

Université de Montpellier II
Licence de Mathématiques.
Quelques notions d'Analyse Numérique.

Pascal Azerad ©

24 novembre 2009

Table des matières

1	Interpolation.	5
1.1	Interpolation polynomiale	5
1.1.1	Existence et unicité du polynôme d'interpolation.	5
1.1.2	Interpolation et approximation	6
1.1.3	Phénomène de Runge	7
1.2	Polynômes de Lagrange.	9
1.3	Polynômes de Newton.	11
1.4	Splines cubiques naturelles.	12
1.4.1	Spline= tige souple	12
1.4.2	Calcul des splines cubiques	13
2	Quadrature.	15
2.1	Méthodes élémentaires	15
2.1.1	Principe d'intégration numérique : poids et noeuds.	15
2.1.2	Rectangle, Trapèze, Simpson	15
2.1.3	Ordre et Formules d'erreur	17
2.2	Méthodes à pas multiples	20
2.2.1	Principe des méthodes composées.	20
2.2.2	Accélération de la convergence. Extrapolation de Romberg- Richardson.	21
2.2.3	Méthode de Montecarlo	23
2.3	Méthode de Gauss et polynômes orthogonaux.	23
2.3.1	Introduction.	23
2.3.2	Méthode de Gauss.	23
3	Méthode des moindres carrés.	29
3.1	Présentation du problème. Régression linéaire.	29
3.1.1	Droite des moindres carrés.	29
3.2	Interprétation géométrique et écriture matricielle.	31
3.2.1	Moindres carrés pondérés.	34
3.3	Algorithme de résolution numérique.	35

Préambule

Ce fascicule volontairement succinct poursuit les buts suivants :

- être assimilable par un étudiant de licence en sept séances.
- servir de tremplin pour les cours de Master et de vade-mecum pour les capétiens.

Ce document est une première version, l'auteur est très reconnaissant pour toute erreur, coquille ou commentaire qu'on voudra bien lui adresser à

`azerad@math.univ-montp2.fr`

Chapitre 1

Interpolation.

1.1 Interpolation polynomiale

1.1.1 Existence et unicité du polynôme d'interpolation.

Le problème de l'interpolation est le suivant : Etant donné $n + 1$ points distincts (x_i, y_i) , $i = 0 \dots n$, construire une courbe passant par ces points. On désire de plus que la courbe soit *lisse* et donnée par une formule simple, du type $y = f(x)$. Ce problème intervient dans de nombreuses situations industrielles.

Exemple. génie mécanique : construction de carrosserie par machines à commande numérique. On rentre un certains nombres de points de contrôle (x_i, y_i) et la machine calcule et usine la forme passant par ces points.

Exercice. Montrer que s'il existe une fonction f dont le graphe contient les points (x_i, y_i) , $i = 0 \dots n$ alors les abscisses x_i sont toutes distinctes.

Une façon simple de résoudre le problème est de chercher la fonction f sous forme d'un *polynôme* (qui est toujours C^∞). L'algèbre linéaire donne immédiatement le résultat d'existence et d'unicité du polynôme d'interpolation.

Théorème 1 *Etant donnés $n + 1$ abscisses distinctes x_0, x_1, \dots, x_n et $n + 1$ valeurs quelconques y_0, y_1, \dots, y_n , il existe un unique polynôme P de degré au plus n tel que $P(x_j) = y_j$, $j = 0 \dots n$.*

Preuve. Notons $\mathbb{R}_n[X] = \{\sum_{k=0}^n a_k X^k, a_k \in \mathbb{R}\}$ l'espace vectoriel des polynômes de degré inférieur ou égal à n . Soit l'application linéaire :

$$F : \mathbb{R}_n[X] \rightarrow \mathbb{R}^{n+1} \\ P \mapsto (P(x_0), P(x_1), \dots, P(x_n))$$

Le noyau de F se réduit au polynôme nul $P = 0$. En effet un polynôme de degré inférieur ou égal à n a au plus n racines, sauf s'il s'agit du polynôme nul. L'application F est donc injective. D'autre part la dimension de $\mathbb{R}_n[X]$ est $n + 1$, donc F est un isomorphisme. ■

On pourrait penser que ce théorème résout complètement la question de l'interpolation. En fait, le résultat précédent est *théorique*, il reste à construire le polynôme P , de façon simple, précise et flexible : c'est l'objet du reste du chapitre.

1.1.2 Interpolation et approximation

On peut aussi se poser la question de la précision de l'interpolation polynômiale pour approcher une fonction donnée f sur un segment $[a, b]$. Etant donnée une fonction f , quel erreur commet-on si on la remplace par un polynôme prenant les mêmes valeurs en des points fixés ?

Soit f une fonction définie sur le segment $[a, b]$. Soit $n + 1$ points distincts x_0, x_1, \dots, x_n de $[a, b]$. D'après le théorème 1, il existe un unique polynôme P_n de degré $\leq n$ tel que

$$P_n(x_j) = f(x_j), \quad j = 0 \dots n.$$

On l'appelle *polynôme d'interpolation* de f aux points $x_i, i = 0, \dots, n$.

Théorème 2 (des accroissements finis généralisés) *Soit f une fonction de classe C^{n+1} sur le segment $[a, b]$. Soit $n+1$ points distincts x_0, x_1, \dots, x_n de $[a, b]$. Soit P_n le polynôme de degré $\leq n$ qui interpole f aux x_i . Il vérifie pour tout x de $[a, b]$:*

$$f(x) - P_n(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!} (x - x_0)(x - x_1) \dots (x - x_n)$$

où ξ_x est un point de $[a, b]$ ¹.

Preuve. La preuve repose sur le théorème de Rolle appliqué « en rafale ». Considérons la fonction

$$\varphi(t) = f(t) - P_n(t) - \lambda \cdot (t - x_0)(t - x_1) \dots (t - x_n)$$

où λ est un paramètre réel qui sera fixé plus tard. Soit maintenant $x \in \mathbb{R}$ différent des $x_j, j = 0 \dots n$. La fonction φ s'annule en x_0, x_1, \dots, x_n . Choisissons maintenant λ de sorte que φ s'annule également en $t = x$. C'est possible car $(x - x_0)(x - x_1) \dots (x - x_n) \neq 0$. Ainsi la fonction φ s'annule en x_0, x_1, \dots, x_n et x donc au moins $n + 2$ fois sur l'intervalle $[a, b]$. Par application du théorème de Rolle sur chaque intervalle séparant deux zéros consécutifs de φ , on obtient que la fonction dérivée φ' s'annule donc au moins $n + 1$ fois sur $[a, b]$. De même la dérivée seconde φ'' s'annule donc au moins n fois sur $[a, b]$. En réitérant, on obtient que $\varphi^{(n+1)}$ s'annule au moins une fois sur $[a, b]$. Notons ξ_x ce point tel que $\varphi^{(n+1)}(\xi_x) = 0$. Mais la dérivée $(n + 1)$ -ème de $\varphi(t)$ se calcule aisément (P_n étant de degré n disparaît !):

$$\varphi^{(n+1)}(t) = f^{(n+1)}(t) - \lambda(n+1)!$$

D'où l'on tire que

$$\lambda = \frac{f^{(n+1)}(\xi_x)}{(n+1)!}.$$

Ecrivons alors le fait que φ s'annule en $t = x$.

$$0 = f(x) - P_n(x) - \lambda \cdot (x - x_0)(x - x_1) \dots (x - x_n).$$

En remplaçant λ par sa valeur on obtient le résultat.

¹dépendant évidemment de x

■

Remarque. Lorsque $n = 0$, on retrouve le théorème des accroissements finis $f(x) - f(x_0) = f'(\xi)(x - x_0)$. □

Ce théorème précise l'erreur d'*approximation* entre f et P_n .

$$|f(x) - P_n(x)| \leq \frac{M_{n+1}}{(n+1)!} |(x - x_0)(x - x_1) \dots (x - x_n)|$$

où $M_{n+1} = \max_{t \in [a, b]} |f^{(n+1)}(t)|$ qui est bien défini puisque par hypothèse $f^{(n+1)}$ est continue sur $[a, b]$.

Supposons que les x_i soit régulièrement espacés, i.e.

$$x_i = a + i \cdot h, \quad h = (b - a)/n, \quad i = 0, \dots, n$$

on peut alors calculer aisément le maximum de $\prod_{0 \leq i \leq n} (x - x_i)$.

Exercice. Montrer que

$$|f(x) - P_1(x)| \leq \frac{1}{8} M_2 h^2$$

$$|f(x) - P_2(x)| \leq \frac{\sqrt{3}}{27} M_3 h^3$$

$$|f(x) - P_3(x)| \leq \frac{3}{128} M_4 h^4$$

En conclusion, nous voyons que la qualité de l'interpolation dépend de $\varphi(x) = (x - x_0)(x - x_1) \dots (x - x_n)$. Cette fonction s'annule aux x_i et vu les résultats particulier des exercices ci-dessus, on pourrait croire que

$$|f(x) - P_n(x)| = O(h^{n+1})$$

donc que la qualité de l'approximation augmente avec le nombre de points. Or il n'en est rien, comme le montre la figure 1.1 qui trace la fonction φ pour 9 points x_i équidistants sur $[-1, 1]$ (le pas $h = 0.25$). Le maximum de φ est environ 0.02 et non pas $h^9 \approx 4.10^{-6}$. Ce phénomène est général avec des points x_i équidistants : lorsqu'on augmente le nombre de points, l'approximation est très mauvaise aux extrémités. Le polynôme d'interpolation oscille avec une grande amplitude vers les extrémités.

1.1.3 Phénomène de Runge

Soit la fonction $t \mapsto 1/(1 + t^2)$. Considérons son polynôme d'interpolation aux points $-5, -4, \dots, 4, 5$. Il se calcule aisément, par exemple avec MAPLE.

$$t \mapsto -\frac{1}{44200} t^{10} + \frac{7}{5525} t^8 - \frac{83}{3400} t^6 + \frac{2181}{11050} t^4 - \frac{149}{221} t^2 + 1$$

Si on trace les graphes des fonctions, on voit le problème des oscillations aux extrémités², voir fig. 1.2

²On peut résoudre ce problème en prenant des x_i non équidistants, en raffinant la subdivision i.e. en prenant plus de points vers les extrémités, voir [2]

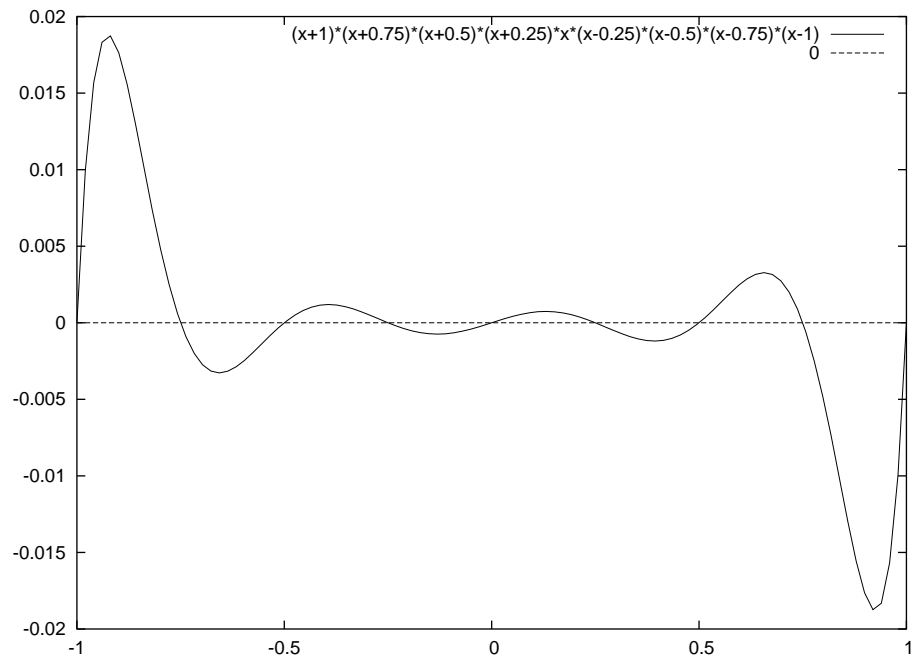


FIG. 1.1 – Erreur interpolation

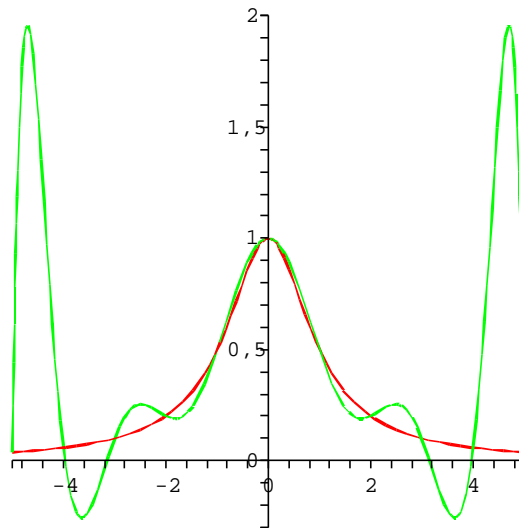


FIG. 1.2 – Phénomène de Runge

1.2 Polynômes de Lagrange.

Soient $n+1$ points distincts x_0, x_1, \dots, x_n . Nous allons donner une base de $\mathbb{R}_n[X]$ particulièrement adaptée à l'interpolation aux points x_i . Pour $j = 0, \dots, n$, notons

$$L_j(X) = \frac{\prod_{i \neq j} (X - x_i)}{\prod_{i \neq j} (x_j - x_i)}$$

Remarque. Attention, c'est bien un polynôme, et non une fraction rationnelle : l'inconnue X ne figure qu'au numérateur. Le dénominateur est une constante de *normalisation* prévue pour que $L_j(x_j) = 1$. \square

Exemple. Fixons $n = 2$ et explicitons les polynômes de Lagrange.

$$L_0(X) = \frac{(X - x_1)(X - x_2)}{(x_0 - x_1)(x_0 - x_2)}$$

$$L_1(X) = \frac{(X - x_0)(X - x_2)}{(x_1 - x_0)(x_1 - x_2)}$$

$$L_2(X) = \frac{(X - x_0)(X - x_1)}{(x_2 - x_0)(x_2 - x_1)}$$

Exercice. Visualisez quelques polynômes de Lagrange avec un logiciel graphique. (voir fig 1.3).

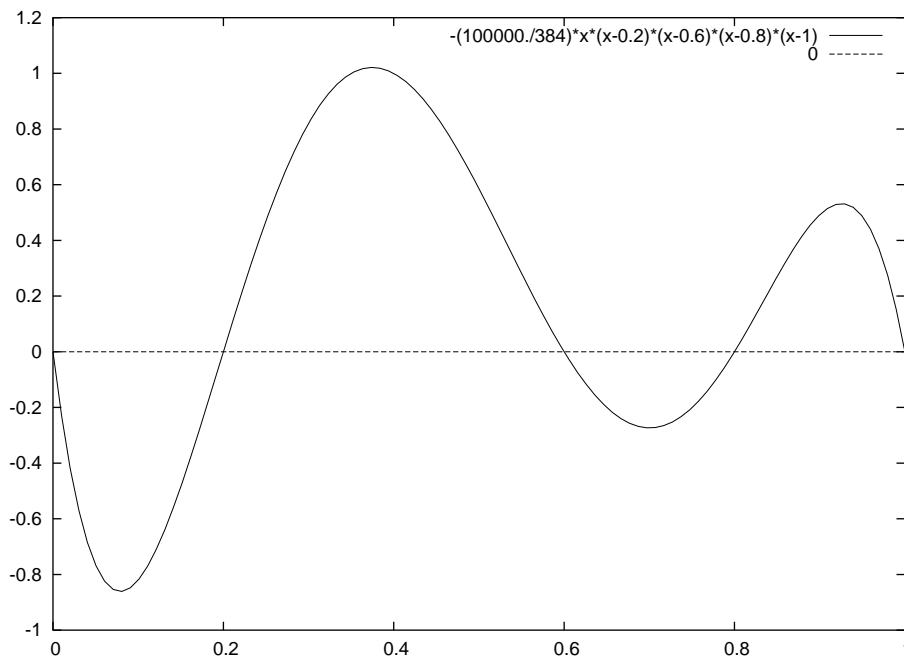


FIG. 1.3 – Un Polynôme de Lagrange de degré 5.

Propriétés 1 L_j est le polynôme de degré n tel que $L_j(x_i) = 0$ pour $i \neq j$ et $L_j(x_j) = 1$:

$$L_j(x_i) = \delta_{i,j}$$

Preuve. Il suffit de faire $X \leftarrow x_i$ dans la définition de L_j . ■

Proposition 1 La famille $(L_j)_{0 \leq j \leq n}$ est une base de $\mathbb{R}_n[X]$.

Preuve. Comme la famille contient exactement $n+1$ éléments, il suffit de prouver qu'elle est libre. Soit une combinaison linéaire

$$\sum_j \lambda_j L_j(X) = 0$$

Faisant $X \leftarrow x_i$ dans cette égalité, on obtient un seul terme non nul :

$$\lambda_i = 0.$$

En faisant cela pour chaque i , on obtient que la combinaison linéaire est triviale. ■

Propriétés 2 (formule de Lagrange) Etant donnés $n+1$ abscisses distinctes x_0, x_1, \dots, x_n et $n+1$ valeurs quelconques y_0, y_1, \dots, y_n , le polynôme d'interpolation est donné par l'expression :

$$P(X) = \sum_{0 \leq j \leq n} y_j L_j(X)$$

Preuve. Il suffit de remarquer que $\sum_j y_j L_j(x_i) = y_i L_i(x_i) = y_i$. Un seul terme de la somme est non nul ! ■

Exercice. Formule barycentrique. Soit $w_i = \frac{1}{\prod_{j \neq i} (x_i - x_j)}$ et $\mu_i(t) = \frac{w_i}{t - x_i}$. Montrer que

$$P_n(t) = \frac{\sum_i y_i \mu_i(t)}{\sum_i \mu_i(t)}.$$

Cette formule est utile pour le calcul pratique (Cf TP). On remarquera que les coefficients $\mu_i(t)$ ne sont pas positifs en général et sont à recalculer pour chaque point t .

Exercice. Montrer que $\sum_i w_i = 0$ (Indication : interpoler la fonction constante 1 et considérer le coefficient dominant).

Les polynômes de Lagrange conduisent à des formules élégantes. On verra en TP l'algorithme de Neville qui permet un calcul rapide du polynôme d'interpolation. Cependant la formule de Lagrange souffre d'un handicap : l'ajout d'un point nécessite de recalculer tous les polynômes.

1.3 Polynômes de Newton.

Un base de l'espace $\mathbb{R}_n[X]$ des polynômes de degré inférieur ou égal à n particulièrement adaptée à l'interpolation est formée des polynôme de Newton :

$$1, (X - x_0), (X - x_0)(X - x_1), \dots, (X - x_0)(X - x_1)(X - x_2) \dots (X - x_{n-1})$$

Les polynômes $N_k(X) = \prod_{j \leq k-1} (X - x_j)$ sont appelés polynômes de Newton. C'est une famille étagée, i.e. $\deg N_k = k$, donc c'est une base de $\mathbb{R}_n[X]$. De plus les racines de N_k sont exactement x_0, x_1, \dots, x_{k-1} . Le polynôme d'interpolation s'exprime suivant cette base par une formule élégante, qui se démontre par récurrence (exercice).

$$P_n(x) = f[x_0] + f[x_0, x_1](x - x_0) + \dots + f[x_0, x_1, \dots, x_n](x - x_0)(x - x_1) \dots (x - x_{n-1}). \quad (1.1)$$

où les $f[x_0, x_1, \dots, x_k]$ appelées *différences divisées* sont définies par récurrence :

$$f[x_0] := f(x_0), \quad f[x_0, x_1] := \frac{f[x_1] - f[x_0]}{x_1 - x_0}$$

$$f[x_0, x_1, \dots, x_k] := \frac{f[x_0, x_1, \dots, x_{k-2}, x_k] - f[x_0, x_1, \dots, x_{k-1}]}{x_k - x_{k-1}}$$

Exercice. si f est de classe \mathcal{C}^k montrer qu'il existe ξ tel que $f[x_0, x_1, \dots, x_k] = \frac{f^{(k)}(\xi)}{k!}$.

Remarque. Dans la formule de Newton 1.1, on peut faire *coïncider* des points : Si par exemple $x_0 = x_1$, x_0 est un point "double" la différence divisée devient

$$f[x_0, x_0] = \lim_{x_1 \rightarrow x_0} \frac{f[x_1] - f[x_0]}{x_1 - x_0} = f'(x_0)$$

Le polynôme d'interpolation vérifie donc en x_0 $P(x_0) = f(x_0)$ ET $P'(x_0) = f'(x_0)$. On peut faire évidemment coïncider plusieurs points (points triples, pour faire coïncider les dérivées secondes, etc. Lorsqu'on veut interpoler une fonction par un polynôme en faisant coïncider les dérivées premières et seconde, on parle d'interpolation de Hermite. \square

Remarque. La formule reste vraie si les x_j ne sont pas rangés par ordre croissant. \square

Remarque. On peut voir la formule de Newton comme une généralisation de la formule de Taylor. En effet, si les x_i sont tous égaux à x_0 , et si on généralise la notion d'interpolation de la façon suivante : on cherche un polynôme de degré n tel que

$$P^{(j)}(x_0) = f^{(j)}(x_0) \quad j = 0 \dots n.$$

\square

L'intérêt des polynômes de Newton est que la formule de Newton se comporte bien lorsqu'on ajoute un point supplémentaire x_{n+1} , il suffit de corriger la formule 1.1 en ajoutant le terme $f[x_0, x_1, \dots, x_{n+1}](x - x_0)(x - x_1) \dots (x - x_n)$.

1.4 Splines cubiques naturelles.

1.4.1 Spline= tige souple

Pour des raisons pratiques, on aimerait interpoler une fonction avec une méthode – qui soit flexible, en ce sens qu'elle gère facilement l'ajout d'un point supplémentaire, – qui ne génère pas d'oscillations du type phénomène de Runge aux extrémités. Une méthode, relativement récente (les années 1970) et maintenant universellement adoptée, est celle des splines cubiques, que nous présenterons succinctement. La figure 1.4 illustre l'efficacité de la méthode. On a interpolé la fonction $t \mapsto 1/(1+t^2)$ aux points $-5, -3, -1, 0, 1, 3, 5$ avec le polynôme d'interpolation de degré 6 et aussi avec une fonction spline cubique (calculée avec MAPLE).

$$s := \begin{cases} -\frac{31}{1040} - \frac{621}{5200}t - \frac{33}{1040}t^2 - \frac{11}{5200}t^3, & t < -3, \\ \frac{7567}{5200} + \frac{7101}{5200}t + \frac{2409}{5200}t^2 + \frac{11}{208}t^3, & -3 \leq t < -1, \\ 1 - \frac{1173}{5200}t^2 - \frac{523}{5200}t^3, & -1 \leq t < 0, \\ 1 - \frac{1300}{1300}t^2 + \frac{1300}{523}t^3, & 0 \leq t < 1, \\ \frac{7567}{5200} - \frac{7101}{5200}t + \frac{2409}{5200}t^2 - \frac{11}{208}t^3, & 1 \leq t < 3, \\ -\frac{31}{1040} + \frac{621}{5200}t - \frac{33}{1040}t^2 + \frac{11}{5200}t^3, & t \geq 3 \end{cases}$$

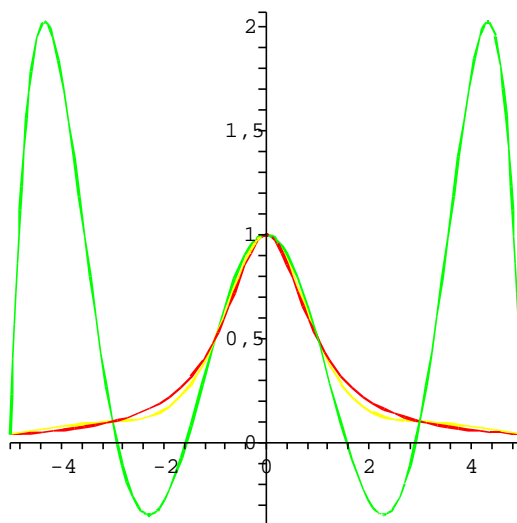


FIG. 1.4 – Spline cubique vs interpolation de Lagrange

Les fonctions spline ne sont plus des polynômes, elles sont seulement polynomiale *par morceaux*. Elles sont caractérisées par les trois propriétés

1. polynomiales par morceaux, de degré trois (d'où le terme cubique) sur chaque (x_i, x_{i+1})

2. elles sont de classe \mathcal{C}^2 . En chaque x_i , les dérivées premières et secondes à gauche et à droite coïncident.
3. la dérivée seconde s'annule aux deux extrémités x_0 et x_n .

Les fonctions spline sont en fait solution d'un problème mécanique simple : on imagine une tige élastique ou tringle souple ³ passant par les points $(x_i, f(x_i))$. La forme $x \mapsto s(x)$ adoptée par la tringle doit minimiser l'énergie potentielle (énergie élastique de déformation) qui s'exprime (voir cours d'élasticité) :

$$J = \frac{1}{2} \int_{x_0}^{x_n} s''(x)^2 dx$$

On démontre (et cela déborde du cadre de ce cours) que la fonction spline cubique s est l'unique solution du problème *variationnel* : s minimise $J(s) = \frac{1}{2} \int_{x_0}^{x_n} s''(x)^2 dx$ parmi toutes les fonctions vérifiant les propriétés

- (i) $s(x_i) = y_i, \quad i = 0, \dots, n$
- (ii) s est de classe \mathcal{C}^1 sur $[x_0, x_n]$
- (iii) s est de classe \mathcal{C}^4 sur chaque $[x_i, x_{i+1}]$

1.4.2 Calcul des splines cubiques

A partir des propriétés 1, 2, 3 il est facile de calculer la fonction s . Soit $[x_i, x_{i+1}]$ un sous-intervalle. Posons $h_i = x_{i+1} - x_i$ et cherchons la restriction de s à cet intervalle sous la forme :

$$s_i(x) = a_i(x - x_i)^3 + b_i(x - x_i)^2 + c_i(x - x_i) + d_i.$$

On obtient :

$$\begin{aligned} s_i(x_i) &= d_i &= y_i \\ s_i(x_{i+1}) &= a_i h_i^3 + b_i h_i^2 + c_i h_i + d_i &= y_{i+1} \\ s'_i(x_i) &= c_i \\ s'_i(x_{i+1}) &= 3a_i h_i^2 + 2b_i h_i + c_i \\ s''_i(x_i) &= 2b_i &= y''_i \\ s''_i(x_{i+1}) &= 6a_i h_i + 2b_i &= y''_i \end{aligned}$$

d'où l'on tire (exercice élémentaire) :

$$\begin{aligned} a_i &= \frac{1}{6h_i}(y''_{i+1} - y''_i) \\ b_i &= \frac{1}{2}y''_i \\ c_i &= \frac{1}{h_i}(y_{i+1} - y_i) - \frac{h_i}{6}(y''_{i+1} + 2y''_i) \\ d_i &= y_i \end{aligned}$$

On voit donc que les valeurs des dérivées secondes y''_k (en plus des valeurs y_k de la fonction à interpoler) déterminent complètement la spline. Pour les déterminer, nous allons utiliser la continuité de la dérivée première. On obtient

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

³c'est la traduction de l'anglais "spline"

En faisant $i \leftarrow i - 1$, On obtient la condition de continuité $s'_{i-1}(x_i) = s'_i(x_i)$

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

En exprimant cette condition en x_1, x_2, \dots, x_{n-1} on obtient un système linéaire en les inconnues $y''_1, y''_2, \dots, y''_{n-1}$ dont la i^e ligne est

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

Exercice. Assembler la matrice A du système obtenu et montrer qu'elle est symétrique, tridiagonale, et à diagonale dominante. En déduire l'existence et l'unicité de la spline cubique.

On peut enfin prouver que A est bien conditionnée :

Exercice. Montrer que la plus grande (resp. petite) valeurs propre de A vérifie

$$\lambda_{max} \leq \max_i \{2h_0 + 3h_1, 3(h_i + h_{i+1}), 3h_{n-2} + 2h_{n-1}\}$$

$$\lambda_{min} \geq \min_i \{2h_0 + h_1, h_i + h_{i+1}, h_{n-2} + 2h_{n-1}\}$$

On peut donc calculer le conditionnement de A :

$$\kappa(A) = \frac{\lambda_{max}}{\lambda_{min}} \leq C^{te} \frac{\max h_i}{\min h_i}$$

Le calcul numérique de la spline cubique est donc rapide, fiable et précis.

Mentionnons pour conclure que d'autres types de splines existent :

- on peut relaxer la condition de nullité des dérivées secondes aux extrémités, cela est intéressant si la fonction à interpoler présente une grande courbure aux extrémités.
- on peut monter en degré, mais on constate que les splines de degré supérieur oscillent davantage. On préfère utiliser les splines cubiques pour l'interpolation.

Chapitre 2

Quadrature.

2.1 Méthodes élémentaires

2.1.1 Principe d'intégration numérique : poids et noeuds.

Soit f une fonction continue sur un intervalle $[a, b]$. Le but de ce chapitre est de donner des méthodes de calcul approché de

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

Bien sûr, lorsqu'on connaît une primitive F de f , la valeur exacte est

$$I(f) = F(b) - F(a).$$

Cependant, ce cas est très rare . . .

Remarque. le nombre $I(f)$ s'interprète comme l'aire sous la courbe représentative de f , d'où le nom *quadrature*, de quarrer, en ancien français, calculer une aire. \square

Commençons par rappeler les méthodes élémentaires.

2.1.2 Rectangle, Trapèze, Simpson

Si on interpole $y = f(x)$ par la fonction constante $y = f((a + b)/2)$ i.e. si on calcule l'aire du rectangle de hauteur $f((a + b)/2)$, on obtient l'approximation

$$I(f) \approx (b - a) \cdot f\left(\frac{a + b}{2}\right).$$

C'est la *méthode du point milieu*.

Si on interpole $y = f(x)$ par la fonction affine $y = f(a) + (x - a) \frac{f(b) - f(a)}{b - a}$ prenant les mêmes valeurs aux deux extrémités a et b , i.e. si on calcule l'aire du trapèze, on obtient l'approximation

$$I(f) \approx (b - a) \cdot \frac{f(a) + f(b)}{2}.$$

C'est la *méthode du trapèze*.

Si on interpole $y = f(x)$ par le polynôme du second degré passant par $(a, f(a))$, $(\frac{a+b}{2}, f(\frac{a+b}{2}))$, $(b, f(b))$ i.e. si on calcule l'aire sous la parabole, on obtient l'approximation

$$I(f) \approx (b-a) \cdot \left\{ \frac{1}{6}f(a) + \frac{4}{6}f\left(\frac{a+b}{2}\right) + \frac{1}{6}f(b) \right\}.$$

C'est la *méthode de Simpson*.

Si on interpole $y = f(x)$ par le polynôme du troisième degré passant par $(x_j, f(x_j))$ où $x_j = a + j \cdot \frac{b-a}{3}$, $j = 0 \dots 3$ on obtient l'approximation

$$I(f) \approx (b-a) \cdot \left\{ \frac{1}{8}f(a) + \frac{3}{8}f\left(\frac{2a+b}{3}\right) + \frac{3}{8}f\left(\frac{a+2b}{3}\right) + \frac{1}{8}f(b) \right\}.$$

C'est la *Règle des 3/8 de Newton*.

Donnons pour terminer ce premier panorama une formule plus sophistiquée. Si on interpole $y = f(x)$ par le polynôme du quatrième degré passant par $(x_j, f(x_j))$ où $x_j = a + j \cdot \frac{b-a}{4}$, $j = 0 \dots 4$ on obtient l'approximation

$$I(f) \approx (b-a) \cdot \left\{ \frac{7}{90}f(a) + \frac{32}{90}f\left(\frac{3a+b}{4}\right) + \frac{12}{90}f\left(\frac{a+b}{2}\right) + \frac{32}{90}f\left(\frac{a+3b}{4}\right) + \frac{7}{90}f(b) \right\}.$$

C'est la *méthode de Villarceau*.

En conclusion, pour donner une valeur approchée de $I(f)$, on utilise donc un tableau de valeurs de f , i.e. la donnée de $f(x_j)$ en certains points x_j appelés *noeuds* et on calcule ensuite une moyenne pondérée de ces valeurs, le coefficient affecté à chaque valeur étant évidemment appelé *poids*.

$$Q(f) = (b-a) \cdot \sum_j \omega_j f(x_j)$$

Faisons maintenant quelques remarques importantes.

Remarque. La quantité

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

correspond à la moyenne¹ de f sur (a, b) . Or toutes les formules précédentes sont du type

$$\frac{\int_a^b f(x) dx}{b-a} \approx \sum_j \omega_j f(x_j)$$

avec les poids ω_j positifs et tels que $\sum_j \omega_j = 1$. Ainsi on a remplacé la moyenne *continue* $\frac{\int_a^b f(x) dx}{b-a}$ par une moyenne *discrète* $\sum_j \omega_j f(x_j)$. \square

Remarque. L'expression $f \mapsto Q(f) = (b-a) \cdot \sum_j \omega_j f(x_j)$ est une forme linéaire, comme l'était $f \mapsto I(f) = \int_a^b f(x) dx$. \square

Remarque. L'expression $f \mapsto I(f) = \int_a^b f(x) dx$ étant positive il est capital que $f \mapsto Q(f) = (b-a) \cdot \sum_j \omega_j f(x_j)$ le soit également.² Cela impose que les

¹ou aussi à l'espérance de $f(X)$ lorsque X suit la loi uniforme sur (a, b)

²sans quoi on pourrait obtenir une intégrale approchée négative pour certaines fonctions positives!

ω_j soient *positifs*. Nous verrons que l'interpolation par des polynômes de degré supérieur ou égal à huit conduit à certains poids négatifs, ce qui fait que les formules de quadrature correspondantes sont dangereuses et inutilisées. \square

Remarque. Les formules précédentes possèdent une symétrie intéressante : elles sont invariantes par le changement $x \leftarrow a + b - x$, i.e. la symétrie par rapport au point milieu du segment, qui échange a en b . les poids ω_j respectent cette symétrie. Ainsi $\check{f} : x \mapsto f(a + b - x)$ et $f : x \mapsto f(x)$ donnent la même intégrale approchée $Q(f) = Q(\check{f})$. Evidemment on a également $I(f) = I(\check{f})$ (calcul par changement de variable.) \square

On voit ainsi que $Q(f)$ conserve les principales propriétés de $I(f)$. C'est une qualité extrêmement agréable de la quadrature numérique.

Remarque. Une propriété de l'intégrale sera cependant perdue irrémédiablement : la stricte positivité. Prenons par exemple la fonction $f : x \mapsto \prod_j (x - x_j)^2$. On a évidemment $I(f) > 0$ alors que $Q(f) = 0$ \square

2.1.3 Ordre et Formules d'erreur

Proposition 2 *La méthode du point milieu est exacte pour f polynôme de degré inférieur ou égal à un. La méthode du trapèze est exacte pour f polynôme de degré inférieur ou égal à un. La méthode de Simpson est exacte pour f polynôme de degré inférieur ou égal à trois. La méthode de Newton est exacte pour f polynôme de degré inférieur ou égal à trois. La méthode de Villarceau est exacte pour f polynôme de degré inférieur ou égal à cinq.*

Preuve. la preuve repose sur la linéarité. Pour alléger les calculs on se ramène aussi à l'intervalle $[-1, 1]$ par changement de variable $t \mapsto x = \frac{a+b}{2} + t\frac{(b-a)}{2}$ qui applique $[-1, 1]$ sur $[a, b]$. L'intégrale se transforme par changement de variable :

$$\int_a^b f(x) dx = \frac{(b-a)}{2} \int_{-1}^1 \varphi(t) dt$$

où

$$\varphi(t) := f(x) = f\left(\frac{a+b}{2} + t\frac{(b-a)}{2}\right)$$

Il suffit alors de prouver que

$$\int_{-1}^1 \varphi(t) dt = 2 \cdot \sum_j \omega_j \varphi(x_j)$$

pour $\varphi(t) = 1, t, t^2, \dots$. Pour la constante $\varphi(t) = 1$, l'égalité est équivalente à la somme des poids valant 1.

$$\sum_j \omega_j = 1.$$

Les autres calculs sont facilités par l'imparité. Pour la méthode du point milieu, par imparité $I(f) = Q(f) = 0$ pour $\varphi(t) = t$. Pour Simpson, par imparité $I(f) = Q(f) = 0$ pour $\varphi(t) = t^3$. Pour Villarceau, lorsque $\varphi(t) = t^5$ on trouve $I(f) = Q(f) = 0$, ainsi il suffit de vérifier l'exactitude pour t^2, t^4 . On voit aussi

que l'imparité permet de gagner un degré, on a intérêt à utiliser des formules avec un nombre *impair* de noeuds. ■

Remarque. On voit aussi que l'imparité permet de gagner un degré, on a intérêt à utiliser des formules avec un nombre *impair* de noeuds. □

Cela motive la définition suivante.

Définition 1 *On dit que le degré de précision d'une quadrature est m si elle est exacte pour tous les polynômes de degré inférieur ou égal à m , mais pas pour tous les polynômes de degré $m + 1$*

Exercice. Formule de Newton-Cotes. Soit $a \leq x_0 < \dots < x_n \leq b$. Soit $(L_j)_j$ les polynômes de Lagrange (voir chapitre précédent) correspondant à cette subdivision. Montrer que $Q_n(f) = (b-a) \sum \omega_j f(x_j)$ où les poids sont donnés par $\omega_j = \frac{\int_a^b L_j(x) dx}{b-a}$ est l'unique règle de quadrature exacte pour les polynômes de degrés inférieur ou égal à n .

Propriétés 3 *On a les formules d'erreur suivantes :*

$$\int_a^b f(x) dx = (b-a) \cdot f\left(\frac{a+b}{2}\right) + \frac{(b-a)^3}{24} f''(\xi), \quad (2.1)$$

$$\int_a^b f(x) dx = (b-a) \cdot \frac{f(a) + f(b)}{2} - \frac{(b-a)^3}{12} f''(\xi), \quad (2.2)$$

$$\int_a^b f(x) dx = (b-a) \cdot \left\{ \frac{1}{6} f(a) + \frac{4}{6} f\left(\frac{a+b}{2}\right) + \frac{1}{6} f(b) \right\} - \frac{(b-a)^5}{2880} f^{(4)}(\xi), \quad (2.3)$$

$$\int_a^b f(x) dx = (b-a) \cdot \left\{ \frac{1}{8} f(a) + \frac{3}{8} f\left(\frac{2a+b}{3}\right) + \frac{3}{8} f\left(\frac{a+2b}{3}\right) + \frac{1}{8} f(b) \right\} - \frac{(b-a)^5}{6480} f^{(4)}(\xi), \quad (2.4)$$

$$\int_a^b f(x) dx = (b-a) \cdot \left\{ \frac{7}{90} f(a) + \frac{32}{90} f\left(\frac{3a+b}{4}\right) + \frac{12}{90} f\left(\frac{a+b}{2}\right) + \frac{32}{90} f\left(\frac{a+3b}{4}\right) + \frac{7}{90} f(b) \right\} - \frac{(b-a)^7}{1935360} f^{(6)}(\xi), \quad (2.5)$$

pour f de classe \mathcal{C}^2 (resp. $\mathcal{C}^4, \mathcal{C}^6$) sur $[a, b]$ et $\xi \in [a, b]$ (évidemment dépendant de la formule utilisée).

Preuve. Les preuves s'appuient par exemple sur la formule de Taylor avec reste intégral (qui n'est rien d'autre que la formule d'intégration par partie appliquée en rafale) et le lemme de la moyenne que nous rappelons.

Lemma 1 *Soit $g(x)$ une fonction mesurable positive sur $[a, b]$. Soit f continue sur $[a, b]$. Alors il existe $\xi \in [a, b]$ tel que*

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

Démontrons par exemple la formule des trapèzes 2.2. Par une intégration par parties,

$$\int_a^b f(x) dx = - \int_a^b (x-c)f'(x) dx + [(x-c)f(x)]_a^b.$$

Pour raison de symétrie, il est judicieux de choisir $c := \frac{a+b}{2}$

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

En intégrant à nouveau par parties :

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

L'erreur $I(f) - Q(f)$ est donc exactement

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

Le lemme de la moyenne s'applique car $g(x) := \frac{(x-a)(b-x)}{2} \geq 0$. On obtient alors

$$I(f) - Q(f) = -f''(\xi) \int_a^b g(x) dx$$

Or $g(x)$ est un polynôme de degré 2, on peut utiliser Simpson qui est exacte pour calculer $\int_a^b g(x) dx = (b-a)\{1/6 \times 0 + 4/6 \times (b-a)^2/8 + 1/6 \times 0\} = \frac{(b-a)^3}{12}$. Les autres formules peuvent se démontrer de manière analogue. ■

Remarque. On peut aussi utiliser le théorème des accroissements finis généralisés 2 qui donne la différence entre le polynôme d'interpolation et la fonction. Cependant, il faut prendre garde aux changements de signe de la fonction $(x-x_0)(x-x_1)\dots(x-x_n)$ aux points x_j . □

On voit dans ce résultat que la précision des formules de quadrature croît avec le nombre de noeuds, et qu'on a intérêt à prendre des nombres de noeuds impairs. Cependant, les termes d'erreurs font intervenir les dérivées d'ordre élevé de la fonction f à intégrer.

- Si la fonction est peu régulière, on ne doit utiliser que les formules les plus simples.
- Si la fonction oscille beaucoup, il y a un risque que les dérivées d'ordre élevé soient très grandes, penser par exemple à $\sin(nx)$. Là aussi il faut éviter d'utiliser les formules d'ordre élevé.

Les formules d'ordre élevé ne présentent un intérêt que si l'intégrand f est très régulier. De plus, les problèmes liés à l'interpolation d'ordre élevé en des points équidistants constatés au chapitre précédent (oscillation de grande amplitude aux extrémités) se traduisent ici par l'apparition de poids négatifs, à partir de $n = 8$. On n'utilise donc ces méthodes que jusqu'au degré 7.

Dans tous les cas, il faut faire attention à la largeur de l'intervalle $(b-a)$ qui doit être toujours inférieure à 1, à cause du terme $(b-a)^n$, c'est pour cette raison qu'on préfère utiliser les méthodes composées qui font l'objet de la section suivante.

2.2 Méthodes à pas multiples

2.2.1 Principe des méthodes composées.

On subdivise l'intervalle $[a, b]$ en n morceaux : $a = x_0 < x_1 < \dots < x_n = b$. Lorsque le pas de la subdivision est constant³ : $x_j = a + j \cdot (b-a)/n$ $j = 0, \dots, n$. La relation de Chasles donne :

$$\int_a^b f(x)dx = \sum_j \int_{x_j}^{x_{j+1}} f(x)dx.$$

Puis on applique une quadrature sur chaque intervalle de la subdivision. Pour la méthode des trapèzes, on obtient ainsi :

$$\int_a^b f(x)dx = \sum_j \left\{ \frac{(b-a)}{n} \cdot \frac{f(x_j) + f(x_{j+1})}{2} - \frac{(b-a)^3}{12n^3} f''(\xi_j) \right\}$$

En remarquant que, hormis les extrémités a et b , chaque point x_j apparaît dans deux termes

$$\int_a^b f(x)dx = \frac{(b-a)}{n} \left\{ \frac{f(a)}{2} + \sum_{a < x_j < b} f(x_j) + \frac{f(b)}{2} \right\} - \frac{(b-a)^3}{12n^2} \left\{ \frac{\sum_j f''(\xi_j)}{n} \right\}$$

Le terme $\frac{\sum_j f''(\xi_j)}{n}$ peut s'exprimer comme une moyenne de valeurs de f'' . Si f est supposée de classe \mathcal{C}^2 , on peut utiliser une forme discrète du lemme de la moyenne, très facile à prouver $\frac{\sum_j f''(\xi_j)}{n} = f''(\xi)$ avec $\xi \in (a, b)$. On obtient ainsi :

$$\int_a^b f(x)dx = \frac{(b-a)}{n} \left\{ \frac{f(a)}{2} + \sum_{a < x_j < b} f(x_j) + \frac{f(b)}{2} \right\} - \frac{(b-a)^3}{12n^2} f''(\xi).$$

Notons

$$T_n(f) = \frac{(b-a)}{n} \left\{ \frac{f(a)}{2} + \sum_{a < x_j < b} f(x_j) + \frac{f(b)}{2} \right\}$$

$$T_n(f) = (b-a) \left\{ \frac{f(a)}{2n} + \sum_{a < x_j < b} \frac{f(x_j)}{n} + \frac{f(b)}{2n} \right\}$$

On remarque que $T_n(f)$ est (encore) une moyenne des valeurs de f aux points x_j , où les extrémités a et b sont affectées d'un poids deux fois moindre que les points intérieurs. Finalement

$$\int_a^b f(x)dx = T_n(f) - \frac{(b-a)^3}{12n^2} f''(\xi).$$

³On peut aussi prendre un pas adapté à la fonction, plus fin là où elle oscille beaucoup et plus grossier là où elle varie peu. Ce sont les méthodes adaptatives dont l'étude dépasse la portée de ce modeste fascicule.

Plus simplement on retiendra

$$\int_a^b f(x)dx = T_n(f) + O\left(\frac{1}{n^2}\right)$$

Le même principe, appliqué à la méthode des rectangles donne :

$$R_n(f) = \frac{b-a}{n} \left\{ \sum_j f(x_{j+1/2}) \right\}$$

Les points $x_{j+1/2} = \frac{x_j + x_{j+1}}{2}$ correspondent aux milieux des segments (x_j, x_{j+1}) . Remarquer que $R_n(f)$ est la moyenne arithmétique des $f(x_{j+1/2})$. On a également l'approximation du même ordre :

$$\int_a^b f(x)dx = R_n(f) + O\left(\frac{1}{n^2}\right)$$

Pour la méthode de Simpson :

$$S_n(f) = \frac{(b-a)}{n} \left\{ \sum_j \frac{1}{6}f(x_j) + \frac{4}{6}f(x_{j+1/2}) + \frac{1}{6}f(x_{j+1}) \right\}$$

En regroupant les poids par noeud de chaque type :

$$S_n(f) = (b-a) \left\{ \frac{1}{6n}f(a) + \sum_{0 < j < n} \frac{2}{6n}f(x_j) + \sum_j \frac{4}{6n}f(x_{j+1/2}) + \frac{1}{6n}f(b) \right\}$$

Il s'agit évidemment toujours d'une formule de moyenne car la somme des coefficients est : $1/6n + 2(n-1)/6n + 4n/6n + 1/6n = 1$. En utilisant à nouveau le lemme de la moyenne, on prouve la formule d'erreur :

$$\int_a^b f(x)dx = S_n(f) - \frac{(b-a)^5}{2880 n^4} f^{(4)}(\xi).$$

Plus simplement on retiendra

$$\int_a^b f(x)dx = S_n(f) + O\left(\frac{1}{n^4}\right).$$

2.2.2 Accélération de la convergence. Extrapolation de Romberg-Richardson.

Le principe de l'accélération de la convergence est simple et peut s'appliquer à toute suite dont la vitesse de convergence est contrôlée explicitement. Voyons cela sur un exemple, celui du calcul approché d'intégrales par la méthodes des trapèzes. Notons

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

On peut démontrer (formule d'Euler-Maclaurin voir TP 4) que

$$T_n = I + \frac{\alpha}{n^2} + O\left(\frac{1}{n^4}\right)$$

Si on double le nombre de noeuds la précision est alors

$$T_{2n} = I + \frac{\alpha}{4n^2} + O\left(\frac{1}{n^4}\right)$$

En combinant judicieusement ces deux approximations, on peut en construire une *beaucoup plus précise* :

$$T'_{2n} := \frac{4T_{2n} - T_n}{3} = I + O\left(\frac{1}{n^4}\right)$$

Supposons (c'est encore une conséquence d'Euler-Maclaurin) que

$$T'_{2n} = I + \frac{\beta}{n^4} + O\left(\frac{1}{n^6}\right)$$

On réitère alors le procédé :

$$T'_{4n} = I + \frac{\beta}{16n^4} + O\left(\frac{1}{n^6}\right)$$

$$T''_{4n} := \frac{16T'_{4n} - T'_{2n}}{15} = I + O\left(\frac{1}{n^6}\right)$$

etc... En pratique on dispose les calculs ainsi

	T_n		
	T_{2n}	T'_{2n}	
	T_{4n}	T'_{4n}	T''_{4n}
précision	$O\left(\frac{1}{n^2}\right)$	$O\left(\frac{1}{n^4}\right)$	$O\left(\frac{1}{n^6}\right)$

Evidemment on peut continuer, mais au delà d'un certain nombre d'accélération, les erreurs d'arrondis gâtent l'amélioration escomptée.

Remarque. Le procédé de base consiste à *extrapoler* à la limite. En effet T'_{2n} est un barycentre de T_n et T_{2n} avec des coefficients 4 et -1 . Cela revient à placer sur une droite T'_{2n} à l'extérieur du segment $(T_n T_{2n})$ dans une proportion bien choisie. Faire un dessin. □

Remarque. on voit aisément que

$$T_{2n} = \frac{T_n + R_n}{2}$$

ainsi on fait la moyenne arithmétique des trapèzes et des rectangles. De plus

$$T'_{2n} = \frac{(b-a)}{n} \left\{ \sum_j \frac{1}{6} f(x_j) + \frac{4}{6} f(x_{j+1/2}) + \frac{1}{6} f(x_{j+1}) \right\} = S_n$$

qui redonne Simpson.⁴ □

⁴comparer T''_{4n} et Villarceau.

2.2.3 Méthode de Montecarlo

Une méthode particulièrement simple est la méthode de Montecarlo. On tire au hasard les noeuds suivant une loi uniforme sur $[a, b]$, en utilisant par exemple un générateur de suite aléatoire (**rand** en scilab). On approche l'intégrale par la moyenne arithmétique des valeurs au noeuds

$$I(f) \approx (b - a) \cdot \left\{ \frac{\sum_{1 \leq j \leq n} f(\xi_j)}{n} \right\}$$

Evidemment le résultat est un variable aléatoire qui dépend de l'échantillon des noeuds tirés. On montre en cours de probabilités que

$$(b - a) \left\{ \frac{\sum_{1 \leq j \leq n} f(\xi_j)}{n} \right\}$$

est un estimateur sans biais de $I(f)$ et on peut donner un intervalle de confiance de largeur $O(1/\sqrt{n})$, grace au théorème central limite. Certes la convergence est bien plus lente que les méthodes déterministes, mais on ne fait aucune hypothèse de régularité sur f . D'autre part la vitesse en $O(1/\sqrt{n})$ est conservée lorsque on calcule des intégrales multiples par cette méthode, alors que la convergence des méthodes déterministes se dégrade lorsque la dimension dépasse 4.

2.3 Méthode de Gauss et polynômes orthogonaux.

2.3.1 Introduction.

Nous avons construit des formules du type de sorte qu'elles soient exactes pour des polynômes de degré le plus élevé possible. Jusqu'à présent, on a placé les noeuds de façon équidistante. On choisissait ensuite les poids de façon à avoir la précision maximale. Pour une formule à n points, on avait un degré de précision n (cas n impair) ou $n - 1$ (cas n pair).

Remarque. La précision maximum théorique est $2n - 1$. En effet, soit le polynôme $P(x) := \prod_j (x - x_j)^2$ où les noeuds sont les x_j . Le degré de P est $2n$. Le polynôme P est positif non identiquement nul donc on a évidemment $\int_a^b f(x) dx > 0$ alors que $Q(f) = 0$. \square

La précision maximale théorique est donc $(2n - 1)$. Les méthodes que nous avons construites jusqu'à présent sont de précision maximum n . Nous allons maintenant construire effectivement une méthode de précision $2n - 1$ en *optimisant le placement des noeuds*.

2.3.2 Méthode de Gauss.

Théorème 3 *Il existe une et une seule formule de quadrature*

$$Q(f) := (b - a) \left(\sum_k \omega_k f(x_k) \right)$$

exacte pour les polynômes de degré $\leq 2n - 1$. Sur l'intervalle $[-1, 1]$ elle s'exprime ainsi :

$$\int_{-1}^1 f(x) dx \approx \sum_k w_k f(x_k)$$

où les x_k sont les racines du n^e polynôme de Legendre

$$P_n(x) := \frac{1}{2^n n!} \frac{d^n}{dx^n} \{(x^2 - 1)^n\}$$

et

$$w_k = 2\omega_k = \int_{-1}^1 \prod_{j \neq k} \left(\frac{x - x_j}{x_k - x_j} \right)^2 dx.$$

Remarque. La formule sur $[a, b]$ se déduit de celle sur $[-1, 1]$ par changement de variable affine. Voir proposition 2. En particulier, comme la longueur de l'intervalle $[-1, 1]$ est 2, on a $\sum_k w_k = \sum_k 2\omega_k = 2$. \square

Avant de démontrer le théorème, donnons quelques cas particuliers.

$$P_0 = 1, P_1(x) = x, P_2(x) = \frac{1}{2}(3x^2 - 1), P_3(x) = \frac{1}{2}(5x^3 - 3x), P_4(x) = \frac{1}{8}(35x^4 - 30x^2 + 3)$$

Cela donne pour $n = 2$, $x_k = \pm \frac{1}{\sqrt{3}}, \quad w_k = 1$

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

Pour $n = 3$

$$\int_{-1}^1 f(x) dx \approx \frac{5}{9} f\left(-\sqrt{\frac{3}{5}}\right) + \frac{8}{9} f(0) + \frac{5}{9} f\left(\sqrt{\frac{3}{5}}\right)$$

Testons la précision de ces méthodes sur une exemple simple. Prenons

$$f(x) = \frac{1}{1 + x^2}$$

,

$$\int_{-1}^1 f(x) dx = \frac{\pi}{2} \approx 1.57$$

Simpson donne

$$\int_{-1}^1 f(x) dx \approx 2 \left\{ \frac{1}{6} \cdot \frac{1}{2} + \frac{4}{6} \cdot 1 + \frac{1}{6} \cdot \frac{1}{2} \right\} = \frac{5}{3} = 1.66$$

Gauss (2 points) donne

$$\int_{-1}^1 f(x) dx \approx \frac{1}{1 + \frac{1}{3}} + \frac{1}{1 + \frac{1}{3}} = \frac{3}{2} = 1.5$$

Gauss (3 points) donne

$$\int_{-1}^1 f(x) dx \approx \frac{5}{9} \frac{1}{1 + \frac{3}{5}} + \frac{8}{9} + \frac{5}{9} \frac{1}{1 + \frac{3}{5}} = \frac{19}{12} = 1.56$$

Sur cet exemple, avec un nombre de point réduit la méthode de Gauss à 3 points est précise à 10^{-2} près alors que Simpson est précise à 10^{-1} près. Pour $n \geq 5$ les

noeuds x_k ne s'expriment pas à l'aide de radicaux, ils sont cependant tabulés et disponibles dans les logiciels évolués.

Preuve. Existence.

Soit

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

P_n est un polynôme de degré n exactement. En effet, il est obtenu en dérivant n fois le polynôme $(x^2 - 1)^n$ qui est de degré $2n$. Montrons maintenant que $P_n(x)$ a exactement n racines simples sur $] -1, 1[$. Nous allons appliquer le théorème de Rolle en rafale. Le polynôme $P(x) = (x^2 - 1)^n$ prend la même valeur en $x = -1$ et $x = +1$, donc il existe $c \in] -1, 1[$ tel que $P' = \frac{d}{dx}(x^2 - 1)^n$ s'annule en c . Comme -1 et 1 sont racines de multiplicité n de P , -1 et 1 sont encore racine de P' mais de multiplicité $n - 1$. Le polynôme P' prend donc la même valeur en $-1, c, 1$. Par le théorème de Rolle, P'' s'annule donc entre -1 et c et entre c et 1 , cela fait donc deux racines en dehors des extrémités. Comme -1 et 1 sont encore racine de P'' de multiplicité $n - 2$, on peut réitérer : à l'étape n

$$\frac{d^n}{dx^n} \{(x^2 - 1)^n$$

s'annule donc n fois sur $] -1, 1[$. Comme $P_n(x)$ est de degré n cela fait donc exactement n racines simples (de multiplicité un) sur $] -1, 1[$.

Montrons maintenant que la famille P_n est une famille orthogonale de $L^2(]-1, 1[)$ muni du produit scalaire

$$\langle f, g \rangle := \int_{-1}^1 f(x)g(x)dx.$$

Il s'agit de prouver que $\langle P_n, P_m \rangle = 0$ si $n \neq m$. Cela se montre très simplement par récurrence sur $m - n$ en intégrant par parties (exercice).

Soit maintenant un polynôme $P(x)$ quelconque de degré $\leq 2n - 1$. Effectuons la division euclidienne de P par le n^e polynôme de Legendre P_n .

$$P = P_n \cdot Q + R. \quad (2.6)$$

Comme $\overline{\deg R} < \overline{\deg P_n} = n$, le degré de Q est $\leq n - 1$. La famille P_0, P_1, \dots, P_{n-1} est étagée, donc elle est libre et c'est une base de l'espace des polynômes de degré $\leq n - 1$. Or P_n est orthogonal à P_0, P_1, \dots, P_{n-1} , donc par linéarité $P_n \perp \text{vect}\langle P_0, P_1, \dots, P_{n-1} \rangle = \mathbb{R}_n[X]$ et $P_n \perp Q$:

$$\int_{-1}^1 P_n(x)Q(x) dx = 0.$$

Avec Eq. (2.6) il vient

$$\int_{-1}^1 P(x) dx = \int_{-1}^1 R(x) dx.$$

D'autre part la formule de quadrature donne :

$$\sum_k w_k P(x_k) = \sum_k w_k (P_n(x_k)Q(x_k) + R(x_k)) = \sum_k w_k R(x_k)$$

car les x_k sont justement les racines de P_n .

Il reste à montrer que

$$\int_{-1}^1 R(x) dx = \sum_k w_k R(x_k)$$

pour tout polynôme R de degré $\leq n-1$. Ceci ne pose aucune difficulté, il suffit de bien choisir les poids w_k .

Lemma 2 Soit $a \leq x_1 < x_2 < \dots < x_n \leq a$. Il existe un unique w_1, w_2, \dots, w_n tel que

$$\int_{-1}^1 f(x) dx = \sum_k w_k f(x_k)$$

pour tout polynôme R de degré $\leq n-1$.

Preuve. Il suffit de prendre $w_k = \int_a^b L_k(x) dx$ où L_k est le k^e polynôme de Lagrange. ■

Avec ce choix de w_k , la formule de quadrature est exacte jusqu'au degré $2n-1$ puisque

$$\int_{-1}^1 P(x) dx = \int_{-1}^1 R(x) dx = \sum_k w_k R(x_k) = \sum_k w_k P(x_k)$$

Le problème dans le lemme c'est que rien ne garantit la positivité des poids : $w_k > 0$.⁵ Mais ici, un miracle se produit. Comme la formule de quadrature est exacte pour P de degré $\leq 2n-1$, elle l'est pour $P := L_k^2$ qui est de degré $2n-2$. Rappelons que

$$L_j(X) = \frac{\prod_{i \neq j} (X - x_i)}{\prod_{i \neq j} (x_j - x_i)}.$$

Calculons

$$\int_{-1}^1 L_k(x)^2 dx = \sum_j w_j L_k^2(x_j) = w_k$$

car dans la somme un seul terme est non nul. On obtient finalement

$$w_k = \int_{-1}^1 L_k(x)^2 dx > 0$$

Unicité.

Soit une autre formule de quadrature à n noeuds définie par les noeuds x'_k et les poids w'_k , $k = 1, \dots, n$. Les poids sont évidemment supposés non nuls, sinon le noeud x'_k ne compte pas dans la quadrature

$$Q'(f) := \sum_k w'_k f(x'_k).$$

Soit L'_k le polynôme de Lagrange construit sur les noeuds x'_j . Il est de degré $n-1$ donc $P_n \perp L'_k$:

$$\int_{-1}^1 P_n(x) L'_k(x) dx = 0$$

⁵En fait, pour $n \geq 8$ avec des noeuds x_k équidistants, certains poids peuvent être négatifs.

Or en utilisant la formule de quadrature qui est exacte car $\deg P_n \cdot L'_k \leq 2n - 1$, on obtient

$$\int_{-1}^1 P_n(x)L_k(x) dx = \sum_j w'_j P_n(x'_j)L'_k(x_j) = w'_k P_n(x'k).$$

D'où l'on tire que

$$P_n(x'k) = 0$$

donc que les noeuds $x'k$ sont les racines de P_n donc, à permutation éventuelle près, ils sont égaux aux x_k . Les deux quadratures ont donc les mêmes noeuds, elles ont donc aussi les mêmes poids correspondants (intégrer par exemple les polynômes de Lagrange avec les deux formules). ■

Chapitre 3

Méthode des moindres carrés.

3.1 Présentation du problème. Régression linéaire.

3.1.1 Droite des moindres carrés.

Soient n observations t_i, y_i . On cherche une droite $y = at + b$ passant « au mieux » par ces points au sens suivant :

$$y_i = at_i + b + \epsilon_i$$

avec $\sum \epsilon_i^2$ minimum.

Exemple. On mesure la position d'une fusée en route pour Mars. y_i position à l'instant t_i . a représente la vitesse de l'engin, b sa position initiale, ϵ_i l'erreur de mesure.

Exemple. On mesure le poids d'un échantillon de personnes en fonction de leur taille. Dans ce cas l'expérience montre que la courbe ressemble à une parabole $y = at^2$. En passant au logarithme, on se ramène au cas précédent :

$$\ln y_i = \ln a + 2 \ln t_i + \epsilon_i$$

L'indice de masse corporelle est le rapport $poids/taille^2$. Pour plus d'exemples et des simulations interactives voir le site de demo de Mathematica.

<http://demonstrations.wolfram.com/LinearAndQuadraticCurveFittingPractice/>
Plus généralement, on veut approcher/approximer/ajuster/fitter n observations t_i, y_i par un polynôme de degré fixé $m < n - 1$

$$p(t) = a_0 + a_1t + \dots + a_mt^m$$

de sorte que l'erreur quadratique

$$\sum_i (p(t_i) - y_i)^2$$

soit minimum.

Remarque. Lorsque $m \geq n - 1$ et que les points t_i sont distincts, il suffit de prendre $m = n - 1$ et on se ramène au problème de l'interpolation qui peut être exactement résolu. Le polynôme passe exactement par tous les points t_i, y_i . Voir chapitre interpolation et l'erreur quadratique est nulle donc minimum. \square

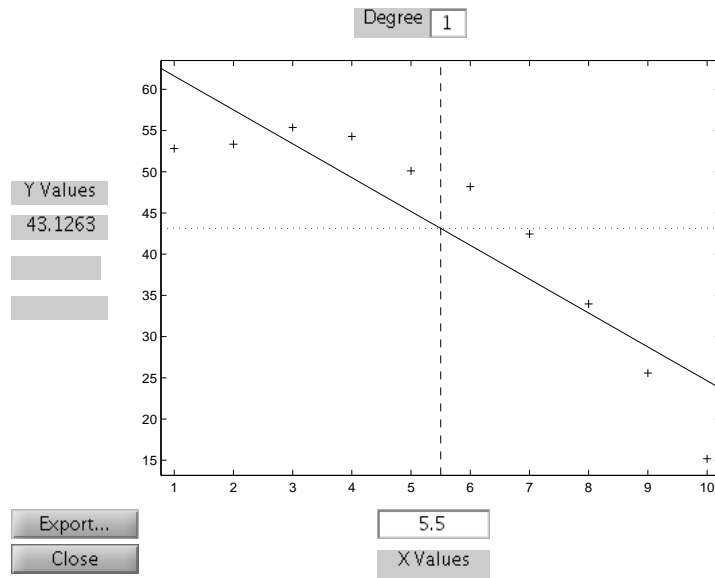


FIG. 3.1 – un exemple de régression linéaire avec Matlab

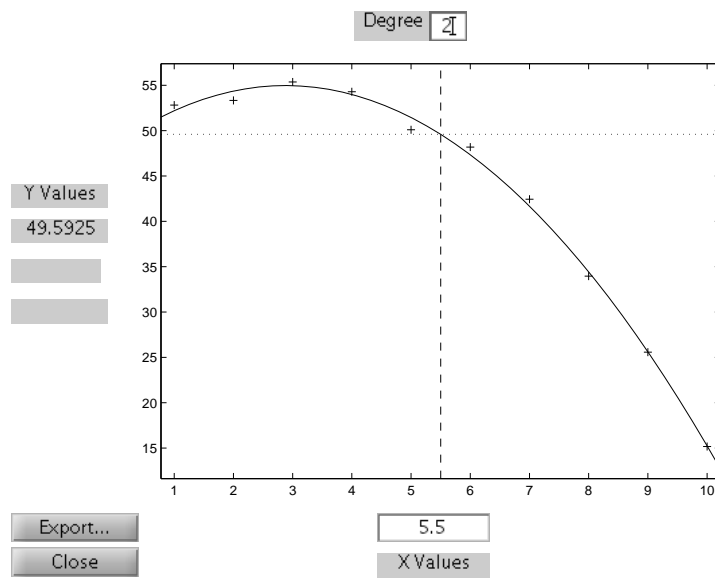


FIG. 3.2 – un exemple de régression quadratique avec Matlab

3.2. INTERPRÉTATION GÉOMÉTRIQUE ET ÉCRITURE MATRICIELLE.31

Lorsque $m < n - 1$, on ne peut pas en général trouver de polynôme p de degré m tel que $p(t_i) = y_i$. Cela se traduit par le système

$$\begin{cases} a_0 + a_1 t_1 + \dots + a_m t_1^m = y_1 \\ a_0 + a_1 t_2 + \dots + a_m t_2^m = y_2 \\ \vdots \\ a_0 + a_1 t_n + \dots + a_m t_n^m = y_n \end{cases}$$

dont les inconnues sont les coefficients du polynôme, i.e. le vecteur $\mathbf{a} = (a_0, a_1, \dots, a_m)^T$ de dimension $m+1$. Ce système est surdéterminé : il y a n équations mais seulement $m+1$ inconnues. En général, il n'a pas de solutions, sauf si le second membre vérifie des conditions de compatibilité particulières.

3.2 Interprétation géométrique et écriture matricielle.

Notons maintenant

$$A = \begin{pmatrix} 1 & t_1 & t_1^2 & \dots & t_1^m \\ 1 & t_2 & t_2^2 & \dots & t_2^m \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & t_n & t_n^2 & \dots & t_n^m \end{pmatrix}$$

la matrice de taille $n \times (m+1)$, appelée matrice de Vandermonde. On constate aisément que pour $\mathbf{a} = (a_0, a_1, \dots, a_m)^T \in \mathbb{R}^{m+1}$ le produit

$$A\mathbf{a} = (P(t_1), P(t_2), \dots, P(t_n))^T \in \mathbb{R}^n$$

Munissons alors \mathbb{R}^n de la norme euclidienne usuelle : pour $\mathbf{z} = (z_1, \dots, z_n)^T \in \mathbb{R}^n$, $\|\mathbf{z}\|^2 = \sum_i z_i^2 = \mathbf{z}^T \mathbf{z}$.

Le problème des moindres carrés se traduit alors :

Etant donné $\mathbf{y} = (y_1, \dots, y_n)^T \in \mathbb{R}^n$ On cherche $\mathbf{a} = (a_0, a_1, \dots, a_m)^T \in \mathbb{R}^{m+1}$ tel que $\|A\mathbf{a} - \mathbf{y}\|^2 = \sum_i (y_i - P(t_i))^2$ soit minimum.

On est ainsi conduit à minimiser la distance

$$\|\mathbf{y} - A\mathbf{a}\|^2 = \min_{\mathbf{c} \in \mathbb{R}^{m+1}} \|\mathbf{y} - A\mathbf{c}\|^2$$

Or lorsque le vecteur \mathbf{c} décrit \mathbb{R}^{m+1} , le vecteur $A\mathbf{c}$ décrit le sous espace vectoriel $\text{Im } A$, image de l'application linéaire de matrice A :

$$\begin{aligned} A : \mathbb{R}^{m+1} &\rightarrow \mathbb{R}^n \\ \mathbf{c} &\mapsto A\mathbf{c} \end{aligned}$$

Il s'agit donc simplement de calculer la distance du vecteur \mathbf{y} au sous espace vectoriel $\text{Im } A$. Or on sait que la distance est réalisée pour la *projection orthogonale* du vecteur $\mathbf{y} = (y_1, \dots, y_n)^T$ sur le sous-espace vectoriel $\text{Im } A$.

Cette **interprétation géométrique** rend le problème très simple à résoudre, car on sait que le sous espace $\text{Im } A$ est *l'espace engendré par les vecteurs colonnes* de A . Nous allons donner la méthode dans le théorème suivant.

Théorème 4 Soit $b \in \mathbb{R}^p$ et soit un système surdéterminé $Ax = b$ avec A matrice $n \times p$, $n > p$. La quantité $\|Ax - b\|^2$ est minimum si et seulement si Ax est la projection orthogonale de b sur l'espace engendré par les colonnes de A , ce qui équivaut encore à

$$A^T Ax = A^T b \quad (3.1)$$

Ce système linéaire s'appelle la forme normale du système $Ax = b$.

Preuve. La norme $\|Ax - b\|^2 = \langle Ax - b, Ax - b \rangle$. Soit $b^* = Ax^*$ la projection orthogonale de b sur l'espace vectoriel $\text{Im } A$.

$$\|Ax - b\|^2 = \langle Ax - b^* + b^* - b, Ax - b^* + b^* - b \rangle$$

développons le carré scalaire :

$$\|Ax - b\|^2 = \|Ax - b^*\|^2 + 2 \langle Ax - b^*, b^* - b \rangle + \|b^* - b\|^2$$

Comme b^* est la projection orthogonale de b sur $\text{Im } A$, on a $(b^* - b) \perp \text{Im } A$ donc

$$\langle Ax - b^*, b^* - b \rangle = \langle Ax - Ax^*, b^* - b \rangle = \langle A(x - x^*), b^* - b \rangle = 0$$

Il reste donc

$$\|Ax - b\|^2 = \|Ax - b^*\|^2 + \|b^* - b\|^2$$

(On reconnaît ici le théorème de Pythagore, faire un dessin.) On en conclut que

$$\|Ax - b\|^2 \geq \|b^* - b\|^2 = \|Ax^* - b\|^2$$

Donc

$$\|Ax^* - b\|^2 = \min_{x \in \mathbb{R}^p} \|b - Ax\|^2$$

Il reste à **caractériser la projection orthogonale** b^* de b sur $\text{Im } A$. Elle est définie par les *deux conditions* :

- (i) $b^* \in \text{Im } A$
- (ii) $(b^* - b) \perp \text{Im } A$

La première condition se traduit par $b^* = Ax^*$ pour un vecteur x^* . Comme $\text{Im } A$ est engendrée par les vecteur colonnes de A , la seconde condition signifie que

$$\begin{aligned} (b^* - b) &\perp A\mathbf{e}_k, \quad k = 1 \dots p \\ \langle (b - Ax^*), A\mathbf{e}_k \rangle &= 0 \quad k = 1 \dots p \end{aligned}$$

En utilisant la transposée de A

$$\langle A^T (b - Ax^*), \mathbf{e}_k \rangle = 0 \quad k = 1 \dots p$$

Donc le vecteur

$$A^T (b - Ax^*) = \mathbf{0}$$

Ce qui donne

$$A^T Ax^* = A^T b$$

■

Il reste à voir si le système sous forme normale (3.1) admet bien une solution unique. C'est l'objet de la proposition suivante.

3.2. INTERPRÉTATION GÉOMÉTRIQUE ET ÉCRITURE MATRICIELLE.33

Proposition 3 Soit A matrice $n \times p$ avec $p \leq n$. La matrice $A^T A$ est inversible si et seulement si A est de plein rang, i.e. $\text{rang}(A) = p$. De plus dans ce cas $A^T A$ est symétrique définie positive.

Preuve. Il est immédiat que $A^T A$ est symétrique :

$$(A^T A)^T = A^T (A^T)^T = A^T A$$

La positivité signifie que

$$\langle A^T A x, x \rangle \geq 0.$$

Or $\langle A^T A x, x \rangle = \langle A x, A x \rangle = \|A x\|^2 \geq 0$. La stricte positivité signifie que

$$\forall x \neq 0, \quad \langle A^T A x, x \rangle > 0$$

Ou encore

$$\langle A^T A x, x \rangle = 0 \Rightarrow x = 0$$

Or

$$\langle A^T A x, x \rangle = \|A x\|^2 = 0 \Rightarrow A x = 0$$

Or $A x = \sum_k x_k A \mathbf{e}_k = \sum_k x_k \mathbf{C}_k$ où $x = (x_1, x_2, \dots, x_p)$ et $\mathbf{C}_k = \mathbf{e}_k$ désigne la k -ème colonne de A . Donc $A x = 0$ pour un $x \neq 0$ si et seulement si les colonnes de A ne sont pas indépendantes. On obtient donc que la matrice $A^T A$ est définie positive ssi les colonnes de A sont indépendantes i.e. $\text{rang}(A) = p$. Enfin la matrice $A^T A$ est **carrée**. Donc elle est inversible ssi son noyau est réduit au vecteur nul.

$$A^T A x = 0 \Rightarrow \langle A^T A x, x \rangle = \|A x\|^2 = 0 \Rightarrow A x = 0$$

et nous venons de voir que $A x = 0$ pour un $x \neq 0$ si et seulement si les colonnes de A ne sont pas indépendantes. ■

Remarque. la condition $\text{rang}(A) = p$ n'est possible que si $p \leq n$, car le rang d'une matrice est à la fois le rang des lignes et le rang des colonnes. □

Nous pouvons maintenant appliquer la théorie au problème de l'approximation polynômiale au sens des moindres carrés. La matrice

$$A = \begin{pmatrix} 1 & t_1 & t_1^2 & \dots & t_1^m \\ 1 & t_2 & t_2^2 & \dots & t_2^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_n & t_n^2 & \dots & t_n^m \end{pmatrix} \quad (3.2)$$

la matrice de taille $n \times (m + 1)$. Le vecteur $b = (y_1, y_2, \dots, y_n)$. On cherche $a = (a_0, a_1, \dots, a_m) \in \mathbb{R}^{m+1}$ tel que $\|y - A a\|^2$ soit minimum. D'après le théorème précédent, a est solution du système

$$A^T A a = A^T y$$

Vérifions que le système a une solution unique

Proposition 4 Soit A la matrice de Vandermonde (3.2). La matrice $A^T A$ est inversible ssi les abscisses t_1, t_2, \dots, t_n sont distincts.

Preuve. Tout d'abord, la condition est nécessaire : il est évident que si les abscisses t_1, t_2, \dots, t_n ne sont pas distinctes, par exemple si $t_1 = t_2$ la matrice A possède au moins deux lignes égales, les lignes 1 et 2.

Réciproquement, montrons que la condition est suffisante. D'après la proposition précédente, il suffit de tester si les colonnes de A sont indépendantes. Soit une relation de dépendance linéaire entre les colonnes : $\sum_k \alpha_k \mathbf{C}_k = 0$ où $\mathbf{C}_k = (t_1^{k-1}, t_2^{k-1}, \dots, t_n^{k-1})^T$. Cette relation s'écrit encore :

$$\sum_k \alpha_k t_i^{k-1} = 0$$

Soit $Q(t) = \alpha_1 + \alpha_2 t + \dots + \alpha_{m+1} t^m$. Ce polynôme de degré au plus m s'annule en t_1, t_2, \dots, t_n . Ces points étant distincts, le polynôme admet donc au moins n racines, or par hypothèse $n > m + 1$ donc Q est le polynôme nul et tous ses coefficients sont nuls : $\alpha_k = 0, \forall k$. Les colonnes de A sont bien indépendantes donc $A^T A$ est inversible. ■

La méthode des moindres carrés est donc bien posée en théorie. Nous allons voir à la section suivante comment résoudre le système (3.1) en pratique. Donnons auparavant un exemple fondamental, la droite de régression.

Exemple. Droite de régression. Soit des mesures y_1, y_2, \dots, y_n correspondant à des points distincts t_1, t_2, \dots, t_n . Montrer que la droite $y = at + b$ qui minimise l'écart quadratique $\sum_i (y_i - (at_i + b))^2$ est donnée par le système linéaire :

$$\begin{pmatrix} n & \sum t_i \\ \sum t_i & \sum t_i^2 \end{pmatrix} \cdot \begin{pmatrix} b \\ a \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum t_i y_i \end{pmatrix}$$

En déduire que la droite des moindres carrés passe par l'isobarycentre du nuage de points :

$$\bar{y} = a\bar{t} + b$$

où \bar{z} désigne $\frac{\sum z_i}{n}$ et que

$$a = \frac{\overline{y \cdot t} - \bar{y}\bar{t}}{\bar{t}^2 - (\bar{t})^2}$$

Indication. Soit

$$A = \begin{pmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_n \end{pmatrix}$$

$$\text{alors } A^T A = \begin{pmatrix} n & \sum t_i \\ \sum t_i & \sum t_i^2 \end{pmatrix}$$

3.2.1 Moindres carrés pondérés.

Supposons que l'on veuille pondérer l'influence des points de mesure (t_i, y_i) , c'est à dire que l'on se donne une suite de poids strictement positifs w_i et que l'on cherche à minimiser l'erreur quadratique *pondérée* :

$$\sum_{1 \leq i \leq n} w_i (p(t_i) - y_i)^2$$

par un polynôme de degré fixé $m < n - 1$

$$p(t) = a_0 + a_1 t + \dots + a_m t^m.$$

Tout ce qui précède peut se généraliser aisément en munissant \mathbb{R}^n du produit scalaire *pondéré* :

$$\langle x, y \rangle_w = \sum_i w_i x_i y_i$$

associé à la norme

$$\|z\|_w = \sum_i w_i z_i^2.$$

La forme normale du système des moindres carrés s'écrit alors

$$A^T W A a = A^T W y$$

avec

$$W = \begin{pmatrix} w_1 & 0 & \dots & \dots & 0 \\ 0 & w_2 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & \dots & 0 & w_n \end{pmatrix}$$

3.3 Algorithme de résolution numérique.

A priori, la matrice du système (3.1) $A^T A$ étant symétrique définie positive, on peut appliquer la méthode de Cholesky pour le résoudre. Il se trouve malheureusement que la matrice $A^T A$ peut être très mal conditionnée. En effet, il est très facile de voir que $\text{cond}(A^T A) = \text{cond}(A)^2$. Dans ce cas, les méthodes usuelles de résolution de système sont très instables numériquement (voir cours d'algèbre linéaire 3, il est préférable de les éviter. Heureusement, il est possible d'appliquer une méthode plus géométrique, la méthode *QR*. Nous ne détaillerons pas ici cette méthode, vue en algèbre linéaire 3, basée sur le procédé d'orthonormalisation de Schmidt. Enonçons simplement la propriété

Proposition 5 *Toute matrice A de taille $n \times p$ se décompose en un produit $A = QR$, où Q est une matrice orthogonale $n \times n$ $Q^T Q = I_n$ et R une matrice triangulaire supérieure de taille $n \times p$.*

¹ A l'aide d'une telle décomposition, la matrice

$$A^T A = (QR)^T QR = R^T Q^T QR = R^T R$$

Mais

$$R^T R = R_1^T R_1$$

où $R = \begin{pmatrix} R_1 \\ 0 \end{pmatrix}$, avec R_1 matrice triangulaire supérieure $p \times p$ à coefficients diagonaux *strictement* positifs. On obtient directement la factorisation de Cholesky $A^T A = R_1^T R_1$. Le système (3.1) s'écrit simplement

$$R_1^T R_1 x = A^T b$$

et se résout par deux systèmes triangulaires : $R_1^T y = A^T b$ puis $R_1 x = y$.

¹pour des exemples et demo voir [4]

Bibliographie

- [1] Gilbert Strang, Introduction to applied mathematics, Wellesley-Cambridge press, 1986.
- [2] H.R. Schwarz, Numerical Analysis, A comprehensive introduction, Wiley, 1989.
- [3] Cleve Moler, Numerical computing with Matlab,
<http://www.mathworks.com/moler/>
- [4] Le Mathematica computational knowledge engine
<http://www.wolframalpha.com/>