_{n+1} - \tilde{u}_n) = h(\tilde{v}_{n+1} - \tilde{v}_n) + h^2(\beta\tilde{f}_{n+2} + (\frac{1}{2}{-}\beta)\tilde{f}_{n+1} - \beta\tilde{f}_{n+1} - (\frac{1}{2}{-}\beta)\tilde{f}_n),$$

d'où avec (3.48) :

$$\begin{aligned} \tilde{u}_{n+2} - 2\tilde{u}_{n+1} + \tilde{u}_n &= h^2((1-\theta)\tilde{f}_n + \theta\tilde{f}_{n+1}) + h^2(\beta\tilde{f}_{n+2} + (\frac{1}{2}{-}2\beta)\tilde{f}_{n+1} - (\frac{1}{2}{-}\beta)\tilde{f}_n) \\ &= h^2\big(\beta\tilde{f}_{n+2} + (\frac{1}{2}{-}2\beta{+}\theta)\tilde{f}_{n+1} + (\frac{1}{2}{-}\beta{-}\theta)\tilde{f}_n\big). \end{aligned}$$

Et si on connaît $\tilde{u}_0$ et $\tilde{u}_1$, ce schéma est explicite pour $\beta = 0$: comme les conditions initiales sont connues, $\tilde{u}_0$ est connu $= u(0)$, et $\tilde{u}_1$ peut-être estimé à l'aide de $u'(0) \simeq \frac{\tilde{u}_1 - \tilde{u}_0}{h}$. ∎

Références

[1] Abramowitz M., Stegun I.A. : Handbook of Mathematical Functions. Dover 1970.

[2] Arnaudiès J.M., Fraysse H. : Cours de mathématiques – 2, Analyse, classes préparatoires. Dunod 1988.

[3] Bernardi C., Maday Y. : Approximations spectrales de problèmes aux limites elliptiques. Springer-Verlag 1992.

[4] Courant R., John F. : Introduction to Calculus and Analysis. Volume 1, 2. Springer-Verlag 1989.

[5] Crouzeix M., Mignot A.L. : Analyse numérique des équations différentielles, Seconde édition, collection mathématiques appliquées pour la maîtrise. Masson 1992.

[6] Derrick W.R., Grossman S.I. : Introduction to Differential Equations with Boundary Value Problems, Third Edition. West Publishing Company 1987.

[7] Fortin A. : Analyse numérique. Ed. de l'École Polytechnique de Montréal (1995).

[8] Hirsch M.W., Smale S. : Differential Equations, Dynamical Systems, and Linear Algebra. Academic Press 1974.

[9] Holmgren R.A. : A First Course in Discrete Dynamical Systems, Second Edition. Springer 1996.

[10] Hubbard J.H., West B.H. : Differential Equations : A Dynamical Systems Approach. Texts in Applied Mathematics 5, Springer 1991.

[11] O'Neil P.V. : Advanced Engineering Mathematics, Fourth Edition. PWS 1995.

[12] Reinhard : *équations différentielles*. Dunod.

[13] Schatzman M. : Analyse numérique, cours et exercices pour la licence. InterEditions 1991.

[14] Strang G. : Calculus. Wellesley Cambridge Press 1991.

[15] Théodor R. : Initiation à l'analyse numérique. CNAM cours A, Masson 1982.