Approximation

Alain Yves LE ROUX, Université Bordeaux 1

1 Approximation d'une fonction

On se donne une fonction f définie sur un intervalle réel I et à valeurs réeles. Cette fonction est caractérisée par une propriété particulière : soit une équation, soit un jeu de données, etc.. mais on ne la connait pas explicitement en général. On cherche à en réaliser une meilleure approximation, dans un sens à déterminer.

Le problème général est le suivant :

Soit $f \in H$, un espace fonctionnel normé.

On cherche $u \in V \subset H$, réalisant

$$\|u - f\| \;=\; \inf_{v \in V} \|v - f\| \quad . \tag{1.1}$$

En analyse numérique, l'espace V est en général de dimension finie ; on peut le noter V_n, où n est sa dimendion, et u sera alors noté u_n. Bien entendu, si n croît, la quantité $\|u_n - f\|$ va décroitre. On parlera d'approximation convergente lorsque

$$\forall\, \epsilon > 0 \quad \exists\, n_0 \in \mathbb{N} \qquad n > n_0 \;\Rightarrow\; \|u_n - f\| \;\leq\; \epsilon \quad . \tag{1.2}$$

1.1 Caractérisation de la meilleure approximation.

On suppose que H est muni d'un prodruit scalaire, noté (f, g), compatible avec la norme, c'est à dire $(f, f) \;=\; \|f\|^2$. Si l'espace H est en plus complet, il s'agit d'un espace de Hilbert.

On suppose le sous espace V_n de dimension finie n. On peut alors construire une base $\{q_1, q_2, .., q_n\}$ et caractériser la solution u_n de (1.1) par sa décomposition sur cette base :

$$u_n \;=\; \sum_{j=1}^{n} \lambda_j^*\, q_j \quad .$$

Le problème est maintenant ramené à un problème sur $\mathbb{R}^n$:

$$\text{chercher } (\lambda_1^*, \lambda_2^*, .., \lambda_n^*) \text{ réalisant } \min_{\lambda \in \mathbb{R}^n} \; \|\sum_{j=1}^{n} \lambda_j q_j \;-\; f\;\|^2 \quad .(1.3)$$

Typeset by the TEX-preprocessor PreTEX

On pose

$$Q(\lambda_1, \lambda_2, .., \lambda_n) \ = \ \| \sum_{j=1}^{n} \lambda_j q_j \ - \ f \ \|^2 \ .$$

On recherche le minimum de cette fonction, qui sera caractérisé par la propriété

$$\forall \, k \in \{1, .., n\} \quad \frac{\partial Q}{\partial \lambda_k} \, (\lambda_1^*, .., \lambda_n^*) \ = \ 0 \ . \tag{1.4}$$

On note $a_{ij} \ = \ (q_i, q_j)$, $b_i = (q_i, f)$, puis $A = ((\ a_{ij}))$, $b = (b_1, .., b_n)^t$, $x^* = (\lambda_1^*, .., \lambda_n^*)^t$.

Théorème 1.1 *Le vecteur x^* est l'unique solution du système linéaire $Ax = b$.*

Démonstration : Il suffit de développer $Q(\lambda_1, .., \lambda_n)$ et de dériver par rapport à chaque λ_k. On obtient

$$Q(\lambda_1, .., \lambda_n) \ = \ \sum_{ij} \lambda_i \lambda_j \, (q_i, q_j) \ - \ 2 \sum_i \lambda_i (q_i, f) \ + \| f \|^2$$

puis

$$\frac{\partial Q}{\partial \lambda_k} \ = \ 2 \, (\sum_{j=1}^{n} a_{kj} \, \lambda_j \ - \ b_k \,)$$

En écrivant que cette dérivée est nulle pour $\lambda = (\lambda_1^*, .., \lambda_n^*)$, on retrouve la $k$ième équation du système linéaire $Ax^* = b$.

Il reste à prouver l'unicité, et pour cela, il suffit de montrer que la matrice A est inversible. Dans le cas contraire, $det(A) = 0$, et on peut exprimer un de ses vecteurs colonnes en fonction des autres, c'est à dire

$$\exists \ \alpha_1, \alpha_2, .., \alpha_n, \text{ non tous nuls, tels que } \sum_{j=1}^{n} \ \alpha_j \, e_j \ = 0, \ avec \ e_j = \begin{pmatrix} (q_j, q_1) \\ (q_j, q_2) \\ ... \\ (q_j, q_n) \end{pmatrix}$$

On a donc

$$\forall \, i \ \in \ \{1, .., n\} \ \sum_{j=1}^{n} \ \alpha_j \, (q_j, q_i) \ = \ 0 \ .$$

On pose $q = \sum \alpha_j q_j$, on multiplie par α_i et on somme, pour obtenir

$$0 \ = \ \sum_{i,j} \ \alpha_i \alpha_j \, (q_j, q_i) \ = \ (\sum_i \alpha_i q_i, \ \sum_j \alpha_j q_j) \ = \ (q, q) \ = \ \| q \|^2$$

On en déduit $q = 0$, d'où $\forall i \ \alpha_i = 0$, et l'hypothèse $det(A) = 0$ est absurde. Il s'en déduit que A est inversible et x^* est unique.□

Le calcul des λ_i^* se ramène donc à l'inversion d'un système linéaire. On peut remarquer que la matrice A est symétrique et définie positive. Cependant, sa structure est très dépendante du choix de la base $\{q_1, q_2, .., q_n\}$ et on peut envisager de nombreux choix.

- des bases locales : les fonctions de base auront le plus souvent des supports disjoints, ce qui va réduire le nombre de diagonales non nulles de A,

- des bases orthonormées, en utilisant par exemple des polynômes orthogonaux,

- et d'autres bases, par exemple la base canonique des polynômes 1, x, x^2,.., x^n, pour laquelle la matrice A est très mal adaptée pour une inversion par les méthodes numériques classiques ; on dit qu'elle est "mal conditionnée".

Exemple : On prend $I =]0,1[$ et l'espace $H = L^2(I)$, espace des fonctions de carré sommable sur I. Pour un entier n, on pose $h = \frac{1}{n}$ puis $M_i =]ih, (i+1)h[$ $(0 \leq i \leq n-1)$, et on introduit le sous espace

$$V_n \ = \ \{ \ v \in H, \ v|_{M_i} \ = \ constante \ \} \quad .$$

On propose la base

$$q_j(x) \ = \ \begin{cases} 1 \ \ si \ x \ \in \ M_i \\ 0 \ \ sinon \end{cases} \quad ,$$

et dans ce cas la matrice A est diagonale : on trouve $A = h^2 I_d$. Son inversion est évidente.

1.2 Exemples de polynômes orthogonaux

On se donne un intervalle réel ouvert I. On peut distinguer trois cas :

- soit I est borné, et par translation et changement d'échelle, on peut se ramener à $I =]-1,1[$,

- soit I n'est borné que d'un seul coté (et va jusqu'à l'infini) de l'autre, et par translation et changement de signe éventuel, on peut se ramener à $I =]0, +\infty[$,

- soit $I = \mathbb{R}$.

1.2.1 Le cas $I =]-1,1[$.

On considère la quantité q_n , pour n entier, définie par

$$q_n(x) \ = \ \frac{1}{w(x)} \ \frac{d^n}{dx^n} \ \{w(x) \ (1-x^2)^n\} \quad , \tag{1.5}$$

où w est une fonction continue et intégrable sur I.

Proposition 1.2 *Soit $n \in \mathbb{N}$ $(n \geq 1)$, $k \in \{0,1,..,n-1\}$, et $w \in C^n(\mathbb{R})$. Alors*

$$\int_{-1}^{1} \ w(x) \ x^k \ q_n(x) \ dx \ = \ 0 \quad .$$

Démonstration : On intègre par parties :

$$\int_{-1}^{1} w(x) \ x^k \ q_n(x) \ dx \ = \ [x^k \ \frac{d^{n-1}}{dx^{n-1}} \ w(x)(1-x^2)^n \ ]_{-1}^{1} - \ k \int_{-1}^{1} \ x^{k-1} \ \frac{d^{n-1}}{dx^{n-1}} (w(x)(1-x^2)^n) dx$$

Or

$$[x^k \frac{d^{n-1}}{dx^{n-1}} w(x)(1-x^2)^n]_{-1}^1 = 0$$

et en réintégrant chaque fois, on aboutit à

$$\int_{-1}^1 w(x)\, x^k\, q_n(x)\, dx \;=\; (-1)^k\, k(k-1)..2.1 \int_{-1}^1 \frac{d^{n-k}}{dx^{n-k}}(w(x)(1-x^2)^n)dx$$

qui est nulle. Notons que nous avons utilisé que le produit $w(x)(1-x^2)^n$ était nul, ainsi que toutes ses dérivées jusqu'à l'ordre $n-1$, en $x=-1$ et en $x=1$.□

Remarque : on vient d'établir que, dans la mesure où w est positive, pour le produit scalaire défini par

$$(p,q) \;=\; \int_{-1}^1 w(x)\, p(x)\, q(x)\, dx\ ,$$

la quantité q_n est orthogonale aux x^k pour $k = 0,1,..,n-1$, et est donc orthogonale à tous les polynômes de degré $\leq n-1$. Si on peut choisir w de telle façon qur q_n soit un polynôme de degré n, on est assuré de construire une famille de polynômes orthogonaux.

Proposition 1.3 *Si w est de la forme suivante, pour $x \in I$,*

$$w(x) \;=\; A\,(1-x)^\alpha\,(1-x)^\beta\ , \tag{1.6}$$

avec $\alpha > -1$, $\beta > -1$, $A > 0$, *la quantité q_n est un polynôme de degré n.*

Remarque : L'hypothèse $\alpha > -1$, $\beta > -1$ assure que w soit intégrable sur I ; l'hypothèse $A > 0$ assure que w soit associé à un produit scalaire.
Démonstration : On a

$$q_n(x) = ((1-x)^{-\alpha}(1+x)^\beta \frac{d^n}{dx^n}\ \{(1-x)^{\alpha+n}\,(1+x)^{\beta+n}\}$$

d'où

$$q_n(x) \;=\; \sum_{k=0}^{n} \binom{n}{k}(n+\alpha)(n+\alpha-1)..(n+\alpha-k)(n+\beta)(n+\beta-1)..(n+\beta-n+k)(1-x)^{n-k}(1+x)^k$$

qui est bien un polynôme de degré n.□

Ces polynômes q_n sont appelés polynômes de Jacobi, dont on distingue quelques cas particuliers, connus surtout pout leur importance en physique, ou des propriétés particulières. Le signe de A ne joue pas un rôle capital, et le choix $A = 1$ est le plus naturel. On peut multiplier q_n par un coefficient $A_n \neq 1$ sans compromettre la propriété d'orthogonalité. On peut citer :

Les polynômes de Legendre $P_n(x)$, pour $\alpha = \beta = 0$;
Les polynômes de Tchébychev $T_n(x)$, pour $\alpha = \beta = -\frac{1}{2}$
Les polynômes de Gegenbauer $C_n^\gamma(x)$, pour $\alpha = \beta = \gamma - \frac{1}{2},\ \gamma > -\frac{1}{2}$.

Remarque : On a

$$T_n(x) \ = \ \sqrt{1-x^2}\ \frac{d^n}{dx^n}\left((1-x^2)^{n-\frac{1}{2}}\ .\right)$$

On a $T_0(x) = 1,\ T_1(x) = -x,\ T_2(x) = 6x^2 - 1, ...$ et cette propriété de parité est conservée pout tout n.

1.2.2 Le cas $I =]0, +\infty[$

On considére encore une expression de la forme

$$q_n(x) \ = \ \frac{1}{w(x)}\ \frac{d^n}{dx^n}\ (x^n\ w(x))\ \ , \tag{1.7}$$

avec w positive et intégrable sur $]0, +\infty[$.

Proposition 1.4 *Soit $n \in \mathbb{N}$ $(n \geq 1)$, $k \in \{0, 1, .., n-1\}$, $w \in C^n(I)$. On suppose $x^n w(x)$ tend vers zéro lorsque x tend vers $+\infty$. Alors :*

$$\int_0^\infty\ w(x)\ x^k\ q_n(x)\ dx\ \ = \ 0\ .$$

Démonstration : On intègre par parties

$$\int_0^\infty w(x)\ x^k\ q_n(x)\ dx\ = [x^k \frac{d^{n-1}}{dx^{n-1}}\ (x^n w(x))]_0^\infty\ - k\int_0^\infty\ x^{k-1}\ \frac{d^{n-1}}{dx^{n-1}}(x^n w(x))\ dx$$

Or, du fait des propriétés de w,

$$[x^k \frac{d^{n-1}}{dx^{n-1}}\ (x^n w(x))]_0^\infty\ \ = \ 0$$

et on réitère k fois l'intégration par parties, pour obtenir finalement

$$\int_0^\infty w(x) x^k q_n(x)\ \ = \ (-1)^k\ k(k-1)..2.1.\int_0^\infty\ \frac{d^{n-k}}{dx^{n-k}}(x^n\ w(x))dx\ \ = \ 0$$

d'où le résultat.□

Remarque : La propriété exigée sur w est une propriété de décroissance rapide, qui signifie que $w(x)$ converge plus vite vers zéro lorsque x tend vers l'infini que n'importe quel polynôme.

Proposition 1.5 *Si $w(x)$ est de la forme $w(x) \ = \ x^\alpha\ e^{-x}$, avec $\alpha > -1$, alors q_n est un polynôme de degré n.*

Démonstration : On calcule

$$q_n(x) = x^{-\alpha}\, e^x\, \frac{d^n}{dx^n}(\, x^{n+\alpha}\, e^{-x}) \;=\; \sum_{k=0}^{n} \binom{n}{k} (-1)^k\, x^{n-k} \;\;,$$

et il s'agit bien d'un polynôme de degré n.□

Remarque : Ces polynômes correspondent aux polynômes de Laguerre. On écrit en général, pour $\alpha > -1$,

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

1.2.3 Le cas $I = \mathbf{R}$.

On prendra cette fois

$$q_n(x) \;=\; \frac{1}{w(x)}\, \frac{d^n}{dx^n}\, w(x)\;, \quad \text{avec } w(x) \;=\; e^{-\frac{1}{2}(x-x_0)^2} \;\;, \tag{1.8}$$

et on obtient poux $x_0 = 0$, les polynômes d'Hermite, généralement notés

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

On vérifie immédiatement qu'il s'agit de polynômes de degré n.

1.2.4 Construction des polynômes orthogonaux.

On peut toujours obtenir une relation de récurrence.

Théorème 1.6 *Soit $\{q_n\}$ une famille de polynômes orthogonaux, le degré de q_n étant n ; on note a_n le coéfficient ($\neq 0$) de x^n dans q_n. Alors, pour tout $n \geq 1$, il existe une relation de récurrence de la forme*

$$q_{n+1}(x)\, (\alpha_n x + \beta_n)\, q_n(x) \;-\; \gamma_n\, q_{n-1}(x) \tag{1.9}$$

avec

$$\alpha_n \;=\; \frac{a_{n+1}}{a_n} \quad,\; \beta_n \text{ réel}, \;\; \gamma_n \;=\; \frac{\alpha_n\, \|q_n\|^2}{\alpha_{n-1}\, \|q_{n-1}\|^2} \;\;.$$

Démonstration : On considère le polynôme $q_{n+1}(x) - \alpha_n x q_n(x)$; on choisit α_n de telle sorte qu'il soit de degré n au plus, d'où $a_{n+1} = \alpha_n a_n$. On peut écrire

$$q_{n+1}(x) \;-\; \alpha_n\, x\, q_n(x) \;=\; \sum_{k=0}^{n} \mu_k\, q_k(x) \;\;,$$

avec des coéficients μ_k à déterminer. On multiplie par q_j, en remarquant que pour $j < n-1$, le produit $(xq_n, q_j) \;=\; (q_n, xq_j) \;=\; 0$, car xq_j est de degré strictement inférieur à n, et pour $j \leq n$, le produit $q_{n+1}, q_j) = 0$. Il reste donc, pour $j \leq n-2$, $\mu_j \|q_j\|^2 = 0$, donc $\mu_j = 0$.

Il ne reste que les termes correspondant à $j = n-1$ et à $j = n$, d'où la formule, avec $\gamma_n = -\mu_{n-1}$ et $\beta_n = \mu_n$. De plus, pour $j = n-1$, on a

$$\alpha_n \, (x q_n, q_{n-1}) \;=\; \gamma_n \, \|q_{n-1}\|^2$$

Or $x q_{n-1}$ est de la forme

$$x q_{n-1}(x) \;=\; \frac{a_{n-1}}{a_n} \, q_n(x) \;+\; \sum_{j=0}^{n-1} \delta_j \, q_j(x)$$

donc

$$(x q_n, q_{n-1}) \;=\; (q_n, x q_{n-1}) \;=\; \frac{a_{n-1}}{a_n} \, \|q_n\|^2 \;=\; \frac{1}{\alpha_{n-1}} \, \|q_n\|^2 \;,$$

et on a obtenu

$$\frac{\alpha_n}{\alpha_{n-1}} \, \|q_n\|^2 \;=\; \gamma_n \, \|q_{n-1}\|^2 \;. \square$$

Remarques : Cette formule permet de construire les polynômes de proche en proche, car β_n et γ_n ne dépendent pas de q_{n+1}, et le choix des a_n, donc des α_n, reste libre. Par exemple, pour le choix $a_n = 1$, on trouve $\alpha_n = 1$.
Il reste à étudier le comportement lorsque n tend vers l'infini, de la représentation d'une fonction f sur une base de tels polynômes. Si on écrit la série

$$f(x) \;=\; \sum_{n=0}^{\infty} \, c_n \, q_n(x) \;\;,$$

on aura, pour tout $k \in \mathbb{N}$,

$$(f, q_k) \;=\; c_k \, \|q_k\|^2 \;\;,$$

d'où les c_k, et il reste encore à vérifier que la somme de la série est bien f... Par exemple pour les polynômes de Tchébychev, on a, à un coéfficient près,

$$T_n(cos\theta) \;=\; cos(n\theta)$$

et donc

$$f(cos\theta) \;=\; \sum_{n=0}^{\infty} \, c_n \, cos(n\theta)$$

c'est à dire le développement de Fourier de $g(\theta) = f(cos\theta)$. On peut utiliser les résultats de convergence obtenus sur l'étude des séries de Fourier.

1.3 Construction d'autres bases

Le choix d'une base orthogonale est évidemment idéal. En prenant par exemple l'espace des polynômes de degré inférieur ou égal à $n-1$ (cet espace est de dimension n), que l'on munit de la base canonique $1,\ x,\ x^2,..,\ x^{n-1}$, les coéfficients de la matrice A sont

$$a_{ij} \ = \ \int_{-1}^{1} \ x^{i+j}\, dx \ = \ \begin{cases} 0 & si\ i+j\ est\ impair \\ \frac{2}{i+j+1} & si\ i+j\ est\ pair \end{cases} .$$

On obtient

$$A \ = \ \begin{pmatrix} \frac{2}{3} & 0 & \frac{2}{5} & 0 & \frac{2}{7} & ..... \\ 0 & \frac{2}{5} & 0 & \frac{2}{7} & 0 & ..... \\ \frac{2}{5} & 0 & \frac{2}{7} & 0 & \frac{2}{9} & ..... \\ & & ..... & & & \\ ..... & & & & \frac{2}{2n+1} & \end{pmatrix}$$

En pratique, cette matrice est difficile à inverser : elle est "*mal conditionnée*", c'est à dire que les méthodes numériques classiques vont créer des situations favorables aux instabilités (des nombres trop grands ou trop petits). De plus la présence de nombreux zéros ne constutue pas ici un avantage : ils sont trop dispersés.

Le problèmde d'instabilité se traduit par le phénomène suivant : une très légère variation des données (ici la fonction f) va provoquer une très forte variation des résultats (ici les coéfficients du polynôme d'approximation). Ce phénomène sera d'autant plus important que n est grand. On ne retient donc le choix d'une base polynomiale que dans le cas où on dispose de polynômes orthogonaux ou lorsqu'on est assuré d'avoir une convergence très rapide, c'est à dire lorsque, dans la formulation

$$f(x) \ = \ \sum_{n=0}^{\infty} \ c_n \, q_n(x) \ ,$$

la suite $\{c_n\}$ converge rapidement vers zéro, et qu'il suffit en pratique de calculer un nombre limité de coéfficients c_n.

Une autre stratégie consiste à construire une base *locale*, c'est à dire constituée de fonctions dont le *support* est minimal. On définit le support d'une fonction φ par

$$supp(\varphi) \ = \ \overline{f^{-1}(\mathbb{R}\backslash\{0\})} \ ,$$

où la barre correspond à l'adhérence. Le principe est le suivant.

Soit N un entier ; on pose $h = \frac{1}{N}$ puis, pour $i \in J_N \equiv \{-N, ..., N-1\}$,

$$K_i \ = \ [ih,\ (i+1)h] \ .$$

On remarque

$$\begin{gathered} K_i \cap K_{i-1} \ = \ \{ih\} \\ K_i \cap K_j \ = \ \emptyset \ \ \text{si} \ |i-j| > 2 \\ \cup_{i \in J_N} K_i \ = \ \overline{I} \end{gathered}$$

Les points $S_i = ih$ sont les *sommets* des *éléments* K_i.

On peut maintenant choisir un entier n (en général n $\leq$ 2 ou 3) et considérer l'espace de dimension finie

$$V_h \;=\; \{u \in E(\overline{I}),\; \forall i \in J_N \;:\; u|_{K_i} \in ?\mathbf{P}_n\} \;\;,$$

où $\mathbf{P}_n$ désigne l'espace des polynômes de degré $\leq n$ et E est un espace de fonctions définies sur $\overline{I}$. La dimension de V_h va dépendre bien entendu de N, mais aussi de la régularité des fonctions de $E(\overline{I})$. Considérons plusieurs exemples.

1 ° - $n = 0$, $E(\overline{I}) = C^0(\overline{i})$; les fonctions de V_h sont des constantes sur chaque éléments, et la contrainte de contituité impose qu'il s'agisse de la *même* constante. En clair : $dim(V_h) \;=\; 1$.

2 ° - $n = 0$, $E(\overline{I}) = L^2(\overline{I})$; il n'y a plus de contrainte de continuité, et $dim(V_h) \;=\; card(J_N) \;=\; 2N$. On peut prendre comme fonctions de base

$$\forall i \in J_N \;\; q_i(x) \;=\; \begin{cases} 1 \;\; si \; x \in K_i \\ 0 \;\; si \; x \notin K_i \end{cases} \;\;,$$

et la matrice de projection sur cette base a pour composantes

$$a_{ij} \;=\; (q_i, q_j) \;=\; \int_I q_i(x) q_j(x) dx \;=\; \int_{K_i \cap K_j} q_i(x) q_j(x) dx \;=\; \begin{cases} h \;\; si \; i = j \\ 0 \;\; si \; i \neq j \end{cases}$$

Remarquons qu'il s'agit d'une base orthogonale.

3 ° - $n = 1$, $E(\overline{I}) = C^0(\overline{I})$. On retrouve la contrainte de continuité et les fonctions de V_h sont affines par élément. Le nombre de degrés de liberté est 2 pour le premier élément, puis 1 pour les suivants, d'où $dim(V_h) = 2N + 1$, c'est à dire exactement le nombre de sommets. Une base locale doit avoir des supports minimaux. En prenant par exemple $q_i(x) = 1$ pour $x = S_i$, on voit qu'il est nécéssaire que le support de q_i recouvre au moins deux éléments si $i \neq -N$ ou $i \neq N$, et bien entendu il peut être réduit à un seul élément si $i = -N$ ou $i = N$. Pour avoir un support minimal, on prendra

$$q_i(x) \;=\; \begin{cases} 1 \;\; si \; x = S_i \\ 0 \;\; si \; x = S_j, \;\; j \neq i \\ affine \; sur \; chaque \; élément \end{cases}$$

On aura $(q_i, q_j) = 0$ si $|i - j| \geq 2$, ce qui conduit à une matrice tridiagonale.

4 ° - $n = 1$, $E(\overline{I}) = L^2(\overline{I})$. Il n'u a pas de contrainte de continuité et on peut associer deux fonctions de base par élément. En prenant

$$q_i^1(x) \;=\; \begin{cases} 1 \;\; si \; x \in K_i, \\ 0 \;\; sinon \end{cases} \;\;,\;\; q_i^2(x) \;=\; \begin{cases} x - \frac{S_i + S_{i+1}}{2} \; si \; x \in K_i \\ 0 \qquad\qquad sinon \end{cases} \tag{1.10}$$

On retrouve une base orthogonale, et $dim(V_h) = 4N$.

5 ° - $n = 1$, $E(\overline{I}) = C^1(\overline{I})$. La contrainte de continuité porte à la fois sur la fonction et sur sa dérivée. On a encore deux choix pour le premier élément, et plus aucun pour les suivants : $dim(V_h) = 2$.

Ces différents exemples montrent l'importance de la contrainte de régularité, c'est à dire du choix de $E(\overline{I})$. Dans les problèmes où la solutuion attendue est très régulière, on aura intérêt à être assez exigeant en régularité, ce qui se traduira par un gain apprécianle sur la taille de la matrice ; d'un autre côté, elle sera plus pleine. C'est par exemple le cas pour la mécanique des structures en faibles déformations. A l'inverse, dans des problèmes où la solution attendue présente de plus fortes variations, il vaut mieux rédure cette exigence de régularité ; la matrice sera alors plus creuse mais de taiile plus grande, et finalement sera assez facile à inverser avec une méthode adaptée.

Bien évidemment, les cas où la dimension de V_h ne tend pas vers l'infini avec N sont à rejeter.

Dans le cas de la dimension deux, les élément K_i deviennent ds triangles et les sommets S_i sont les sommets de ces triangles. Dans le cas où chaque sommet S_i est effectivement sommet de tous les triangles dont il est adjascent, on peut facilement généraliser les remarques précédentes.

1.4 Un théorème de convergence, pour les éléments affines

On se place dans la situation N ° 3 : $n = 1, E(\overline{I}) = C^1(\overline{I})$, et on choisit un produit scalaire de la forme

$$(u, v) \;=\; \int_I \left(a(x)u(x)v(x) + b(x)u'(x)v'(x)\right)\, dx \tag{1.11}$$

avec $a > 0$ et $b \geq 0$. La norme associée est telle que

$$\|u\|^2 \;=\; \int_I \left(a(x)\; u(x)^2 \;+\; b(x)\; v(x)^2\right)\, dx.$$

On suppose les fonctions a et b continues sur $\overline{I}$. On utilise la base q_i proposée en (1.10), dont les valeurs aux sommets S_i correspondent aux degrés de liberté. On note u_h la projection de la donnée f sur V_h :

$$\|u_h - f\| \;=\; \mathrm{Inf}_{v \in V_h} \|v - f\| \;. \tag{1.12}$$

Théorème 1.7 *On suppose $f \in C^1(\overline{I})$. Alors $\|u_h - f\|$ tend vers zéro lorsque N tend vers l'infini (simultanément, h tend vers zéro).*

Démonstration : On va montrer qu'une approximation $v_h \in V_h$ converge, c'est à dire que $\|v_h - f\|$ tend vers zéro ; comme $\|u_h - f\| \leq \|v_h - f\|$, ceci entrainera la convergence de u_h.

On pose

$$v_h(x) \;=\; \sum_{i=-N}^{N} \; f(S_i)\; q_i(x)\;.$$

On évalue maintenant $\|v_h - f\|^2$. Notons que la dérivée de v_h n'existe pas aux sommets S_i, soit un nombre fini de points, mais v_h' reste définie presque partout, constante par morceaux et bornée, donc intégrable. On a

$$\|v_h - f\|^2 = \sum_{j \in J_N} \int_{K_j} \left(a(x)\, |v_h - f|^2 + b(x) |v_h' - f'|^2 \right) dx \ .$$

Soit $\epsilon > 0$. Sur $\overline{I}$ la fonction f' est uniformément continue, donc

$$\exists \delta_1 \, \forall x \in \overline{I}, \ \forall \xi \in \overline{I}, \ |x - \xi| < \delta_1 \ \Rightarrow \ |f'(x) - f'(\xi)| \leq \epsilon \ . \tag{1.13}$$

On se place sur un élément K_j :

$$v_h'(x) \ = \ \frac{f(S_{j+1}) - f(S_j)}{h} \ = \ f'(\xi_j) \ , \quad \text{avec } \xi_j \in K_j \ ,$$

et si $h \leq \delta_1$, on aura, pour tout $x \in K_j$: $|x - \xi_j| \leq h \leq \delta_1$, donc,

$$|v_h'(x) - f'(x)| \leq \ |f'(\xi_j) - f'(x)| \leq \epsilon \ .$$

De même pour f :

$$\exists \delta_0 \, \forall x \in \overline{I}, \ \forall \eta \in \overline{I}, \ |x - \eta| < \delta_0 \ \Rightarrow \ |f(x) - f(\eta)| \leq \epsilon \ . \tag{1.14}$$

On se replace sur un élément K_j :

$$v_h(x) \ = \ \frac{1}{h} \left(f(x_{j+1}) - f(x_j) \right)(x - x_j) \ + \ f(x_j) \ ,$$

qui se situe entre les valeurs $f(x_j)$ et $f(x_{j+1})$. D'après le théorème de la moyenne, il existe $\eta_j \in K_j$ tel que $v_h(x) = f(\eta_j)$, et pour $h \leq \delta_0$,

$$|x - \eta_j| \ \leq \ h \ \leq \ \delta_0 \ \Rightarrow \ |f(x) - f(\eta_j)| \ \leq \ \epsilon \ .$$

Ainsi, on a obtenu, pour $h \ \leq \ \min(\delta_0, \delta_1)$,

$$\|v_h - f\|^2 \ \leq \ \sum_{j \in J_N} \epsilon^2 \int_{K_j} \left(a(x) + b(x) \right) dx \ \leq \ K^2 \, \epsilon^2 \ ,$$

en posant

$$K^2 \ = \ \int_I \left(a(x) + b(x) \right) dx \ .$$

Finalement,

$$\exists \ N_0 \ = \ \frac{1}{\min(\delta_0, \delta_1)} \quad \text{tel que } N > N_0 \ \Rightarrow \ \|v_h - f\| \ \leq \ \epsilon \ .$$

D'où la convergence de $\|v_h - f\|$ donc de $\|u_h - f\|$.□

Application

Soit g une fonction donnée sur I, par exemple $g \in C^0(\overline{I})$. On considère le problème aux limites

$$\text{Chercher } y \in C^0(\overline{I}) \quad \text{tel que } \begin{cases} a(x)\, y \ - \ (b(x) y')' \ = \ g \\ y(-1) \ = \ y(1) \ = \ 0 \end{cases} \tag{1.15}$$

Soit v une fonction continue, telle que v' soit intégrable sur I (elle est par exemple continue par morceaux et bornée). On suppose que $v(1) = v(-1) = 0$, et on multiplie l'équation par v, puis on intègre par parties, pour obtenir :

$$\int_I \left(a(x)\, y\, v \;+\; b(x)\, y'\, v'\right) dx \;=\; \int_I gv\, dx \tag{1.16}$$

On peut maintenant envisager d'approcher la solution y en utilisant la méthode précédente : on remplace y par $u_h \in V_h$, et qu'on représente en utilisant la base $\{q_i\}_{i\in J_N^*}$, où $J_N^* = \{-N+1, ..., N-1\}$. On note

$$u_h(x) \;=\; \sum_{j\in J_N^*} y_j\, q_j(x) \;\;,$$

où les composantes y_j vont en fait approcher les valeurs $y(S_j)$. On obtient, en prenant $v = q_i$,

$$\sum_{j\in J_n^*} \left(\int_I \left(a(x)q_jq_i + b(x)q_j'q_i'\right)dx\right) y_j \;=\; \int_I g\, q_i dx$$

et en posant

$$a_{ij} \;=\; \int_I \left(a(x)q_jq_i + b(x)q_j'q_i'\right)dx \;, \quad b_i \;=\; \int_I g\, q_i\, dx \;\;,$$

ceci se réduit au système linéaire

$$\sum_{j\in J_N^*} a_{ij}\, y_j \;=\; b_i \;.$$

De plus, on remarque que

$$(y, q_i) \;=\; \int_I \left(a(x)\, y\, q_i \;+\; b(x)\, y'\, q_i'\right) dx \;=\; b_i \left(= \int_I g\, q_i\, dx\right) \tag{1.17}$$

c'est à dire que les produits scalaires (y, q_i) sont connus, bien qu'on ne connaisse pas y.

Il s'agit bien du problème de projection précédent, avec un produit scalaire un peu particulier, et qui permet d'approcher la solution d'un problème différentiel aux limites. de plus, le théorème précédent assure la convergence de cette méthode.

1.5 Le théorème de Weierstrass

Soit $f \in C^0([a, b])$, où $[a, b]$ est un intervalle borné. On veut approcher f par un polynôme, au sens de la convergence uniforme

$$\|f\| \;=\; \mathrm{Sup}_{\mathrm{a\le x\le b}} |\mathrm{f(x)}| \;\;.$$

On sait que $C^0([a, b])$, muni de cette norme, est complet.

Théorème 1.8 *Soit $f \in C^0([a,b])$, $\epsilon > 0$. Il existe au moins un polynôme p tel que*

$$\|f - p\| \ \leq \ \epsilon \ . \tag{1.18}$$

Démonstration : Par translation et changement d'échelle, on peut se ramener à $a = 0,\ b = 1$. On pose, pour $n \in \mathbb{N}$,

$$B_n(x,f) \ = \ \sum_{k=0}^{n} \binom{n}{k} \, f(\frac{k}{n}) \, x^k \, (1-x)^{n-k} \ . \tag{1.19}$$

Il s'agit d'un polynôme de degré n, appelé *polynôme de Bernstein*. Dans le cas $f = 1$, on vérifie immédiatement (formule du binôme)

$$B_n(x,1) \ = \ 1 \ .$$

Pour $f(x) = x$, on obtient

$$B_n(x,x) \ = \ \sum_{k=0}^{n} \frac{k}{n} \, \frac{n!}{k! \, (n-k)!} \, x^k (1-x)^{n-k} \ = \ x \sum_{j=0}^{m} \binom{m}{j} x^j (1-x)^{m-j} \ = \ x$$

avec $m = n - 1$. L'approximation est donc exacte pour les polŷnomes de degré ≤ 1. On considère maintenant le cas $f(x) = x^2$:

$$B_n(x,x^2) = \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} \, \frac{k^2}{n^2} \, x^k (1-x)^{n-k} = \frac{n-1}{n} x^2 \sum_{j=0}^{m} \binom{m}{j} x^j (1-x)^{m-j} \ + \ \frac{x}{n}$$

où $m = n - 2$. On a obtenu

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

et on en déduit l'estimation

$$|B_n(x,x^2) - x^2| \ \leq \ \frac{x(1-x)}{n} \ \leq \ \frac{1}{4n} \, .$$

Nous allons utiliser ces calculs de $B_n(x,1),\ B_n(x,x),\ B_n(x,x^2)$ pour établir le résultat. Soit $\epsilon > 0$. Alors, f étant uniformément continue sur $[0,1]$, on a

$$\exists \delta > 0 \ \forall x \in [0,1], k \in \{0,1..n\} \ \ |x - \frac{k}{n}| < \delta \ \Rightarrow \ |f(x) - f(\frac{k}{n})| \ < \ \frac{\epsilon}{2} \ .$$

On pose $M = \|f\|$ et $K(x) = \Big\{ k \in \{0,1,..n\} |\ |x - \frac{k}{n}| < \delta \Big\}$. Alors

$$|f(x) - B_n(x,f)| \leq \sum_{k \in K(x)} \binom{n}{k} |f(x) - f(\frac{k}{n})| x^k (1-x)^{n-k} \ + \sum_{k \notin K(x)} \binom{n}{k} |f(x) - f(\frac{k}{n})| x^k (1-x)^{n-k}$$

d'où

$$|f(x) - B_n(x,f)| \ \leq \ \frac{\epsilon}{2} \ + \ 2 \ M \sum_{k \notin K(x)} \binom{n}{k} x^k (1-x)^{n-k}$$

et comme

$$k \notin K(x) \;\Rightarrow\; 1 \;\leq\; \frac{1}{\delta^2}\,(x - \frac{k}{n})^2$$

il vient

$$|f(x) - B_n(x,f)| \;\leq\; \frac{\epsilon}{2} \;+\; \frac{2\,M}{\delta^2} \sum_{k \notin K(x)} \binom{n}{k} (x - \frac{k}{n})^2\, x^k (1-x)^{n-k}$$

puis, en étendant cette somme à tout $k \in \{0, 1, ..n\}$, il vient

$$|f(x) - B_n(x,f)| \;\leq\; \frac{\epsilon}{2} \;+\; \frac{2M}{\delta^2}\,(x^2 + \frac{x(1-x)}{n} \;- 2x^2 + x^2) \;\leq\; \frac{\epsilon}{2} \;+\; \frac{2M}{4n\delta^2}\,.$$

Prenons maintenant $N = \frac{M}{\epsilon\delta^2}$ et il vient

$$n \geq N \;\Rightarrow\; |f(x) - B_n(x,f)| \;\leq\; \frac{\epsilon}{2} \;+\; \frac{M}{2N\delta^2} \;=\; \epsilon\,. \square$$

Ce résultat induit d'autres résultats de convergence. Consédérons par exemple le problème de la recherche de la meilleure approximation par un polynôme, associée à la forme

$$\|f\|^2 \;=\; \int_0^1 a(x)\,|f(x)|^2\,dx$$

avec $a \in C^0([0,1]),\; a > 0$. On a immédiatement un résultat de convergence :

Corollaire 1.9 *Il y a convergence de la meilleure approximation par un polynôme.*

Démonstration : On note p_N la meilleure approximtion par un polynôme de degré N. Alors

$$\|f - p_N\|^2 \;\leq\; \|f - B_N(.,f)\|^2 \;\leq \int_0^1 a(x)\,|f(x) - B_N(x,f)|^2\,dx \;\leq\; \epsilon^2 \int_0^1 a(x)dx$$

d'où la convergence.$\square$

1.6 Approximation au sens des moindres carrés

On dispose d'une suite de données (ou mesures) expérimentales (x_i, y_i) pour $i = 1, 2, ..n$, et on se propose de déterminer une relation (loi physique) de la forme

$$y \;=\; f(x) \quad , \tag{1.20}$$

à partir de ces informations. On n'aura pas exactement en général $y_i = f(x_i)$ pout tout i, compte tenu d'inévitables ereurs de mesures expérimentales, et par ailleurs, si f est par exemple un polynôme, on voit mal son degré dépendre du nombre d'expériences réalisées... En général, la formulation de f est très simpe, et implique un faible nombres de paramètres : un polynôme de faible degré, un élément d'un espace vectoriel de faible dimension, etc...

Soit V un espace vextoriel de dimension finie. Le problème consiste à chercher $f \in V$ rélisant

$$\mathrm{Inf}_{f\in V} \sum_{i=1}^{n} |f(x_i) - y_i|^2 \quad , \tag{1.21}$$

c'est à dire que f est la meilleure approximation des données, dans V et au sens des moindres carrés.

Remarque : La quantité $\Big(\sum_i |f(x_i)-y_i|^2\Big)^{\frac{1}{2}}$ n'est pas la norme d'une expression de la forme $f(x)-y$. En effet, si $q \in V$ et vérifie $q(x_i) = 0$ pour tout i, les quantités $f+\lambda q$ seront également des solutions, ceci quelquesoit λ réel. Il n'y aura pas unicité sauf si la dimension de V est suffisamment réduite pour éviter l'existence d'une telle solution.

Théorème 1.10 *On se donne n points (x_i, y_i), $1 \le i \le n$, $n \ge 2$ avec $x_i \neq x_j$ si $i \neq j$. Alors, il existe une droite unique $f(x) = ax + b$, réalisant (1.21), appelée droite des moindres carrés.*

Démonstration : On considère

$$J(a,b) \;=\; \frac{1}{2} \sum_{i=1}^{n} |ax_i + b - y_i|^2$$

et on écrit que J réalise son minimum en un point (a^*, b^*). La solution correspondra ainsi à la droite $y = a^*x + b^*$ $(= f(x))$. Ce minimum existe car J est à valeurs dans $[0, \infty[$, et son ensemble de valeurs admet un élément minimal, qui est effectivement réalisé par une valeur prise par J, qui est continue.

En un point minimal, on aura

$$\frac{\partial J}{\partial a}(a^*, b^*) \;=\; \frac{\partial J}{\partial b}(a^*, b^*) \;=\; 0 \quad ,$$

ce qui conduit à deux équations :

$$\frac{\partial J}{\partial a}(a^*, b^*) \;=\; 0 \;\Rightarrow\; a^* \left(\sum_{i=1}^{n} x_i^2\right) \;+\; b^* \left(\sum_{i=1}^{n} x_i\right) \;=\; \sum_{i=1}^{n} x_i y_i \tag{1.22}$$

$$\frac{\partial J}{\partial b}(a^*, b^*) \;=\; 0 \;\Rightarrow\; a^* \left(\sum_{i=1}^{n} x_i\right) \;+\; n\, b^* \;=\; \sum_{i=1}^{n} y_i \tag{1.23}$$

Il s'agit d'un système linéaire, de déterminant

$$\Delta \;=\; n \sum_{i=1}^{n} x_i^2 \;-\; \left(\sum_{i=1}^{n} x_i\right)^2 .$$

Or, par l'inégalité de Cauchy Schwartz,

$$\left(\sum_{i=1}^{n} x_i\right)^2 \;\le\; \left(\sum_{i=1}^{n} 1\right) \left(\sum_{i=1}^{n} x_i^2\right) \;\le\; n \sum_{i=1}^{n} x_i^2 \;,$$

et donc $\Delta \ge 0$. De plus, l'égalité n'est acquise que lorsque les deux vecteurs $(1, 1, ..1)^t$ et $(x_1, x_2, ..x_n)^t$ sont colinéaires, c'est à dire lorsque tous les x_i sont égaux, ce qui est exclus par l'hypothèse. On a donc $\Delta > 0$, d'où une solution unique pour le système linéaire. Cette unicité assure qu'il s'agit bien d'un minimum global, et non local.□

Remarques : En pratique, on calcule a^* et $bU*$ en résolvant le système sonstitué des équations (1.22) et (1.23).

Il existe une autre technique, permettant d'établir que (a^*, b^*) correspond bien à un minimum global. On effectue un développement de Taylor à l'ordre deux, qui est ici exact, car J est quadratique. On a

$$J(a^*+h, b^*+k) \;=\; \left\{ \begin{array}{c} J(a^*,b^*) \\ +\, h\, \frac{\partial J}{\partial a}\,(a^*,b^*) \;+\; k\, \frac{\partial J}{\partial b}\,(a^*,b^*) \quad (=0) \\ +\, \frac{h^2}{2}\, \frac{\partial^2 J}{\partial a^2} \;+\; hk \frac{\partial^2 J}{\partial a \partial b} \;+\; \frac{k^2}{2}\, \frac{\partial^2 J}{\partial b^2} \end{array} \right.$$

ceci pour tout h et k réels. Il reste effectivement

$$J(a^*+h, b^*+k) \;=\; J(a^*,b^*) \;+\; \frac{1}{2} \sum_{i=1}^{n} \left(h^2 x_i^2 + 2hkx_i + k^2 \right) \;\geq\; J(a^*,b^*) \;\; .$$

Il s'agit bien d'un minimum, et cette démonstration assure la convexité de J.

On peut faire la même démarche à partir de deux fonctions φ et ψ. On définit V comme l'espace de dimension 2 $V = \{u = \alpha\varphi + \beta\psi\}$ et on aura une solution unique, à condition que

$$\sum_{i=1}^{n} \varphi(x_i)^2 \sum_{i=1}^{n} \psi(x_i)^2 \;\neq\; \left(\sum_{i=1}^{n} \varphi(x_i)\psi(x_i) \right)^2 \;\; ,$$

c'est à dire que les vecteurs $(\varphi(x_1), ..\varphi(x_n))^t$ et $(\psi(x_1), ..\psi(x_n))^t$ ne sont pas colinéaires. Ceci est exclu si effectivement la dimension de V est égale à 2.

1.6.1 Polynômes des moindres carrés

La démarche précédente peut être étendue au cas où V est l'espace des polynômes de degré $\leq p$ (qui est de dimension $p+1$). On se donne $n \geq p+1$, et on cherche f de la forme

$$f(x) \;=\; a_0 + a_1 x + ... + a_p x^p \;\; .$$

On note $a = (a_0, a_1, ..a_n)^t$. Le problème revient à minimiser

$$J(a) \;=\; \frac{1}{2} \sum_{i=1}^{n} \left| a_0 + a_1 x_i + a_2 x_i^2 + ... a_p x_i^p \;-\; y_i \right|^2$$

On écrit que le minimum est réalisé lorsque la gradient de J est nul, c'est à dire qu'à ce point, noté a^*, on aura

$$\forall k \in \{0, 1, ..p\} \qquad \frac{\partial J}{\partial a_k}\,(a^*) \;=\; 0 \;\; .$$

Comme

$$\frac{\partial J}{\partial a_k}\,(a) \;=\; \sum_{i=1}^{n} (a_0 + a_1 x_1 + .. + a_p x_i^p - y_i) x_i^k \;\; ,$$

en posant

$$S_q \;=\; \sum_{i=1}^{n} x_i^q \;\; , \qquad v_k \;=\; \sum_{i=1}^{n} \; x_i^k \, y_i \quad \;\; ,$$

il vient

$$\begin{pmatrix} S_0 & S_1 & ... & S_p \\ S_1 & S_2 & ... & S_{p+1} \\ & ..... & & \\ S_p & S_{p+1} & ... & S_{2p} \end{pmatrix} \begin{pmatrix} a_0^* \\ a_1^* \\ .. \\ a_p^* \end{pmatrix} = \begin{pmatrix} v_0 \\ v_1 \\ .. \\ v_p \end{pmatrix} \quad . \tag{1.24}$$

On introduit la matrice rectangulaire (n lignes et $p+1$ colonnes)

$$A = \begin{pmatrix} 1 & x_1 & x_1^2 & ... & x_1^p \\ 1 & x_2 & x_2^2 & ... & x_2^p \\ & & ..... & & \\ 1 & x_n & x_n^2 & ... & x_n^p \end{pmatrix} .$$

Proposition 1.11 *On a* $S = A^t A$.

Démonstration : il suffit de le vérifier en effectuant le produit $A^t A$.□

Proposition 1.12 *Le système linéaire (1.24) admet une solution unique.*

Démonstration : Dans le cas contraire, $\exists u \neq 0$ tel que $Su = 0$. Or, on a

$$\|Au\|^2 = (Su, u) \ ,$$

et donc $Au = 0$. Ceci se traduit par

$$\forall j \in \{1, ..n\} \qquad u_0 + u_1 x_j + ... + u_p x_j^p = 0$$

et donc le polynôme $q(x) = ud0 + u_1 x + ... + u_p x^p$ admet $n \geq p+1$ racines tout en étant de degré p. Il est donc identiquement nul.
Il reste à vérifier qu'il s'agit bien d'un minimum. On a, pour $h \in \mathbb{R}^n$,

$$J(a^* + h) = J(a^*) + \sum_{j=0}^{p} \frac{\partial J}{\partial a_j}(a^*) h_j + \frac{1}{2} \sum_{j=0}^{p} \sum_{k=0}^{p} \frac{\partial^2 J}{\partial a_j \partial a_k} h_j h_k$$

Or le gradient de J est nul en $a = a^*$, et

$$\frac{\partial^2 J}{\partial a_j \partial a_k} = \frac{\partial}{\partial a_j} \left(\sum_{i=1}^{n} (a_0 + a_1 x_i + ... + a_p x_i^p - y_i) x_i^k \right) = \sum_{i=1}^{n} x_i^{j+k} \ ,$$

qui sont les éléménts de la matrice S. On peut donc écrire

$$J(a^* + h) = J(a^*) + \frac{1}{2} \|Ah\|^2 \geq J(a^*) \, ,$$

qui montre que a^* est bien le minimum.□

Remarques : Le calcul des composantes de a^* correspond à inverser le système linéaire (1.24), et on verra plus loin que la matrice S est mal conditionnée, c'est à dire qu'une faible variation des données (ici les v_k)peut provoquer une forte variation du résultat. ce phénomène est d'autant plus important que p est grand, pour devenir incontrolable si par exemple $p > 10$.

On peut généraliser la même démarche au cas

$$f(x) \;=\; \sum_{j=1}^{p} \; a_j \; \varphi_j(x) \;\;,$$

avec les mêmes inconvénient.. On exigera que les fonctions φ_j soient linéairement indépendantes. En pratique, des fonctions φ_j de type exponentielles sont souvent utilisées, en biologie notamment, pour représenter des phénomènes qui s'amortissent rapidement à l'infini.